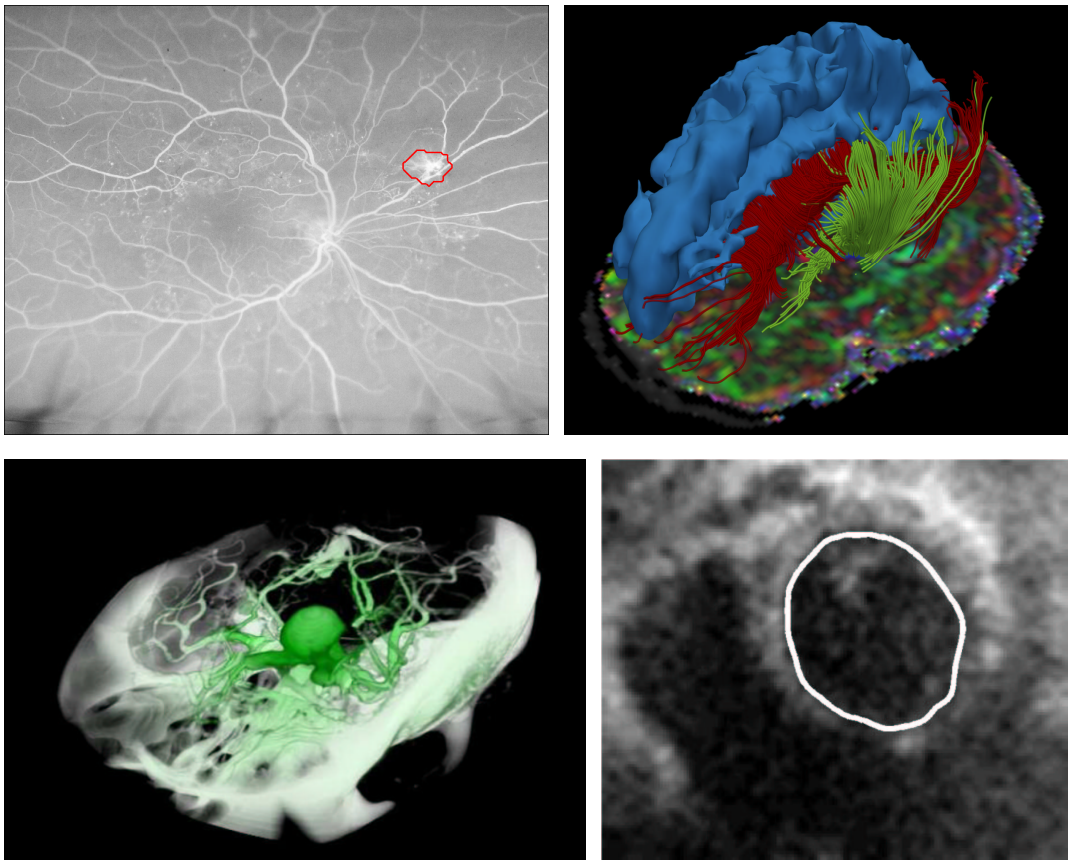
Medical Image Understanding and Analysis 2007

University of Wales Aberystwyth, 17-18th July



Reyer Zwiggelaar, Frédéric Labrosse

Editors Reyer Zwiggelaar and Frédéric Labrosse are employees of the University of Wales Aberystwyth.

ISBN 1 901725 33 2

Preface

This is the 11th conference on Medical Image Understanding and Analysis. Since 1997 MIUA has served as a forum for presentation and discussion of scientific research in the area of medical imaging and the analysis of medical images among researchers coming from a range of disciplines — engineering, physics, computer science, clinical practice and fundamental bioscience. While researchers in each of these communities have their own outlets for publishing their work, MIUA provides a focus for their common interests. MIUA's role is communication amongst researchers in the UK. However, it has had an increasing participation from overseas, and that trend continues this year with contributions from Belarus, Croatia, Finland, Germany, Pakistan, Romania and Spain. It should be noted that this 11th is the first MIUA to be held in Wales.

Since the remit of MIUA is communication amongst members of the UK community, the principal criteria are scientific soundness and interest, rather than originality. Work can, and has, been accepted for MIUA that has been published elsewhere, provided it merits presentation to the UK community.

Although there are two names at the end of this preface, organising MIUA 2007 would have been much harder work if it were not for the following persons: Simon Jackson who provided CAWS support, Edward Hadley for setting up the webpages, Jim Graham (co-chair MIUA 2006) who was able to answer all questions about running a conference, Meinir Davies who has been running all local (except for the scientific) aspects to make MIUA 2007 happen, and all PhD students who provided support during MIUA 2007. The conference would not have happened if it were not for the efforts of the reviewers who provided detailed comments by the specified deadline and as organisers we are grateful for this. The resulting set of papers are expected to provide interesting oral and poster presentations. Finally, thanks to all who submitted papers and attended MIUA 2007, we hope you enjoyed it.

Reyer Zwiggelaar, Frédéric Labrosse

Table of Contents

Podium Session 1: Fusion and Registration

Poster Session 1

Podium Session 2: Shape and Topological Modelling

Invited Lecture 1

Podium Session 3: Oncology

Poster Session 2

Podium Session 6: Segmentation

Simultaneous Segmentation of Groups of Medical Images

V Petrović *and T Cootes and A Mills and C Taylor[a]

[a]Imaging Science and Biomedical Engineering, University of Manchester, Oxford Road, Manchester, M13 9PT

Abstract. We describe a system for simultaneous segmentation of groups of images. We assume that we have multiple images showing the same anatomy from different subjects, and wish to segment all of them. Rather than segment each independently, we seek to take advantage of the information in the group. Each image is assumed to be constructed from a number of different tissue types, and the segmentation seeks to recover the proportion of each type present in each pixel. The approach is to use a groupwise registration algorithm, which simultaneously registers each image to a reference frame, whilst using a separate model for the distribution of each tissue type in each image to estimate tissue fractions. The model in the reference frame is of tissue fractions, rather than raw intensities, and evolves during the process. On completion we can project the estimate of the mean fraction into each target image, effectively segmenting it. We describe the approach in detail, and show the results of experiments on MR images of the brain, in which the estimate of the mean fraction is combined with a more data driven estimate for each individual image.

1 Introduction

Medical image segmentation and in particular segmentation of MR imagery has been the subject of considerable research in recent years [1]. Segmenting MR brain images manually for parameters such as the cortical thickness, involves huge amounts of effort and concentration on the part of experts but can still lead to highly subjective and error prone results. As numbers of acquired volumes grow, demand for computational segmentation methods and methods for segmenting groups of images simultaneously grows accordingly.

For direct segmentation of medical images into different tissues methods based directly on pixel intensity and gradient information [2], methods using hidden random Markov fields [3] and more advanced deformable structures [4,5] have been proposed. A more recent approach [6] trains a kNN classifier to optimally separate the different tissue classes while in [7] this method is extended to segment also White Matter Lesions. To train their classifiers, these last two approaches rely on a common approach in brain volume segmentation referred to as atlas propagation. Namely, a single image is manually segmented by an expert(s) and this region information is propagated into all the images that need to be segmented via a deformation field derived by registering these to the atlas image. The advantage of this approach is that region information can be efficiently propagated into a large number of images, particularly if the registration is performed in an automated manner [8]. However, this approach depends highly on the limitations of the registration algorithm which lead to inaccuracies in volume segmentations and can seriously affect any measurements derived from such results.

In our previous work [9] we proposed an automated framework for the analysis of groups of medical images containing a common deformable structure, outlined in Section 2. It builds an explicit composition model of the common structure which is capable of representing tissue fraction distributions across an anatomy. This paper extends this work by proposing a method to derive a hard segmentation of the analysed structure using information from the entire group of images (the model) and effectively delineate different tissues within each image, Section 3. We report an explicit analysis of the segmentation ability of the proposed automated analysis approach and find that good segmentation accuracy can be achieved by the framework in an automated manner, Section 4.

2 Automated Structure Analysis

An overview of the proposed approach to automated analysis of structure in groups of images is illustrated in Figure 1. A set of N images T_i, $i = 1...N$, (the training set) is assumed to contain a common structure that consists of M distinct components whose content is defined according to some composition model F and whose intensities obey some specific distribution model with parameters θ_i. Furthermore, for the entire set, a spatial correspondence with a reference (model) frame, and implicitly with each other, is assumed through a set of spatial deformation fields defined for each example in the training set, $W_i()$. Deformations are initialised as identity transformations and true correspondences $W_i()$ along with the intensity distribution model parameters θ_i and the structure composition model F are then estimated incrementally across the set in an iterative procedure as follows:

*v.petrovic@manchester.ac.uk

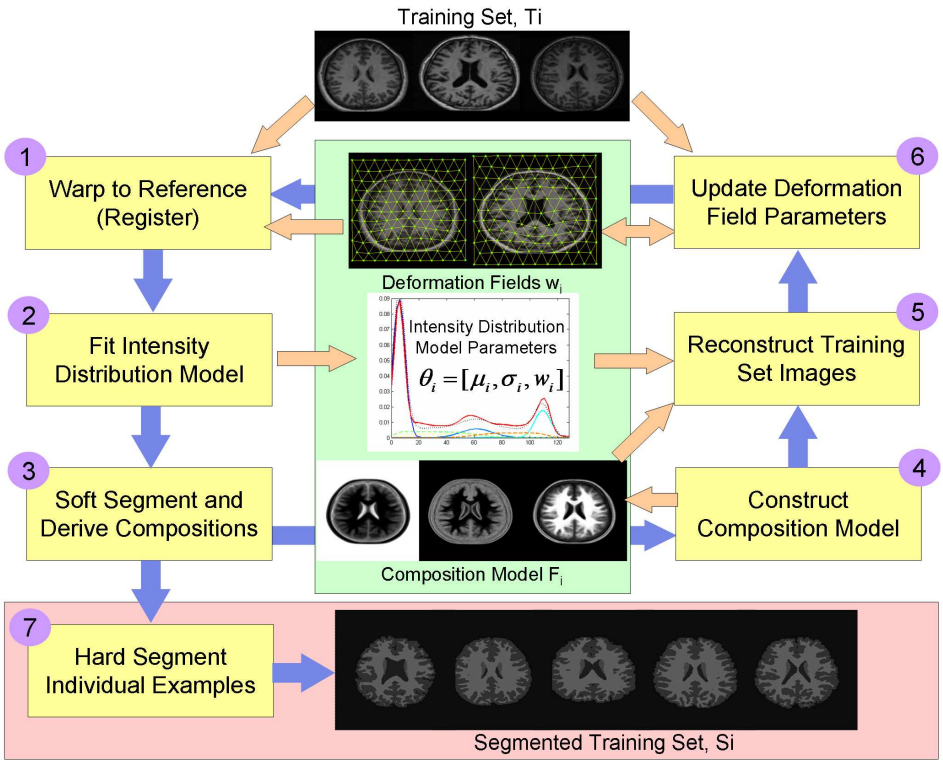


Figure 1. Outline of the proposed structure analysis algorithm: dark arrows indicate the progress of the algorithm, light arrows flow of data, the central box contains the structural information derived from the data set, and the proposed extension that performs explicit image segmentation is at the bottom

1. Warp each training image T_i into the reference frame, constructed as the mean of all W_i, using the current estimate of the deformation field. $T_i' = W_i^{-1}(T_i)$.
2. Fit the intensity distribution model (IDM) to each image and extract parameters (means, SDs and weights) for each of the M components encoded in $\theta_i = \{\mu_{ij}, \sigma_{ij}, w_{ik}\}$, as well as distributions due to mixtures of components.
3. Use the resulting distributions to estimate the most probable composition of each pixel, and encode a set of fraction images $F_i^{(j)}$, $j = 1...M$ for each training example.
4. Combine the fraction images from all examples to construct a single composition model for the common structure, $\{\hat{F}^{(1)}...\hat{F}^{(M)}\}$.
5. Synthesize a reconstruction of each training set image R_i using the current estimates of intensity distribution parameters θ_i (μ_{ij}) and the current composition model $\hat{F}$, $R_i = \sum_{j=1}^{M} \mu_{ij}\hat{F}^{(j)}$
6. Update current estimate of W_i to best register R_i onto T_i, minimising a suitable similarity measure, $D_{im}(T_i, W_i(S_i))$.

We repeat the above steps until deformation field optimisation and the composition model converge. Initial identity deformation fields will contain a considerable misalignment of the examples resulting in a fuzzy composition model. However, as the algorithm progresses and correct correspondences become established both the composition model and the model shape will converge to a true, crisp representation of the underlying structure. Finally hard segmentation is performed using the converged composition model and final IDM fit.

3 Groupwise Segmentation

3.1 Soft Individual Segmentation

A soft initial segmentation of the common structure is achieved by first fitting an intensity distribution model (IDM) and then deriving the most likely composition of tissues of each pixel in each example. In this early stage we used a relatively simple mixture of Normal distributions model, $p_j(g) = N(g : \mu_j, \sigma_j^2)$ for the MR brain images with three classes (consistent with white matter, grey matter and cerebro-spinal fluid/background tissue types), fitted using only the intensity histogram of the data in the reference frame. In general we follow [2] in assuming that each pixel in the structure is either due to one of M different tissue types or a fractional mixture of at most two different ones. Although mixtures of more than two classes are theoreticaly possible, they are significantly less numerous than two class mixtures

that themselves form a small portion of the volume. Furthermore, if we know the distributions of intensities for pure components, we can construct the distribution for a particular fractional distribution by convolution [2,9]. In this case it can be shown that the distribution for a partial volume with fraction f of tissue type i and $1-f$ of type j is given by

$$p_{ij}(g|f) = N(g : f\mu_i + (1-f)\mu_j, f\sigma_i^2 + (1-f)\sigma_j^2). \quad (1)$$

The distribution over all partial volumes containing i and j is given by

$$p_{ij}(g) = \int_{f=0}^{f=1} p_{ij}(g|f)p(f)df = \int_{f=0}^{f=1} p_{ij}(g|f)df \quad (2)$$

where we assume all values of f in the range $[0,1]$ are equally likely ($p(f) = 1$). As any pixel contains at most 2 different tissue types, we need only consider M pure tissue classes with distributions $p_k(g)$, $k = 1..M$, and $M(M-1)/2$ partial tissue classes (enumerated $p_k(g)$,$k = (M+1)..M_t = M(M+1)/2$). Thus the measured image intensity distribution, $h(g)$, can be approximated as a weighted sum

$$p(g : \theta) = \sum_{i=1}^{M_t} w_i p_i(g) \quad (3)$$

where $\theta = \{\mu_i, \sigma_i, w_k\}$ ($i = 1..M, k = 1..M_t$). We thus perform an optimisation to estimate the parameters θ which optimise $D_p(p(g:\theta), h(g))$, where $D_p(p,q)$ is a suitable measure of divergence between distributions. Having estimated the probability that a pixel with intensity g belongs to class k is given by $P_k(g) = w_k p_k(g) / \left(\sum w_k p_k(g)\right)$ (see Figure 2) that pixel can then be classified as belonging to class $k_c = \arg\max_k P_k(g)$.

However, we are actually interested in the estimate of the fraction of each pure class tissue (f_i,$i = 1..M$), in the pixel, not the probability of each class. If $k_c \leq M$ then the pixel is a pure tissue, so we define $f_{k_c} = 1$ and $f_{i \neq k_c} = 0$. If $k_c > M$ then the pixel is classified as a partial volume, containing two tissues, say of type i and type j. In this case we wish to find the most likely value of the fractions for each tissue. We define

$$\begin{aligned} f_i &= \arg\max_f p_{ij}(f|g) \\ &= \arg\max_f p_{ij}(g|f)p(f)/p(g) \\ &= \arg\max_f p_{ij}(g|f) \end{aligned} \quad (4)$$

where $p_{ij}(g|f)$ is defined above in Equation 1. We then set $f_j = 1 - f_i$ and $f_{k \neq i,j} = 0$. Using this approach we compute M images, $\{F_i^{(1)}, ...F_i^{(M)}\}$, recording the fraction of each tissue type at each pixel in the normalised version of image i (projected into the reference frame).

3.2 Hard Segmentation

A single composition (tissue fraction) model is constructed from the M fractional images from each of the N images in our set. Though more detailed statistical models (eg PCA based methods) are possible, in this preliminary study we simply compute the mean of the fraction images,

$$\{\hat{F}^{(1)}...\hat{F}^{(M)}\} = \frac{1}{N}\sum_i \{F_i^{(1)}...F_i^{(M)}\}. \quad (5)$$

As the process converges $\hat{F}$ defines how much of each of the components is present at any location within the common structure, estimated using the information from all the images. This is effectively a soft segmentation of the common structure and can be projected to each example. In practice however, optimised deformation fields will never capture all the spatial variation in the examples and the final hard segmentation of each image S_i is obtained by combining the final tissue fraction estimate for that image (using appropriate constraints for partial/true volumes) with the model using a linear combination, dependent on the parameter alpha and taking the class with the highest fraction in each pixel:

$$S_i = \arg\max_m [(1-\alpha)F_i^{(m)} + \alpha\hat{F}^{(m)}]. \quad (6)$$

4 Results

We applied the proposed method initially to a set of 37 near equivalent 2D slices of magnetic resonance (MR) images of the brain (Figure 2). The images are annotated by experts with tissue type labels including WM, GM, ventricles and caudate tissue. Our system used sum of absolute differences for both the image similarity, $D_{im}()$, and Bhattacharya distance as the divergence between intensity distributions, $D_p()$. As the brain is a true 3D structure an exact correspondence will never be possible between 2D slices, no matter how carefully chosen. However these images have

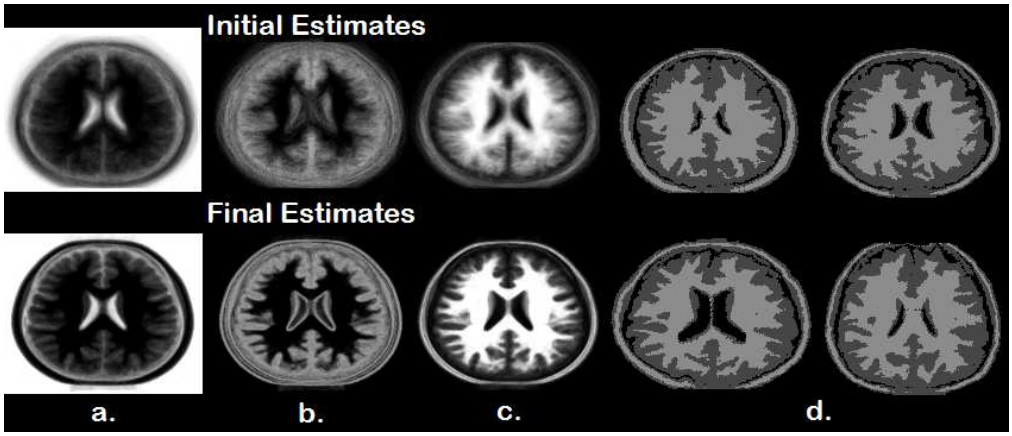


Figure 2. Results of analysis of brain images: a.-c. composition model estimates for the three tissue classes, d. a set of hard segmented input set images

TO	$min(TO)$	$max(TO)$	$median(TO)$	σ_{TO}
AAF Segmentation	-	0.7091	-	-
PW-rand Atlas Propagation	0.4340	0.6954	0.6566	0.0554
PW-opt Atlas Propagation	0.4560	0.7500	0.6691	0.0585
GW Atlas Propagation	0.4264	0.7324	0.7049	0.0673
Fluid Atlas Propagation	0.4685	0.6888	-	-
AAF Atlas Propagation	0.4904	0.7649	0.7342	0.0587

Table 1. Label overlap scores for various segmentation approaches

representative shape and intensity variations and provide a proof of concept. Figure 2 a.-c. shows tissue fraction models in the brain as they are derived by the proposed method. The model starts from a very fuzzy estimate which becomes more accurate as the images are registered and ends with all three tissues sharply delineated and in close agreement with their anatomical distribution.

Quantitative evaluation of the hard segmentation results was obtained by measuring the overlap of tissue class labels produced by the proposed automated analysis framework (AAF segmentation) for each image in the set and manually defined labels for WM, GM+caudate and ventricles using the Tanimoto overlap (TO) metric [10], Table 1. The simplistic IDM model adopted in this initial stage uses no spatial information and is unable to distinguish between the fat, skull and GM for example, so a mask of the marked labels was applied to all S_i produced by the system in order to avoid a bias in TO measurements.

The method was compared with the atlas propagation approach where label information was propagated from a segmented image (atlas) into each of the images in the set using deformation fields obtained from a number of registration algorithms (including the described analysis framework, AAF Atlas Propagation). We tested each system by using all input images as atlases and the minimum, maximum and median values of mean tissue label overlaps are given in Table 1. Registration schemes used included i) pairwise registration where each image in the set is registered to a common reference image selected either randomly PW-random or closest to the mean of the set PW-opt (a benchmark in atlas propagation), ii) groupwise registration (GW) where the set is registered to its progressively sharper intensity mean [11] (all using 24x24 point piece-wise affine deformation field and sum of absolute differences objective function) and iii) fluid flow registration (Fluid) [12], using a deformation field defined at each pixel and sum-squared difference objective function.

Despite having no manual intervention the proposed approach achieves an average label overlap of 70.9% comparable with the best of the atlas propagation systems and close to the limit of the deformation field measured on registering label images directly and producing an overlap of 71.7%. Of the atlas propagation methods it is the automated framework once again that provides best performance with maximum TO of 76.5% and median value of 73.4%. As expected, the performance of all atlas based schemes varies significantly with the choice of the atlas image. Using the automated framework avoids this problem and ensures reasonably accurate performance is achieved even with a very simplistic intensity model that uses no explicit spatial information. Finally, Figure 3 shows the results of TO for various α values used in Equation 6. There is a clear optimum around $\alpha = 0.7$ which suggests information from the entire group embodied in the model improves performance of individual segmentations (62% at $\alpha = 0$) significantly.

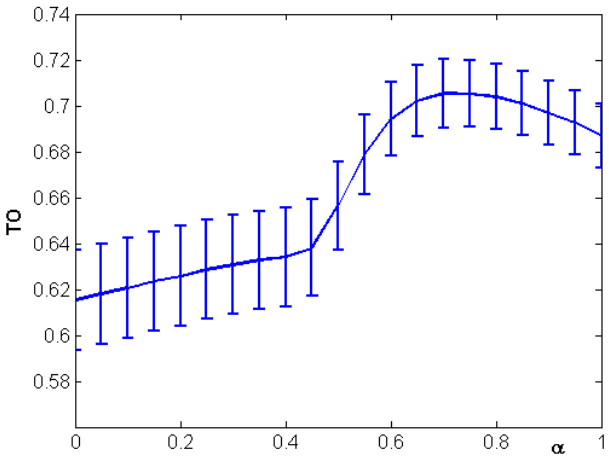


Figure 3. α vs TO for the proposed hard segmentation

5 Conclusions

We have presented an effective technique for segmentation of groups of medical images that operates as a part of a broader analysis framework in an automated manner. The method constructs an explicit model of the composition of the analysed anatomy and fuses this model with individual information extracted from each image to achieve accurate delineation of different tissue types. The results presented indicate that the method achieves accuracy of over 70% on segmentation of tissue types in MR images of the brain using information from the entire group, even when fitting relatively simplistic intensity distribution models and no spatial information. Further work on this will explore the use of more powerful distribution models as well as techniques that would perform automated model selection thereby generalising the framework to operate on any type of medical image data.

Acknowledgements

The work reported here has been funded by the MIAS IRC, EPSRC GR/N14248/01 and we would like to thank David Kennedy of the Center for Morphometric Analysis, Boston for providing the MR imagery used in this paper.

References

1. D. Pham, C. Xu & J. Prince. "Current methods in medical image segmentation." *Annu. Rev. Biomed. Eng.* **02**, pp. 315 – 337, 2000.
2. N. A. T. M. Pokric & A. Jackson. "The importance of partial voluming in multi-dimensional medical image segmentation." In *Proceedings of Information Processing in Medical Imaging (IPMI)*, volume 2208 of *Lecture Notes in Computer Science*, pp. 1293–1294. Springer, 2001.
3. D. R. C. Xiaohua, M. Brady. "Simultaneous segmentation and registration for medical image." In *Proceedings of MICCAI 2004*, number 3216 in Lecture Notes in Computer Science, pp. 663 – 670. 2004.
4. Y. Zhang, M. Brady & S. Smith. "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm." *IEEE Transactions on Medical Imaging* **20**, pp. 45 – 57, 2001.
5. A. Yezzi, L. Zollei & T. Kapur. "A variational framework for inregrating segmentation and registration through active contours." *Medical Image Analysis* **7**, pp. 171–185, 2003.
6. C. Cocosco, A. Zijdenbos & A. Evans. "A fully automatic and robust brain mri tissue classification method." *Medical Image Analysis* **7**, pp. 513 – 527, 2003.
7. R. de Boer, F. van den Lijn, H. Vrooman et al. "Automatic segmentation of brain tissue and white matter lesions in mri." In *Int. Symp. on Biomedical Imaging (ISBI)*, pp. 652–655. 2007.
8. C. J. Twining, T. F. Cootes, S. Marsland et al. "A unified information-theoretic approach to groupwise non-rigid registration and model building." In *Proceedings of Information Processing in Medical Imaging (IPMI)*, volume 3565 of *Lecture Notes in Computer Science*, pp. 1–14. Springer, 2005.
9. C. T. V. Petrović, T. Cootes & C. Taylor. "Simultaneous registration, segmentation and modelling of structure in groups of medical images." In *Int. Symp. on Biomedical Imaging (ISBI)*, pp. –. 2007.
10. W. Crum, O. Camara & D. Hill. "Generalised overlap measures for evaluation and validation in medical image analysis." *IEEE Trans. Medical Imaging* **25**, pp. 1451 – 1461, 2006.
11. T. Cootes, C. Twining, V. Petrovic et al. "Groupwise construction of appearance models using piece-wise affine deformations." In *Proceedings of the* 16^{th} *British Machine Vision Conference (BMVC)*, volume 2, pp. 879–888. 2005.
12. G. Christensen, R. Rabbitt & M. Miller. "Deformable templates using large deformation kinematics." *IEEE Trans. Medical Imaging* **5**, pp. 1435 – 1447, 1996.

Semi-automated strategy for quantifying longitudinal brain changes using Boundary Shift Integral

A Douiri[a]*, M Siddique[a], R Boyes[a], J Barnes[a], R Scahill[a], D Hill[b] and N Fox[a]

[a]DRC, UCL, 8-11 Queen Square London WC1N 3BG
[b]CMIC, UCL, Malet Place, London WC1E 6BT

Abstract.
Atrophy quantification using volumetric longitudinal magnetic resonance imaging (MRI) is increasingly being considered as an outcome in trials in dementia. Methods that require manual delineation of the brain in the baseline and the repeated scans are time consuming. In this study, we suggest and validate a novel semi-automatic methodology to measure atrophy rates using Boundary Shift Integral (BSI). This method uses only the baseline brain regions and employs affine registration followed by morphological operations to identify brain boundary delineations in the repeated scans. Using the registration transformation and the BSI, we deduce the atrophy rate on these two aligned brain regions. To show the validity of the proposed method and its consistency with the gold standard manual method, we have compared them using the statistical agreements on clinical data from three different sites, involving 150 subjects with Alzheimer's disease. Our results confirm that it is possible to measure brain atrophy rates with similar levels of success to those achieved with the manual method.

1 Introduction

Alzheimers disease (AD) is the most common cause of dementia and is the major public health problem. Early diagnosis is difficult, so too is monitoring progression for trials. Brain atrophy is known to occur early in AD on a spatial scale large enough to be detectable using serial MRI. Assessing the progression of cerebral atrophy may improve early diagnosis and enable monitoring of clinical therapies. Thus, it is important to measure atrophy precisely in order to be useful.

Quantifying cerebral atrophy can be performed using an accurate delineation of brain from non-brain tissue on a volumetric MRI scan. Multiplying the voxel volume by the number of voxels that lie within this segmented mask gives an estimate of the brain volume; the difference between volumes on serial images can approximate the tissue loss. While this method is suitable for approximating cerebral atrophy rates, most segmentation methods are prone to errors which can lead to inaccurate atrophy measurement. A widely used technique that aims to improve on that described above is the Boundary Shift Integral (BSI) [1]. This uses the segmentations described above and then registers the repeat scan to the baseline and uses the same transformation to map the repeat segmentation to the baseline. The algorithm then creates a logical exclusive OR (XOR) or 'between borders' region of the two registered segmentations. Differences in the scans that lie between the bounds of a normalized intensity window are then integrated over this XOR region, approximating the volume traversed by the brain/CSF boundary over time.

Brain segmentation is an important stage for the quantification analysis. Numerous algorithms have been suggested to perform brain extraction. Several attempts have been made to propose semi-automated segmentation methods that reduce manual segmentation labour requiring user-interaction [1]. Most are devised to work on T1-weighted MRI data, with several exceptions to other modalities [2]. Various methodologies are used to achieve a semi-automated [3] or fully automated [4] separation of brain from non-brain tissue. Atlas registration techniques for segmentation transfer brain labels to an individual subject [5]. Most perform well when defining deep structures of the brain but may fail at the cortical surface due to the large degree of inter-subject variability in sulcal and gyral morphology. Other methods have been published on whole-brain segmentation from volume MRI data [6]. For example, Sandor and Leahy [7] describe a brain labeling method, the first stage of which consists of a morphological algorithm for extraction of the cortex. The algorithm relies on automatic edge detection followed by morphological operations, but it requires a human observer to select a seed point inside the brain by visual inspection and therefore is not fully automatic. Atkins and Mackiewich [8] have presented a fully automatic method that can segment the brain in MRI datasets of different contrasts. However, the algorithm is not fully three-dimensional and does not work on sagitally acquired T1-weighted data. Lemieux *et al.* [9] presented an approach that is based on automatic foreground thresholding and morphological operations. To make the algorithm independent of scan orientation and to ensure the full connectivity of all parts of the segmented brain, all operations are performed in 3D.

*Email: douiri@drc.ion.ucl.ac.uk

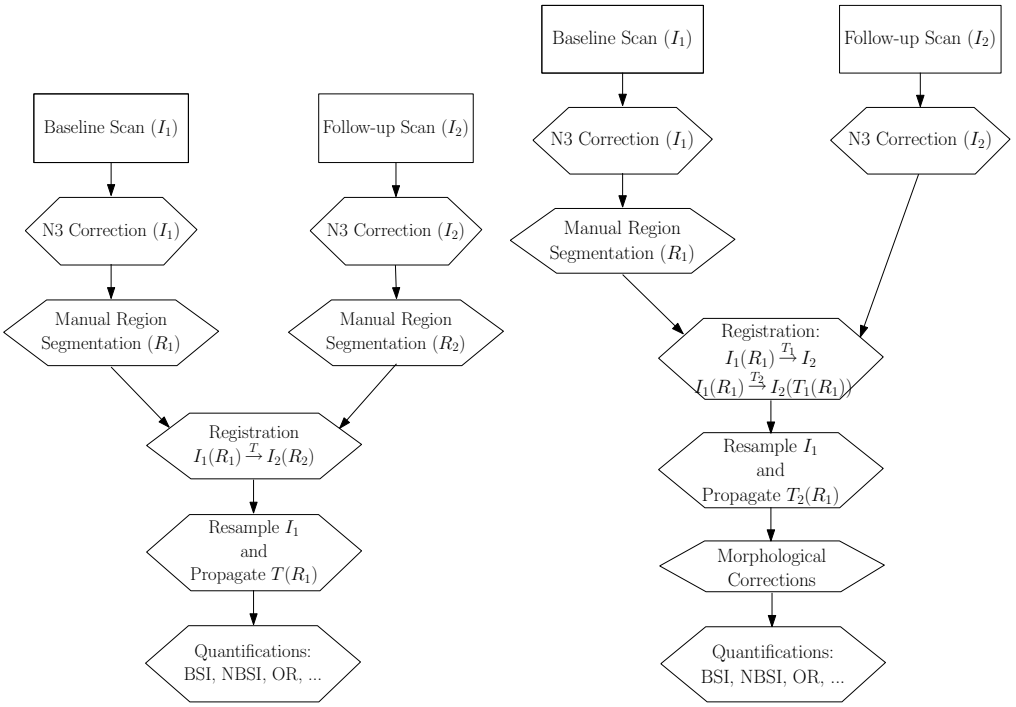


Figure 1. Flow chart of the gold standard manual method (left) and the proposed method (right).

In this study, we propose a fully automated technique for propagation of a baseline segmentation to subsequent repeat scans. This propagated segmentation is used as input to the BSI.

2 Materials and methods

2.1 Data-sets and MRI Acquisition

Three data-sets from different studies were used to validate our method. The first data-set includes 20 AD subjects taken at five time points, the second data-set includes 43 AD subjects taken at three time points and 57 AD subjects taken at two time points and the third data-set contain 30 AD subjects taken at two time points. In total we have examined 150 AD subjects from three different data-sets. All scans were T1-weighted volumetric studies acquired coronally on 1.5T scanners; multiple scanners were used however each subject's serial scans were acquired on the same scanner.

2.2 Methodology: semi-automatic strategy

Firstly, the bias-field artefact in the scans were removed using the non-parametric non-uniform intensity normalization, commonly referred to N3 method [10].

We wanted to obtain a mask on repeat scans given a baseline scan of the same subject that had already been segmented using MIDAS [1]. We proceeded by registering the repeat scan to the stripped baseline. Inverting the transformation, the baseline mask was transformed to the repeat scan using trilinear interpolation and threshold of the 50% level. We then performed morphological corrections to this propagated segmentation to take into account any cerebral change. The segmentation was eroded once to correct any excess due to the registration. We then applied two conditional dilations (within 60% to 160% of mean brain intensity) to slightly expand the voxels to fill the brain. Finally, The BSI was calculated using the baseline and propagated repeat scans and segmentations as described in [1]. As far as we know, this is the first time that morphological corrections to the registration were used to propagate the brain regions.

Figure 1 shows the flowchart of BSI calculation using the gold standard method and the proposed method.

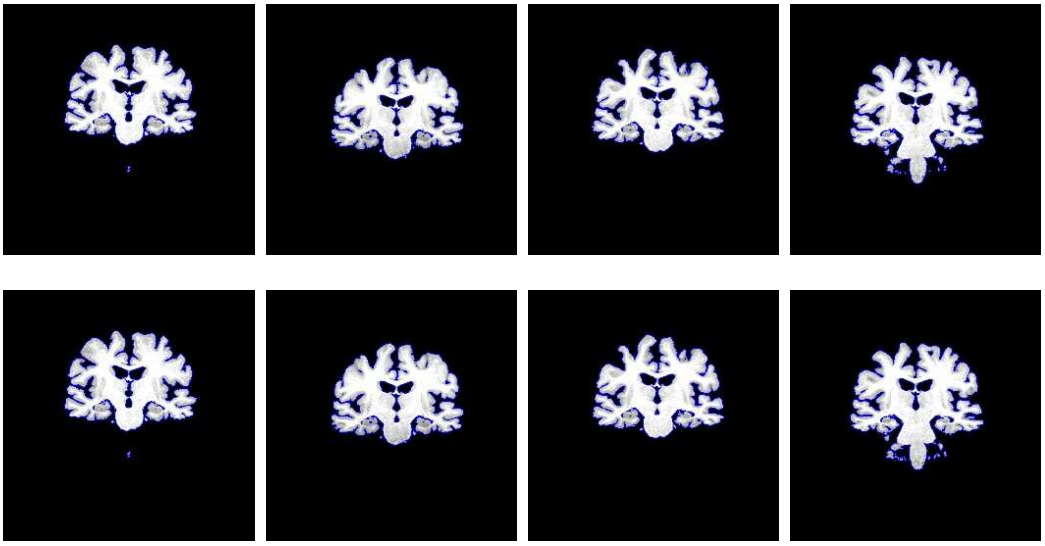


Figure 2. Brain regions extracted from gold standard (top row) and proposed method (bottom row).The columns correspond to follow-up time points from left (1st time point) to right (4th time point).

3 Experiments and results

The atrophy analysis used in this validation was based on boundary shift integral (BSI). We recall that the BSI determines the total volume through which the boundaries of a given cerebral structure have moved and, hence, the volume change, directly from voxel intensities. Figure 2 presents images of brain regions extracted using the conventional method and the proposed method. The BSI calculation was performed using repeat masks obtained manually and using the proposed automated method. Figure 3 presents the results comparing atrophy in terms of BSI values at different time points on the three data-sets used for our validation.

The statistical analysis is based on the quantification of the limits of agreement between the proposed and manual methods with the brain Boundary Shift Integral (BSI) measure using Bland and Altman agreement criteria [11]. Figure 3, middle row, shows the result of the two methods with a line of equality graph, and the bottom row the graphical presentation of agreement of the two methods with the three data-sets. Because we do not know the true volume changes we are measuring, we use the mean of the quantities of the two methods as our best estimate (Figure 3, bottom) [11]. The calculation of 95% agreement, estimated by mean differences ± 1.96 standard deviations of the differences of the two methods. Such differences, in all likelihood, follow a normal distribution. As we can observe, there is compatibility between the two methods. There are, in fact, less than 5% of BSI values outside the limits of agreements (8/241 BSIs, $\propto 3\%$). We do not want to exclude the values outside the limits. Revisiting the corresponding scans, we have found that these outside limits values were an inter-segmentor bias. But as the error was less than (5%) and we have largely a 95% of agreement, the statistical agreement is sufficient to conclude that the two methods are very similar for BSI quantifications.

4 Conclusion

The method described here only involves user input in the delineation of the brain on the baseline scan. We have shown it is consistent with the established method requiring brain delineation on each scan. There was no significant difference between the results of the two methods. The proposed method considerably reduces the manual effort for quantification of brain atrophy and inter-segmentor bias. This semi-automatic strategy can be extended to other applications where measurement of the brain volume change from serial MRI is important.

Acknowledgements

This work was undertaken at UCLH/UCL who received a proportion of funding from the Department of Healths NIHR Biomedical Research Centres funding scheme, Engineering and Physical Sciences Research Council and Medical Research Council. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health.

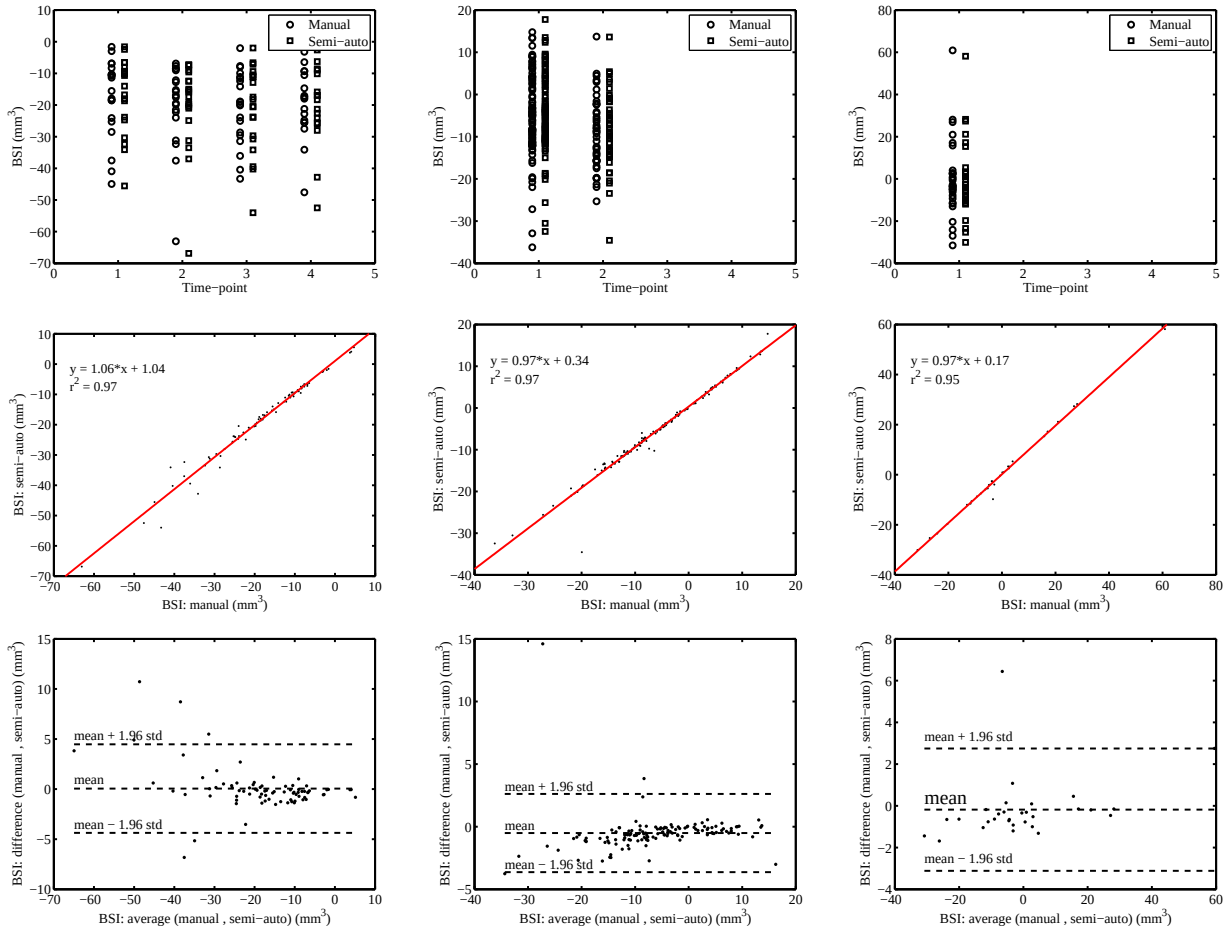


Figure 3. BSI measure with the three data-sets (top), conventional vs proposed observed by line of equality graphs (middle) and difference of the two methods vs average with 95% limits of agreements(bottom)

References

1. P. A. Freeborough, N. C. Fox & R. I. Kitney. "Interactive algorithms for the segmentation and quantitation of 3-d mri brain scans." *Computer Methods and Programs in Biomedicine* **53(1)**, pp. 15–25, 1997.
2. B. J. Bedell & P. A. Narayana. "Volumetric analysis of white matter, gray matter, and csf using fractional volume analysis." *Magnetic Resonance in Medicine* **39(6)**, pp. 961–969, 1998.
3. K. H. Hohne & W. A. Hanson. "Interactive 3d segmentation of mri and ct volumes using morphological operations." *Journal of Computer Assisted Tomography* **16(2)**, pp. 285–294, 1992.
4. S. M. Smith, Y. Y. Zhang, M. Jenkinson et al. "Accurate, robust, and automated longitudinal and cross-sectional brain change analysis." *Neuroimage* **17(1)**, pp. 479–489, 2002.
5. D. L. Collins, A. P. Zijdenbos, W. F. C. Baare et al. "Animal+insect: Improved cortical structure segmentation." In *Information Processing in Medical Imaging, Proceedings*, volume 1613 of *Lecture Notes in Computer Science*, pp. 210–223. 1999.
6. M. E. Brummer, R. M. Mersereau, R. L. Eisner et al. "Automatic detection of brain contours in mri data sets." *Ieee Transactions on Medical Imaging* **12(2)**, pp. 153–166, 1993.
7. S. Sandor & R. Leahy. "Surface-based labeling of cortical anatomy using a deformable atlas." *Ieee Transactions on Medical Imaging* **16(1)**, pp. 41–54, 1997.
8. M. S. Atkins & B. T. Mackiewich. "Fully automatic segmentation of the brain in mri." *Ieee Transactions on Medical Imaging* **17(1)**, pp. 98–107, 1998.
9. L. Lemieux, G. Hagemann, K. Krakow et al. "Fast, accurate, and reproducible automatic segmentation of the brain int(1)-weighted volume mri data." *Magnetic Resonance in Medicine* **42(1)**, pp. 127–135, 1999.
10. J. G. Sled, A. P. Zijdenbos & A. C. Evans. "A nonparametric method for automatic correction of intensity nonuniformity in mri data." *Ieee Transactions on Medical Imaging* **17(1)**, pp. 87–97, 1998.
11. J. Bland & D. Altman. "Measuring agreement in method comparison studies." *Statistical Methods in Medical Research* **8**, pp. 135–160, 1999.

Topology-Preserving Breast MRI Registration using Simultaneously Computed Local and Global Deformations

Tobias Boehler* and Heinz-Otto Peitgen

MeVis Research, Universitaetsallee 29, 28359 Bremen, Germany

Abstract.
Dynamic contrast-enhanced (DCE) breast MRI facilitates the quantitative analysis of contrast agent kinetics and its relation to suspicious breast lesions. Due to tissue variability of the breast and involuntary patient movement, subtraction images employed for this task are often perturbed by motion artifacts. Thus, the correction of patient motion is essential for a valid image analysis, removing these artifacts, while preserving lesion topology. We present an integrated combination of linear and non-linear registration into one single algorithm, eliminating the need for a prior linear registration step and reducing the computational costs. To ensure topological consistency, we incorporate a volume-preserving term, which serves as an additional force inducing a deformation that preserves consistent image regions. We demonstrate that our proposed method reliably reduces prominent motion artifacts in both synthetic images and dynamic 3-D clinical breast MRI data, and facilitates an efficient usage in a clinical environment. Furthermore, the impact of the additional volume constraint on registration quality is demonstrated on artificial and clinical image data.

1 Introduction

For the qualitative and quantitative analysis of dynamic contrast-enhanced MR mammography, motion correction of the original image time series is a fundamental prerequisite to a valid pharmacokinetic analysis. Individual image time-points are acquired during the uptake and diffusion of the injected contrast agent. Resulting intensity differences are visualized employing subtraction images of different time-points, so that present lesions are more readily detected and the computation of parameter mappings is facilitated. However, motion artifacts induced by involuntary patient motion or muscle relaxation during image acquisition affect this analysis and complicate a visual and quantitative evaluation. Correction of such image artifacts is achieved by the application of image registration algorithms to the whole dynamic image series, prior to the actual image analysis. For DCE breast MRI, appropriate registration methods effectively reduce data variability and facilitate improved detection and staging of present breast lesions [1].

Several registration algorithms using various similarity measures have been proposed for breast MRI motion correction [2]. Some approaches consider intensity variations, using enhancement models [3] or explicitely modeling local signal variations [4]. The entropy-based normalized mutual information measure has proven to be a robust error measure for breast MRI registration [5]. Normalized cross-correlation measures can be determined from discretized correlation coefficients, with local correlation being comparable to mutual information in some respects [6] and proposed for breast MRI registration [7].

For breast registration, the computed transformation must reflect the elastic deformability and tissue of the breast. For general non-linear image registration, viscous fluid and elastic registration are frequently employed [8, 9]. For breast MRI registration, a variety of algorithms has been proposed [2]. Physically motivated models can be employed in the deformation [10]. Transformations based on the deformation of B-Splines or Thin-Plate-Splines [11, 12] have been used extensively for the registration of X-ray and MR mammographies. Furthermore, optical flow based registration methods [13], or curvature-minimizing regularizations of the displacement field [14] have been proposed. Consistency and reduced divergence of the computed deformation is often forced by the incorporation of volume-preservation constraints [12, 15] into the similarity functional.

Although patient movement is restricted by the breast MR coil, involuntary motion may cause considerable artifacts. Thus, non-linear registration methods typically require a preceding linear registration, since most of these algorithms cannot establish correspondence between remote image regions. The curvature registration method is one exception, with affine linear transformations not being penalized by the regularization operator [14]. A very common combination of affine linear and B-spline transformations has been proposed by Rueckert *et al.* [11]. However, this approach does not employ an elastic deformation model, and the initial affine linear registration is computed only prior to the non-rigid scheme. To a certain degree, global motion can also be compensated for by employing general multi-resolution schemes [13], but the resulting deformation field would contain vectors of significant length. A different, potential

*T. Boehler: E-mail: boehler@mevis.de, Telephone: +49 421 218 3539

registration approach constrains the maximum rigidity of the computed deformation [16], but has not yet been evaluated on dynamic breast MRI.

Unlike common motion correction algorithms, we employ only a single iteration scheme, combining local and global deformations and using an intensity-based similarity measure. Thus, a prior linear registration is no longer required. This approach is similar to that of Haber and Modersitzki [17]. However, in [17], linear and non-linear deformations are completely decoupled. In our approach, we employ dynamic weighting functions to guide the transition from linearity to non-linearity [18]. Additionally, for both the linear and non-linear parts, we employ a combined first-order gradient descent, whereas Haber und Modersitzki use a multigrid scheme for the non-linear constrained optimization, and a second-order optimization for the linear part. The presented method has the advantage of reusing already calculated data, so that the overall computational costs (e.g. for the warping) are reduced. The employed voxel-wise similarity measure can be implemented using efficient SSE/SSE2 instructions. Furthermore, because the relationship between local and global transformation is adjusted by weighting functions, this combination potentially allows for a variety of image transformations. To ensure consistency of the deformation, we incorporate the additional volume-preserving constraint proposed by Rohlfing *et al.* [12]. The correctness and performance of the algorithm are demonstrated on both synthetic and clinical images, underlining the neccessity of volume-preservation and fast computational schemes.

2 Methods

Let $\mathbf{R}(\mathbf{x})$ be the image intensity at voxel position $\mathbf{x} = (x, y, z)^T$ of reference image $\mathbf{R}$ and $n = |\mathbf{R}|$. Furthermore, let $\mathbf{T}$ be the template image to be registered to $\mathbf{R}$. Extending the common sum of squared intensity differences (SSD) to

$$e(\mathbf{p}) = \frac{1}{2n} \sum_{i=0}^{n-1} \left[\mathbf{R}(\mathbf{x}_i) - \mathbf{T}(\mathbf{g}(\mathbf{x}_i, \mathbf{p}) + \mathbf{u}(\mathbf{x}_i))\right]^2 \tag{1}$$

yields the error measure to be minimized [17, 18]. Image deformation is determined by a global linear transformation $\mathbf{g}(\mathbf{x}, \mathbf{p}) = \mathbf{A_p}\mathbf{x} + \mathbf{t_p}$, $\mathbf{g} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$, and a local displacement $\mathbf{u}(\mathbf{x}) = (u_x, u_y, u_z)^T$, $\mathbf{u} : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ for each grid position $\mathbf{x}$. Here, $\mathbf{A_p}$ denotes a matrix defining a linear transformation determined by the parameters $\mathbf{p} = (p_1, \ldots, p_9)^T$, whereas $\mathbf{t_p} = (p_{10}, p_{11}, p_{12})^T$ gives the corresponding translation. Thus, the combined global and local transformation is defined by $\mathbf{w}(\mathbf{p}) = \mathbf{g}(\mathbf{x}, \mathbf{p}) + \mathbf{u}(\mathbf{x})$.

Minimization of equation (1) is facilitated by performing a first-order gradient descent, computing parameter updates $\Delta\mathbf{p}^{(t)}$ and $\Delta\mathbf{u}^{(t)}(\mathbf{x})$ in each time step and weighting them by a dynamic factor α adjusted in each temporal iteration t:

$$\mathbf{p}^{(t+1)} = \mathbf{p}^{(t)} + (1 - \alpha)\Delta\mathbf{p}^{(t)} \qquad \text{and} \qquad \mathbf{u}^{(t+1)}(\mathbf{x}) = \mathbf{u}^{(t)}(\mathbf{x}) + \alpha\lambda\Delta\mathbf{u}^{(t)}(\mathbf{x}). \tag{2}$$

For α, we define a strictly increasing function $\alpha = 1 - |e_t - e_{t-1}|/|e_0|$, where e_t denotes the error value computed in iteration $t = 1 \ldots n$, and e_0 the initial error. Weighting factor $\alpha \in [0, 1]$ guides the transition from linear ($\alpha = 0$) to non-linear ($\alpha = 1$) transformations, while λ additionally weights elastic deformations.

The required first-order derivatives of the similarity measure can be computed analytically. The descent direction is defined by the gradient $\nabla e(\mathbf{p})$ of equation (1) w.r.t. global parameter changes and local deformations, respectively. Since motion artifacts remote to the actual breasts are less relevant to the analysis, Dirichlet boundary conditions are employed. To increase the performance of the method, an efficient multi-resolution approach based on a Gaussian image pyramid was integrated into the implementation. For the computation of non-linear image deformations, the method is related to several approaches that perform similar computations, such as the "demons" algorithm [19], optical flow approaches [13], as well as the regularized gradient flow [20]. To reduce local variation, the deformation field is regularized using a discretized Gaussian kernel, shown to be a rough approximation to the linear elastic potential and the corresponding Navier-Lamé equation [19].

2.1 Volume Preservation

To ensure image consistency despite of high intensity variations typical to contrast-enhanced MR, we postulate a curl-free deformation field with reduced divergence. Consequently, a common additional functional term is incorporated into equation (1), penalizing degenerating transformation Jacobians [12]. Strictly spoken, since a gradient descent method is employed instead of a constrained optimization, the additional term serves as a penalizer rather than a genuine functional constraint. We define

$$v(\mathbf{p}) = \sum_{i=0}^{n-1} (|\Im_{\mathbf{w}}(\mathbf{x}_i)| - 1) \tag{3}$$

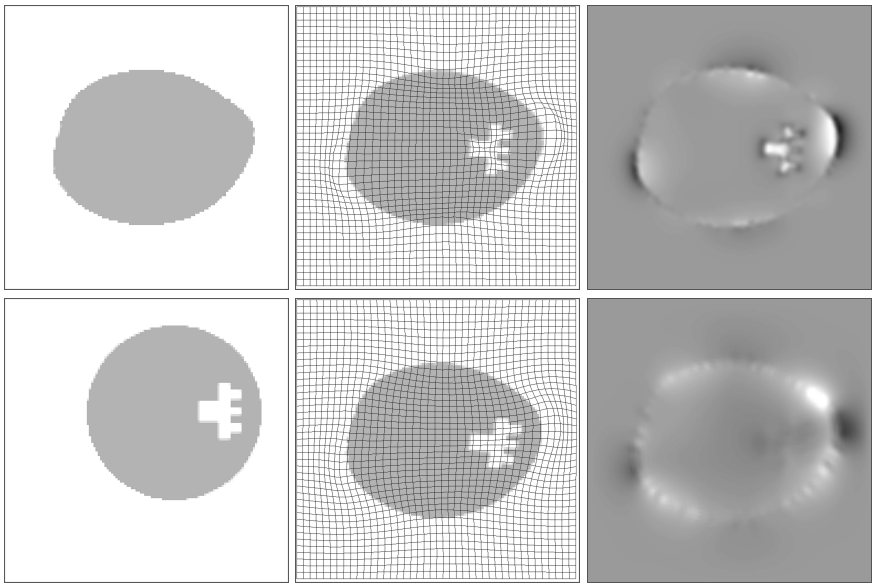


Figure 1. Synthetic 2-D images. Left column: Reference (top) and template (bottom) images. Middle: Warped template without (top) and with (bottom) volume-preservation, and corresponding deformation grids overlayed. Right: Scaled value of the determinant of the non-rigid transformation Jacobian. Employing the preservation, in particular the intensity variation has a significantly smoother deformation, and the associated determinant is close to unity.

as an appropriate term, where $|\Im_{\mathbf{w}}(\mathbf{x})|$ denotes the determinant of the Jacobian of the transformation $\mathbf{w}(\mathbf{x})$. Since global rigid transformations preserve areas of consistent image regions, we simplify (3) by only considering the Jacobian of the non-rigid local part. Incorporating this into the error function (1) yields

$$e(\mathbf{p}) = \frac{1}{2n}\sum_{i=0}^{n-1}\left[\mathbf{R}(\mathbf{x}_i) - \mathbf{T}(\mathbf{g}(\mathbf{x}_i,\mathbf{p}) + \mathbf{u}(\mathbf{x}_i))\right]^2 + \gamma\sum_{i=0}^{n-1}[|\Im_{\mathbf{u}}(\mathbf{x}_i)| - 1]. \tag{4}$$

The factor γ additionally weights the influence of the volume-preserving term. Hence, the descent update $\nabla e(\mathbf{p})$ of equation (4) is extended by adding the derivative of the new term w.r.t. local displacements, e.g. the derivative for the x-direction is given by

$$\frac{\partial}{\partial x}|\Im_{\mathbf{u}}(\mathbf{x})| = |\mathbf{I} - \Im_{\mathbf{u}}(\mathbf{x})|\; tr[(\mathbf{I} - \Im_{\mathbf{u}}(\mathbf{x}))^{-1}\frac{\partial}{\partial x}\Im_{\mathbf{u}}(\mathbf{x})]. \tag{5}$$

To ensure second-order differentiability, the update (5) is computed from the explicitely regularized deformation in each iteration.

3 Results

The algorithm was implemented in the `C++` programming language, using the Intel® SSE/SSE2 instruction set for some components (e.g. warping). Testing scenarios for the registration were computed on common personal computer hardware (single-core CPU, 2.41GHz, 2GB RAM). Assessment of the proposed algorithm was facilitated using two different types of data. First, the implemented algorithm was applied to synthetic images to verify that the method works properly. Subsequently, it was executed on MR data, demonstrating the method's ability for clinical settings. For our experiments, we set $\lambda = \gamma = 1$, and employed 3 multi-resolution levels for the registration.

3.1 Synthetic Images

The effectiveness of the volume-preserving algorithm was evaluated on 2-D images of size 128×128. To simulate contrast-agent uptake, a local intensity pattern was embedded into the template image. The test data simulates global and local non-linear distortions, created by manually applied linear and B-spline transformations. For the 2-D registration, convergence was reached in $0.76s/2.18s$ employing $70/125$ iterations on 3 levels of the Gaussian multi-resolution pyramid for the unconstrained and constrained variants, respectively. The resulting deformations were visualized by corresponding deformation grids superimposed on the warped template image (see Figure 1). Noticably, the grids are

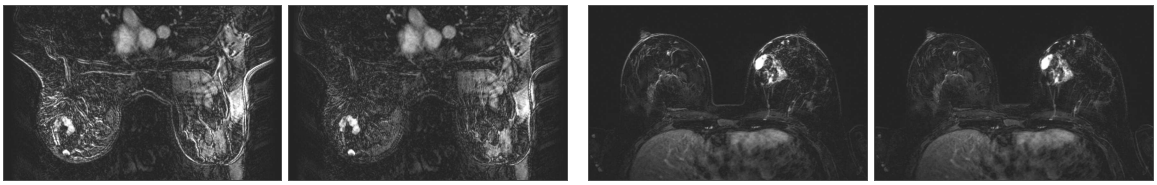


Figure 2. Subtraction images for two registered breast MR datasets (each before (l.) and after (r.) registration), showing the reduction of artifacts. All MR images courtesy of J. Wiener, Boca Raton Community Hospital, FL.

substantially different at the image region containing the intensity alteration. Whereas the unconstrained algorithm causes the grid to contract, the constrained algorithm displays a smooth deformation, resulting in a more consistent contrast-enhanced region. Visualization of the determinant of the transformation Jacobian underlines this result, since only low-frequency variations of the deformation remain. Computed quantitative measurements validate the effect of the additional volume-preservation: The contrast-enhanced region of the warped images was segmented automatically, and the resulting volume was determined. For the unconstrained version, the enhanced region retained only 68.09% of the original volume (257 voxels), whereas the constrained variant retained 90.66%. Considering interpolation artifacts perturbing the segmentation of this relative small region, an improvement of approximately 20% is a sound result.

3.2 Clinical Breast MRI

Evaluation of the method under more realistic circumstances was performed by application of the proposed technique to actual clinical data (see Fig. 2). The algorithm was applied to a set of 21 dynamic MR images with varying image dimensions. Typically, an image consisted of four time-points with dimensions of $512 \times 512 \times 80$ voxels. Voxel size was approximately $0.6mm \times 0.6mm \times 2mm$. Timings for the motion correction must be compared per time-point registration, since the datasets contained different numbers of time-points. On average, the computation for the constrained, combined linear and elastic registration of an individual time-point took $236.91s$ and required 24 iterations on 3 multi-resolution levels. Consequently, application for clinical diagnostics is feasible, since an ordinary DCE MR examination takes about $10 - 15$ minutes. On average, the SSD and normalized cross correlation similarities for the registration of the first timepoint were increased by $14.52\%(\pm 21.93)$ and $1.82\%(\pm 3.49)$, as well as $17.49\%(\pm 25.3)$ and $2.11\%(\pm 4.32)$ for the unconstrained version.

For images with present lesion shapes and motion artifacts, detailed views of the deformation were examined. As with synthetic images, the algorithm effectively reduced motion in the patient data. The inherent linear transformations reduced the amount of global image variation, such that the resulting elastic deformation contains mostly local variation. Figure 3 shows two exemplary registration results computed with the proposed method. As visible in the deformation grid, the constrained version of the algorithm penalizes local intensity changes, while the overall elastic deformation is unaffected and motion artifacts were significantly reduced. Measurement of lesion volume changes to further assess the influence of volume preservation is currently under way.

4 Conclusion

A dedicated method for motion correction of breast MR mammography was proposed, combining local and global models into one single iterative scheme. Furthermore, the method was enhanced by the addition of a volume-preserving force to conserve the topology of present lesions. The combined algorithm sufficiently compensates for larger image deformations induced by involuntary patient movement, and local elastic deformations caused by individual breast tissue motion. The presented method was evaluated on synthetic and clinical images, demonstrating successful motion correction and the impact of the additional force. Though employing only an intensity-based similarity measure, the constrained registration results are promising and demonstrate the potential of the algorithm for a clinical application.

5 Future Work

To enhance performance, we are currently extending usage of the employed Intel® SSE/SSE2 instruction set, and will additionally utilize recent GPU hardware for registration tasks. Furthermore, a modification and evaluation of the algorithm w.r.t. advanced similarity measures, such as local cross correlation or normalized mutual information, as well as image noise will be conducted. Finally, computed results need to be studied more thoroughly, for instance, by qualitative and quantitative assessment of lesion volumetry changes, as well as evaluation of lesion displacements.

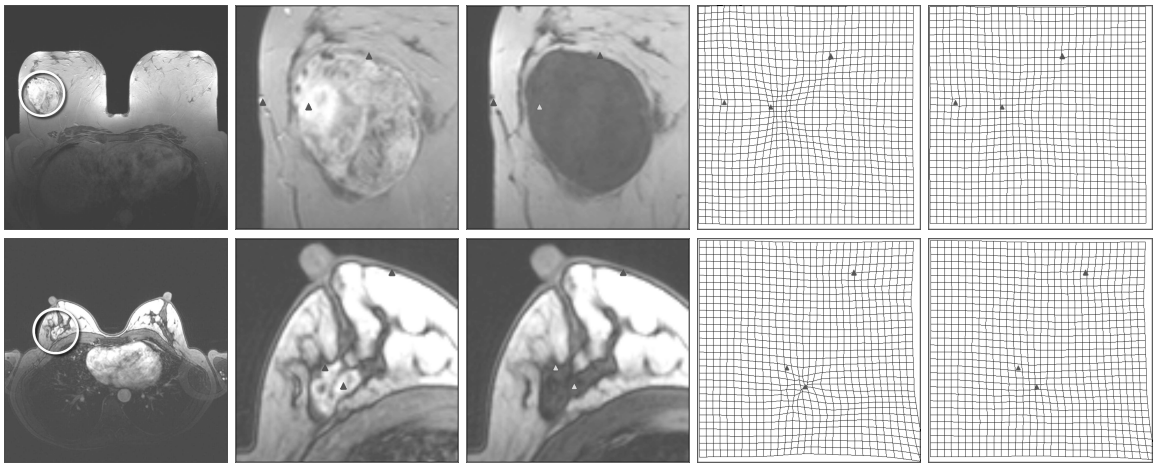


Figure 3. Application of the proposed algorithm to lesion cases (enlarged views). From left to right: Entire reference image, enlarged reference image (time-point 1), enlarged template image (time-point 0), deformation grid for the unconstrained registration, deformation grid for the volume-preserving registration. Magnitudes of the deformation fields were enlarged by the same factor for visualization purposes. Corresponding positions are indicated by markers.

References

1. P. M. Hayton, M. Brady, L. Tarassenko et al. "Analysis of Dynamic MR Breast Images Using a Model of Contrast Enhancement." *Medical Image Analysis* **1**, pp. 207–224, 1997.
2. Y. Guo, R. Sivaramakrishna, C.-C. Lu et al. "Breast Image Registration Techniques: A Survey." *Medical and Biological Engineering and Computing* **44**, pp. 15–26, 2006.
3. P. Hayton, M. Brady, S. Smith et al. "A Non-Rigid Registration Algorithm for Dynamic Breast MR Images." *Artificial Intelligence* **114**, pp. 125–156, 1999.
4. S. Periaswamy & H. Farid. "Elastic Registration in the Presence of Intensity Variations." *IEEE Transactions on Medical Imaging* **22(7)**, pp. 865–874, 2003.
5. D. Rueckert, C. Hayes, C. Studholme et al. "Non-rigid Registration of Breast MR Images Using Mutual Information." In *Proceedings of MICCAI*, pp. 1144–1152. 1998.
6. T. Netsch, P. Roesch, A. van Muiswinkel et al. "Towards Real-Time Multi-Modality 3-D Medical Image Registration." In *International Conference on Computer Vision (ICCV'01), Volume 1: Medical Imaging*, pp. 718–725. 2001.
7. L. Moy, L. Bogoni, G. Hermosillovaladez et al. "Evaluation of a 3D Motion Correction Algorithm for Breast MR." *European Radiology Suppl. MRM* **16**, pp. 105–107, 2006.
8. R. Bajcsy & S. Kovačič. "Multiresolution Elastic Matching." *Comp. Vision Graph. Im. Proc.* **46(1)**, pp. 1–21, 1989.
9. G. E. Christensen, R. D. Rabbitt & M. I. Miller. "Deformable Templates Using Large Deformation Kinematics." *IEEE Transactions on Image Processing* **5(10)**, pp. 1435–1447, 1996.
10. C. Tanner, A. Degenhard, J. A. Schnabel et al. "A Comparison of Biomechanical Breast Models: A Case Study." In *Proceedings of SPIE Medical Imaging*, pp. 1807–1818. 2002.
11. D. Rueckert, L. Sonoda, C. Hayes et al. "Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images." *IEEE Transactions on Medical Imaging* **18**, pp. 712–721, 1999.
12. T. Rohlfing, C. M. Jr., D. Bluemke et al. "Volume-Preserving Nonrigid Registration of MR Breast Images Using Free-Form Deformation with an Incompressibility Constraint." *IEEE Transactions on Medical Imaging* **22**, pp. 730–741, 2003.
13. R. Kumar, J. Asmuth, K. Hanna et al. "Application of 3D Registration for Detecting Lesions in Magnetic Resonance Breast Scans." *Proceedings of SPIE Medical Imaging* **2710**, pp. 646–656, 1996.
14. B. Fischer & J. Modersitzki. "Curvature Based Registration with Applications to MR-Mammography." In *Computational Science - ICCS 2002*. 2002.
15. E. Haber & J. Modersitzki. "Volume-Preserving Image Registration." In *Proceedings of MICCAI*, pp. 591–598. Springer, 2004.
16. S. Keeling & W. Ring. "Medical Image Registration and Interpolation by Optical Flow with Maximal Rigidity." *Journal of Mathematical Imaging and Vision* **23(1)**, pp. 47–65, 2005.
17. E. Haber & J. Modersitzki. "COFIR: Coarse and Fine Image Registration." Technical Report TR-2004-006-A, Department of Mathematics and Computer Science, Emory University, Atlanta GA 30322, June 2004.
18. T. Boehler, S. Wirtz & H.-O. Peitgen. "A Combined Algorithm for Breast MRI Motion Correction." *Proceedings of SPIE Medical Imaging* **6514**, 2007.
19. X. Pennec, P. Cachier & N. Ayache. "Understanding the Demon's Algorithm: 3D Non-Rigid Registration by Gradient Descent." In *Proceedings of MICCAI*, pp. 597–605. 1999.
20. R. Strzodka, M. Droske & M. Rumpf. "Image Registration by a Regularized Gradient Flow. A Streaming Implementation in DX9 Graphics Hardware." *Computing* **73(4)**, pp. 373–389, 2004.

Evaluation of local and global atrophy measurement techniques with simulated Alzheimer's disease images

Oscar Camara[a]*, Rachael I Scahill[b], William R Crum[a],
Julia A Schnabel[a], Gerard R Ridgway[a], Derek LG Hill[a] and Nick C Fox[b]

[a]Centre for Medical Image Computing (CMIC), Department of Medical Physics and Bioengineering,
University College London, Malet Place Engineering Building, London WCE1 6BT, UK
[b]Dementia Research Centre, Institute of Neurology,
University College London, Queen Square, London WC1N 3BG, UK

Abstract. The main goal of this work was to evaluate several well-known methods which provide global (BSI and SIENA) or local (Jacobian integration) estimates of atrophy in brain structures using Magnetic Resonance images. For that purpose, we have generated realistic simulated Alzheimer's disease images in which volume changes are modelled with a Finite Element thermoelastic model, which mimic the patterns of change obtained from a cohort of 19 real controls and 27 probable Alzheimer's disease patients. SIENA and BSI results correlate very well with gold standard data (BSI mean absolute error $< 0.29\%$; SIENA $< 0.44\%$). Jacobian integration was guided by both fluid and FFD-based registration techniques and resulting deformation fields and associated Jacobians were compared, region by region, with gold standard ones. The FFD registration technique provided more satisfactory results than the fluid one. Mean absolute error differences between volume changes given by the FFD-based technique and the gold standard were: sulcal CSF $< 2.49\%$; lateral ventricles $< 2.25\%$; brain $< 0.36\%$; hippocampi $< 1.42\%$.

1 Introduction

Atrophy measurements in some key brain structures, obtained from structural Magnetic Resonance (MR) images, can be used as biomarkers for neurodegenerative diseases in clinical trials [1], giving complementary information to cognitive tests. Computational anatomy methods [2] have been developed to analyse longitudinal and cross-sectional MR data, including quantification of atrophy.

Until recently, the evaluation of these methods has been extremely difficult since there was no reliable gold standard. Furthermore, semi-automatic or manually traced measurements of regions of interest suffer from lack of reproducibility and sensitivity, as well as being labor-intensive. Recently, Karacali et al. [3] and Camara et al. [4] proposed two different approaches [1] aiming to answer this question. The first technique is based on the generation of topology-preserving deformation fields with Jacobian determinants matching the desired volumetric changes on a specific region of interest. The main drawback of this technique is that it does not take into account the interrelation of different structures. In Camara et al. [4], we presented a technique in which atrophy in brain structures is simulated with a thermoelastic model of tissue deformation. It requires a set of segmented structures to build the input of the FEM solver, unlike Karacali's method, which does not necessarily need a segmentation step prior to simulation (the region of interest can be a sphere centered on a manually selected point in the image). On the other hand, in [4], the biomechanical readjustment of structures is modelled, using conventional physics-based techniques based on biomechanical tissue properties.

In this work, we have evaluated some well-known atrophy measurement methods with a set of realistic simulated Alzheimer's disease (AD) images, providing very valuable information for their use in a clinical context or investigate their use in drug trials. To the best of our knowledge, this is the first time that such an assessment study, using plausible simulated brain atrophy to compare several techniques, has been performed. The gold standard data has been created using an improved version of the methodology presented in [4]. Two popular methods, SIENA [5] and BSI [6], that provide global estimates of brain atrophy, and Jacobian integration guided by two different nonrigid registration methods (B-Spline FFD [7] and fluid-based [8] one), that gives local volume changes, were evaluated. A cohort of pairs of MR scans corresponding to 19 controls and 27 probable Alzheimer's disease patients was used to guide the generation of the gold standard data.

2 Atrophy simulation

The methodology for the generation of simulated images involves four main steps: generation of a reference labelled 3D mesh; its adaptation to every subject anatomy; running the FEM solver to simulate regional volumetric change on

*Corresponding author: o.camara-rey@ucl.ac.uk

[1]Both authors provide tools or gold standard data at the following addresses: `https://www.rad.upenn.edu/sbia/`; `http://www.ixi.org.uk`

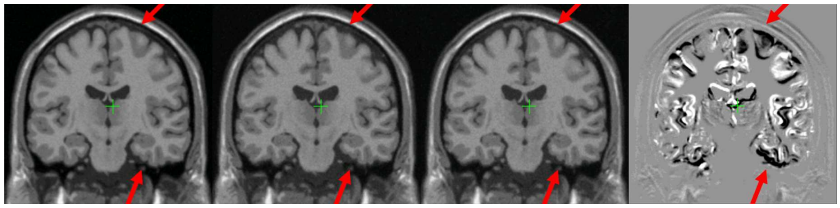


Figure 1. From left to right: original MNI Brainweb atlas; image with simulated atrophy with old; and new boundary conditions; subtraction between simulated AD images with old and new boundary conditions.

every subject-specific mesh; and application of the resulting deformation fields to the corresponding baseline MRI.

2.1 Reference volumetric mesh

A reference tetrahedral mesh (868404 elements) was built using the procedure detailed in [4]. In our work, we have employed a different set of labels, according to the information we had from the cohort of real images (both hippocampi and whole brain) and the boundary conditions imposed on the FEM model, as explained in Section 2.3. Therefore, we used labels for the whole brain (Grey Matter and White Matter together), the lateral ventricles, left and right hippocampus, the subtentorial area, extra-sulcal and sulcal cerebrospinal fluid (CSF), the last three being relevant for boundary condition purposes. The separation of the extra-ventricular CSF into two different classes was obtained by applying a brain cortex segmentation tool, available in Brainvisa [2], on the grey-level version of the MNI Brainweb [3] atlas. The outer interface of the resulting segmentation reaches the brain hull, therefore it includes sulcal CSF, which is isolated using the GM and WM labels of the MNI atlas.

2.2 Subject-specific meshes

In order to generate a cohort of simulated images with neuroanatomical variation representative of the population, the atlas-based 3D mesh has been adapted to the cohort of real images cited above, thus building a set of corresponding subject-specific meshes, which will be subsequently introduced into the FEM solver. T1-weighted volumetric MR images acquired on a 1.5 Tesla Signa unit (General Electric, Milwaukee) using a 256*256 matrix to provide 124 contiguous 1.5mm coronal slices through the head were available for every subject.

The mesh adaptation was performed by applying a Mesh-Warping (MW) technique [9], in which the transformation resulting from a fluid registration [8] between the grey-level atlas image and every subject MR scan is applied to the atlas-based mesh. The classical Jaccard overlap measure of semi-automatically obtained brain-masks [10] after fluid registration was (mean$\pm$STD) 0.855$\pm$0.024 for probable ADs and 0.886$\pm$0.018 for controls, demonstrating a good fitting of the subject-specific meshes.

2.3 Finite-Element deformation model

2.3.1 Model

The subject-specific meshes were introduced into the FEM solver, in which volume changes were simulated with a thermoelastic model of soft tissue deformation, based on the TOAST package, which is freely available [4]. After defining the elastic material properties of every region, a set of thermal coefficients, one per structure, that will result in differential regional volume changes to be applied, is computed. These were based on semi-automatically obtained segmentations [10] for both hippocampi and the whole brain between the pairs of MR scans of the cohort described above.

Homogeneous Dirichlet boundary conditions were introduced using a Payne-Irons method to suppress the displacement of the mesh nodes corresponding to the surface of the mesh and the subtentorial area. In Camara et al. [4], there was not any restriction on the labels composed of CSF, since they should fill the space left by brain atrophy. This approach could result in unrealistic shifts at the top and the bottom of the brain, as illustrated in Figure 1. We addressed this problem in a different way, by separating the extra-ventricular CSF into sulcal and extra-sulcal CSF, and applying Dirichlet boundary conditions to the latter, notably improving the realism of the simulated images (see Figure 1). This

[2] http://brainvisa.info
[3] http://www.bic.mni.mcgill.ca/brainweb/
[4] http://www.medphys.ucl.ac.uk/~martins/toast/index.html

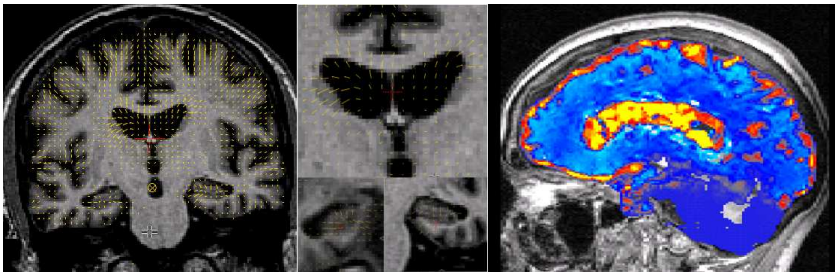


Figure 2. Gold standard deformation field (left) and zoom into the lateral ventricles (middle, top) and both hippocampi (middle, bottom). The corresponding Jacobian map (right) is also shown (yellow: volume gain; red: volume loss; blue: small volume change).

fact raises the question about the behaviour of sulcal and extra-sulcal CSF and their interrelation when atrophy occurs, which is, to the best of our knowledge, not investigated in the literature.

2.3.2 Gold standard data

The FEM solution consists of a deformation field, defined at each node of the mesh, which produces the desired volume changes. It can be directly applied to the input mesh or introduced to an interpolation step to generate a voxel-by-voxel defined deformation.

In [4], the volumetric gold standard data was obtained by integrating over every region the volume differences between corresponding elements of the original and warped meshes. Therefore, it did not take into account the interpolation step needed to apply the FEM solution to a grey-level image. In this work, the volumetric gold standard data is obtained directly from the voxel-by-voxel deformation fields after this interpolation step since they are the ones used to generate the simulated grey-level images, i.e. they are a more accurate gold standard. The determinants of the Jacobians of these dense deformation fields were computed and integrated over partial volume Regions of Interest (ROI) defined with available information about the percentages of every tissue on each mesh element. Figure 2 shows an example of a gold standard deformation field and its corresponding Jacobian map for a simulated AD subject.

3 Atrophy measurement techniques

3.1 Global techniques (BSI/SIENA)

Freeborough and Fox [6] proposed the Boundary Shift Integral (BSI) technique which computes volume change via the amount by which the boundaries of a given cerebral structure have moved. A region around these boundaries is defined with a series of morphological operations, and, subsequently, volume loss is approximated by integrating the differences in intensities between both MR scans over the defined region, normalized by image intensity means and bounded by predefined upper and lower intensity values. A rigid registration and a semi-automatic segmentation step to delineate the targeted structures are needed as pre-processing steps.

The SIENA (Structural Image Evaluation, using Normalization, of Atrophy) technique, proposed by Smith et al. [5], automatically extracts the brain from a pair of MR images, aligns the brain masked images with an affine transformation constrained by outer skull segmentations and finally estimates atrophy based on the movement of brain edges. The latter step is based on finding all brain surface edges in both MR scans, integrating the distance between the closest matching edge points, multiplied by voxel volume and normalized by the number of points found.

3.2 Local techniques (Jacobian integration)

The main purpose of Jacobian integration strategies is to capture volume changes within the deformation fields resulting from applying a high-dimensional registration technique between two pairs of MR scans. The analysis of the deformation fields is usually achieved at a voxel-wise level by computing the determinant of their Jacobian matrix, which gives a point-estimate of volumetric change. Additionally, if regions of interest are available, an integration of the Jacobians over these regions gives an estimate of their local volume change. Therefore, the accuracy will depend on the performance of the registration methodology chosen to cope with the deformations between the images. In this work, we have evaluated, for cerebral atrophy measurement purposes, two well-known and widely used registration techniques: B-Spline Free-Form Deformations [11] and a particular implementation of a fluid-based method [8].

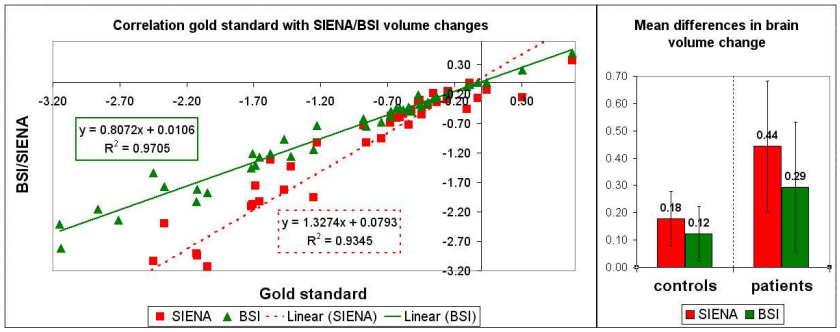


Figure 3. Correlation (left) and mean differences (right) between gold standard and SIENA/BSI results for simulated controls and patients.

The former deforms an image volume by manipulating an underlying mesh of control points. Displacements are then interpolated using the 3D cubic B-spline tensor. For this work, a multi-level FFD was used [7], by first deforming an isotropic FFD mesh of 5mm resolution, followed by a further refinement step using a 2.5mm resolution mesh. The meshes were adapted to exclude deformations outside of the reference brain masks. The similarity measure of choice was Normalized Mutual Information. For each level, gradient descent optimization for a maximum of 20 iterations, for 4 steps and initial step sizes of 2mm and 1mm, respectively, was performed.

In fluid registration the transformation is modelled as a viscous flow which warps the source image to match the target image. The driving force was derived from Intensity Cross-Correlation (ICC). The registration was run at half image-resolution, without any masking of structures, for up to 400 iterations subject to the ICC improving globally at each iteration. Further details on the implementation can be found in [8].

4 Results

4.1 Global techniques (BSI/SIENA)

Figure 3 shows a good correlation between the gold standard volume changes and SIENA/BSI results, both for controls and patients. Average absolute differences in brain volume change with respect to the gold standard (-0.39±0.74 for controls and -1.45±0.94 for patients) were small, both for controls (BSI of 0.14±0.10; SIENA of 0.18±0.14) and probable ADs (BSI of 0.29±0.24; SIENA of 0.44±0.48).

SIENA and BSI results provided similar accuracy but their behaviour was different since SIENA tended to overestimate brain volume change whereas BSI underreported atrophy consistently, as illustrated in Figure 3 (left). Another difference can also be observed between the performance of both methods in controls and probable ADs respectively (see Figure 3, right), the latter being more challenging due to a higher amount of brain volume change.

4.2 Local techniques (Jacobian integration)

Volume change values (mean±STD of percentage of the baseline structure volume), for simulated controls and patients, from the gold standard and those provided by the FFD and fluid-based techniques, are shown in Table 1.

The FFD-based method provided extremely good accuracy in whole brain and both hippocampi, with larger errors occurring in the sulcal CSF, as was expected due to the complexity of such a region.

Table 1. Volume changes (mean±STD of percentage of the baseline structure volume) provided by gold standard, FFD and fluid techniques.

Structures	Controls			Patients		
	Gold	FFD	Fluid	Gold	FFD	Fluid
Extra-sulcal CSF	1.43±3.12	0.95±2.65	0.72±1.57	4.52±2.87	3.27±2.92	1.87±1.60
Sulcal CSF	1.95±4.20	1.13±2.50	0.60±1.32	6.11±4.07	3.61±2.94	1.71±1.48
Ventricles	2.24±5.22	1.96±4.91	2.04±4.30	6.02±4.13	5.09±3.94	5.08±3.85
Left Hippocampi	-0.03±1.11	-0.18±0.56	-0.11 ±0.31	-4.00±2.97	-2.99±2.13	-1.90±1.29
Right Hippocampi	-0.39±1.17	-0.35±0.57	-0.20±0.54	-4.05±3.05	-2.80±2.08	-1.71±1.33
Subtentorial Area	0.09±0.19	0.03±0.18	0.09±0.17	0.30±0.20	0.23±0.20	0.28±0.21
Whole Brain	-0.47±1.01	-0.40±0.78	-0.32±0.61	-1.71±1.12	-1.36±1.02	-1.06±0.82

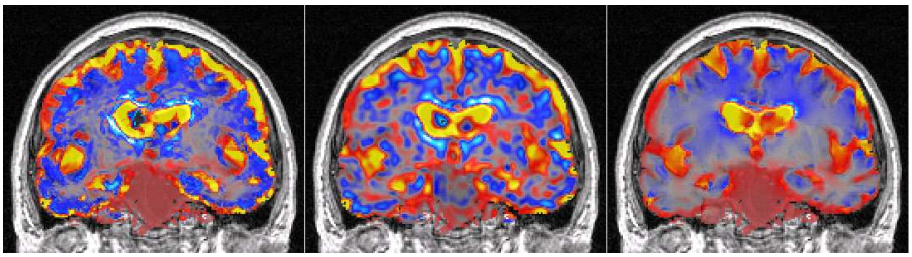


Figure 4. Jacobian images of gold standard (left) , FFD (centre) and fluid (right) deformation fields (yellow: volume gain; red: volume loss; blue: small volume change).

The fluid algorithm provided less accurate estimates of volume change for all structures involved in our experiment. An additional difference with respect to the FFD-based technique is the smoothness of the fluid-derived Jacobian map, as shown in Figure 4. This figure allows a visual comparison, for one probable AD case, between the Jacobian maps computed from the gold standard, the FFD and fluid deformation fields. The smoothness of the fluid-based Jacobian map is due to its intrinsic distribution of the volume change in homogeneous regions.

5 Conclusions

To the best of our knowledge, this work presents the first evaluation of some well-known atrophy measurement techniques with realistic simulated Alzheimers' disease images. Both global techniques analysed in this paper have been extensively employed for the neuroimaging community and results presented here confirm the appropriateness of both of them for global volume change estimation purposes. Regarding local methods, the FFD-based one performed better than the fluid registration technique, demonstrating results accurate enough for being used in studies such as clinical/drug trials. Future work will be focused on the development of a thermodiffusion model to model brain atrophy.

Acknowledgements

OC acknowledges support of the EPSRC GR/S48844/01. WRC acknowledges support of the EPSRC GR/N14248/01 and the UK Medical Research Council Grant No. D2025/31. JAS acknowledges support of the EPSRC GR/S82503/01. RIS and NCF acknowledge support from the UK Medical Research Council, G90/86 and G116/143 respectively.

References

1. N. Fox, R. Black, S. Gilman et al. "Effects of Aβ immunization (AN1792) on MRI measures of cerebral volume in Alzheimer's disease." *Neurology* **52**, pp. 1687–1689, 1999.
2. J. Ashburner, J. Csernansky, C. Davatzikos et al. "Computer-assisted imaging to assess brain structure in healthy and diseased brains." *Lancet Neurology* **2**, pp. 79–88, 2003.
3. B. Karacali & C. Davatzikos. "Simulation of tissue atrophy using a topology preserving transformation model." *IEEE Transactions on Medical Imaging* **25**, pp. 649–652, 2006.
4. O. Camara, M. Schweiger, R. Scahill et al. "Phenomenological model of diffuse global and regional atrophy using finite-element methods." *IEEE Transactions on Medical Imaging* **25**, pp. 1417–1430, 2006.
5. S. Smith, N. DeStefano, M. Jenkinson et al. "Normalised accurate measurement of longitudinal brain." *Journal of Computer Assisted Tomography* **25**, pp. 466–475, 2000.
6. P. Freeborough & N. Fox. "The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI." *IEEE Transactions on Medical Imaging* **16(5)**, pp. 623–629, 1997.
7. J. Schnabel, C. Tanner, A. Castellano-Smith et al. "Validation of Nonrigid Image Registration Using Finite-Element Methods: Application to Breast MR Images." *IEEE Transactions on Medical Imaging* **22(2)**, pp. 238–247, February 2003.
8. W. Crum, C. Tanner & D. Hawkes. "Anisotropic multi-scale fluid registration: evaluation in magnetic resonance breast imaging." *Physics in Medicine and Biology* **50**, pp. 5153–5174, 2005.
9. O. Camara, W. Crum, J. Schnabel et al. "Assessing the quality of Mesh-Warping in normal and abnormal neuroanatomy." In *Medical Image Understanding and Analysis (MIUA'05)*, pp. 79–82. 2005.
10. P. Freeborough, N. Fox & R. Kitney. "Interactive algorithms for the segmentation and quantitation of 3-D MRI brain scans." *Computer Methods and Programs in Biomedicine* **53**, pp. 15–25, 1997.
11. D. Rueckert, I. Somoda, C. Hayes et al. "Nonrigid Registration Using Free-Form Deformations: Applications to Breast MR Images." *IEEE Transactions on Medical Imaging* **18(8)**, pp. 712–721, 1999.

Exploring the limits of CT image intensity for discriminating lung tumours and the atelectasis

Vassili Kovalev[a], Maria Petrou[b] and Serguey Khoruzhik[c] *

[a]Biomedical Image Analysis Group, United Institute of Informatics Problems,
Belarus National Academy of Sciences, Surganova, 6, 220012 Minsk, Belarus
[b]Communications and Signal Processing Group, Department of Electrical and Electronic Engineering,
Imperial College, Exhibition Road, South Kensington, London SW7 2AZ, UK
[c]Department of Diagnostic Imaging, N.N. Alexandrov Research Institute of Oncology
and Medical Radiology, P.O. Lesnoj-2, 223040 Minsk, Belarus

Abstract. Computed tomography (CT) is the primary modality for imaging the lung cancer patients. However, the differentiation of lung atelectasis and malignant tumours is hardly possible due to a very similar visual appearance. In this paper, we explore the limits of the usefulness of CT image intensity information for discriminating the lung atelectasis and tumour regions. The statistical significance of intensity differences as a function of voxel sample size was assessed on CT scans of 40 lung cancer patients using unpaired *t*-test. The classification accuracy was evaluated with the help of Hierarchical Clustering, Support Vector Machines, and Random Forests methods using 44000 training and test sets of voxels sampled at random. Visualisation of sample data vectors was performed using a Multidimensional Scaling technique. The Hierarchical Clustering algorithm was found to be the best suited for segmentation purposes with its potential segmentation accuracy of 4.1 mm for 2D and 2.0 mm for 3D cases.

1 Introduction

The atelectasis term denotes the collapse of all or part of a lung due to bronchial plugging or the chest cavity being opened to atmospheric pressure. This can happen when the vacuum between the lung and chest wall is broken, allowing the lung to collapse within the chest (e.g., pneumothorax), when the lung is compressed by masses in the chest, or when an airway is blocked, leading to slow absorption of the distal air into the blood without replenishment. In this work we were dealing with the bronchial compression caused by lung cancer tumours, the most common cause of the atelectasis.

Computed tomography (CT) is the primary modality for imaging lung cancer patients. However, on CT scans the lung regions with the atelectasis and malignant tumours have quite similar attenuation values. Therefore the visual discrimination and separation of the atelectasis and tumours is hardly possible [1], [2]. Yet an accurate tumour segmentation is strongly necessary by the following two reasons. First, the correct tumour localisation, segmentation, and precise measurement of tumour diameter play a crucial role in the therapy planning and choosing suitable surgery technique. Second, if the radiation therapy is prescribed, an exact tumour border is required for precise targeting and accurate delivery of the ionising radiation exactly to the tumour but not to the surrounding tissue [3].

This work should be considered as a part of more general project the ultimate goal of which is to develop methods and software solutions for interactive discrimination and segmentation of cancerous tissue from the atelectasis. In this paper, we exploring the limits of the usefulness of CT image intensity information alone for differentiation of lung atelectasis and malignant tumours using statistical and pattern recognition methods. All the analyses were performed intra-subjectly (eg, the training and test sets of image voxels were sampled from the *same* patient). This is because we were mainly interested in performing differentiation of atelectasis and tumour regions in a gradient-like manner [4] for each particular patient but *not* in the evaluating existing inter-subject distinctions.

2 Materials

In this study we used 40 CT images of the chest of patients with lung cancer and atelectasis of a portion of the lung as diagnosed by a qualified radiologist and confirmed histologically. Thirty seven of them were males and remaining three were females. The age of patients ranged from 41 to 80 years with the mean value of 61.1 and STD of 9.2 years.

CT scanning was performed on a multi-slice Volume Zoom Siemens scanner with the standard clinical kV and mA settings during the one-breath hold. The voxel size was equal to 0.68 mm in the axial image plane with the slice thickness equal to the inter-slice distance of 7 mm. No intravenous contrast agent was administered before the collection of scan data. Typical example of original CT image slice and mean intensity values are shown in Fig. 1.

*vassili.kovalev@gmail.com (corresponding author), maria.petrou@imperial.ac.uk, skhoruzhik@nld.by

Atelectasis and tumour regions were segmented manually in each image slice of each patient by a radiologist using institutional software package called Voligator. Depending on the particular patient, the atelectasis occupied from 4 to 27 axial slices. A narrow manually-adjusted intensity window was used during the outlining atelectasis and tumour borders as it helps to slightly accentuate the intensity difference between the two region types. Generally, the tumours were located in the lung more centrally (ie, closer to the mediastinum) while the atelectasis often was located more peripherally. It was also noticed that the tumour in-plane contours tend to have more bulging shape while the atelectasis regions were rather bounded by "piecewise-straight" lines. Finally, it should be pointed out that the manual segmentation was performed intuitively, solely based on an extensive clinical experience. Nevertheless, in some cases the correct delineation of the "true" boundaries may not be guarantied due to the above-mentioned reasons.

3 Methods

The approach followed in this study was to sub-sample image voxels from two types of lung regions at random and to evaluate the significance of the intensity differences as a function of the sample size. This was done for each patient separately. The voxels were sampled *without* replacement. The training and test sets do not overlap. In order to ease the interpretability of the results, the sample sizes were selected so that they correspond to the number of voxels in square-shaped image slice patches with the side size of 3, 4, ..., 10, 15, 20, and 30 voxels that is 9, 16, ..., 100, 225, 400, and 900 sample voxels respectively. This does not means that the analysis methodology we developing is 2D-oriented, though. All statistical and pattern recognition analyses described in this work were performed using R, a language and environment for statistical computing which is available for free [5]. The atelectasis and tumour classes were compared by various ways to eliminate possible bias of one singe method. First, the significance of intensity differences between the two classes was assessed statistically using a two-tailed unpaired *t*-test with the significance level of *t*-statistics set to $p < 0.05$. The resultant *t*-values, which depend on the degree of freedom (sample size) were converted into $z-$scores to enable direct comparison of statistical significance obtained in different experiments as well as to calculate the mean significance scores over all 40 patients correctly. For each patient and each sample size the procedure consisting of random voxel sub-sampling and performing *t*-test was replicated 100 times in order to obtain reliable results.

At the second stage, the atelectasis and tumour voxel samples (ie, the vectors of voxels sorted in descending order and treated as features) were clustered using Hierarchical Clustering [6], Support Vector Machines [7], and Random Forests [8] methods. For each sample size and each patient the classifiers were trained on a training sets containing 10 atelectasis and 10 tumour samples and tested on the datasets of the same size. Training and test sets were sampled independently. There was no voxels included in both training and test sets simultaneously. The three classifiers were run on exactly the same data. Each test was replicated 100 times in order to obtain statistically representative estimates of the classification accuracy. The classification accuracy was corrected for agreement by chance using the kappa statistic as implemented in `classAgreement` function provided with R. For two classes this particularly means that the minimal accuracy value was 0 but not 50%. The corrected classification accuracy was used as a measure of the dissimilarity of two lung regions as well as the basic value for estimating possible image segmentation accuracy. The total number of performed classification tests was: 40 patients $\times$ 11 sample sizes $\times$ 3 methods $\times$ 100 replications = 132000.

Finally, the Cailliez′s version [9] of Multidimensional Scaling (MDS) method [10] was utilized for reducing the dimensionality of the space of voxel sample vectors down to two dimensions in order to produce conventional voxel sample scatterplots. Note that the multidimensional scaling provides an approximate solution, which is mostly suitable for visual examination of overlapping object classes. The total computational time (including all overheads) for 132000 classification tests and 40 $\times$ 11 = 440 runs of the multidimensional scaling algorithm was 109 minutes on a personal computer equipped with 1.86 GHz Intel Core 2 Duo processor.

4 Results

Results of statistical assessment of the significance of intensity differences between the atelectasis and tumour regions of lung CT scans of 40 patients are reported in Fig. 2. As it can be seen from the figure, the fraction of significantly different voxel samples and the mean significance scores varied considerably depending on the patient. For instance, for one patient the percentage of significantly different samples exceeds notable 60% already on 9 voxels and achieves 100% with the sample size as little as 36 voxels (see the left panel of Fig. 2) while in other it starts close to zero with 9 voxels and finishes at about 10% only. Similarly, for some patients the mean $z-$score achieves the significance threshold $z > 1.96$ which is equivalent to $p < 0.05$ on the sample size of 9–25 voxels (see the right panel of Fig. 2)

while for others these values remain insignificantly low even on reasonably large samples consisting of 400-900 voxels.

On the contrary, the voxel sample classification results demonstrate much more consistent behaviour (see Fig. 3). As it can be revealed from the figure, a very useful property of the classification approach for separating the atelectasis and tumour regions is that the results are converged to 90–100% of the classification accuracy for relatively large samples *in each* patient. As for the comparative efficiency of the three classification methods, it is easy to see from Fig. 3 that the Hierarchical Clustering algorithm outperforms both SVM and Random Forests for each voxel sample size. Moreover, in case of Hierarchical Clustering, the classification accuracy corrected for the agreement by chance starts from the value above 50% almost for each patient and achieves 90% on the sample size of 225 voxels for all 40 patients except for 2 outliers. The mean and standard deviation values of the classification accuracy computed over 40 patients (see the bottom right plot of Fig. 3) make the superiority of Hierarchical Clustering method evident and renders other two as almost identical in the voxel sample classification task. Considering that the one possible segmentation technique could be based on a direct voxel sample classification using sliding window of suitable size, the mean accuracy threshold should be set to a reasonably high value, say 95%. If so, the minimal sample size should be set to approximately 100-200 voxels. This corresponds to the window size of about 12×12 voxels (ie, the half window size is 4.1 mm) for 2D and less than $6\times6\times6$ voxels (2.0 mm) for 3D case.

Fig. 4 shows an example of separation of voxel samples in multidimensional space. It should be stressed that this figure illustrates one single case out of 440 and thus may not be as conclusive as the results represented in Fig.3.

5 Conclusions

In this work we have documented results of patient-wise assessment of CT image intensity differences between the lung atelectasis and malignant tumour regions. Our results suggest that it is unlikely that the use of statistical significance scores for separating lung atelectasis and tumour regions would produce good quality discrimination for all the patients. However, the recent clustering algorithms demonstrate some encouraging classification accuracy on the CT intensity samples consisting of few hundred voxels. The Hierarchical Clustering method was found to be better suited for highlighting the border between the two types of regions in a gradient-like way [4] comparing to SVM and Random Forests classifiers. This is in agreement with other studies where classes overlap in feature space substantially (eg, [11], [4]). The voxel sample classification accuracy potentially allows to reliably discriminate atelectasis and tumour regions using relatively small sliding window of 12×12 voxels (ie, the half window size is 4.1 mm) in 2D and no more than $6\times6\times6$ voxels (2.0 mm) in 3D case.

Acknowledgements

This work was supported by the International Joint Project grant 2006/R1 from British Royal Society .

References

1. J. Mira, G. Fullerton, J. Ezekiel et al. “Evaluation of computed tomography numbers for treatment planning of lung cancer.” *Int J Radiat Oncol Biol Phys.* **8(9)**, pp. 1625–1628, 1982.
2. J.-M. Kuhnigk, H. Hahn, M. Hindennach et al. “Lung lobe segmentation by anatomy-guided 3D watershed transform.” In M. Sonka & J. Fitzpatrick (editors), *Medical Imaging 2003*, volume 5032, pp. 1482–1490. SPIE, USA, May 2003.
3. P. Bowden & F. Richard. “Measurement of lung tumor volumes using three-dimensional computer planning software.” *Int J Radiat Oncol Biol Phys.* **53(3)**, pp. 566–573, 2002.
4. V. Kovalev & M. Petrou. “The classification gradient.” In *Int Conf on Pattern Recognition (ICPR’06)*, volume 3, pp. 830–833. IEEE Comp Society, Hong Kong, 20-24 August 2006.
5. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.
6. L. Breiman, J. Friedman, R. Olshen et al. *Classification and Regression Trees*. Chapman & Hall, New York, 1984. 384 p.
7. C. Cortes & V. Vapnik. “Support-vector network.” *Machine Learning* **20**, pp. 1–25, 1995.
8. L. Breiman. “Random forests.” *Machine Learning* **45(1)**, pp. 5–32, 2001.
9. F. Cailliez. “The analytical solution of the additive constant problem.” *Psychometrika* **48**, pp. 343–349, 1983.
10. T. F. Cox & M. A. A. Cox. *Multidimensional Scaling*. Monographs on Statistics and Applied Probability 88. Chapman & Hall/CRC Press, London, second edition, 1994. 309 p.
11. V. Kovalev, N. Harder, B. Neumann et al. “Feature selection for evaluating fluorescence microscopy images in genome-wide cell screens.” In *Computer Vision and Pattern Recognition (CVPR’06)*, pp. 276–283. IEEE Comp Society, New York, 17-22 June 2006.

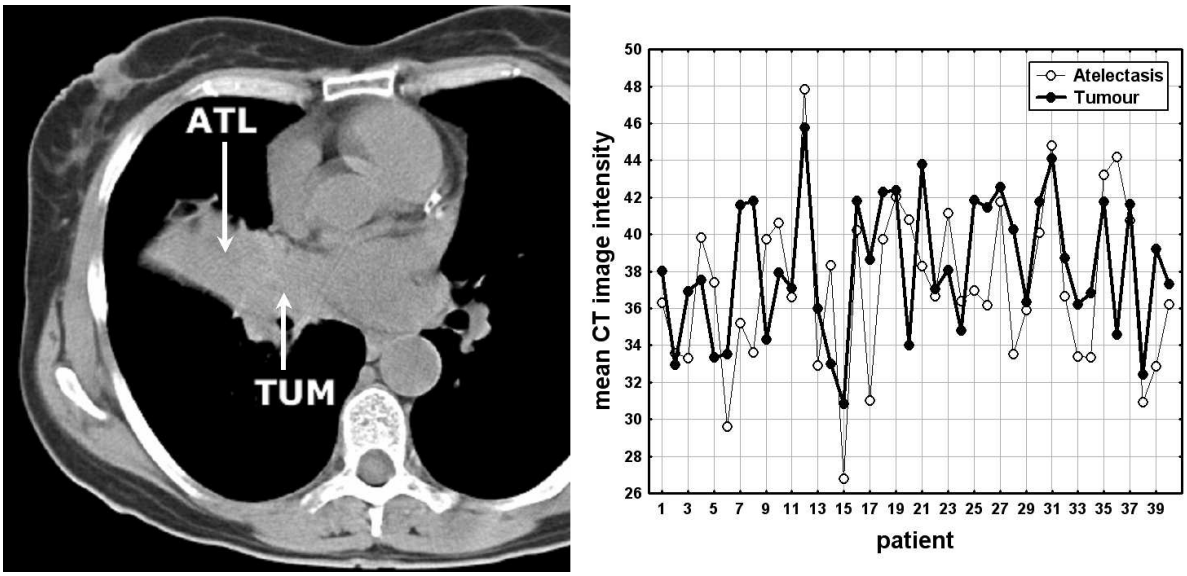


Figure 1. Example slice of typical lung CT image with atelectasis (ATL) and tumour (TUM) regions (left panel) and plots of mean intensity values of the atelectasis and tumour regions for 40 patients (right panel).

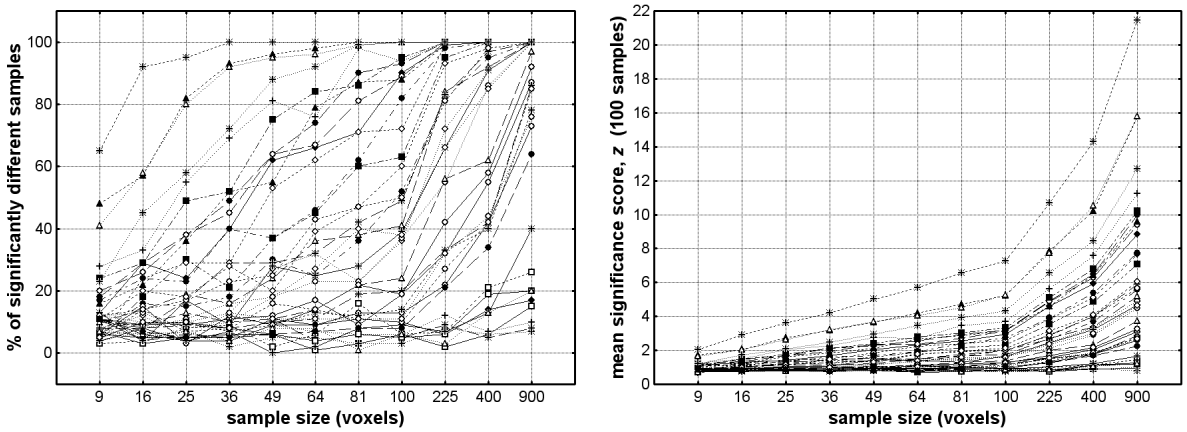


Figure 2. Significance of the intensity differences of lung atelectasis and tumour voxel samples for 40 patients (curves) as a function of the voxel sample size. Left panel: percentage of voxel samples for which the intensity difference is statistically significant at $p < 0.05$. Right panel: the mean value of significance score z. In both occasions image voxels were sampled from atelectasis and tumour regions at random and each measurement is replicated 100 times.

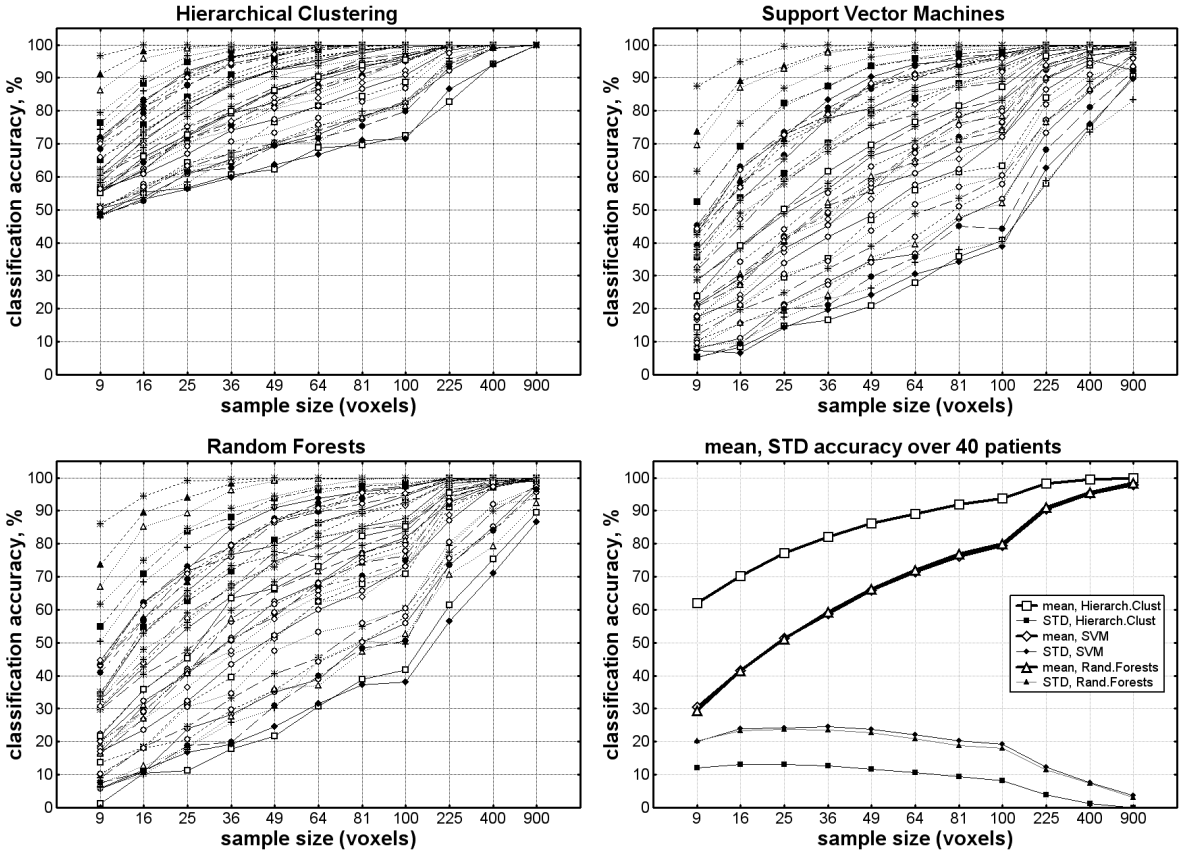


Figure 3. Dependence of the classification accuracy on sample size of lung atelectasis and tumour voxels for 40 patients (curves) when using Hierarchical Clustering (top left plot), Support Vector Machines (top right plot), and Random Forests (bottom left plot) clustering methods. Each test was replicated 100 times for the reliability of results. The mean and standard deviation accuracy computed over 40 patients is given on the bottom right panel.

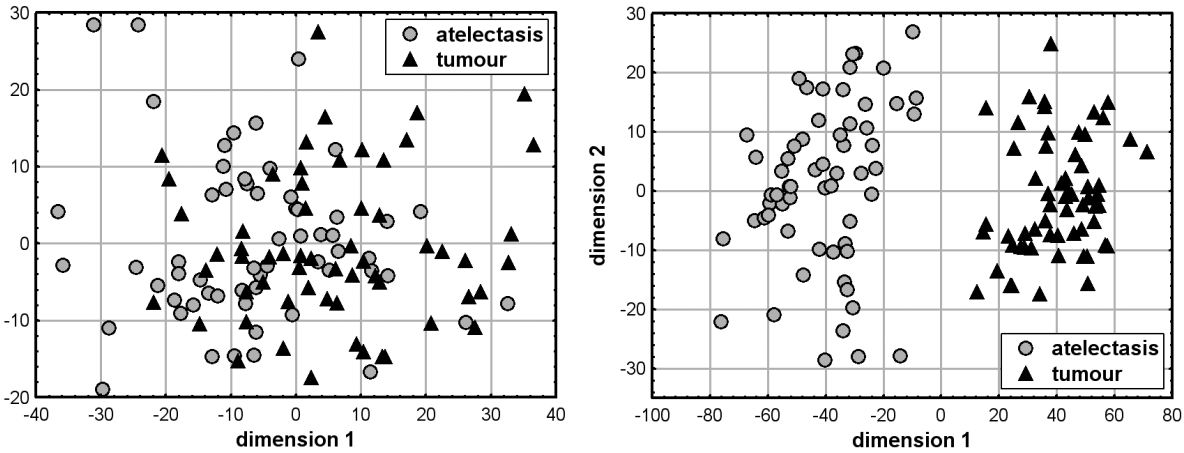


Figure 4. Example of separation of 120 image voxel vectors (dots) sampled from the lung atelectasis and tumour regions of a patient with the sample size of 9 (left panel) and 400 (right panel) voxels. In both cases the sample points spanning the N-dimensional space (N is equal to 9 and 400 respectively) are projected to a plane with two conditional dimensions using the Multidimensional Scaling method. The scatterplots represent two replications taken at random.

Improving Accuracy and Efficiency of Registration by Mutual Information using Sturges' Histogram Rule

Philip A. Legg[a*], Paul L. Rosin[a†], David Marshall[a‡] and James E. Morgan[b§]

[a]School of Computer Science, Cardiff University. [b]School of Optometry, Cardiff University.

Abstract. Mutual Information is a common technique for image registration in the medical domain, in particular where images of different modalities are to be registered. In this paper, we wish to demonstrate the benefits of applying a common method known in statistics as Sturges' Rule for selecting histogram bin size when computing Entropy as a part of the existing Mutual Information algorithm. Although Sturges' Rule is well known in the field of statistics it has received little attention in the Computer Vision community. By augmenting Mutual Information with Sturges' Rule, we show that this offers an improvement to both the runtime of the algorithm and also the accuracy of the registration. Our results are demonstrated on images of the eye, in particular, Fundus images and SLO (Scanning Laser Ophthalmoscopy) images.

1 Introduction

Mutual Information is a widely used measure for performing image registration in the medical imaging domain, due to its ability to register images of different modalities [1]. Mutual Information relies greatly on a measure known as entropy, which can be thought of as the amount of information an event provides when it occurs [2]. Mutual Information is defined as $I(A, B) = H(A) + H(B) - H(A, B)$, where $H(A)$ is the entropy of the template image, $H(B)$ is the entropy of the section of the reference image at which the template image is currently located and $H(A, B)$ is the joint entropy of the two. We wish to find the registration transformation that maximises $I(A, B)$.

Computation of entropy is based on the probability of the values within the data set, defined as $\sum_{i=0}^{n} -p(i) \log_2 p(i)$ where $p(i)$ is the probability of intensity i occuring within the data set n. One possible approach to finding this probability distribution is by using a histogram. There are alternative methods that exist, such as using a Parzen Window [3], B-splines or k-Nearest Neighbours [4], along with more recent techniques such as that described in [5], although using a histogram tends to be the most popular choice due to its simplicity and computational efficiency.

In the Computer Vision literature relating to Mutual Information, very little mention has been made regarding the selection of histogram bin size. Most papers use a fixed number of bins either equal to or less than the possible data range [6], but this means that no consideration is given to the data being organised. In [7], they state that no method exists for predicting the exact number of bins to use for a histogram, which is clearly not the case, as we shall demonstrate in our work. Even in comprehensive reviews of Mutual Information such as [2], there is no mention of how bin size should be selected and how this could affect the performance of Mutual Information based registration. Nevertheless, bin size is a *crucial* parameter. Excessive quantisation caused by too large a bin size will result in important information being lost. On the other hand, too small a bin size may result in many bins becoming sparsely populated, and consequently making the probability density estimates unreliable.

Sturges' Rule is a well known method used in statistics for histogram binning [8]. Sturges' Rule is one possible technique for determining the size of each group that the data should be separated into, to try give the optimum group size. Another common method in the statistics literature for estimating bin size is Scott's Rule [9]. Scott's Rule is thought to be an improvement over Sturges' Rule as Sturges' Rule can over-smooth a histogram which may be problematic in some applications [10]. We shall consider Scott's Rule within our study and compare how this performs alongside with Sturges' Rule and using 256 bins.

Applying Sturges' Rule can provide a two fold benefit. Reducing the number of bins will reduce computational cost and so improve runtime. It can also be seen as cleaning up an image that may contain irrelevant and distracting detail, consequently improving the accuracy and robustness of the registration process. Sturges' Rule is derived on the assumption of normally distributed data, through the application of a binomial approximation. It is not intended for

*E-mail: P.A.Legg@cs.cardiff.ac.uk
†E-mail: Paul.Rosin@cs.cardiff.ac.uk
‡E-mail: Dave.Marshall@cs.cardiff.ac.uk
§E-mail: morganje3@cardiff.ac.uk

use on data that, for example, features several, well-separated peaks. However, our area of research is targeted towards retinal images, in particular, Fundus images and SLO images [11], where such features of the histogram are not present. In this paper, we shall demonstrate the effects that Sturges' Rule has on these images, and evaluate the performance that this technique offers in Mutual Information based registration in comparison to using 256 histogram bins or using Scott's bin size rule.

2 Method

We will begin by stating the formulae for Sturges' Rule and Scott's Rule. We will then demonstrate the effects that Sturges' Rule can have on an image and its associated histogram. A comparison of number of bins given by Sturges' Rule and Scott's Rule can be found in Section 3.1. We will also discuss Histogram Equalization as a method for bin size selection.

2.1 Sturges' Rule Definition

Sturges' Rule is defined as $w = \frac{r}{1+log_2(n)}$ where r is the range of values within the data set, and n is the number of elements within the data set [8]. The result will give the ideal bin width, w, to be used for the histogram (i.e. the range for each group of values). To find the number of bins for an image, we simply use r/w.

2.2 Scott's Rule Definition

Scott's Rule is defined as $w = 3.49\sigma n^{-1/3}$ where σ is the standard deviation of the data set, n is the number of elements within the data set [9], and w is the bin width.

2.3 Sturges' Rule in Practice

In Figure 1, we present a typical 8 bit Fundus photograph of the eye, along with its associated histogram. The effect that Sturges' Rule has had on the histogram is clear to see in Figure 2. Although the histogram contains just 16 bins, the requantized image has retained all the significant detail from the original in Figure 1.

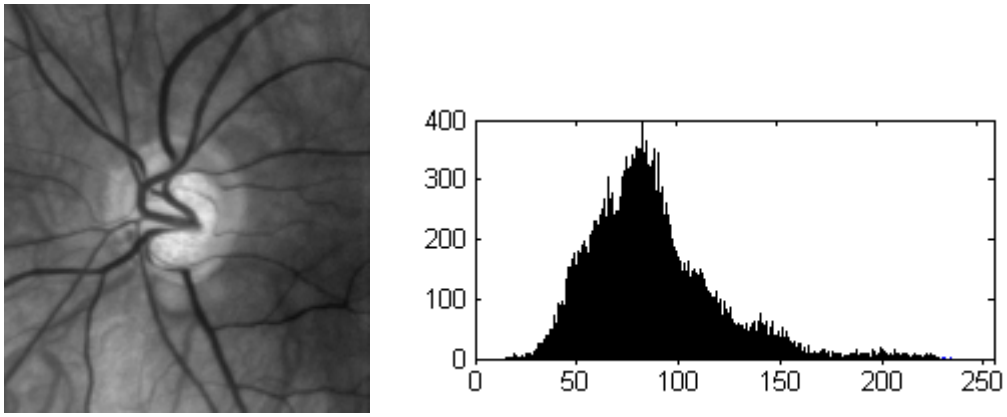


Figure 1. Fundus image with associated histogram

It can be noticed that the new histogram does not capture the exact shape of the original histogram and distinct peaks have been lost due to binning which may lead to important information being lost. Comparing the actual images shown in Figures 1 and 2, we can see the difference that Sturges' Rule has made. Although the images are still very similar in what they represent, areas that were originally smooth have now become solid areas, with more noticable steps between intensity changes. This may be useful in some situations where, like here, the background intensities are unclear and highly varied which may have an adverse effect when it comes to processing the image.

Applying Scott's rule, the Fundus image is represented using 145 bins. Although this has reduced the number of bins slightly, there would be much less of a difference to the original image compared with that of the Sturges' Rule image. The distinct intensity changes seen in the background of the Sturges' Rule image would not be evident on the Scott's Rule image.

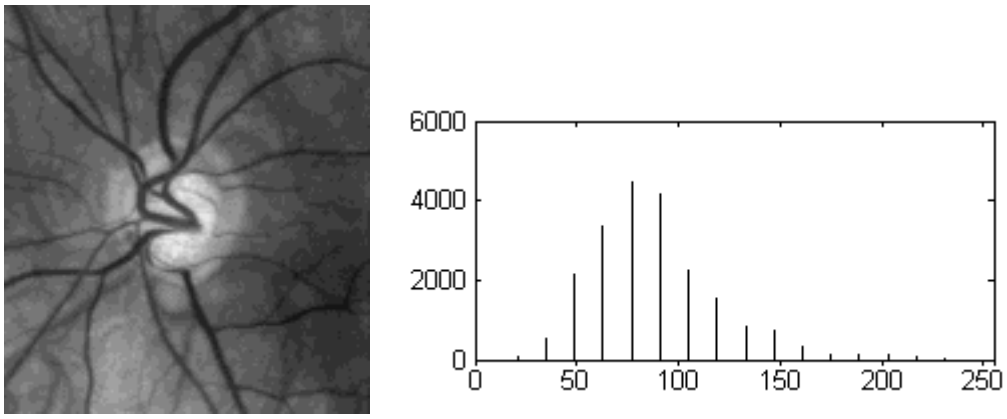


Figure 2. Fundus image with associated histogram, after applying Sturges' Rule

2.4 Histogram Equalization

One other method that is related to bin size estimation is histogram equalization [12]. This normally applies a monotonic remapping of intensity values to make the intensity histogram approximately flat, with the aim of improving visibility of features. As a consequence, adjacent sparsely populated bins are merged, thereby improving probability estimates and reducing the number of bins. The effects of histogram equalization in essence allow for variable bin width within a histogram. We shall investigate the effects of histogram eqaulization on registration in our paper.

3 Testing

Mutual Information can be used to effectively perform image registration on two images of different modalities, by transforming the template image on to the reference image, such that it maximises Mutual Information. Our aim is to successfully register our image data correctly, with a high rate of accuracy that is also time efficient.

For our work, we have two images captured from the eye, a Fundus photograph which will be our reference image, and an SLO (Scanning Laser Ophthalmoscopy) image which is our template image. It can be seen in Figure 3 that the images are of the same source but have different appearances due to the information captured by each camera. For the benefit of this paper, the images have been rotated and scaled appropriately beforehand, as we will just report on the effect of Sturges' Rule on the accuracy and runtime of estimating translation. Similar effects were found for the remaining transformation parameters.

To find where the correct registration occurs, we have adopted two approaches, exhaustive search and hill climbing. Exhaustive search will attempt to match the template image to every possible position on the reference image. This method can cause our search to check areas where we do not wish our images to match at, for instance, in our data we know that the registration will occur around the centre of the reference so searching the edges is not necessary. However, it provides a fair result for the image as a whole that does not rely on having any previous knowledge of the data being registered. Hill climbing is a more common search technique that will start at a given point (in our case, this point will be the centre of the reference image) and try to improve on the existing result by testing local neighbouring positions. This provides a much faster search method, although may not give the true result if caught in a local maximum which differs from the global maximum.

For our testing, we use twenty-six 8 bit greyscale image pairs which we shall perform registration on, using the traditional 256 bin representation, Sturges' Rule, Scott's Rule and Histogram Equalization. The number of bins used for computing the joint entropy is found by the number of bins used in template image $\times$ the number of bins used in reference image segment, since this is computed by means of a 2-D histogram. The dimensions for the images are 153×137 for the reference images, and 50×51 for the template images. In each case, we shall attempt to register the images using the exhaustive search and hill climbing search methods.

3.1 Results

Following our testing of Mutual Information using our four approaches, we wish to quantify the alignment with respect to the ground truth manually determined by a clinician. To do this, we have calculated the error of the translation. The

tables below shows the average translation error, the average number of bins used (represented by (A,B) where A is the number of bins used for the reference image and B is the number of bins used for the template image), the average runtime and the number of successful registrations (where a match is found within a 2 pixel radius of the ground truth).

	$SturgesRule$	$256Bins$	$ScottsRule$	$HistogramEqualization$
Average Translation Error	17.15	24.92	18.96	20.46
Average Number of Bins	(14,14)	(256,256)	(134,79)	(60,64)
Average Runtime (secs)	7.38	19.74	13.21	10.54
Successful Registrations	13	2	10	8

Table 1. Results for performing Mutual Information using Exhaustive Search

	$SturgesRule$	$256Bins$	$ScottsRule$	$HistogramEqualization$
Average Translation Error	10.73	14.15	11.34	12.53
Average Number of Bins	(14,14)	(256,256)	(112,78)	(59,64)
Average Runtime (secs)	0.119	0.239	0.172	0.137
Successful Registrations	11	5	8	7

Table 2. Results for performing Mutual Information using Hill Climbing

As can be seen in Table 1, Sturges' Rule gives the lowest translation errors, compared to the other methods used, and manages to successfully register 13 of the 26 test images. Histogram equalization provides a fair improvement over 256 bins by successfully registering 8 of the 26 images. We note that Scott's Rule can give a large difference between the number of bins used in the reference image and the template image. In some cases, Scott's Rule would give an estimated bin size greater than 256. This is not necessarily a problem, but the results of registration are worse than Sturges' Rule. Scott's Rule manages to register 10 of the images and 256 bins only manages to register 2. It is surprising that using 256 bins gives such a low success rate, although in comparison with the other methods presented here, it is likely due to having very sparsely populated bins, especially so in the joint histogram. The joint histogram would have 65536 possible bins, where as there are at most only 2550 points to be binned (the true value would most likely be less as this assumes each pixel and its corresponding pixel in the other image be a unique combination).

Comparing the runtime of performing exhaustive search, using 256 bins takes 19.74 seconds compared to Sturges' Rule which takes 7.38 seconds. As expected, this reduction is due to the fewer number of bins that the entropy formula has to be calculated for. It can be seen that Scott's Rule and Histogram Equalization also offer an improvement to runtime compared to 256 bins.

Table 2 shows that when using Hill Climbing, Sturges' Rule manages to successfully register 11 of the 26 images. Although this is less than the number of registrations by exhaustive search, it can be seen that the translation error has reduced. This anomaly shows that although Hill Climbing may not register the images exactly, the results are much closer to the ground truths than when exhaustive search fails to register. This is because exhaustive search can potentially place the template far away from the desired position due to the nature of the search technique, which is a common occurrence on the failed registrations. Scott's Rule and Histogram Equalization experience a similar situation to Sturges' Rule when using Hill Climbing, which suggests that the algorithms are caught by local maxima within the search space. In comparison, when using 256 bins, we notice an increase in successful registrations. This is due to being caught by local maxima which, in these cases, has been the correct registration. This improvement is purely coincidental and could not be guaranteed when registering other sources. The runtime of Hill Climbing is reduced greatly due to the limited nature of the search, with Sturges' Rule taking 0.119 seconds.

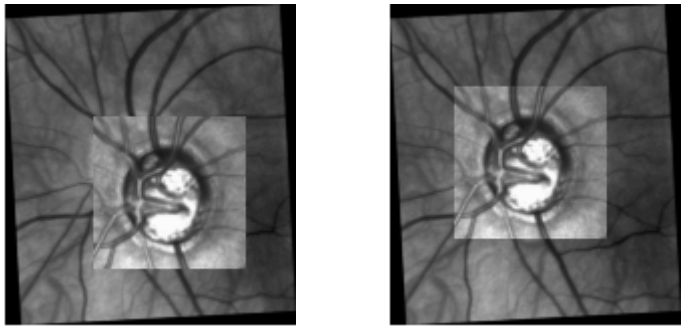


Figure 3. Results of registration using Exhaustive search (256 Bins Vs. Sturges Rule)

Figure 3 shows a comparison between using 256 bins and using Sturges' Rule for selecting the number of bins. It is clear to see that when 256 bins are used, the registration is incorrect. This is likely to be caused by the noisy background that is detracting from the data that we are actually concerned with (the blood vessels and optic disc). When using Sturges' Rule to perform registration, we can see that it is aligned correctly with all corresponding blood vessels matching up. By using Sturges' Rule to reduce the complexity of an image, it is shown that we can achieve better results for registration.

4 Discussion

In this paper we have demonstrated the benefits of using Sturges' Rule in Mutual Information. It is a common method in statistics, but has only ever been briefly mentioned with regards to computing entropy. Our testing shows that it has a large effect as to whether Mutual Information can actually perform the registration correctly, along with improving the runtime of the algorithm.

In the statistics literature, there are concerns that Sturges' Rule can smooth the histogram too much [10], and that Scott's Rule is a better approach to estimate bin size. In the context of image registration for our data we have found this *not* to be the case since Sturges' Rule consistently outperformed Scott's Rule, although both methods are still better than using a traditional 256 bin representation.

Bin size selection is just one aspect that can affect the performance of Mutual Information. We can clearly see from the results that although Sturges' Rule offers an improvement on registration, there is still much scope for developing the algorithm further to give satisfactory results for our data. One drawback of the standard Mutual Information measure is that it is calculated on a pixel by pixel basis, so much spatial information is lost. Existing techniques have attempted to resolve this by performing Mutual Information over the neighbourhood of each individual pixel [13] or by combining the standard Mutual Information measure with local gradient information from the image [14]. We wish to develop on these methods further to allow for successful registration of our image data.

It has been shown that Sturges' Rule offers a simple yet effective way to depict the original image histogram that can be used for determining the entropy result and improve upon the Mutual Information measure, for both accuracy and efficiency.

References

1. P. A. Viola and W. M. Wells III. Alignment by maximization of mutual information . In *ICCV*, pages 16–23, 1995.
2. J. P. W. Pluim, J. B. Antoine Maintz, and M. A. Viergever. Mutual information based registration of medical images: A survey. *IEEE Trans. Med. Imaging*, 22(8):986–1004, 2003.
3. N. Kwak and C. Choi. Input feature selection by mutual information based on Parzen Window. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(12):1667–1671, 2002.
4. N. Nicolaou and S. J. Nasuto. Mutual information for EEG analysis. *Proc. 4th IEEE EMBSS UKRI Postgraduate Conference on Biomedical Engineering and Medical Physics (PGBIOMED'05)*, pages 23–24, 2005.
5. T. Kadir and M. Brady. Estimating statistics in arbitrary regions of interest. *Proceedings of British Machine Vision Conference (BMVC)*, pages 6–7, 2005.
6. Y. Zhu and S. M. Cochoff. Influence of Implementation Parameters on Registration of MR and SPECT Brain Images by Maximization of Mutual Information. *J Nucl Med*, 43(2):160–166, 2002.
7. A. Rajwade, A. Banerjee, and A. Rangarajan. Continuous image representations avoid the histogram binning problem in mutual information based image registration. In *ISBI*, pages 840–843, 2006.
8. H. A. Sturges. The choice of a class interval. *J. American Statistical Association*, pages 65–66, 1926.
9. D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.
10. M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51(1):59–60, 1997.
11. P. L. Rosin, D. Marshall, and J. E. Morgan. Multimodal retinal imaging: new strategies for the detection of glaucoma. In *ICIP (3)*, pages 137–140, 2002.
12. M. Sonka, V. Hlavac, and R. Boyle. Image Processing, Analysis, and Machine Vision. *Brooks and Cole Publishing*, 1998.
13. D. B. Russakoff, C. Tomasi, T. Rohlfing, and C. R. Maurer Jr. Image similarity using mutual information of regions. In *ECCV (3)*, pages 596–607, 2004.
14. J. P. W. Pluim, J. B. Antoine Maintz, and M. A. Viergever. Image registration by maximization of combined mutual information and gradient information. *IEEE Trans. Med. Imaging*, 19(8):809–814, 2000.

A Model-Independent, Multi-Image Approach to MR Inhomogeneity Correction

P. A. Bromiley* and N.A. Thacker

Imaging Science and Biomedical Engineering, Stopford Building,
University of Manchester, Oxford Road, Manchester, M13 9PT.

Abstract. We present a technique for model-independent estimation of the inhomogeneity effect across multiple different acquisitions of the same subject, on the assumption of equal inhomogeneities in each image. The technique operates by integration of smoothed estimates of the local derivative of the bias field. Combination of evidence from multiple images is achieved at the derivative estimation stage via statistical weighting. Quantitative analyses of the approach on simulated data, and qualitative analyses on genuine MR images of the foot and hip, demonstrate the superiority of the combined image approach over the equivalent algorithm applied to individual images. The technique has the potential for application to situations where a bias field correction is needed, but likely to be unstably constructed for the acquisition of interest.

1 Introduction

Magnetic resonance images typically exhibit artefacts in the form of non-anatomic spatial intensity variations, known as the inhomogeneity effect, bias field, or gain field. Several sources contribute to this artefact, including B_0 (static) field inhomogeneity introduced both by the main magnet itself and by susceptibility effects at tissue boundaries, and B_1 (radio-frequency, RF) field inhomogeneity in the receiver coil [1]. Many computer-aided image analysis techniques developed for MR, such as intensity-based segmentation techniques, assume that each tissue can be characterised by a well-defined mean intensity. Therefore, the correction of significant inhomogeneity artefacts in the images under analysis is an important pre-processing step in the application of such techniques. The magnitude of the inhomogeneity effect is dependent on a variety of parameters, including both the physical properties of the object being imaged and the scanning parameters, including echo time and repetition time [1]. This makes accurate modelling of the effect for arbitrary objects extremely problematic, and has led to the development of post-hoc correction algorithms. A wide variety of algorithms have been developed for this task ([2] and references therein). One significant problem for such algorithms is that their ability to measure the inhomogeneity is dependent on the signal-to-noise ratio in the image. Thus, the correction is more difficult in regions of the image containing low average intensities, with the bias field being estimated here via a process that is effectively a smooth interpolation. In order to increase the statistical power of the data available to estimate the inhomogeneity effect, several researchers have investigated methods utilising multiple images e.g. [3, 4], on the assumption of equivalent anatomy across the images. However, in some circumstances it may be possible to acquire multiple images with approximately equivalent inhomogeneity fields. For example, in surface coil images (and to a lesser extent birdcage coils used in brain imaging) the main source of inhomogeneity is the RF field inhomogeneity of the receiver coil itself. Therefore, through alteration of the imaging parameters (echo time and inversion time) it is possible to acquire multiple images of the same object with differing intensities for each tissue, but with substantially equivalent inhomogeneity effects. This provides a novel method for dealing with the problems of estimating the inhomogeneity effect in tissues of low average intensity and thus low SNR: a companion image with high intensities in those regions can also be acquired, and the inhomogeneity estimated from the image pair.

In previous work we have presented a model-independent algorithm (i.e. not based upon a particular assumption of the tissues present and their expected distributions), for the estimation of inhomogeneity effects in single images. The issue of model-independence is an important one, since techniques which assume particular distributions may bias the later interpretation of pathological tissue. Our technique is based on the assumption that an image is composed predominantly of regions of homogenous tissue, separated by distinct step boundaries. The algorithm therefore estimates a (statistically weighted) local relative image derivative, taking care to first eliminate boundaries, and integrates this data to generate the bias field. In this paper, we demonstrate that the approach can be extended to simultaneously estimate the inhomogeneity effects present in multiple images of the same subject, on the assumption that they are approximately equivalent. The technique is applied to simulated chequerboard images, and simulated MR brain images acquired from Brainweb [5], and the superiority of the multiple-image approach over the application of the equivalent algorithm to individual images is demonstrated. Qualitative results from genuine MR images of the foot and hip are also presented. We conclude with some observations on the potential applicability of the proposed approach.

*E-mail: paul.bromiley@man.ac.uk

2 Method

We assume an image formation process in which a mean regional tissue grey level g_t is corrupted by a smoothly varying multiplicative gain $G(x, y)$ and additive noise $n(x, y)$, where x and y are the image plane coordinates

$$I(x, y) = g_t G(x, y) + n(x, y). \tag{1}$$

If the multiplicative distortion can be reliably calculated in well-measured regions, it is a relatively straightforward task to interpolate the observed trends across the entire image. The first step in the algorithm involves estimation of the image noise σ_I from the width of zero-crossings in horizontal and vertical gradient histograms [6], followed by application of a 3x3 median filter to remove outlier noise. A $(-1, 1)$ template is then used to estimate the intensity shift across adjacent voxels $\Delta_k(x, y)$ where $k = x$ or y, and normalised to remove tissue dependency, g_t, within homogeneous regions e.g. in the x direction we define

$$\Delta_x^{rel}(x, y) = \frac{\partial G(x, y)/\partial x}{G(x, y)} = \frac{2\Delta_x(x, y)}{I(x, y) + I(x-1, y)} = \frac{2(I(x, y) - I(x-1, y))}{I(x, y) + I(x-1, y)}.$$

Error map images for the above calculations are produced using standard error propagation e.g.

$$\sigma_x^{-2}(x, y) = \frac{(I(x, y) + I(x-1, y))^4}{16\sigma_I^2(I(x, y)^2 + I(x-1, y)^2)} \tag{2}$$

A conservative approach to tissue boundaries and regions of low statistical accuracy is taken: the inverse error is set to 0 for voxels of low signal $I(x, y) < \sigma_I$, or those with edges $> 3\frac{\sigma_I}{\sqrt{2}}$. The Δ_k^{rel} images are passed through a smoothing filter S in order to remove the effects of image structure and thus obtain the low spatial frequency variation i.e. the multiplicative gain map. A smoothing kernel of 5% of the image size is a good rule of thumb across the images studied. The appropriately weighted mean estimate of smooth local relative gradient is given by

$$\Delta_x^S(x, y) = \frac{S \otimes \Delta_x^{rel}\sigma_x^{-2}(x, y)}{S \otimes \sigma_x^{-2}(x, y)}. \tag{3}$$

where $\otimes$ represents a 2D convolution. The weighting reduces the significance of poorly measured gradients during convolution, after which they are removed by division. For an image with uniform intensity in homogeneous tissue, equation 3 is zero for all voxels in a slice. This can be used as the basis for a regularisation term to provide increased mathematical stability. Δ_x^S becomes

$$\Delta_x^S(x, y) \Rightarrow \frac{S \otimes (\Delta_x^{rel}\sigma_x^{-2}(x, y) + 0_{reg})}{S \otimes (\sigma_x^{-2}(x, y) + \sigma_{reg}^{-2})}. \tag{4}$$

The factor σ_{reg} is calculated at $|I(x, y) - I(x-1, y)| = 3\sigma_I$. This upper limit on the error constrains the effect of any voxels with large errors in their gradient calculations. However, the bias introduced towards zero gradient in regions dominated by noise necessitates a limited number of iterations in order to converge on the correct value of local gain.

At this stage, since error maps for the relative gradient maps are available, the gradient maps from any number of images with approximately equivalent inhomogeneity effects can be combined in a simple weighted averaging process, and the error map for the combined data calculated. A 2D map of the inhomogeneities can then be obtained by re-integrating the x and y gradient maps, using the centre of the image (which is within the region of interest for most MR images) to provide a relative point of unity gain. In order to prevent "integral wind-up" of errors, redundant information available from both directions of re-integration is utilised as described in [7]. The inverse of the exponential of the resulting inhomogeneity map provides the multiplicative correction.

Evaluation of the algorithm comprised two quantitative and two qualitative stages. First, two chequerboard images with opposite intensities were prepared (Fig. 1) and a 40% multiplicative bias field consisting of a pair of Gaussians was applied, followed by the addition of 1% random Gaussian noise. The inhomogeneity correction algorithm was applied to each image individually, and also to the images as a pair. Coefficients of variation (i.e. the ratio of the intensity standard deviation to the mean) were calculated for the light squares in each image, and used to calculate the percentage of the original inhomogeneity remaining after correction. An equivalent experiment was then performed using simulated T1 and T2 brain MR images from Brainweb [5]. Slice 160 of Brainweb inhomogeneity field A was applied multiplicatively to slice 27 from each volume, followed by the addition of 3% Gaussian random noise, and the same evaluation procedure used with the chequerboard images was followed, using the Brainweb tissue phantom to

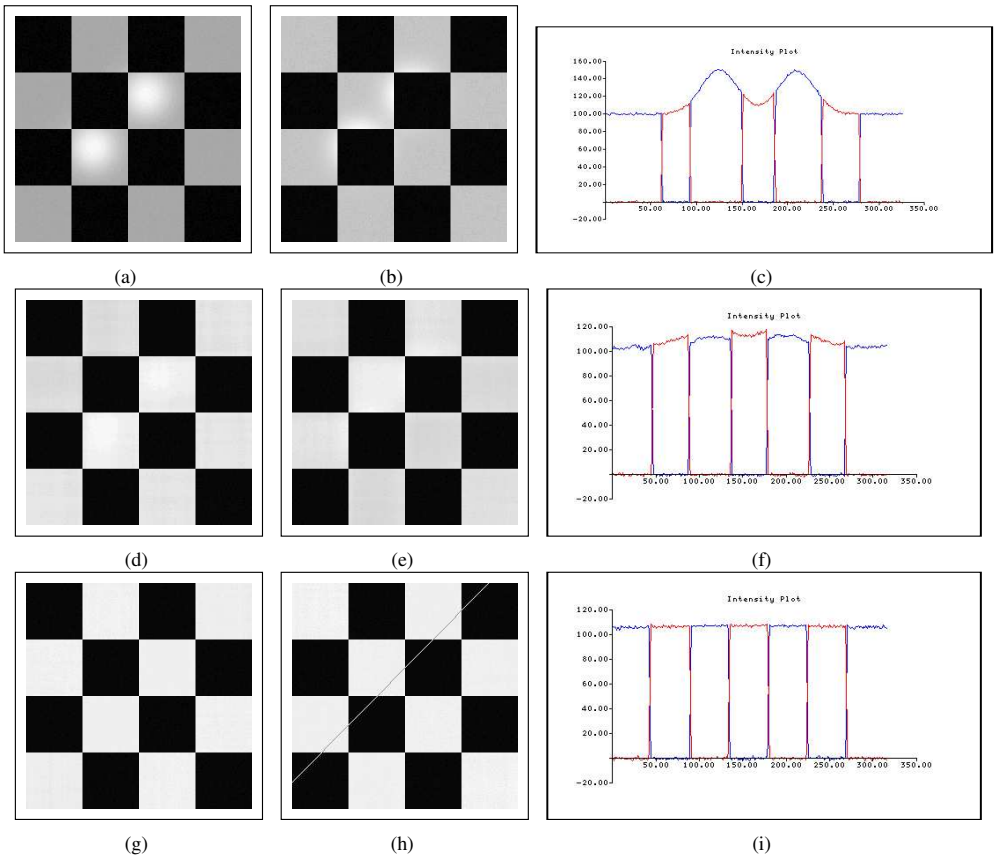


Figure 1. The original chequerboard images (a,b), individually corrected images (d,e), jointly corrected images (g,h) and intensity profiles (c, f, i) for each pair along the line shown in h.

identify white matter (WM) and grey matter (GM) in the corrected images. Finally, the algorithm was applied to T1 and T2 MR image pairs of the foot and hip, to demonstrate qualitatively the operation of the algorithm on real data.

Correction Method	T1 WM	T1 GM	T2 WM	T2 GM
Individual	47.5	89.0	17.0	28.8
Joint	0.5	8.3	0	2.5

Table 1: Inhomogeneity (as percentage, $\pm \approx 5\%$) remaining in the WM and GM after correction in the simulated Brainweb images.

3 Results

Figure 1 shows the result of individual and joint inhomogeneity correction of the simulated chequerboard images. As can be seen from the intensity variation along the profiles, the individual corrections fail to remove all of the inhomogeneity: 31% of the inhomogeneity originally present in Figure 1a and 74% of that in Figure 1b remain after correction: the performance is worse on the second image as the peaks of the bias field lie predominantly within the dark squares were the SNR, and so the inhomogeneity after multiplication, is zero. However, when corrected as a pair the inhomogeneity is completely removed to within the noise level. Equivalent results were obtained with the Brainweb images (Fig. 2, Table 1). The SNR is lower in the GM of the T1 image and the WM of the T2 image, and correction of the GM is particularly problematic since it occupies a relatively thin band in the image. Therefore, whilst the original algorithm corrects the majority of the inhomogeneity in most regions, it removes only 10% of the inhomogeneity in the GM of the T1 image. In contrast, the combined correction removes all the inhomogeneity to within errors due to the higher statistical power available in the combined data. Finally, Fig. 3 and Fig. 4 show applications of the correction individually and jointly to T1 and T2 pairs of hip and foot images. These data are presented only for qualitative evaluation; however, the joint correction is visibly more stable. This is particularly clear in the upper-left region of Fig. 4: joint correction removes more inhomogeneity (Fig 4f vs. 4e) with less noise enhancement (Fig. 4c vs. 4b).

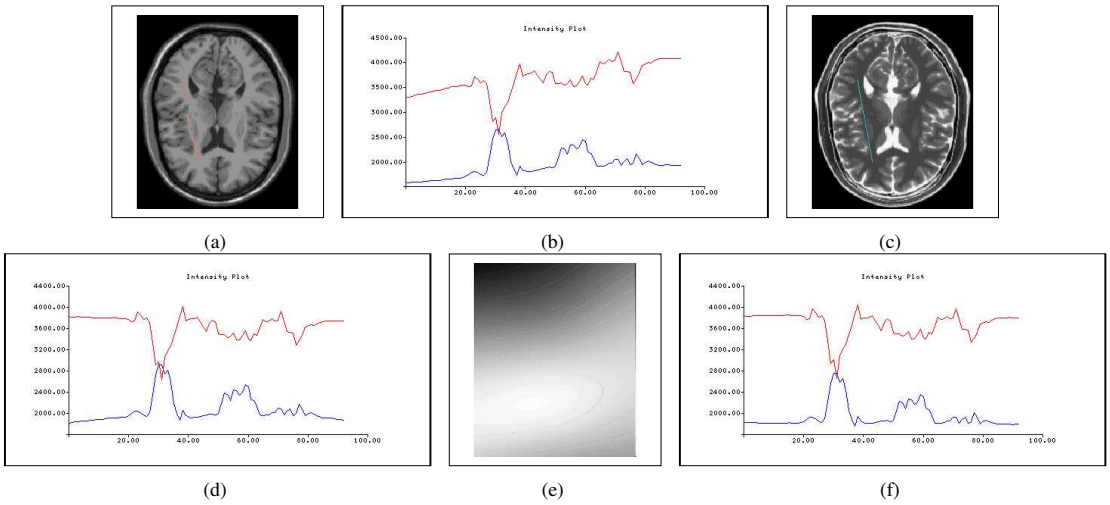


Figure 2. The T1 (a) and T2 (c) Brainweb images, the multiplicative bias field (e), and the profiles along the line shown in (c) for the uncorrected (b), individually corrected (d) and jointly corrected (f) images.

4 Conclusion

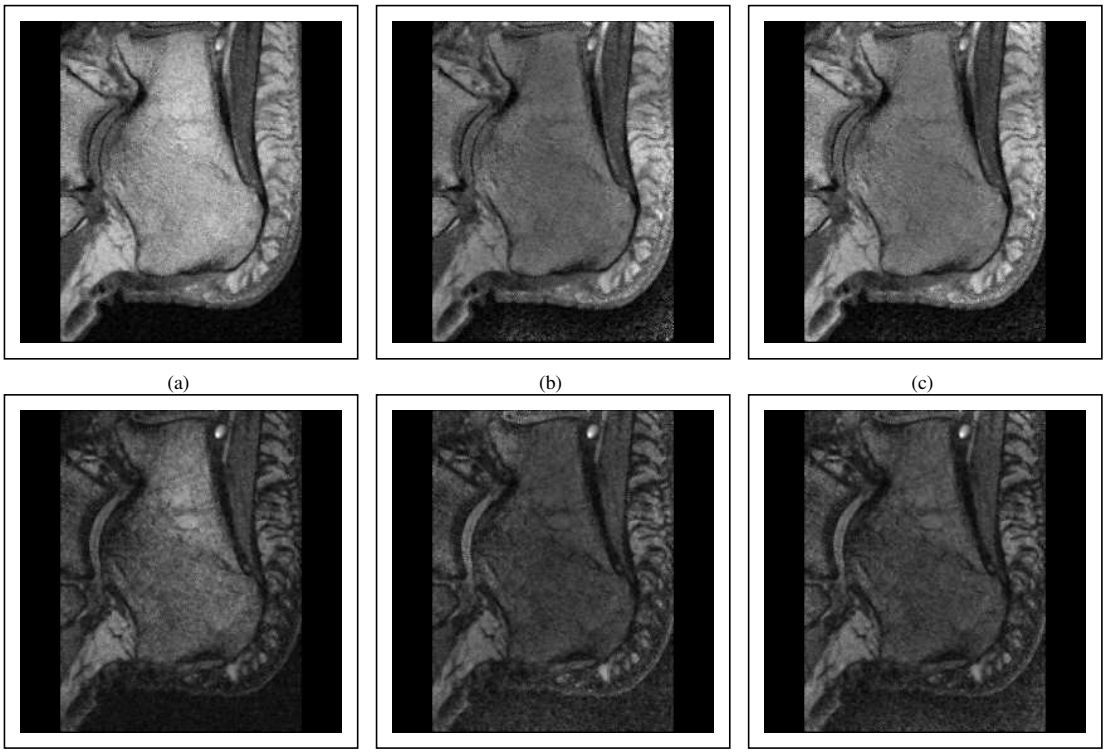
In previous work [7] we presented a model-free inhomogeneity correction algorithm. Like other model-free techniques, its successful application is dependent upon having good quality data. In many real-world situations this may prove a real impediment to use of the technique. This work has demonstrated an extension to simultaneous correction of multiple images on the assumption of equivalent inhomogeneity. This assumption is approximately valid under certain circumstances, for example when scanning is performed using a surface coil, where the inhomogeneity artefact is dominated by the RF inhomogeneity of the coil itself and so is independent of the scanning parameters to a large extent. Under such circumstance multiple images of the same subject can be acquired with varying intensities in each tissue through variation of the scanning parameters, without greatly affecting the inhomogeneity, ensuring that high SNR is present in all tissues across the image set. This increases the statistical power of the data to determine the inhomogeneity effect, and thus results in superior inhomogeneity correction. The resultant increase in performance has been demonstrated quantitatively on simulated data and qualitatively on real MR images. We anticipate that this approach may also be applicable to some modern MR scanning protocols (e.g. [8]) where multi-contrast images are acquired simultaneously with some sharing of K-space, and so some sharing of inhomogeneity effects.

Acknowledgements

This work was supported by the the MIAS (Medical Images and Signals) IRC under EPSRC grant no. GR/N14248/01 and the MRC grant no. D2025/31. The software used in this study is freely available from www.tina-vision.net.

References

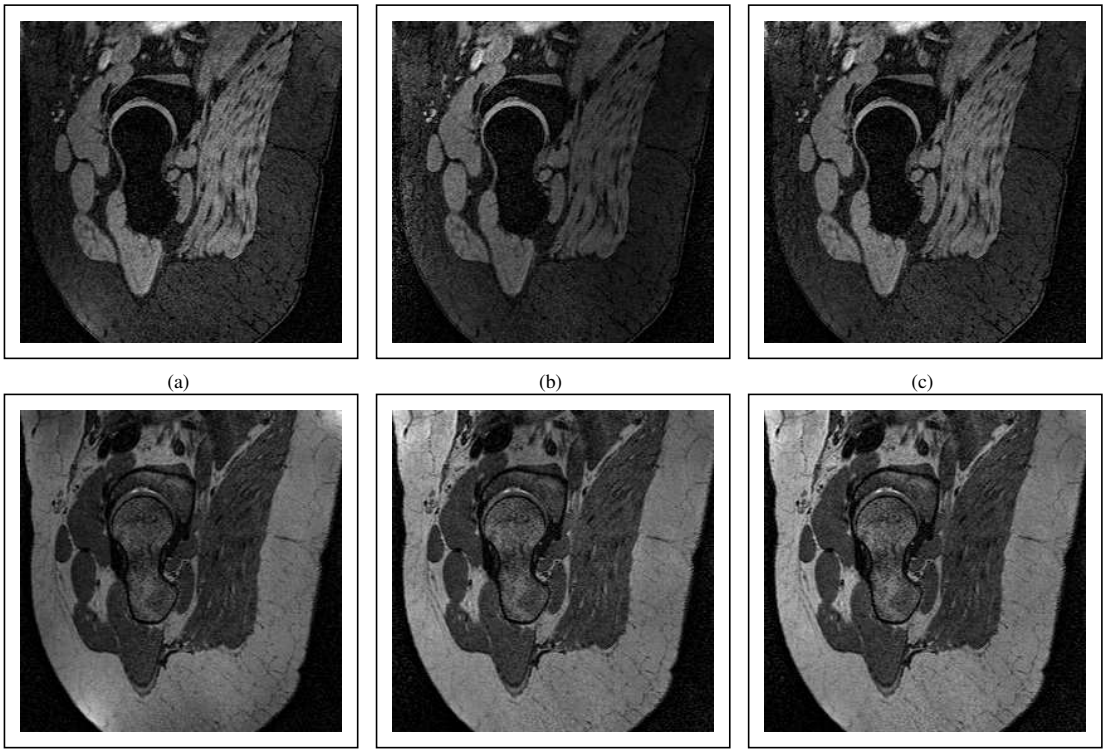
1. A. Simmons, P. S. Tofts, G. J. Barker et al. “Sources of intensity nonuniformity in spin echo images at 1.5T.” *Magnetic Resonance Imaging* **32**, pp. 121–128, 1994.
2. Z. Hou. “A review on MR image intensity inhomogeneity correction.” *International Journal of Biomedical Imaging* **2006**, pp. 1–11, 2006.
3. E. G. Learned-Miller & P. Ahammad. “Joint MRI bias removal using entropy minimisation across images.” *Neural Information Processing Systems* **17**, pp. 761–768, 2005.
4. S.-H. Lai & M. Fang. “A dual image approach for bias field correction in magnetic resonance imaging.” *Magnetic Resonance Imaging* **21**, pp. 121–125, 2003.
5. C. A. Cocosco, V. Kollokian, R. K.-S. Kwan et al. “Brainweb: Online interface to a 3D MRI simulated brain database.” *NeuroImage* **5(4)**, pp. S425, 1997.
6. S. I. Olsen. “Estimation of noise in images: An evaluation.” *CVGIP: Graphical Models and Image Processing* **55**, pp. 319–323, 1993.
7. E. Vokurka, N. A. Thacker & A. Jackson. “A fast model independent method for automatic correction of intensity non-uniformity in MRI data.” *Journal of Magnetic Resonance Imaging* **10**, pp. 550–562, 1999.
8. R. Mekle, A. F. Laine & E. X. Wu. “Combined MR data acquisition of multicontrast images using variable acquisition parameters



(a) (b) (c)

(d) (e) (f)

Figure 3. The original (a,d), individually corrected (b,e) and jointly corrected (c,f) MR foot images.



(a) (b) (c)

(d) (e) (f)

Figure 4. The original (a,d), individually corrected (b,e) and jointly corrected (c,f) MR hip images.

and k-space data sharing." *Magnetic Resonance Imaging* **21**, pp. 121–125, 2003.

Rigid –Registration of MRI to Ultrasound Kidney Images for Use During HIFU Treatments

Tara Yates[a], Penny Probert-Smith[a], Alison Noble[a], Tom Leslie[b], Anthony McIntyre[c], Rachel Phillips[c]

[a] Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, [b]The HIFU Unit, Churchill Hospital, Headington, Oxford, OX3 7JL, [c]Department of Radiology, Chruchill Hospital, Headington, Oxford, OX3 7JL

Abstract

A method of rigid registration of an MR volume to a 2D ultrasound image of the kidney to be used as an aid during HIFU therapy is presented. In both images, kidney boundary points are first segmented to which an ellipse is fitted. Manual segmentation is used in the case of the ultrasound image, whereas an automatic statistical intensity based method is applied to each slice of the MR volume. The MR volume is reduced to an appropriate 2D slice and the geometric transformation parameters between this slice and the ultrasound image are determined using the relevant parameters of the fitted ellipses.

1. Introduction

High intensity focused ultrasound (HIFU) is an emerging non-invasive therapy technique for treatment of a variety of tumours including the liver and kidney [1]. Tumour tissue is destroyed through the heating of tissue at a cigar shaped focus of the ultrasound energy, approximately 12mm in length and 2mm in diameter. Currently the most common intra-operative imaging modality is 2-D diagnostic ultrasound (US). Although 2D diagnostic ultrasound has the advantage of being compatible with therapy delivery and is relatively inexpensive it is inferior to magnetic resonance imaging (MRI) in terms of its ability to show tissue differentiation and to perform temperature mapping. In order for HIFU to be successful in a clinical setting accurate knowledge of the position of the focus in relation to the tumour boundaries is required. In the planning stages of HIFU therapies at the Churchill Hospital, Oxford, both pre-operative MRI and US images are acquired. The registration of the pre-operative MRI volume to the intra-operative real-time 2D US images would provide extra structural information for the interventionalist during treatment. In this paper we present a method of rigid registration between a pre-operative MRI volume and a 2D US image based on the shape of the kidney. Due to the necessity to include a water bath between the imaging head and the patient skin as well as interference from the treatment field and temperature changes the quality of the images acquired during treatment is poorer than in diagnosis. Thus shape is the dominant feature of the images that can be utilised in image registration. This study focuses on the registration of images of the kidney. Work on renal volume suggests that the kidney outline may be approximated to an ellipse [2]. In this paper an ellipse is fitted to points on the kidney boundary separately for each slice in an MR volume and a 2D ultrasound image. Points are found by automatic segmentation for the MR slice and manual segmentation is used for the ultrasound image. Ellipse parameters for each slice in the MR volume and the US image are then compared to find the best geometric transformation parameters.

2. Data Acquisition

In this paper, we used pre-treatment MR and ultrasound images due to the availability of this data. However the methods used are directly transferable to ultrasound images acquired during HIFU treatment. Images were acquired on patients who were to undergo HIFU treatment at the Churchill Hospital, Oxford. The ultrasound images were acquired using an Acuson Sequoia system with a 4 MHz abdominal probe whereas the MRI images were acquired using a GE Medical Systems, Genesis Signa 1.5T model.

3. Segmentation

3.1 Ultrasound segmentation

In this paper, manual segmentation was used for the ultrasound images to provide a proof of principle of the registration method. Future work will look into automating this step. However, it is possible that manual segmentation will provide the best results in this case, due to the poor quality of the images acquired during HIFU treatment, and the ability for the surgeon to remain in control of the process. Pre-processing to reduce speckle was

found to improve point selection and estimation. The aim of the pre-processing is to reduce speckle, while preserving the organ boundaries. The "sticks" filter proposed by Czerwinski [3] was used to reduce speckle prior to point selection, see Fig. 1. This was found to give better result than selecting points directly on the original images. In this study around 35 points were selected manually as shown in Fig 1c). Preliminary investigations suggested that the number of points could be reduced to 10-20, without significant change to the fitted ellipse. This number of points would be practical for use in HIFU surgery.

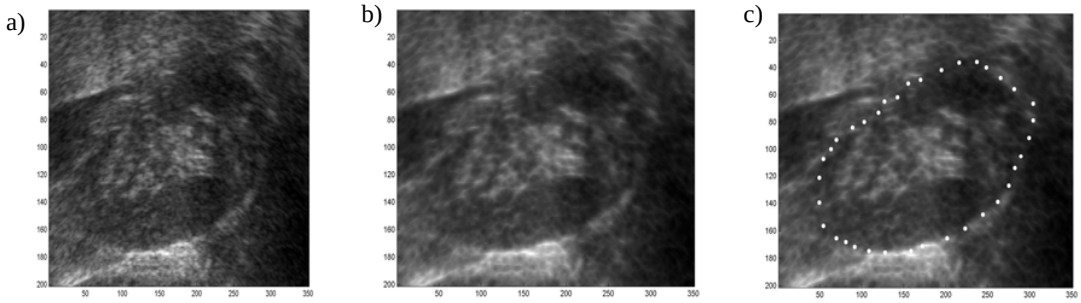


Figure 1: The application of the Sticks Algorithm to reduce speckle in an ultrasound image of the kidney a) original ultrasound image b) result of the sticks algorithm c) the manually selected points

3.2 MRI segmentation

The kidney boundary was determined from a set of MR images automatically. Segmentation of MR images was done using a statistical intensity based method [4]. This method applies Ostu's thresholding method [5] to separate the image into distinct classes. The segmentation was then further refined using a Markov random field Expectation- Maximization approach as described in [6]. The results for one MR slice are shown in Fig. 2. The algorithm produces a binary image of the segmented regions from which the boundary points can be selected.

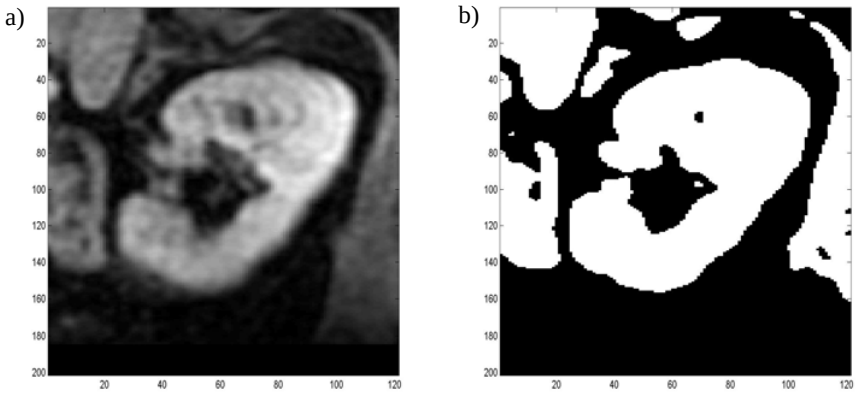


Figure 2: Segmentation of MR image for one slice through the kidney

4. Rigid registration

The registration of a 2D slice, as acquired from a standard ultrasound image, to a 3D volume, the MR image volume, is non-trivial. To avoid a full 3D-2D registration, a 2D oblique slice from the MR volume is selected to reduce the problem to a 2D-2D registration. This step reduces the computational time required for the registration which is an important requirement for real-time application, but requires the MR slices to be taken in a direction consistent with the imaging plane of the ultrasound as defined by the radiographer. Slice selection and ultimately the image registration is achieved through the use of 2D shape matching. The segmentation steps described in section 3 identify boundary points for the kidney in both imaging modalities. An ellipse is fitted through each set of boundary points using a least squares method and the major and minor axes are calculated. An ellipse was deemed an appropriate approximation for the shape of the kidney due to its extensive use in determining kidney volume in ultrasound [2]. The fitting of an ellipse to a variety of kidney images, including intra-operative images, showed that

this assumption was valid in most cases. An example of an ellipse fitted through the boundary points of the segmented kidney from both ultrasound and MR images is shown in Figure 3.

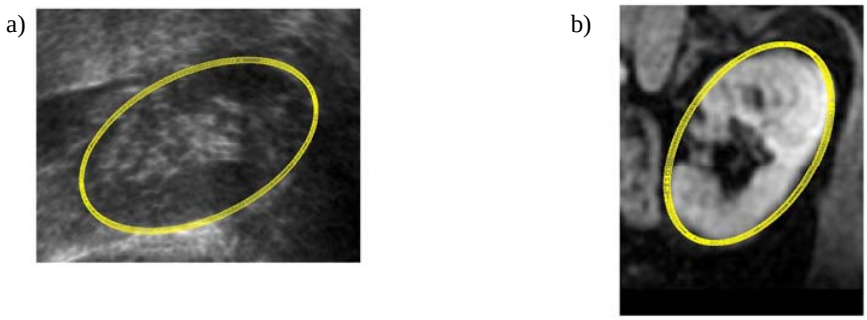
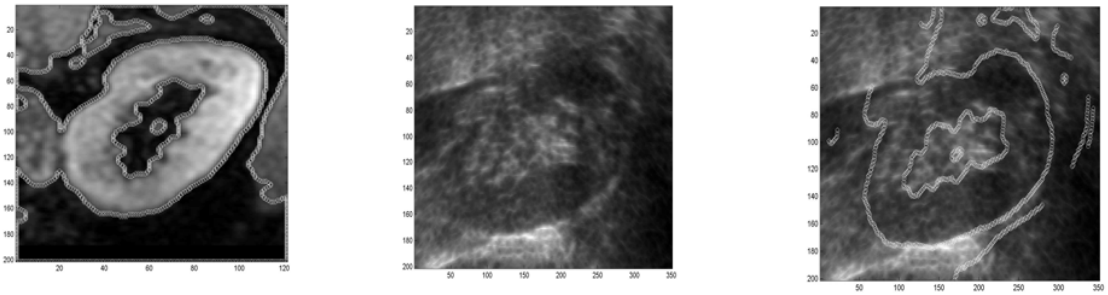


Figure 3: Ellipse fitted to a) US and b) MR data

It is assumed in this work that the shape of the kidney undergoes little deformation in subsequent scans and is independent of scan position. From this we can use the shape of the fitted ellipse to select the appropriate MR slice from the kidney volume. This was achieved by selecting the MR slice with the closest major-minor axis ratio to the ellipse fitted to the ultrasound data. Once the MR slice was selected rigid registration of the two images was achieved using the parameters of the ellipse, namely the position of the ellipse centre, the angle between the major axis and the x-axis and the size of the major and minor axis, to provide the transformation parameters, transformation (Tx,Ty), rotation (Rx,Ry) and scale (Sx,Sy). The results for one patient study are shown in Fig 4.

An independent measure is needed to determine whether the method does indeed select the correct slice and to assess the accuracy of the registration. The calyceal system appears in both images as a contrasting intensity central region. Figure 5 shows the overlay of the registered MR outline on the ultrasound image of the selected slice (Fig. 5 c)) and the two previous and the two subsequent MR slices from the volume. In each case the region of the calyceal system has been segmented from the MR images. A visual inspection shows that this region is mapped most closely in Fig. 5 c) as expected. However a close correspondence is also seen in Fig.s 5 b) and 5 d). It is encouraging to note that the greatest deviation is seen in Fig.s 5 a) and 5 d) which are the illustrated slices furthest from the selected slice.

Registration time (excluding segmentation) is approximately 1 second showing that the registration technique provides an extremely fast registration method, which would be applicable to HIFU therapy.



a) MR image of the kidney with segmented boundary indicated

b) ultrasound image of the kidney after the application of a “sticks” filter

c) representation of the final registration; ultrasound image with MR boundary superimposed

Figure 4: Registration of an MR image to an ultrasound image of the kidney. The boundary segmented from the MR image is shown as an overlay on the ultrasound image.

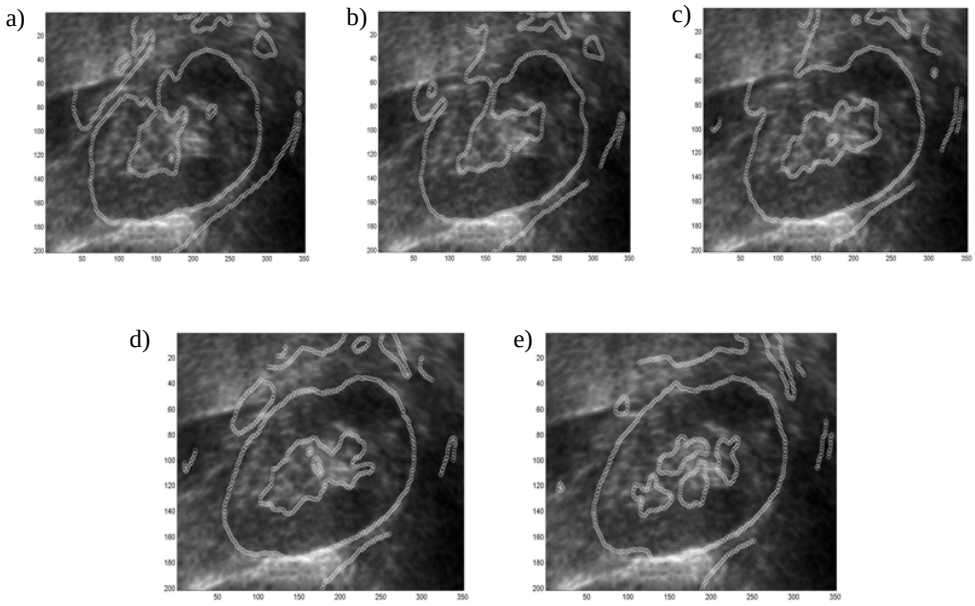


Figure 5: The registration of MR slices around the slice c) selected as the optimal match by shape matching. The shape of the high intensity central region (the calyceal system) varies across MR images with the closest match to the US image being observed in c).

5. Discussion

A method of rigid registration of an MR volume and a 2D ultrasound image of the kidney has been presented and applied to a single patient dataset to show proof of principle of the method. The outline of the selected MRI slice provides enhanced structural information to the ultrasound images which would be of use during HIFU treatment. In particular the calyceal system within the kidney, is enhanced in the ultrasound image through the overlay of the MR boundary. This region can also be used to assess the quality of the registration method. Although this shows the slice selected from the registration algorithm provides the closest fit to this region, its shape in the previous and subsequent slices is not significantly different suggesting further studies are necessary. It is envisaged that a combination of registration from the ellipse parameters and further information from the texture and shape of the calyceal system may lead to improved results.

Further validation of the method is required through application to a greater number of sample images. It is clear that the shape of the boundary provides a good basis for registration and more general shapes could be investigated (for example based on fractal measures) that may be applicable to registration of other organs treated in HIFU such as the liver.

Although it is difficult to quantify the error involved in the registration at this stage there are several sources of likely error. First, the reliance on radiographer expertise to provide a good initial direction for MR slice sampling is likely to introduce errors into the registration. It is possible that this could be overcome by searching over the whole volume of the MRI for a close fit to the ellipse obtained from the ultrasound image or through the use of a system to determine the co-ordinates of the ultrasound probe exactly. Both these approaches would add computational cost. It is likely that this second method could be achieved when the registration is applied to the real-time ultrasound acquired during therapy as the imaging transducer has a fixed geometry in this case. Second, the assumption that the kidney is rigid, though widely accepted, may lead to errors in the registration by this method. If so, it may be necessary to refine the method using a non-rigid approach. This would also come at the cost of computational time.

With further validation the speed of the registration and the ability to process the majority of the MR data prior to surgery would realistically allow the application of this method to the real-time ultrasound images acquired during HIFU treatment.

References

1. J.E. Kennedy, G. R Ter Haar, D. Cranston “Review article: High Intensity Focused Ultrasound: Surgery of the Future?”, *The British Journal of Radiology*, **76,** pp 590-599, 2003
2. T.B. Jones, L.R. Riddick, M.D. Harpen, R.L Dubuisson., and D. Samuels, “Ultrasonographic determination of renal mass and renal volume”, *Journal of Ultrasound in Medicine*, **2(4)**, pp 151-154, 1983
3. R.N. Czerwinski, D.L Jones and W.D. O’Brien, “Detection of Lines and Boundaries in Speckle Images – Application to Medical Ultrasound”, *IEEE Transactions on Medical Imaging*, **18(2)**, pp126-136, February 1999
4. C. Hsieh, P. Probert-Smtih, J. Kennedy, T. Leslie, “Monitoring Cavitation in HIFU as an aid to Assisting Treatment”, *ISTU6 Proceedings, Oxford*, Sept. 2006
5. N. A. Ostu, “Threshold Selection Method from Grey-Level Histograms”, *IEEE Trans. Syst. Man. Cybern., SMC-9*, **(1)**, pp 62-66, 1979
6. K.Van Leemput, F.Maes, D. Vandermeulen, and P. Suetens, “Automated Model-based Tissue Classification of MR Images of the Brain”, *IEEE Transactions on Medical Imaging*, **18(10)**, pp 897-908, October 1999

Multispectral MRI De-noising Using Non-Local Means.

J. V. Manjón[1] , M. Robles[1] and N.A. Thacker[2]
1: IBIME Group, ITACA Institute, Universidad Politécnica de Valencia, Valencia, Spain.
2: Imaging Science and Biomedical Engineering, University of Manchester.

Abstract.
Clinical MRI data is normally corrupted by random noise from the measurement process which reduces the accuracy and reliability of any automatic analysis. For this reason, de-noising methods are often applied to increase the SNR and improve image quality. Most of these methods work on single channel images by correcting each grey level using an implicit model of the surrounding region, but without taking into consideration the potential multispectral nature of MR images. In this paper we present an extension of a recently proposed filter to reduce random noise in multispectral MR images and test it on synthetic and real images. We compare performance to a multispectral approach based upon the imaging physics and published previously at this conference using real data. We conclude from our results that these methods can be used for de-noising of MR data.

Introduction

MRI de-noising is a common pre-processing step in many MR image processing and analysis tasks, such as segmentation or registration. Many filtering methods are based on the signal averaging principle which uses the spatial redundancy in the image. In this sense, Gaussian filters have been largely used in some applications such as fMRI but they have the disadvantage of blurring edges due to averaging of non similar patterns. In order to avoid this problem many edge preserving filters have been proposed. Probably the best known is the Anisotropic Diffusion Filter (ADF) (Guerig et al,1992; A. Samsonov and C. Johnson, 2004). Such filters respect edges by averaging pixels in the orthogonal direction of the local gradient. However such filtering can erase small features and may change image statistics. Wavelet based filters have also been applied to MRI de-noising (R. Nowak, 1999; J. C. Wood and K. M. Johnson,1999) but such filters tend to introduce characteristic artifacts that can be very problematic for the clinicians.

Most existing filters work on single channel images without taking into consideration the potential multispectral nature of MR images. There are few methods that use the multispectral information as basis for the de-noising process. One of the first attempts to use this information was the multispectral ADF proposed by Gerig (Gerig et al,1992). In their work they propose using the gradient information of the different channels to conduct the diffusion process. A wavelet base de-noising technique for multispectral images exploiting interscale and interchannel correlations has also been proposed (Scheunders, 2004). This technique is demonstrated to outperform single channel wavelet thresholding techniques. Finally, a partial volume segmentation based approach has been recently proposed in Thacker (Thacker et al, 2004) where the filtering is performed using multidimensional data and a partial volume data density model (MPVM). This approach abandons altogether local smoothness constraints and achieves noise filtering by enforcing agreement between measured data using underlying tissue proportions computed from a physics based image formation model. This work also introduced a new way of characterising noise filters, which permits direct comparison of filtering techniques on real (as opposed to simulated) data, thereby obtaining more meaningful measures without the need for a gold standard. This multi-spectral filtering method was previously evaluated in comparison to single image filtering approaches. At the time a state of the art multispectral noise filter was not implemented for comparison.

In the present work we extend the application of a new filter recently proposed by Buades (Buades et al 2005) know as Non-local means (NL-means) to de-noise multispectral MR images. The technique is first evaluated in the conventional way using simulated data sets, and then compared to the MPVM noise filter using our previously suggested evaluation method on real data.

Methods

The NL-means filter is an evolution of the Yaroslavsky filter (Yaroslavsky, 1985) which averages similar image pixels defined according to their local intensity similarity. The main difference between the NL-means and this filter is that the similarity between pixels has been made more robust to noise by using a region comparison, rather

than pixel comparison and also that matching patterns are not restricted to be local. That is, pixels far from the pixel being filtered are not penalised.

Given an image Y the filtered value at a point p using the NL-means method is computed as a weighted average of neighbouring pixels N_p in the image following this formula:

$$NL(Y(p)) \;=\; \sum_{\forall q \in N_p} w(p,q)Y(q) \quad with \quad 0 \leq w(p,q) \leq 1 \quad and \quad \sum_{\forall q \in N_p} w(p,q) \;=\; 1 \qquad (1)$$

where p is point being filtering and q represents any other image pixel[1]. The weights $w(p,q)$ are based on the similarity between the neighbourhoods N_p and N_q of pixels p and q. N_i is defined as a square neighbourhood window centred around pixel i with a user-defined radius R_{sim}. Theoretically, noise filtering performed must be considered as an estimation task. Therefore, the process of linear weighting and the weight factors $w(p,q)$ can be regarded as the calculation which computes the most likely noise free grey-level value of the selected pixel, on the basis of the measured evidence. A simple example which leads to this form of solution can be derived using Likelihood, on the assumption that the measurements form each image ($Y(q)$ and $Y(p)$) can be taken as an independent estimate of the noise free value, drawn from a Gaussian distribution with variance $1/w(p,q)$. The process of selecting the most effective noise filtering algorithm can be considered as a way of optimising the match between the assumed computational form and the statistical distributions in the data.

In the case of NL-means, $w(p,q)$ is calculated as:

$$w(p,q) \;=\; \frac{1}{z(p)} exp(-\frac{\sum_i \sum_j G_\rho(i,j)(Y(p_{ij}) - Y(q_{ij}))^2}{h^2}) \;=\; \frac{1}{z(p)} exp(-d(p,q)/h^2) \qquad (2)$$

$z(p)$ is the normalising constant, h is a exponential decay control parameter and G_{rho} is a normalised Gaussian weighting function with zero mean and standard deviation ρ (generally set to 1), so that d is a Gaussian weighted squared Euclidian distance of all the pixels of the neighbourhood. The normalisation factor is defined by:

$$z(p) \;=\; \sum^{N_p} exp(-d(p,q)/h^2) \qquad (3)$$

The weighting process penalises pixels far from the centre of the neighbourhood window giving more weight to pixels near the centre[2]. To eliminate over-weighting in (1), the original NLM method w(p,p) was calculated as:

$$w(p,p) = max(w(p,q), \forall q \neq p) \qquad (4)$$

However, this correction can have the disadvantage of blurring singular points (i.e. pixels with no similar patches, like image corners and peaks or valleys) by averaging them with non similar patches[3].

As the acquisition of multispectral sequences is common in clinical practice the above method can be extended to be used on a multichannel framework. Effectively, the similarity measure can be better obtained by combining information of various channels. Therefore, we propose to use a multispectral similarity function described as follows:

$$w(p,q) \;=\; \frac{1}{z(p)} exp(\frac{-\sum_i^C d(p_i,q_i)/h_i^2}{C}) \qquad (5)$$

where $z(p)$ is the appropriate normalisation factor, C is the number of channels and h_i parameter is closely related with the noise standard deviation of each channel. We will refer to this multispectral method as MNLM.

The NLM algorithm has three parameters and the filter results depend highly on their setting. The first parameter is the radius of the search window enclosing N_P. The second parameter, R_{sim}, is the radius of the neighbourhood window used to find the similarity between two pixels. If the value of R_{sim} is increased the similarity measure

[1] In principle N_p can include the whole image, though efficiency requirements normally prevent this.

[2] The center pixel of the Gaussian weighting window is set to the same value that the pixels at a distance 1 to avoid over-weighting effects.

[3] To overcome this situation, we apply equation 4 only if the maximum w is above to a fixed threshold. We have fixed this value to 0.002 which corresponds to pixels far more than 2.5 times the image noise standard deviation. This modification is especially effective on low noise conditions.

will be more robust but fewer similar neighbourhoods will be found. The third parameter, h, is related to the decay of the exponential curve and controls the degree of smoothing. If h is too small, little noise will be removed while if h is set too high, the image will be blurred. In our experiments we use an 11x11 search window, which seems a reasonable value for medical images. The best setting for R_{sim} and h under different noise levels was estimated to be $R_{sim} = 1$ and h is set to the estimated standard deviation of image noise.

To conduct the experiments we used a simulated T1-weighted, PD-weighted and T2-weighted, 1mm3 voxel resolution images (8 bit quantization) from the Brainweb phantom (C. Cocosco et al.,1997) (fig 1) [4]. All experiments were performed using MATLAB 7.0 (Mathworks Inc.).

Experiments and Results

The accuracy of the filter over typical levels of MR image noise (1 % to 9 %) was evaluated with the mono and multispectral version of the filter to show the effect of adding multiple channels in the image de-noising process.

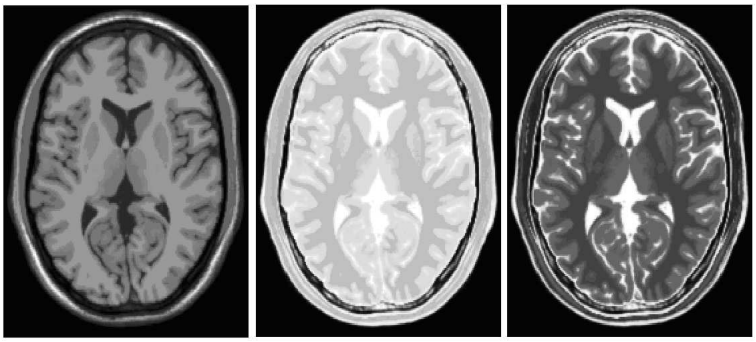
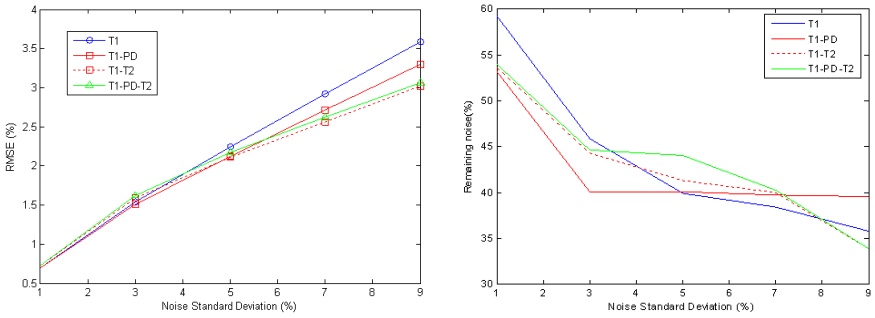


Figure 1. From left to right. T1-Weighted synthetic noise free MR image, PD and T2.

Four different methods of the NLM were evaluated, one mono-spectral and three multispectral. The Root Mean Square Error (RMSE) was used to evaluate the differences between the de-noised image (T1) and the original without noise. Results are summarised on fig 2. As can be seen, to include multiple channels improves filter



(a) (b)

Figure 2. Fractional removal of RMSE for the different methods evaluated.

performance. The best results were obtained with the combination T1-T2 probably due to a better tissue contrast.

Estimates of noise removal using simulated data have a number of restrictions. The main one of these is that we can never be sure that a simulation accurately matches the problems seen in real data. How do we know that the the filter is doing something appropriate at all locations? Secondly, performance figures at low noise levels are diffciult to interpret, as most simulation data is derived at some point from real data and contains residual noise which will inevitably be removed by the noise filter, preventing a low RMSE. In figure 2 we might argue that results below 3% noise are underestimating the true capabilites of the filter, and beyond 3 % the performance can be considered as effectively constant, as expected for an averaging based scheme.

For this reason it has been suggested that noise filters can be assessed using real image data using two complementary performance measures. The first, the residual outlier measure (ROM) simply measures the number of

[4] Although it is well known that magnitude MRI data is Rician distributed, for signals away from zero such as in the region of the brain, it can be well approximated by a Gaussian.

image values which have been changed unreasonably by the application of the filter (ie: by more than 3 standard deviations of the estimated image noise). The second evaluates the efficiency of the noise filter by measuring the proportion of additional noise which passes the filter. These measures are now given for the multispectral NLM filter.

To evaluate effectiveness of the multispectral approach, four real images were used (IRTSE, PD,T2 and FLAIR top row in Figure 3). The noise level in each image was estimated using the LNE technique (Thacker et al, 2003). The proposed method has been compared with the multispectral partial volume modelling (MPVM) method proposed by Thacker (Thacker et al, 2003). Although the proposed method does not remove as much noise as the MPVM method it is less destructive.

	LNE	MPVM	MNLM	Gaussian Filter
IRTSE	58.76	0.22 (1689)	0.50 (125)	0.27 (2405)
PD	64.06	0.20 (1804)	0.55 (45)	0.26 (3127)
T2	58.2	0.17 (938)	0.44 (43)	0.26 (1909)
FLAIR	52.4	0.13 (4971)	0.56 (194)	0.27 (3827)

Table 1. Table 1: Monte-Carlo estimate of fraction of remaining noise following Filtering and data lying beyond 3 S.D. of original value following Filtering (brackets). The results for standard Gaussian smoothing with a kernal width of 1 pixel are provided for comparison.

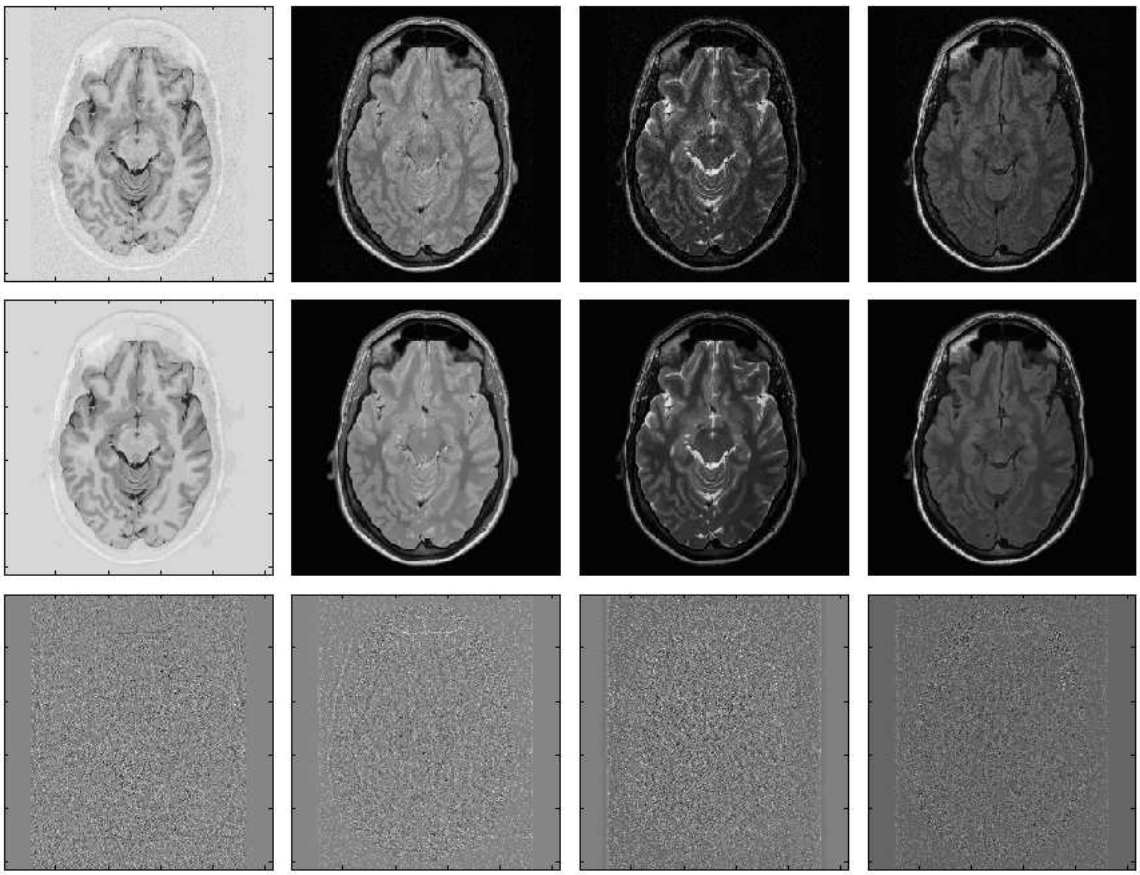


Figure 3. MNLM results (original, de-noised and residuals)

Images of the filtered images are shwn in Figure 3. Although the MPVM method is seen to have structural patterns in the estimated noise which are due to the constraints imposed on the estimated tissue proportions across the 4 input images, the MNLM method shows little evidence for tissue dependant changes.

Conclusion

In this paper we have explained how, for a noise filter to be statistically valid, its algorithmic form should be consistent with the statistical distributions present in real data. In practice, these distributions are often not measured

but instead algorithms are tuned by directly adjusting for best performance. This process has been applied here for the application of the method of an NLM noise filter to multispectral data.

The new method of multispectral filtering has been evaluated in the standard way using simulated data. In addition, given the limitations of such an approach, we have compared the performance of the filter to other approaches on the basis of measures computed from real data. There appears to be sufficient evidence to conclude that empirically based filters, such as NLM, have superior performance to simple noise filters such as median and Gaussian filtering, and that also this performance is enhanced if data are combined across multiple MR sequences, in the form of a multispectral filter. The method can be thought of as a bootstrap filtering scheme which filters on the basis of recurring stucture in a given image, as opposed to say Bayesian (or any model based) methods which pre-suppose the same prior expectation for all images. The performance of this filter also compares well with a physics based approach (MPVM) which reconstructs the most likely image data from estimates of tissue proportions. Though it is slightly less efficient at removing noise, it seems to be less destructive, particularly for FLAIR images, in which the MPVM filter also corrects image processes which are not described by a linear model of image formation, such as flow artefacts. For the chosen control parameters, the level of data modification by the NLM is less than that predicted by the measured noise level of the image for a perfect noise filter (ie: for Gaussian noise we expect a value which is the integral of a Gaussian beyond 3 S.D. $\approx 1,200$) . Such under-correction errs on the side of caution for medical data where image details play an important role in medical diagnosis. As the MNLM filter does not transform the remaining noise on correlated noise (noise-to-noise principle (Buades,2005)) the filter can potentially be applied iteratively to remove further noise without introducing a significant number of outliers. However, this claim must be moderated by the observation that any averaging process is likely to suppress the appearance of unique patterns in the data.

The main advantages of NLM are probably the simplicity of this model free approach (certainly in comparison to MPVM) and its non destructive nature, though execution time may be an issue for some implementations and setting up the filter is still a process of trial and error. For situations where the image formation process matches a linear process and unique locations need to be preserved, the physics based approach (MPVM) has built in parameter estimation and superior noise filtering characteristics. These issues may yet be addressed with more research.

References

1. Paul Scheunders. Wavelet Thresholding of Multivalued Images, IEEE TRANSACTIONS ON IMAGE PROCESSING, 13(4), 475-483, 2004.
2. G. Gerig, O. Kubler, R. Kikinis, and F. A Jolesz. Nonlinear Anisotropic Filtering of MRI Data. IEEE Trans. Med Imag., vol. 11, pp. 221-232, 1992.
3. M.Pokric, N.A.Thacker and D.C. Williamson, Noise Filtering and Testing for MR Using a Multi-Dimensional Partial Volume Model. Proc. MIUA, 21-24, London, 2004.
4. A. Buades, B. Coll, J.M Morel, "A review of image denoising algorithms, with a new one" , Multiscale Modeling and Simulation (SIAM interdisciplinary journal), Vol 4 (2), pp 490-530, 2005.
5. C.A. Cocosco, V. Kollokian, R.K.-S. Kwan, A.C. Evans : "BrainWeb: Online Interface to a 3D MRI Simulated Brain Database". NeuroImage, vol.5, no.4, part 2/4, S425, 1997 – Proceedings of 3-rd International Conference on Functional Mapping of the Human Brain, Copenhagen, May 1997.
6. A. Samsonov and C. Johnson. Noise-Adaptive Nonlinear Diffusion Filtering of MR Images With Spatially Varying Noise Levels. Magnetic Resonance in Medicine 52:798-806, 2004.
7. J. C. Wood and K. M. Johnson. Wavelet Packet Denoising of Magnetic Resonance Images: Importance of Rician Noise at Low SNR. Magnetic Resonance in Medicine 41:631-635, 1999.
8. R. D. Nowak, Wavelet-based Rician noise removal for magnetic resonance imaging, IEEE Transactions on Image Processing, 8:10, 1408 -1419, 1999.
9. L.P. Yaroslavsky, Digital Picture Processing - An Introduction, Springer Verlag, 1985.

Skin Pattern Analysis for Lesion Classification Using Gradient of Skin Line Direction

Zhishun She [a, b], Y Liu [a] and J Tatler [a]

[a] School of Science and Technology, NEWI, University of Wales, Wrexham, LL11 2AW, U.K.
[b] School of Electronic & Information Engineering, Southwest University, 400715, China.

Abstract: Skin line direction is extracted by skin pattern analysis as a measure of skin pattern disruption caused by the lesion. Mathematical morphological processing is used to determine the internal and external sub-images of lesion boundary. Gradient variability of skin line direction is calculated over the lesion boundary. The mean of this gradient is chosen for lesion classification. Test results show that the gradient of skin line direction is promising for distinguishing malignant melanoma from benign lesions.

1 Introduction

As the incidence of skin cancer is rapidly increasing throughout the world, computer automatic diagnosis (CAD) of skin lesions using early indicators is crucial to primary clinical care. This is because that detection of malignant melanoma at an early stage could reduce the risk of mortality and increase the chance of prognosis considerably. In order to achieve this, a set of features enabling accurate differentiation between benign and malignant skin lesions needs to be investigated.

In the CAD system it is known that the existed features for lesion classification are ABCD features, that is, Asymmetry, Border irregularity, Colour variegation and Diameter of lesion [1]. However the observation that skin patterning tends to be disrupted by malignant but not by benign skin lesions suggests that measurements of skin pattern disruption on simply-captured white light optical skin images could be a useful contribution to a diagnostic feature set [2]. Previous work using skin line direction for lesion classification was encouraging [3]. The mean of skin line direction over lesion area is compared to that of normal skin to measure the skin pattern disruption. However the skin line direction may vary locally if the lesion area is large. Therefore it is more appropriate to measure the gradient variability of skin line direction across the lesion boundary.

In the work described in this paper two sub-images which are inside and outside the lesion boundary are extracted by mathematical morphological processing. Then the difference of skin line directions between inside and outside lesion border is calculated locally across the lesion boundary. The mean of these differences is chosen for classification. The result of a classification test on a set of clinical skin lesions including 8 malignant and 14 benign lesions is encouraging.

2 Direction of Skin Line

There are three steps to generate skin pattern [3]. Firstly the skin image is smoothed by convolving with a 9×9 window with a value of 1/81 and then this smoothed image is subtracted from the original. The result is further enhanced by histogram equalization (histogram linearization) and finally the output is inverted so that the skin lines are seen as high intensity.

From skin pattern image, the gradient vector is calculated. Then an averaged tensor is computed on a local patch centred at pixel (i, j) with a size of $M \times N$ pixels. Skin line direction in this local area is determined by the eigenvector corresponding to minimum eigenvalue of this tensor [3].

As two examples, Figure 1 shows, from top to bottom, the optical image, skin pattern and skin line directions. The left is that of a benign compound naevus. The right is that of a malignant melanoma with regression. The resolution of image is 230×350 pixels and the pixel size is 0.04mm×0.04mm. They are chosen to have a reasonable skin area surrounding the lesions. Skin line directions were calculated over a

Corresponding author: z.she@newi.ac.uk

local patch with 16×16 pixels. It indicates that there is no obvious disruption of skin line direction in Figure 1 (e) and a disturbance of skin line direction across lesion boundary in Figure 1 (f). However this qualitative description can not be used in the CAD system and a quantitative measure for skin pattern disruption needs to be developed.

3 Gradient of Skin Line Direction across Lesion Boundary

A snake-based edge detection technique is used to determine the lesion boundary [4]. The detected boundary segments the image into skin area A_s and lesion area A_l. After the boundary is detected, a binary image, $A(k,l)$ is generated which satisfies $A(k,l)=1$ if $(k,l)\in A_l$ and $A(k,l)=0$ elsewhere. Then the mathematical morphology is applied to extract the two sub-images which are outside and inside the lesion boundary.

3.1 Extraction of External Sub-image

Dilation is an expansion operation. It is used to determine the external sub-image. If B denotes a structuring element, then the external sub-image is determined by

$$D = A \oplus B - A \tag{1}$$

where $\oplus$ represents the dilation.

3.2 Extraction of Internal Sub-image

Erosion is the dual process of dilation. It is a shrinking operation and is utilized to extract the internal sub-image. The internal sub-image is expressed as

$$E = A - A \otimes B \tag{2}$$

where $\otimes$ is the erosion.

3.3 Gradient Calculation

After the internal and external sub-images are extracted, the nearest neighbour patch E_j in the external sub-image is found for each patch I_i in the internal sub-image. Then the gradient of skin line direction across the lesion boundary is computed locally by

$$\nabla\phi_i = |\phi_{I_i} - \phi_{E_j}| \qquad \forall i \tag{3}$$

where ϕ_{I_i} and ϕ_{E_j} are the skin line directions over patches I_i and E_j, respectively. The mean of this gradient over the lesion boundary is chosen for lesion classification.

As an example, the optical image of the benign compound naevus as shown in Figure 1 (a) was segmented using a fast snake algorithm [4] and its binary image is displayed in Figure 2 (a). The binary image was dilated and eroded with a structuring element of disk. The diameter of disk is 16. The difference between the dilated image and the binary image, that is, the external sub-image is shown in Figure 2 (b) and the difference between the binary image and the eroded image which is the internal sub-image is given in Figure 2 (c). The gradients of skin line directions were calculated for the two cases shown in Figure 1. They were 11.62^o for the benign compound naevus and 21.34^o for the malignant melanoma with regression, indicating a quantitative measure of skin line disruption by a malignant rather than a benign lesion.

4 Classification Results

The image set used in the experiment of this technique contains examples of several types of lesion including 8 melanomas and 14 compound or junctional naevi. The gradients of skin line direction were calculated and the distribution of gradient is shown in Figure 3 (a), showing some degree of separation between the benign and malignant lesions. There is a tendency to a smaller gradient of skin line direction in the benign lesion images compared to that in the malignant melanoma images. Figure 3 (b) shows the receiver operating characteristic (ROC) curve and the area under the curve is about 0.86, which is a positive improvement over the previous classification results [3].

Feature of skin line direction was combined with that of skin line intensity [5] to enhance the classification accuracy. The scatter-plot of 22 lesions in the two-dimensional feature space is given in Figure 4 (a) which demonstrates that malignant lesions usually have greater gradients in skin line direction and intensity across the lesion border and thus they can be discriminated from benign lesions. A ROC curve was plotted using a straight line (a linear classifier) with a slope of -0.36. Figure 4 (b) shows the ROC curve and the area under the curve is approximately 0.90, indicating an encouraging classification result.

5 Conclusions

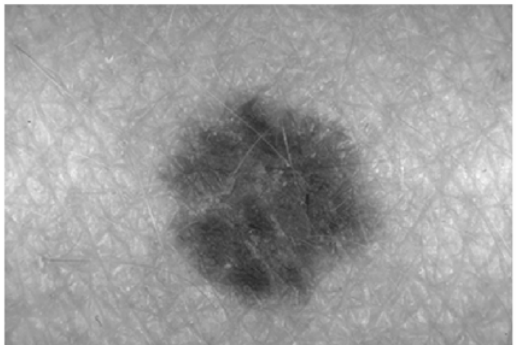
A new skin pattern characterisation, gradient of skin line direction across the lesion boundary has been developed and suggested as a feature to measure the disruption of skin pattern caused by lesions. The mathematical morphological processing has been used to extract the sub-images which are inside and outside the lesion border for gradient calculation. Classification results indicate that the gradient of skin line direction tends to be small for benign naevus but not for malignant melanoma leading that this is a promising feature for lesion classification. Future work includes to investigate the performance of the proposed classifier using a large image set.

Acknowledgements

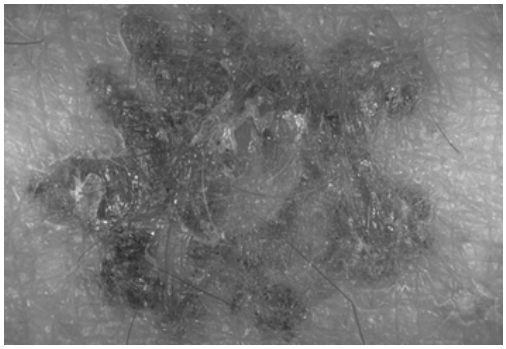
The optical skin lesion images were provided by M.Dickson, V.Wallace and Dr. J.Bamber of the Physics Department, Clinical Research Centre, Royal Marsden Hospital, Sutton. Permission to use them is gratefully acknowledged. This work was funded by EPSRC grants GR/M72371 and GR/M72289. Recently it has been supported by natural science foundation project of CQ CSTC.

References

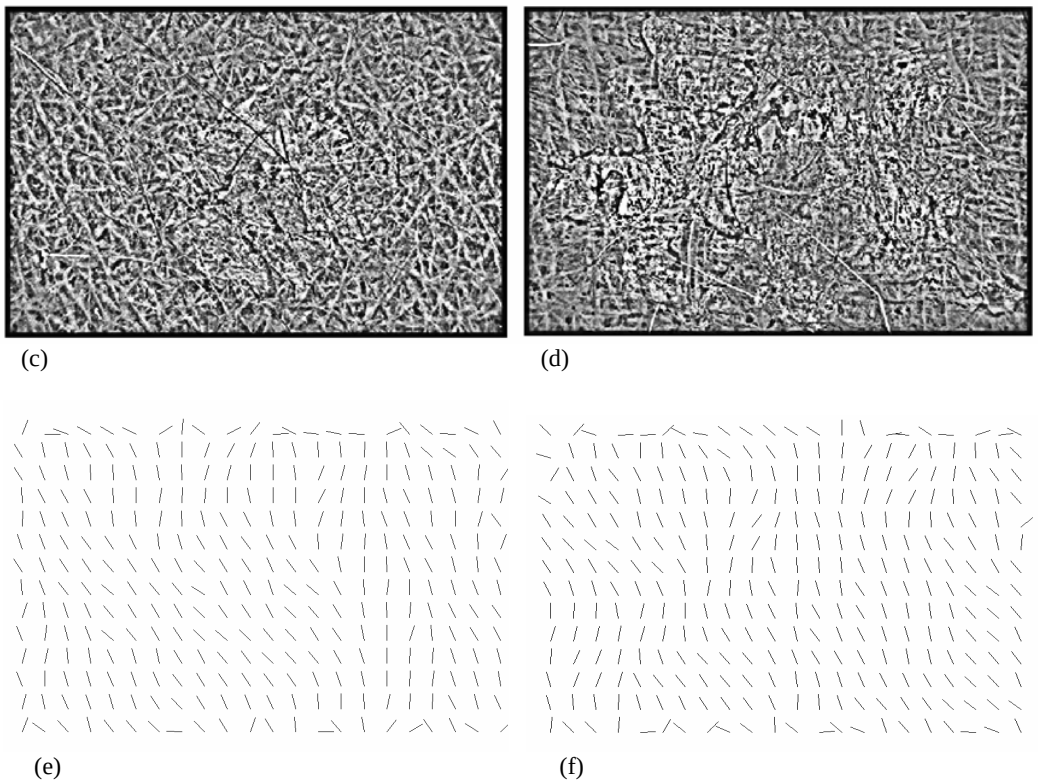
1. R.M. Mackie, *An illustrated guide to the recognition of early malignant melanoma*, Blackwood Pillans & Wilson Ltd., Edinburgh and Department of Dermatology, University of Glasgow, 1986.
2. P.N.Hall, "Clinical diagnosis of melanoma", *Diagnosis and management of melanoma in clinical pratice*, pp. 35-52, Springer-Verlag, New York, 1992.
3. Z. She & P.Fish, "Analysis of skin line pattern for lesion classification", *Skin Research and Technology*, **9**, pp. 73-80, 2003.
4. Z. She & P.Fish, "Boundary detection of skin lesion using a fast snake algorithm", In *Proceedings of 16th biennial international EURASIP Conference*, pp. 295-297, 2002.
5. Y. Liu & Z. She, "Skin pattern analysis for lesion classification using skin line intensity", In *Proceedings of Medical Image Understand and Analysis*, pp.207-210, 2005.



(a)



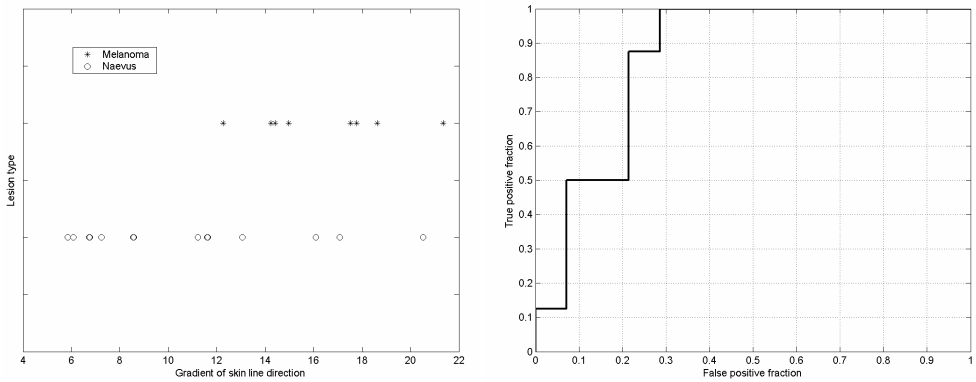
(b)



(c) (d)

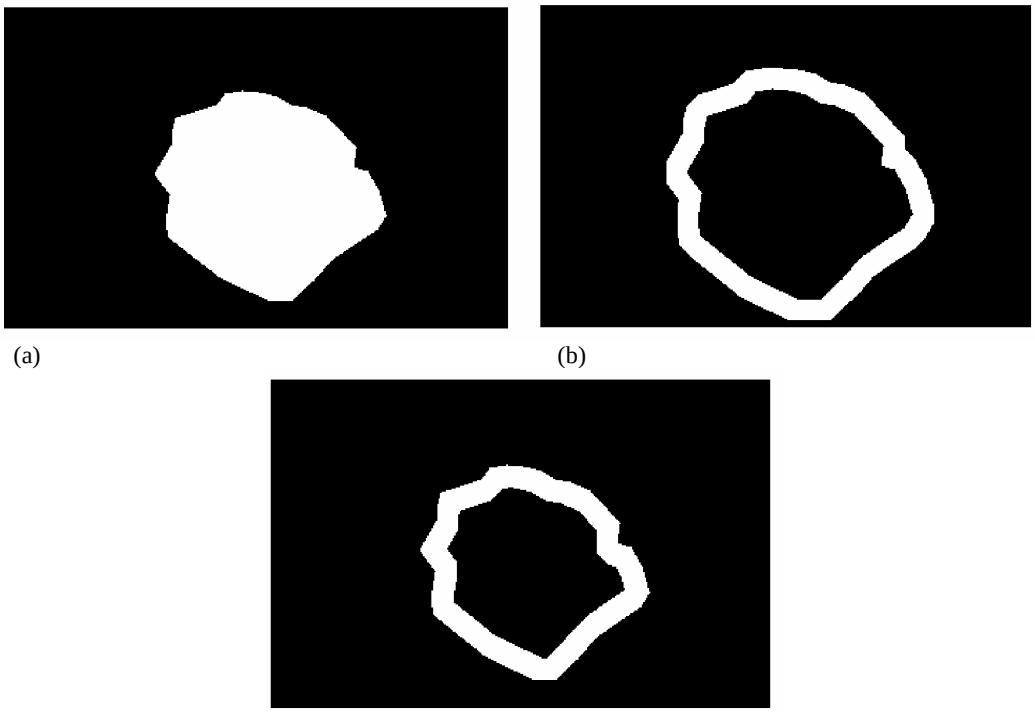
(e) (f)

Figure 1 Top to bottom: optical, skin pattern and skin line direction image. Left: a compound benign naevus. Right: a malignant melanoma with regression.



(a) (b)

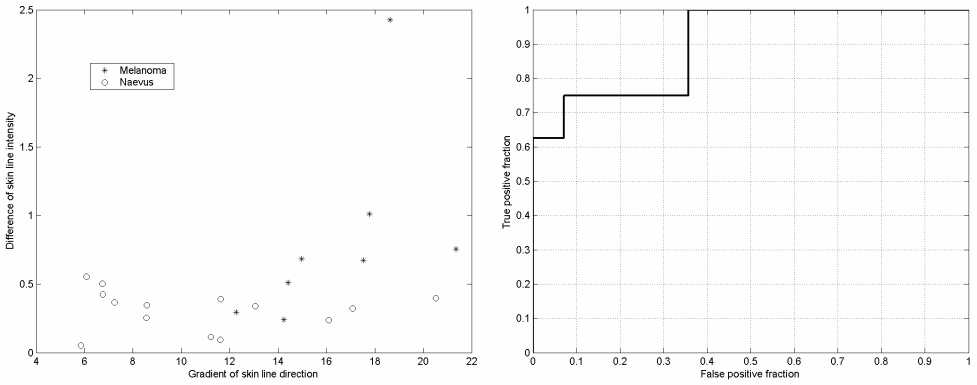
Figure 3 Lesion classification using gradient of skin line direction: (a) Scatter plot of skin lesions and (b) ROC curve for lesion classification.



(a) (b)

(c)

Figure 2 Morphological processing: (a) segmented binary image, (b) sub-image outside the lesion border and (c) sub-image inside the lesion border.



(a) (b)

Figure 4 Lesion classification using combined features: (a) Scatter plot of skin lesions and (b) ROC curve for lesion classification.

Correspondence Points Identification by Using a Minimum Entropy Approach

Zihua Su, Tryphon Lambrou and Andrew Todd-Pokropek

Department of Medical Physics & Bioengineering, University College London, Malet Place Engineering Building, Gower Street, London, UK, WC1E 6BT

Abstract. Statistical Shape Models (SSMs) have wide applications in image processing analysis both for image segmentation and registration. A main problem in shape modelling has been the need for identification and/or placing landmarks. This paper is mainly inspired by the Minimum Description Length (MDL) approach to build SSMs automatically. We propose a new approach for automatic landmark identification by optimizing an Entropy based cost-function. The algorithm is tested on different 2D datasets and compared with models constructed by the original MDL method by using the "one master example" approach. It can be seen from our preliminary results that the new proposed Minimum Entropy approach conveys better results on the measurement of Generalization Ability and Specificity. Also, the new proposed algorithm shows good potential in solving the so-called "Pile Up" problem in MDL.

1 Introduction

Statistical Shape Modelling (SSM) has turned out to be a very successful and effective tool in image segmentation and image interpretation [1][2]. The basic idea behind SSM is that from a given training set of known shapes to be able to describe new formerly unseen shapes, which still are realistic. The shape is traditionally described using landmarks on the shape boundary. Therefore, a major drawback for SSM is that a solid correspondence between the shapes must be established. Manual landmark placement by an expert is sometimes subjective and error prone. During the last few years, different researchers have attempted different approaches for tackling this problem [3][4]. The most successful method is developed by Davies and Cootes, *et al* [5]. They developed a technique by using the Minimum Description Length (MDL). They argued that a "simple" model preserved the correspondence between datasets. Therefore, MDL is used to measure how much information is used to describe the model and the "simplest" model is found during optimization. Later, a method was developed by Ericsson [6], which, by deriving the gradient of MDL cost function, can speed up the optimization efficiently. Inspired by the above authors, we propose a new method to finding correspondence points across datasets automatically by using another branch of information theory. Also, similar to the approach proposed by Ericsson, we report an approximation of the gradient to optimize our cost function.

This paper is organised as follows. In section 2, the necessary background on shape models, MDL and our proposed Minimum Entropy Model (MEM) are given. The gradient of our method is derived and an algorithm to minimize the cost function is proposed too. In section 3, we show some results based on the measurement of Generalization Ability, and Specificity between our proposed method and MDL. Compactness is not introduced here, as it is an application based measurement. An in-depth analysis on the results of facial profiles is conducted, since in this case, MDL encountered the so-called "Pile Up" problem. The conclusions and future work is shown in section 4.

2 Shape Model

2.1 Preliminaries

The standard shape model used in e.g. [1] describes the variations of a set of points as linear combinations of a set of basis vectors found by Principal Component Analysis (PCA).

$$\phi = \bar{\phi} + Qb \tag{1}$$

Equation 1 means that the new shape (ϕ) is represented by the mean shape $(\bar{\phi})$ plus the shape variations (Qb). Where, Q is the eigen-vector captured by PCA and b is a weighting component vector.

Zihua Su , email: z.su@medphys.ucl.ac.uk

2.2 MDL

In 2D case, our shape is represented by a number of parameterised curves $c_i : [0,1] \to R^2$. Therefore, finding the correspondence among the shapes is equivalent to re-parameterising the curves. MDL is a concept derived from information theory and is, in simple words, the effort that is needed to send the model bit by bit. The basic idea is to minimise the length of the message (the description length) required to transmit a full description of the training set. The cost function is composed by the simplified version of total description length [7]:

$$DL = \frac{n_s}{2}\left(\sum_{i=1}^{n_g} \log\left(\frac{\lambda_i}{\lambda_c}\right) + 1 + \sum_{j=n_g+1}^{n_g+1+n_s} \frac{\lambda_j}{\lambda_c}\right) \tag{2}$$

Where, λ_i are the eigen-values of the SSM, λ_c is a cut off constant, n_s is the number of samples and n_g is the number of eigen-values which are lager than λ_c. By optimizing this cost function, the correspondence is preserved in a minimum status. From Equation 2, we can see that the gradient of the description length is directly related to the derivatives of λ_i. A stunning inference was calculated by Ericsson [6], who gives out the derivatives of λ_i with respect to the *n*th control node θ_{mn} on m_{th} sample by using Single Value Decomposition (SVD).

$$\frac{\partial \lambda_k}{\partial \theta_{mn}} = 2 S_k U_{mk} V_K \frac{\partial \phi_m}{\partial \theta_{mn}}, \quad \frac{\partial DL}{\partial \theta_{mn}} = \sum_{\lambda_k \geq \lambda_c} \frac{1}{\lambda_k} \frac{\partial \lambda_k}{\partial \theta_{mn}} + \sum_{\lambda_k < \lambda_c} \frac{1}{\lambda_c} \frac{\partial \lambda_k}{\partial \theta_{mn}} \tag{3}$$

Here S, U and V are output of SVD applied onto shape covariance matrix. For example, the SVD of ϕ gives $\phi = USV^T$. If ϕ is zero centred, V will correspond to Q in Equation 1 and the diagonal $S^T S$ gives the eigen-values λ_k. In the implementation, $\frac{\partial \phi_m}{\partial \theta_{mn}}$ can be estimated using differential approximation. By searching the steepest down hill direction, the cost function can be optimized more efficiently than the previously used simplex method in Davies [8].

2.3 Minimum Entropy Approach

Rather than using the MDL to control the process of shape re-parameterization, we propose using Entropy which is another branch of information theory. In Physics, the minimum Entropy status has attracted many scientists with its "mystery" properties. Physicists describe the minimum Entropy status as the most "ordered" status. We argue that this "ordered" status generalizes the correspondence. The method can also be interpreted in another way. It can be very useful to think of finding correspondence as trying to maximize the amount of shared information in all images within datasets. In a qualitative sense, we may say that if images with correct correspondence are correctly aligned then the corresponding structure will overlap furthest. On the other hand, if their correspondence is poor, the images will be out of alignment, in which case, we will have duplicate information on the aligned images. According to Cootes *et al* [1], the probability of the element in weighting vector *b*, in Equation 1, complies the form of a Gaussian Distribution. Therefore, the probability of *b* and Entropy *H* is in Equation 4. Where *n* is the number of modes used in Equation 1.

$$P_i = \frac{1}{\sqrt{2\pi\lambda_i}} e^{-b_i^2/2\lambda_i}, \; H = -\sum_{i=1}^{n} P_i \log P_i \tag{4}$$

As opposed to Independent Component Analysis [9], PCA assumes variations which are orthogonal to each other but not necessarily independent to each other. Thus, the cost function can not be composed of the sum of all *H*s directly. We assign a weighting component onto each *H*. As a result, the cost function is shown in equation 5.

$$CostFunction = \sum_{i=1}^{n} \lambda_i H_i \tag{5}$$

Where, the weighting parameter is the corresponding eigen-value.

Since *H* only changes in a small interval, if the control points on the shape boundary are condensed enough. We assume that *H* is "constant" during the optimization process. Consequently, the derivative of the cost function with respect to control points can be approximated with the form of Equation 6.

$$\frac{\partial(CostFunction)}{\partial\theta_{in}} \approx \sum_{i=1}^{n} H_i \frac{\partial\lambda_i}{\partial\theta_{in}} = \sum_{i=1}^{n} H_i 2s_n u_{in} V_n \frac{\partial\phi_i}{\partial\theta_{in}} \tag{6}$$

Where, H_i is the Entropy extracted from the i_{th} variable on each weighting vector b .

3 Experiments and discussion

In experiments, initial landmarks are placed equally along the boundary in a hierarchically manner. For example, the level $i+1$ node is in between level i nodes. For all experiments are optimization is done up to level 3 and the rest of landmarks are equally placed between control nodes. For validating this algorithm, experiments were performed on different benchmarking 2D datasets from Ericsson [10]. The datasets are metacarpals, femurs, silhouettes of heads and some shark shapes. Due to lack of ground truth, it is quite difficult to measure the accuracy of the corresponding points. Next we are going to introduce some "Benchmark" comparison criteria which are developed by Davies *et.al* in [8]. They are the Generalization Ability, and Specificity. By using these two objective measurements, researchers can compare different approaches objectively and effectively.

3.1 Comparison Methods

Briefly, the Generalization Ability (*G*) of a model measures its capability to represent unseen instances of the class of the object model. The *G* is measured from training sets using leave-one-out reconstruction. The reconstruction error for each model is $\varepsilon_i^2(M)$ with *M* is the number of the modes used in the model. Here, n_s is the number of training sets, σ is the sample standard deviation of *G(M)*.

$$G(M) = \frac{1}{n_s}\sum_{i=1}^{n_s}\varepsilon_i^2(M), \sigma_{G(M)} = \frac{\sigma}{\sqrt{n_s - 1}} \tag{7}$$

Specificity (S) is the ability to measure if the model can generate instances of the object that are close to those in the training set.

$$S(M) = \frac{1}{n_s}\sum_{i=1}^{N}\left|\phi_i(M) - \phi_i'\right|, \sigma_{G(M)} = \frac{\sigma}{\sqrt{n_s - 1}} \tag{8}$$

Where ϕ_i is a shape example generated by the model (choosing the value of the weighting parameter randomly in the range over the training set), ϕ_i' is the closest member in the training set to ϕ_i, σ is the sample standard deviation of *S(M)*and *N* is the number of samples, in our cars, *N* is 100000.

We can conclude that if a model *A* is better in G and S, it will achieve lower value on that ability measurement, for more details, it is recommended to read [8].

3.2 Pile up problem

This problem was first introduced in [8][10]. The problem is that during optimization, points can possibly move into one location that will actually attain a meaningless minimum avoiding describing other parts of the shape. Davies tries to use a fixed master example to avoid piling up. Hans [7] has argued that one fixed example shape would not always hold right, since the statistic weight of other examples can overweight the single master example easily in some case with large number of dataset. Therefore, Hans has tried to add a compensate component, extracted from curvature, onto the MDL cost function to avoid this pile up problem. However, assigning weighting parameters for this external cost function according to different applications is not an easy task and no evidence has been shown how to solve this problem.

3.3 Experimental Results

For validating this algorithm, experiments were performed on four different datasets; they are Metacarpals, Femurs, Silhouettes and Shark contours. In each experiment, Both MDL with one master example and MEM are implemented and they all use its corresponding gradient for acceleration purpose. In the experiments of Metacarpals, Femur and Sharks contours, correspondence points captured by MEM are shown in Figure 1; comparisons results on Generalization ability and Specificity are shown in Figure 2.

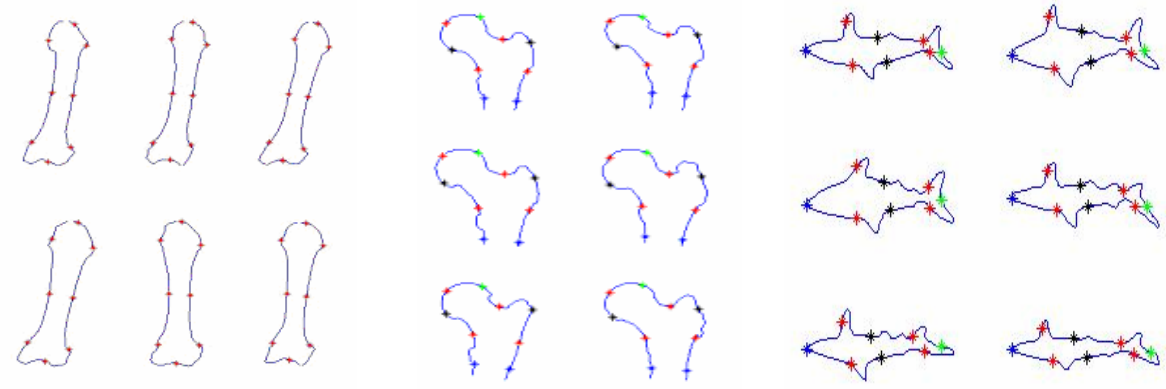


Figure 1. From left to right shows the correspondence points captured by MEM, here only control nodes are shown, landmarks between nodes are omitted. According to different application, the nodes can be optimized hierarchically.

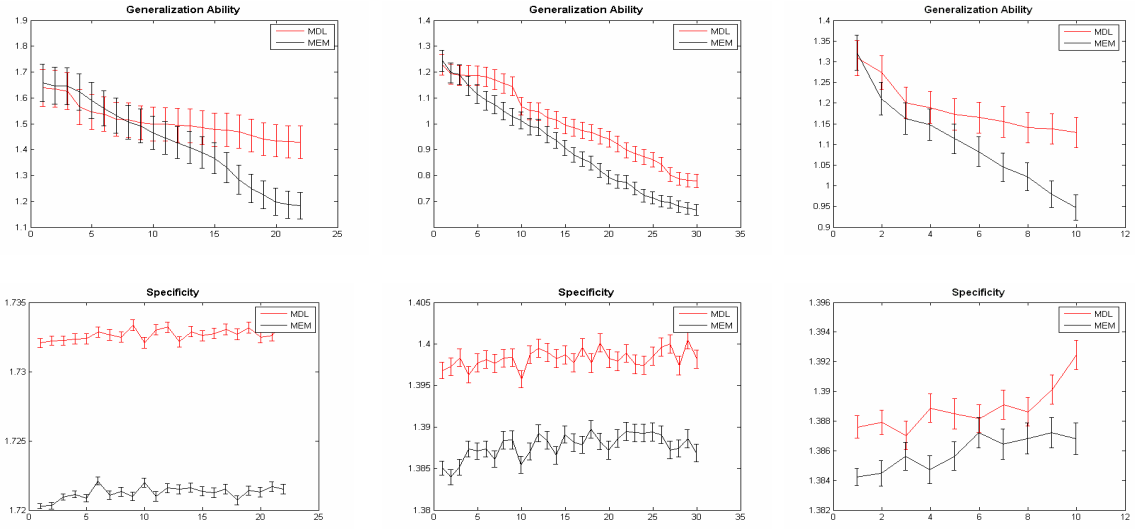


Figure 2. From left to right shows the comparison results on the datasets of metacarpals, femurs and shark shapes respectively. From top to bottom shows Generalization ability and Specificity. X Axis represents the number of mode used in SSM and Y Axis represents error level of normalized shapes

When applying MDL in the dataset of silhouettes, we encountered the problem of "Pile Up". Two control nodes at the bottom of the chin area piled up when the MDL cost function finally converged. We show both the MDL and MEM results in Figure 3. The inherent difference between MDL and MEM can be derived from the above comparison. When piling up occurred during optimization, the MDL cost function will achieve a minimum either local of global because the MDL cost function only describes part of the shape model. On the contrary, "Pile up" status is a Maximum status for MEM, either local or global. For example, when "pile up" happened; the probability (P_i) of each element in vector b will have the trend to be equal to each other. It can be seen from Equation 4 that H will be a maximum when P_i are the same. In this sense, we can conclude that MEM favors a distributed correspondence and MDL prefers a condensed correspondence.

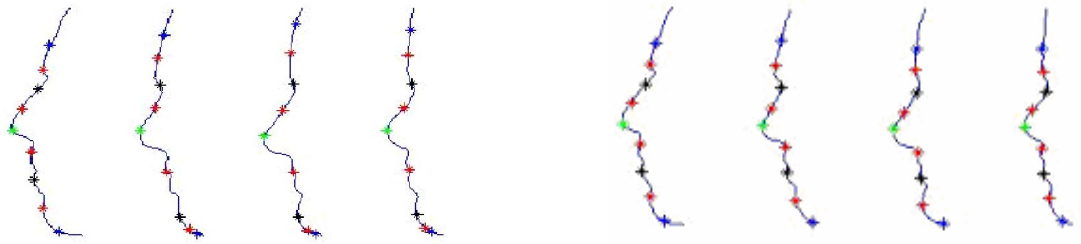


Figure 3. From left to right, the first four results are captured by MDL. Results showed one step before MDL cost function finally converged. It can be seen that even with the first example fixed, still control points at the chin area will piled up eventually. The last four results are from MEM. It can be seen that compared with MDL, MEM "favour" distributed correspondence.

4 Discussions and Conclusions

In this paper, we proposed a new method to compose the homogenous function to build statistical shape model automatically. The method is successfully implemented and applied to different datasets for validation purposes. In the mean time, comparisons with MDL based on the measurement of Generalization ability and Specificity are shown. Our preliminary results showed that our approach have better Generalization Ability and Specificity. Further more analysis showed that, the MDL met "Pile up" problem but our proposed method did not when applied the two methods to dataset of facial profiles. In this sense, we conclude that our proposed method has better potential to fight the "pile up" problem.

In the future, we are going to incorporate more datasets into our algorithm for validation. Our goal is developing a robust automatically shape model building method. Though, moving the method from 2D to 3D is not straight forward, 3D method is still one of our future objectives.

Acknowledgements

We acknowledge Hans for providing MDL comparison method and Ericsson for providing shark contours datasets.

References

1. T.F. Cootes, D. Cooper, C.J. Taylor and J. Graham. "A Trainable Method of Parametric Shape Description." *Proc. British Machine Vision Conference pub. Springer-Verlag*, pp 54-61. 1991.
2. M.A. Walker, J.R. Highley, M.M. Esiri, B. McDonald, H.C. Roberts, S.P. Evans, T.J. Crow: "Estimated neuronal populations and volumes of the hippocampus and its subfields in schizophrenia." *Am J Psychiatry*; Vol 159: pp 821–828, 2002.
3. A.C.W. Kotcheff and C.J. Taylor, "Automatic construction of eigenshape models by direct optimization" *Medical Image Analysis,* pp303-314, 1998.
4. A. Hill and C.J. Taylor. "Automatic landmark generation for point distribution models," *5th British Machine Vision Conference,* pp 429-438, 1994.
5. R. H. Davies, T. F. Cootes, C. J. Twining, and C. J. Taylor. "An information theoretic approach to statistical shape modelling." *In 12th British Machine Vision Conference*, pp 1–10, 2001.
6. A. Ericsson and K. Åström. "Minimizing the description length using steepest descent." *In Proc. British Machine Vision Conference, Norwich, United Kingdom*, vol 2, pp 93–102, 2003.
7. H.H. Thodberg, " A Minimum Description length Approach to statistical shape modelling " *Lecture note in computer science* Vol 2732, pp 525-537, 2003.
8. R.H. Davies, T.F. Cottes, and C.J. Taylor. "A minimum description length Approach to statistical shape modeling" *IEEE transaction medical Imaging*, Vol 21, pp 525-537, 2002.
9. M. Uzumcu, A.F. Frangi, J.H.C. Reiber and B.P.F. Lelieveldt, "Independent component analysis in statistical shape models", *Processing of the SPIE,* PTS 1-3, pp 375-383, 2003.
10. A. Ericsson and J. Karlsson, ``Benchmarking of algorithms for automatic correspondence localisation", *British Machine Vision Conference*, BMVC September 2006.

Segmentation of Colour Microscopy Images using an Adaptive Fuzzy C-Means Algorithm

Daan Zhu[a*] and Yanong Zhu[b†]

[a]National Institute for Medical Research, London, NW7 1AA, UK
[b]School of Computing, University of Leeds, Leeds, LS2 9AU, UK

Abstract.
This paper presents an Adaptive Fuzzy Clustering Algorithm (AFCM) for the segmentation of colour microscopy images. Intelligent location of a range of colours based on a manually selected reference colour is implemented using the fuzzy clustering algorithm. The colours within the range have similar colour properties in terms of hue, saturation and intensity, where the size of the colour range is determined by an adaptive Lagrange parameter λ. Experiments on a small set of microscopy images show that the AFCM algorithm is able to produce a good agreement with the segmentation performed by human experts. Meanwhile, the presented adaptive algorithm gives better segmentation results than the standard non-adaptive FCM algorithm.

1 Introduction

Segmentation of regions of interest is an important and often difficult task in the domain of biomedical image analysis. This is especially true for colour microscopy images, where the regions are often scattered and small in size, and the edges are often poorly defined. Due to such characteristics, the application of model-based segmentation methods (e.g., Active contour models) and traditional thresholding methods (e.g., [1]) are much limited.

Fuzzy clustering algorithms have been successfully applied in various tasks such as colour lip image segmentation [2] and medical diagnostic systems [3]. Deng et al. [4] used a fuzzy algorithm to quantize colour into several colour classes representing the different regions in an image. The original image colours were then replaced pixel by pixel, with their corresponding colour class labels, thus forming a class-map of the image. More colour classes produce a more accurate representation of the original image. The traditional Fuzzy C-Means (FCM) algorithm [5] and most of its variations find the colour centroids and the corresponding colour classes globally. Using this method, at least 30 colour classes were needed to have a reasonable representation of our histological images. Pham and Prince [6] used an Adaptive FCM (AFCM) algorithm to segment gray magnetic resonance images. Dave and Bhaswan [7] detected the feather of ellipses using an adaptive FCM algorithm and the results were compared with those of the Hough transform. However, it appears that there have been a limited number of attempts to apply adaptive fuzzy methods to the segmentation of colour histological images. There are 16 million different colours in a true colour image with three 8-bit colour channels. An efficient algorithm for the classification of the colours of interest is essential for colour image segmentation using adaptive fuzzy methods.

In this paper we present an Adaptive FCM algorithm for the segmentation of colour microscopy images. The presented adaptive algorithm differs from those traditional ones in two ways. Firstly, in our algorithm a local clustering analysis, rather than the global analysis as in most fuzzy clustering applications, is performed to achieve a better specificity. Secondly, we adopt a Lagrange parameter, namely λ, to regularise the manually defined reference colour. Our algorithm approaches optimum cluster centroids through an adaptive step. Under this regularisation only the colour centroids which have similar colour properties to the reference colours are selected.

2 Method and Mathematical Formulation

2.1 Colour models

Colour models are mathematical representations of a set of colours, with the most popular models being Red-Green-Blue (RGB), Hue-Saturation-Intensity (HSI) and Hue-Saturation-Brightness (HSB). The well known RGB model is used by most visualisation systems. It is defined by a cube subspace on a Cartesian coordinate system. In this cube the RGB values are at three corners, while the CMY (Cyan, Magenta, Yellow) are at the others. The HSI model defines a colour space in terms of three constituent components. It is a nonlinear transformation of the RGB colour space. In this

*Email:dzhu@nimr.mrc.ac.uk
†Email:yanong@comp.leeds.ac.uk

model, H describes a pure colour, S measures the degree to which a pure colour is diluted with white light. The HSI colour model is represented by a double prism subspace based on the Maxwell triangle generated by the RGB cube subspace. The third colour model HSB, also known as Hue-Saturation-Value (HSV), is similar to HSI. HSB is also a nonlinear transformation of the RGB colour space. The forward and backward transforms between RGB to HSI (or HSB) are well-defined (see, e.g., [8]). In our work, microscopy images are converted from RGB to HSI or HSB, in which the optimum colour centroids are identified. The inverse transform is applied to covert the image into RGB colour space. Our data indicates that segmentation performed in either the HSB or HSI gives similar results.

2.2 Adaptive Fuzzy C-Means Algorithm

The standard FCM obtains segmentation via fuzzy pixel classification. Unlike hard classification methods which force pixels to belong exclusively to one class, FCM allows pixels to belong to multiple classes with varying degrees of membership. This leads one its major advantages, which is if a local pixel is corrupted by noise the segmentation will only be affected to a fractional extent, while for hard classifications, the entire segmentation may be changed [6]. In the standard FCM algorithm for scalar data, the following objective function is minimised by identifying the optimum membership functions u_k and centroids v_k:

$$J_{FCM} = \sum_{i,j} \sum_{k=1}^{C} u_k(i,j)^q \|x(i,j) - v_k\|^2 \tag{1}$$

where $u_k(i,j)$ is the membership value at pixel location (i,j) for class k such that $\sum_{k=1}^{C} u_k = 1$, $x(i,j)$ is the observed image intensity at location (i,j), and v_k is the centroid of class k. The total number of classes, C, is normally defined in advance. The parameter q ($q > 1$) is a weighting exponent on each fuzzy membership and determines the value of fuzzy classification (in most applications q is set to 2).

Equation (1) denotes the objective function for grey scale image segmentation. To consider the colour channels in colour images, the above equation is extended to

$$J_{FCM} = \sum_{i,j} \sum_{k=1}^{C} u_k(i,j)^q \|x_l(i,j) - v_{l,k}\|^2 \tag{2}$$

where l ($1 \leq l \leq 3$) are the colour channels. Equation (2) is the partition FCM algorithm. The norm operator $\|\cdot\|^2$ represents the standard Euclidean distance that defines the distance of the pixel values (three channels) to the current centroid point.

By introducing an adaptive parameter λ and the reference colour $\tilde{x_l}$ in Equation (2), we have

$$J_{AFCM} = \sum_{i,j} \sum_{k=1}^{C} u_k(i,j)^q \|x_l(i,j) - v_{l,k}\|^2 + \lambda \|\tilde{x_l} - v_{l,k}\|^2 \tag{3}$$

where λ is the Lagrange parameter used to regularise the reference colour in the global classification sets. The selection of λ is an adaptive step (see below). The reference colour is selected by the user before the classification.

2.3 Iterative Solution

The objective function (3) can be minimised in a fashion similar to the standard FCM algorithm. Taking the first derivatives of J_{AFCM} with respect to $u_k(i,j)$, $v_{l,k}$ and $\tilde{x_l}$ and setting them equal to zero results in three necessary conditions for J_{AFCM} to be at a minimum. Using these conditions, the process of image segmentation using our AFCM algorithm can be described as follows:

1 Manually select of a reference colour and convergence tolerance ε.

2 Set the target number of centroids and initial values for the membership matrix, $v_{l,k}, k = 1, \cdots, C$

3 Update the membership matrix:

$$u_k(i,j) = [\sum_{m=1}^{c} \frac{\|x(i,j) - v_{l,k}\|}{\|x(i,j) - v_{l,m}\|}^{\frac{2}{m-1}}]^{-1} \tag{4}$$

for all (i,j) and $k = 1, \cdots, c$; and $l = 1, 2, 3$, m is any real number greater than 1.

4 Update the centroids:

$$v_{l,k} = \frac{\sum_{k=1}^{c} u_k(i,j)^m \cdot x_l(i,j)}{\sum_{k=1}^{c} u(i,j)^m} + \lambda \cdot \frac{\sum_{k=1}^{c} u_k(i,j)^m \cdot \tilde{x_l}}{\sum_{k=1}^{c} u(i,j)^m}$$

where $l = 1, 2, 3$.

5 Estimate the errors:

Two types of errors are involved. The simplest approach is to calculate the iterative difference.

$$Error_{t+1}(v) = \sum_{k=1}^{c} \sum_{l=1}^{3} \|v_{l,k}^{(t+1)} - v_{l,k}^{(t)}\|^2 \tag{5}$$

where t denotes the iterations. Another approach to the convergence error is by calculating the Euclidean distances between the values $x_k(i,j)$ and the centroid $v_{l,k}$ in the same classified class.

$$Error_{t+1}(v) = \sum_{k=1}^{c} \|x_k(i,j)^{(t+1)} - v_{l,k}^{(t+1)}\|^2 \tag{6}$$

6 If the $Error < \varepsilon$, the iteration is converged, then quit. Otherwise, go to step 3.

In practice, the membership matrix u is initialised randomly, and the rows of the matrix must be normalised by their sum. The initial membership matrix may not always guarantee to produce the optimum centroids, which is one of the known disadvantages of using FCM/AFCM. Two different ways are used to estimate the iterative errors in Step 5. In equation 5, the error is estimated by computing the Euclidean distance between the current and previous centroids. This type of error ensures convergence under any conditions. However difficulties arise when equation 6 is used as the convergence criterion. The optimum solution of FCM and AFCM depends on the initial membership matrix u and the number of centroids. In cases that the FCM or AFCM fail to yield optimum centroids, Equation 6 will never reach the convergence tolerance and the algorithm will iterates infinitely. In practice, equation 5 is used to determine the termination of the iterations, and equation 6 can be used to judge whether a set of optimum centroids have been reached.

2.4 Calculation of the Lagrange Parameter

We use a Lagrange parameter, λ, to constrain the colours under the regularisation of the reference colour. AFCM degrades to FCM if λ is zero. Pham and Prince [6] introduced two regularisation terms with two *fixed* Lagrange parameters. In our work, the value of λ is adjusted based on the previous iterations. The calculation of λ is an adaptive step based on the estimated centroids of the previous iteration. This criterion provides the basis for the method used to select λ. With increasing number of iterations the value of λ should converge to a small positive value. Calculation of λ leads to

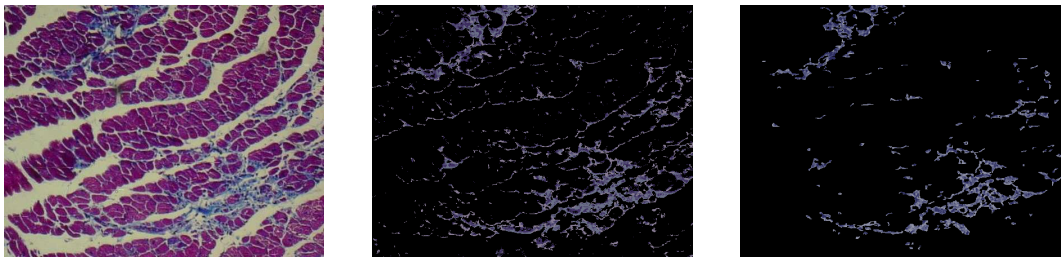
$$\lambda = \frac{\sum_{i,j} \sum_{k=1}^{C} u_k(i,j)^q \|x_l(i,j) - v_{l,k}\|^2}{\|\tilde{x_l} - v_{l,k}\|^2} \tag{7}$$

This calculation is self-adapting. At the beginning of the iterations, the value of λ provides a heavy penalty to the classification. On the contrary, when the iteration converges to the optimum solution, λ contributes less to the reference colour.

3 Experiments and Results

Both FCM and AFCM are implemented on a Pentium 4®, CPU 3.2GHz workstation with 2.0Gb of RAM. Figure 1(a) shows an original image of a pathologic feature in mammalian heart tissue that often needs to be quantified. Hearts were excised from knockout mice known to develop dilated cardiomyopathy [9] characterised by cardiomyocyte cell death and extensive fibrosis (i.e. deposition of collagen in damaged areas) and embedded in paraffin. Histological sections were stained with Massons trichrome that selectively stains collagens (blue), whereas cardiomyocytes and other cellular components assume a reddish colour. The blue regions need to be segmented and the amount needs to be quantified. Figure 1 compares the segmentation results for the purple colour using FCM and AFCM. Figure 1(c) matches more closely the results obtained from medical experts performing the purple colour selection manually.

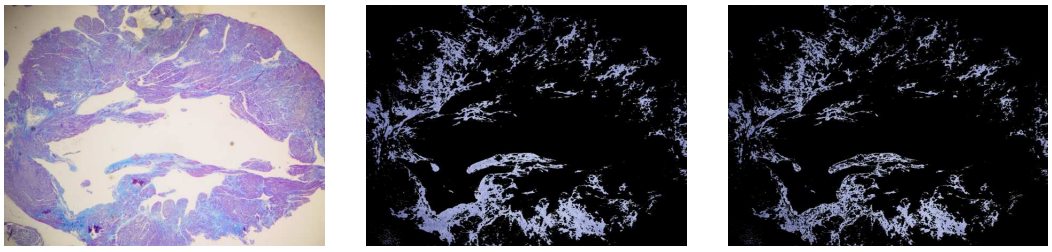
Figure (3) compares the convergence rates between FCM with different number of centroids, and AFCM. The convergence tolerance here is set to 1.0×10^{-3}. Figure 3(a) and 3(b) present the convergence curves of FCM with 30 and



(a) Original image (b) FCM results using 4 colour centroids selected from 30 (c) AFCM segmentation results using 4 centroids

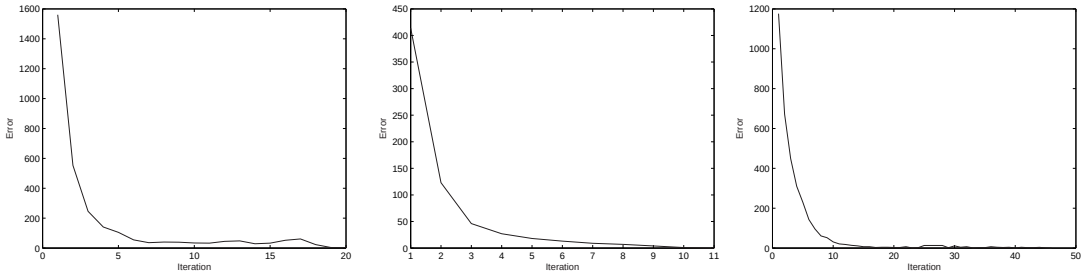
Figure 1. Comparison FCM and AFCM segmentation

5 centroids. After 20 and 11 iterations respectively, FCM with 30 and 5 centroids reach the convergence tolerance. However it takes 46 iterations for AFCM to reach the same convergence tolerance. These suggest that AFCM has a slower convergence rate than traditional FCM.



(a) Original image (6.7a) (b) AFCM segmentation using 5 centroids (c) AFCM segmented result using 7 centroids

Figure 2. Comparison AFCM segmented results with different centroids



(a) Error plot of FCM, 30 centroids (b) Error plot of FCM, 5 centroids (c) Error plot of AFCM, 5 centroids

Figure 3. Comparison of convergence rates between FCM and AFCM

Figure (2a) presents a cross-section of the left ventricle of a pathogenic heart. The blue colour stains the collagens that are characteristic of tissue fibrosis. Figures (2b) and (2c) show the segmentation results of AFCM using 5 and 7 centroids, respectively. The segmented structures in both cases are similar, but it can be observed that by using use of 7 centroids a better definition of the fibrotic network is produced. The results of using 7 centroids relate more closely to the original pathological image and match the manual analysis performed by an experienced pathologist. This is further confirmed in Table (1), which compares the results between manual segmentation performed by a pathologist and our AFCM algorithm for a number of histological images. This comparison indicates that, the AFCM using more centroids produced a better agreement with the expert. However, it should be noticed that increasing the number of centroids will also increasing the computational load correspondingly.

4 Conclusions

We have presented an Adaptive Fuzzy C-Means (AFCM) algorithm to classify colours on biomedical images. In order to efficiently segment the image, a reference colour is manually selected before segmentation. Based on the hue-saturation-intensity of the reference colour, a regularisation factor Lagrange parameter, is further introduced into the AFCM. The experimental results suggest that by using this algorithm, colour centroids are found with similar colour

Table 1. Comparison manual and AFCM segmented results

Images IDs	Manual segmentation	AFCM using 5 centroids	AFCM using 7 centroids
6.7a	17.99%	18.98%	18.03%
4.2a	1.57%	3.8%	2.8%
9.7a	7.31%	3.34%	8.0%
9.7b	4.62%	1.6%	3.56%
SC13.1b	23.1%	21.57%	22.3%

properties. In comparison to the traditional FCM, the convergent rate of the AFCM is slower for the same number of centroids.

Using our AFCM, the colour of the image is classified based on the reference colour instead of a global classification in normal FCM/AFCM methods. Therefore fewer centroids are needed. The computational complexity and space complexity (memory usage) of both AFCM and FCM are $O(n^2)$ and $\Omega(n)$ respectively, where n is the number of centroids. This means using fewer centroids will save on computational time, and this should compensate the relatively lower convergence rate of the presented AFCM algorithm.

A disadvantage of both FCM and the AFCM is that the classified results may rely on the initial membership matrix and the number of centroids. Unfortunately, we never know the optimum number of centroids in advance. Several different algorithms, such as principal component analysis (PCA) and subtractive clustering [10], may be reviewed to address this issue. These may improve the automatic estimation of the optimum number of centroids.

References

1. F. H. Y. Chan, E. K. Lam & H. Zhu. "Adaptive thresholding by variational method." *IEEE Transactions on Image Process* **17(3)**, pp. 468–473, 1998.
2. A.Liew, S.H.Lenung & W.H.Lau. "Segmentation of color lip images by spatial fuzzy clustering." *IEEE Transactions on Fuzzy Systems* **11**, pp. 542–549, 2003.
3. S.Albayrak & F.Amasyali. "Fuzzy c-means clustering on medical diagnostic systems." *The 12th International Turkish Symposium on Artificial Intelligence and Neural Networks* 2003.
4. Y.Deng, B.Manjunath & H.Shin. "Color image segmentation." *IEEE Conference on Computer Vision and Pattern Recognition* 1999.
5. G.Bueno, R.González, J.González et al. "Fuzzy colour c-means clustering for pattern segmentation in histological images." *The 3rd European Medical and Biological Engineering Conference* 2005.
6. D.L.Pham & J.Prince. "Adaptive fuzzy segmentation of magnetic resonance images." *IEEE Transactions on Medical Imaging* **18**, pp. 737–751, 1999.
7. R.N.Dave & K.Bhaswan. "Adaptive fuzzy c-shells clustering and detection of ellipses." *IEEE Transactions on Neural Networks* **3**, pp. 643–662, 1992.
8. R. C. Gonzalez & R. E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001.
9. D. Milner, G. Weitzer, D. Tran et al. "Disruption of muscle architecture and myocardial degeneration in mice lacking desmin." *Journal of Cell Biology* **134**, pp. 1255–1270, 1996.
10. W. Liu, C. Xiao, B. Wang et al. "Study on combining subtractive clustering with fuzzy c-means clustering." *Machine Learning and Cybernetics* **5**, pp. 2659– 2662, 2003.

DIARETDB1 diabetic retinopathy database and evaluation protocol

T. Kauppi[1] V. Kalesnykiene[2] J.-K. Kamarainen[1,3] L. Lensu[1] I. Sorri[2] A. Raninen[2]
R. Voutilainen[2] J. Pietilä[4] H. Kälviäinen[1] H. Uusitalo[2]
[1]Machine Vision and Pattern Recognition Research Group, Lappeenranta University of Technology
[2] Department of Ophthalmology, Faculty of Medicine, University of Kuopio
[3] Centre for Vision, Speech and Signal Processing, University of Surrey
[4] Perimetria Ltd., Helsinki, Finland

Abstract. For a particularly long time, automatic diagnosis of diabetic retinopathy from digital fundus images has been an active research topic in the medical image processing community. The research interest is justified by the excellent potential for new products in the medical industry and significant reductions in health care costs. However, the maturity of proposed algorithms cannot be judged due to the lack of commonly accepted and representative image database with a verified ground truth and strict evaluation protocol. In this study, an evaluation methodology is proposed and an image database with ground truth is described. The database is publicly available for benchmarking diagnosis algorithms. With the proposed database and protocol, it is possible to compare different algorithms, and correspondingly, analyse their maturity for technology transfer from the research laboratories to the medical practice.

1 Introduction

Diabetes has become one of the rapidly increasing health threats worldwide [1].Only in Finland, there are 30 000 people diagnosed to the type 1 maturity onset diabetes in the young, and 200 000 people diagnosed to the type 2 latent autoimmune diabetes in adults [2]. In addition, the current estimate predicts that there are 50 000 undiagnosed patients [2]. Proper and early treatment of diabetes is cost effective since the implications of poor or late treatment are very expensive. In Finland, diabetes costs annually 505 million euros for the Finnish health care,and 90% of the care cost arises from treating the complications of diabetes [3]. These alarming facts promote the study of automatic diagnosis methods for screening over large populations.

Fundus imaging has an important role in diabetes monitoring since occurrences of retinal abnormalities are common and their consequences serious. The eye fundus is sensitive to vascular diseases and therefore fundus imaging is also considered as a candidate for non-invasive screening. The success of this type of screening approach depends on accurate fundus image capture, and especially on accurate and reliable image processing algorithms for detecting the abnormalities. Numerous algorithms have been proposed for fundus image analysis by many research groups (e.g. [4–6]). However, it is impossible to judge the accuracy and reliability of the approaches because there exists no commonly accepted and representative fundus database and evaluation protocol. With a proper protocol, it would be possible to evaluate the maturity and state-of-the-art of the current methods, i.e. produce the best achieved sensitivity and selectivity rates. For example, commonly accepted strict guidelines for the evaluation of biometric authentication methods, such as the FERET and BANCA protocols for face recognition methods [7,8], have enabled the rapid progress in that field, and the same can be expected in medical image processing related to diabetic retinopathy detection.

The main contribution of this work is to report a publicly available diabetic retinopathy database and protocol, DIARETDB1, which contains training and test set images, expert marked finding ground truth, and a tool kit for method evaluation. The protocol is demonstrated by a baseline method included to the tool kit. This study provides the means for the reliable evaluation of automatic methods for detecting diabetic retinopathy.

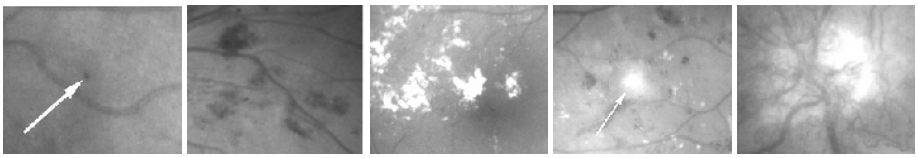
2 Diabetic retinopathy

In the type 1 diabetes, the insulin production in the pancreas is permanently damaged, whereas in the type 2 diabetes, the person is suffering from increased resistance to insulin. The type 2 diabetes is a familial disease, but also related to limited physical activity and lifestyle [1]. The diabetes can cause abnormalities in the retina (diabetic retinopathy), kidneys (diabetic nefropathy), and nervous system (diabetic neuropathy) [9]. The diabetes is also a major risk factor in cardiovascular diseases [9].

The diabetic retinopathy is a microvascular complication of diabetes, causing abnormalities in the retina, and in the worst case, blindness. Typically there are no salient symptoms in the early stages of diabetic retinopathy, but their number and severity predominantly increase with time. The diabetic retinopathy typically begins as small changes in the retinal capillaries. The first detectable abnormalities are mircroaneurysms (Ma) shown in Fig. 1(a), which are local distensions of the retinal capillary and when ruptured, cause intraretinal hemorrhage (H) shown in Fig. 1(b). The disease

severity is classified as mild non-proliferative diabetic retinopathy when the first apparent microaneurysms appear in the retina [10]. With time, the retinal edema and hard exudates (He) shown in Fig. 1(c) appear because of the increased permeability of the capillary walls. The hard exudates are lipid formations leaking from these weakened blood vessels. This state of the retinopathy is called moderate non-proliferative diabetic retinopathy [10]. However, if the above-mentioned abnormalities appear in the central vision area (macula), the condition is called diabetic maculopathy [1]. As the retinopathy advances, the blood vessels become obstructed which causes microinfarcts in the retina. These microinfarcts are called soft exudates (Se) shown in Fig. 1(d). When a significant number of intraretinal hemorrhages, soft exudates, or intraretinal microvascular abnormalities are encountered, the state of the retinopathy is defined as severe non-proliferative diabetic retinopathy [10].

The severe non-proliferative diabetic retinopathy can quickly turn into proliferative diabetic retinopathy when extensive lack of oxygen causes the development of new fragile blood vessels [10]. This is called neovascularisation shown in Fig. 1(e), which is a serious eye sight threatening state. The proliferative diabetic retinopathy may cause sudden loss in visual acuity, or even permanent blindness due to vitreous hemorrhage or tractional detachment of the central retina. After the diabetic retinopathy has been diagnosed, regular monitoring is needed due to the progressive nature of the disease. However, broad screenings cannot be performed due to the fact that the fundus image examination requires attention of medical experts. For the screening, reliable automatic image processing methods must be developed.



(a) (b) (c) (d) (e)

Figure 1. Abnormal diabetic retinopathy caused eye fundus findings (in colour): (a) microaneuryms (marked with an arrow); (b) hemorrhages; (c) hard exudates; (d) soft exudate (marked with an arrow); (e) neovascularization.

2.1 Current evaluation practises

In medical diagnosis, the medical input data is usually classified into two classes, where the disease is either present or absent. The classification accuracy of the diagnosis is assessed using the sensitivity and specificity measures. Following the practises in the medical research, the fundus images related to the diabetic retinopathy are evaluated by using sensitivity and specificity per image basis. Sensitivity is the percentage of abnormal funduses classified as abnormal, and specificity is the percentage of normal fundus classified as normal by the screening. The higher the sensitivity and specificity values, the better the diagnosis. Sensitivity and specificity can be computed as [11]:

$$\text{sensitivity (SN)} = \frac{T_P}{T_P + F_N}, \text{ specificity (SP)} = \frac{T_N}{T_N + F_P} \quad (1)$$

where T_P is the number of abnormal fundus images found as abnormal, T_N is the number of normal fundus images found as normal, F_P is the number of normal fundus images found as abnormal (false positives) and F_N is the number of abnormal fundus images found as normal (false negatives). Sensitivity and specificity are also referred to as the true positive rate (TPR) and true negative rate (TNR), respectively.

3 Evaluation database

A necessary tool for reliable evaluations and comparisons of medical image processing algorithms is a database including a dedicatedly selected set of high-quality medical images which are representative for the problem and have been verified by experts. In addition, information about the medical findings, the ground truth, must accompany the image data. An accurate algorithm should take the images as input, and produce output which is consistent with the ground truth. In the evaluation, the consistency is measured, and algorithms can be compared based on these measures. In the following, we describe the images and ground truth for the diabetic retinopathy database DIARETDB1.

3.1 Fundus images

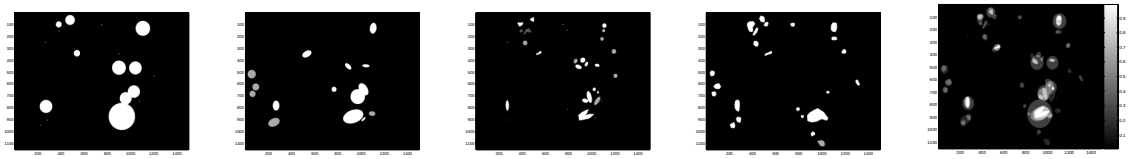
The database consists of 89 colour fundus images of which 84 contain at least mild non-proliferative signs (Ma) according to all experts participated in the evaluation. The images were taken in the Kuopio university hospital. The images were selected by the medical experts, but their distribution does not correspond to any typical population, i.e., the data is biased and no a priori information can be devised from it. The diabetic retinopathy abnormalities in the database are relatively small, but they appear near the macula which is considered to threaten the eyesight.

Images were captured with the same 50 degree field-of-view digital fundus camera [1] with varying imaging settings and without preprocessing. The intensity of the camera flash was controlled by the photographer, but other settings such as shutter speed, aperture, and gain were controlled by the fundus camera. The images contain a varying amount of imaging noise, but the optical aberrations and photometric accuracy (colour or intensity) are the same. Therefore, the photometric variance over the visual appearance of the different retinopathy findings can be considered as small. In the absence of a well-founded procedure for camera characterisation, the data correspond to a good (not necessarily a typical nor the most general) practical situation, where the images are comparable, and can be used to evaluate the general performance of diagnostic methods. The general performance corresponds to the situation where no calibration is performed (no correspondence to the physical measurements), but where the images correspond to commonly used imaging conditions, i.e. the conditions typically encountered in hospitals. This data set is referred to as "calibration level 1 fundus images". A data set taken with several fundus cameras containing different amounts imaging noise and optical aberrations is referred to as "calibration level 0 fundus images". To publish an ultimate tool for the evaluation of diabetic retinopathy, which is the research group's main objective, it is necessary to study the different calibration levels. Therefore, the current database was not aimed to be statistically representative, but as one step in the development process.

3.2 Ground truth

Independent markings from 4 medical experts were collected by using a software tool provided for image annotation. The computer displays used in the collection process were not calibrated. A person with medical education and solid experience in ophthalmology was considered as an expert. The experts were asked to mark the areas related to the microaneurysms, hemorrhages, and hard and soft exudates. The experts were instructed to avoid marking the findings so that the borders of the marked areas contain any pixels belonging to the finding. The experts were further instructed to report their confidence and especially annotate the single most representative point for each finding. The ground truth confidence, $\{< 50\%,\ > 50\%,\ 100\%\}$, represented the certainty of the decision that a marked finding is correct. The experts were taught to use the image annotation tool, but they were not instructed how to make the annotations to prevent a biased scheme; the medical experts learnt their own best practises.

The uninstructed collection process caused significant differences between the medical experts as shown in Fig. 2. Therefore, it was not possible to use the expert information as such as the ground truth. However, using the original data the expert knowledge was fused for a better spatial accuracy and suppression of outliers. The fusion was performed on a pixel basis using the reported confidence levels. Several different approaches for fusing the markings are possible, e.g., voting, minimum, maximum or sum of values. The first three provide binary classifications, but a normalised average provides values in the range $[0, 1]$ (Fig. 2(e)). It should be noted that the markings do not provide any absolute ground truth, but reveal how the medical experts analyse and interpret the retinopathy from the digital fundus images. Not to discard any information, the approach using the average was selected since it provides a linear confidence scale, and in the evaluation, the confidence level can be fixed to one or several values. In DIARETDB1, the confidence level is fixed to $conf_{GT} = 0.75$.



(a) (b) (c) (d) (e)

Figure 2. Example of ground truth (colour decodes the confidence): (a) expert 1; (b) 2; (c) 3 (d) 4; (e) mean.

3.3 Training and test set

89 images were manually assigned into categories representing the progressive states of retinopathy: normal, mild, moderate and severe non-proliferative, and proliferative. Using the categories, the images were divided into the representative training (28 images) and test sets (61 images). In the training set with $conf_{GT} = 0.75$, 18 images contain hard exudates, 6 soft exudates, 19 microaneurysms, and 21 hemorrhages. In the test set with $conf_{GT} = 0.75$, 20 images contain hard exudates, 9 soft exudates, 20 microaneurysms, and 18 hemorrhages

[1] ZEISS FF 450^{plus} fundus camera with Nikon F5 digital camera

4 Evaluation protocol

4.1 Performance measures

In the literature, the sensitivity and specificity values are typically reported since they correspond to the current medical practice and have straightforward interpretations in the medical terms. These values provide the means for analysing how many diseased and how many healthy patients are correctly diagnosed with a provided method. From the method comparison point of view, however, these two values are not feasible since the two distributions overlap and the true accuracy is always a trade-off. Furthermore, in our evaluation protocol we have selected two evaluation principles: 1) the evaluation is image-based and 2) is done separately to different diabetic retinopathy findings (Sec. 2). The first principle is justified by the fact that it corresponds to the medical practice where decisions are "patient-based". It should be noted that spatial area (pixel-wise) based evaluation can be useful in method development. The second principle is due to the practical fact that most researchers concentrate only on one or several finding types. The image acquisition process affects the evaluation protocol only through the ground truth, i.e., the marking accuracy of experts depends on the quality of image acquisition.

4.1.1 ROC

For a proper comparison the sensitivity and specificity values must be combined into a form which can describe the behaviour over different combinations of the values. Receiving operating curve (ROC) is a natural selection due to its popularity and proved applicability in similar computer vision tasks, such as face recognition [12], object class recognition [13] and medical research [14]. The ROC provides a graphical representation for sensitivity (TPR) and 1-specificity (FPR) trade off. The ROC curve provides the means for the optimal analysis when the problem is to find the best method parameters for the task or compare performances irrespective to operating conditions. In our evaluation we adapted the practises from [13], where each method is required to provide a score for each test image. A high score corresponds to a high probability that a finding is present in the image. From provided scores a ROC curve can be automatically generated.

4.1.2 Weighted error rate (WER)

The ROC curve is a reliable method for comparisons, but often method ranking is also needed, where single performance values are necessary. Single valued measures should be derived from the ROC curve, e.g. by computing an equal error rate (EER) ($TPR = TNR$) or a total area under roc curve (AUC). Here we prefer the more interpretable EER. The EER measure however assumes equal penalties for the both false positives and negatives, which is not typically the case in the medical diagnosis. Therefore, we adapt a measure utilised in [7, 8], where the two measures, sensitivity (SN) and specificity (SP), are combined to a weighted error rate defined as

$$WER(R) = \frac{FPR + R \cdot FNR}{1 + R} = \frac{(1 - SP) + R \cdot (1 - SN)}{1 + R} \quad . \tag{2}$$

In (2) $R = \frac{C_{FNR}}{C_{FPR}}$ is the cost ratio between FPR and FNR. In the DIARETDB1 protocol the following values are computed: $\mathbf{WER(10^{-1})}$ (FNR is by order of magnitude less harmful), $\mathbf{WER(1)}$ (equal) and $\mathbf{WER(10)}$ (FNR is by order of magnitude more harmful). The values are computed using the nearest true ROC points with no interpolation.

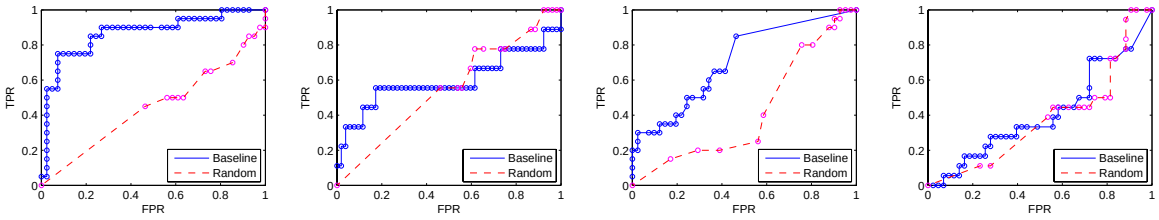
5 Evaluation example

5.1 Baseline method

The baseline method is a straightforward implementation of the method proposed by Schwerdt and Crowley [15] and is based on a principle that the findings can be detected based on their photometric information (colour). The baseline method was implemented to study the discrimination power of colour information, and to demonstrate the protocol's usage. The method utilises the R and G colour channels and intensity normalises them by $R + G + B$. A class description for each finding type, F_i, is formed by computing their colour histograms $h_{F_i}(r, g)$. The finding histograms are computed from the normalised pixels in the 8x8 neighbourhood of the most representative points marked by the experts. The background histogram is formed by computing the total colour histogram, $h_{total}(r, g)$, of each observed image. By using a ratio of the two histograms Schwerdt and Crowley derived a Bayesian formula for classifying pixels. We have manually fixed a posterior threshold for every finding type and the image score is the sum of pixels passing the threshold.

5.2 Results

Evaluation results for the baseline method are shown in Fig. 3 and Table 1. The ROC graphs also include random classification results (random score for each test image). The baseline method performs moderately only for the hard exudates. The results however are the current baseline and every reported method should outperform them.



(a) (b) (c) (d)

Figure 3. The DIARETDB1 ROC baseline: a) hard exudates; b) soft exudates; c) microaneurysms; d) hemorrhages.

Table 1. DIARETDB1 baseline performance measures.

Baseline method									
	FPR	FNR	**WER**			FPR	FNR	**WER**	
He	0.8049	0	0.0732	R = 0.1	Se	1.0000	0	0.0909	R = 0.1
	0.0732	0.2500	0.1616	R = 1		0.1731	0.4444	0.3088	R = 1
	0.0244	0.4500	0.0631	R = 10		0	0.8889	0.0808	R = 10
Ma	0.4634	0.1500	0.1785	R = 0.1	H	0.8837	0.2222	0.2824	R = 0.1
	0.4634	0.1500	0.3067	R = 1		0.1628	0.8333	0.4981	R = 1
	0	0.8000	0.0727	R = 10		0	1.0000	0.0909	R = 10

6 Conclusion

A novel evaluation database and protocol was proposed for evaluating and comparing methods for automatic detection of diabetic retinopathy. All technical documentaton, images, ground truth and Matlab functionality for the baseline and computing performance measures are publicly available on the project's web page (IMAGERET [2], `http://www.it.lut.fi/project/imageret/`). Research groups investigating diabetic retinopathy detection methods are encouraged to report their results using the DIARETDB1 database and according to its evaluation protocol.

References

1. G. von Wendt. *Screening for diabetic retinopathy: Aspects of photographic methods*. Ph.D. thesis, Karolinska Institutet, 2005.
2. Finnish Diabetes Association. "Development programme for the prevention and care of diabetes.", 2001. ISBN 952 5301-13-3.
3. Finnish Diabetes Association. "Programme for the prevention of type 2 diabetes in Finland.", 2003. ISBN 952-5301-36-2.
4. M. Niemeijer, B. van Ginneken, J. Staal et al. "Automatic detection of red lesion in digital color fundus photographs." *IEEE Trans. on Medical Imaging* **24(5)**, pp. 584–592, 2005.
5. A. D. Fleming, S. Philip, K. A. Goatman et al. "Automated microaneurysm detection using local contrast normalization and local vessel detection." *IEEE Trans. in Medical Imaging* **25(9)**, pp. 1223– 1232, 2006.
6. A. Osareh, M. Mirmehdi, B. Thomas et al. "Classification and localization of diabetic-related eye disease." In *Proc. of 7th European Conference on Computer Vision (ECCV)*, pp. 502–516. 2002.
7. P. Phillips, H. Moon, S. Rizvi et al. "The FERET evaluation methodology for face-recognition algorithms." *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22(10)**, 2000.
8. E. Bailliere, S. Bengio, F. Bimbot et al. "The BANCA database and evaluation protocol." In *Proc. of the Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, pp. 625–638. 2003.
9. M. Niemi & K. Winell. "Diabetes in Finland, prevalence and variation in quality of care.", 2006.
10. C. P. Wilkinson, F. L. Ferris, R. E. Klein et al. "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales." *Ophthalmology* **10(9)**, pp. 1677–1682, September 2003.
11. T. Walter, J.-C. Klein, P. Massin et al. "A contribution of image processing to the diagnosis of diabetic retinopathy - detection of exudates in color fundus images of the human retina." *IEEE Trans. on Medical Imaging* **21**, pp. 1236–1243, 2002.
12. P. Grother, R. Micheals & P. Phillips. "Face recognition vendor test 2002 performance metrics." In *Proc. of 4th Int. Conf. on Audio- and Video-based Biometric Person Authentication*. 2003.
13. M. Everingham & A. Zisserman. "The pascal visual object classes challenge 2006 results." Workshop in ECCV06, 2006.
14. T. A. Lasko, J. G. Bhagwat, K. H. Zou et al. "The use of receiver operating characteristic curves in biomedical informatics." *Journal of Biomedical Informatics* **38**, pp. 404–415, 2005.
15. K. Schwerdt & J. Crowley. "Robust face tracking using color." In *Proc. of 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition*. 2000.

[2] Supported from the FinnWell technology program (40430/05, 40039/07) of the Finnish Funding Agency for Technology and Innovation, Tekes.

Evaluation of Internal Carotid Artery Segmentations

D. Stampouli[a], M.R.Varley[a], T. Lambrou[d], C.F. Walshaw[b], A.P. Jones[c], R.W. Bury[b], L.-K. Shark[a]

[a]ADSIP Research Centre, University of Central Lancashire, Preston, UK

[b]Radiology Department, Blackpool Victoria NHS Trust, Blackpool, UK

[c]North Western Medical Physics, Christie Hospital NHS Trust, Manchester, UK

[d]Department of Medical Physics & Bioengineering, University College London, London, UK

Abstract. The measurement of the internal carotid artery (ICA) is one of the most important criteria of assessment and selection of patients for internal carotid endarterectomy. A method is presented and evaluated for segmentation of internal carotid arteries using adaptive thresholding in arterial cross sections. Experimental results have been obtained for 72 cases (six different slice measurements from each of twelve arteries), and the segmentations obtained from the proposed method are evaluated against two experts' manual delineations, using area measurements and Udupa's *et al.* evaluation framework. The comparison of the manual and automated delineations show that there is a good correspondence, thereby confirming the algorithm's performance and validity.

1 Introduction

The measurement of the internal carotid artery (ICA) is one of the most important criteria for assessment and selection of patients for internal carotid endarterectomy, a surgical removal of plaque, which obstructs the passage of blood and is the most common cause of ischemic stroke. The internal carotid artery is located inside the neck and is the main supply of blood from the heart to the brain. Vessel segmentation is the process of effectively distinguishing the vessels of interest in an image from the background and the surrounding objects in order to quantify the severity of the plaque blockage. Contrast-enhanced magnetic resonance angiography (CE-MRA) is widely used for the visualization of the carotid arteries and the quantification of carotid artery disease. In current clinical practice, carotid artery disease is estimated by manual inspections of the MR Angiograms. The determination of carotid arterial area from MR angiograms is an important, though complicated, issue due to several factors; the main one being partial volume effect of contrast containing voxels near the vessel periphery with non-contrast containing vessel wall and adjacent tissues. Several authors have studied the segmentation of the carotid arteries. Yim *et al.* [1] used a semi-automated deformable model to reconstruct and measure the carotid artery. Bemmel *et al.* [2] semi-automatically segmented stenosed ICAs in CE-MRA images via level-set techniques, using a user-defined central axis as the initialization. Kim and Park [3, 4] used thresholding in high contrast axial 3D Time-of-Flight MR angiograms to separate the carotid plaque from the blood flow, and quantify the stenosed area of the artery. This approach assumes that the maximum stenosis is presented perpendicular to the axial plane, ignoring the 3D geometry of the artery. Two major studies have been recently published on methodologies for evaluating medical image segmentation techniques; Crum *et al.* [5] presented overlap measures for evaluating image registration and segmentation techniques and Udupa *et al.* [6] created a framework for evaluating image segmentation techniques. The authors previously developed a segmentation technique that uses adaptive thresholds in cross-sectional planes perpendicular to the central axis of the artery to delineate the artery [7, 8]. In this paper, the segmentation technique is evaluated using Udupa's framework. Results obtained by applying this algorithm to twelve arteries are presented, and the results are evaluated against anonymous experts' manual delineations. The results of this evaluation study indicate a good agreement with expert observers.

2 Image Acquisition and Pre-processing

Blackpool Victoria Hospital provided the MRA images used for this study. The data was obtained using a Siemens 1.0 Tesla Harmony-Siemens MR scanner. The images were acquired in the coronal plane. The original CE-MRA 3D data set was used for this work rather than the post-processed Maximum Intensity Projection (MIP) images. The MR image acquisition parameters were TR = 4.7 ms, TE = 1.8 ms and flip angle = 25°. The 3D data was partitioned into 64 contiguous slices, each of thickness 1 mm. The image acquisition involves an interpolation within Fourier space, prior to data reconstruction, to permit an acquisition time less than normally required for the 64 slices. The image matrix size was 320 × 512 pixels for an image field of view of 187.5 mm × 320 mm, resulting in a voxel size of 0.6 mm × 0.6 mm × 1.0 mm.

In the pre-processing stage from the entire MRA dataset two sub-volumes are automatically defined, each containing a set of the carotids [7]. Within each sub-volume, the artery to be processed (either left or right ICA) is automatically defined and the arterial central axis defined [7, 8]. Having identified the arterial central axis, image consecutive cross-sectional slices are taken perpendicular to it, in order the arterial cross-sectional area to be measured.

3 Methods

Popular image segmentation techniques include threshold-based methods, region-based techniques, and connectivity-preserving relaxation methods. Local pixel information is utilised for threshold-based methods, which can be very successful when the intensity levels of the objects of interest are distinctly different from the background intensity ranges, and a threshold is appropriately selected. Connected regions of the image are generated by region-based methods, by grouping neighbouring pixels that are similar according to a set of predefined criteria. Iterative region growing is a popular technique in this category: however this can produce difficulties in the initialisation and termination rule, and is costly in terms of computational load. Connectivity-preserving relaxation-based segmentation methods, usually referred to as the *active contour models* or *deformable models*, are based on some energy function. These methods start with an initial elastic curve (or surface), internal forces (elastic forces) and external forces (image and constraint forces), and attempt to dynamically conform the curve to object shapes. These forces can be the result of a functional global minimization process or based on local information. In general deformable models can not deal with topological changes, while it has problems with initialisation, speed, and poor convergence to boundary concavities.

Thresholding is a pixel classification technique that separates the pixels in the image into two classes based on a threshold value, with the classes representing the pixels that belong to the object of interest, and those that belong to the background, thereby producing a binary image from (typically) a grey scale image. The selection of an appropriate threshold value is crucial for successful segmentation, and can either be performed interactively or automatically [7]. However, in most real applications there are significant intensity and contrast variations within the input image, and local adaptive thresholding, based on local image characteristics, is preferred. The authors have previously demonstrated the need for adaptive segmentation of the carotid arteries [7], and developed a solution using a local adaptive thresholding method to achieve fine segmentation of the arteries [8]. This paper focuses on validation of the results from this method by comparing results with manual segmentation (assumed here to be the reference method) using pixel counts and Udupa's efficacy measures.

The segmentation is performed in each consecutive cross sectional slice, as taken perpendicular to the arterial central axis, adapting therefore, the thresholding criteria to the local image characteristics. In Udupa's study [6], a framework of evaluation of medical image segmentation techniques is outlined, and this framework is adopted and implemented for additional analysis of the performance of the algorithm. The efficacy of a segmentation method in a specific application domain can be assessed in terms of three factors: precision, accuracy and efficiency. Precision (also called reliability) describes how closely different measurements of the same thing agree with each other, therefore tests repeatability of segmentation taking into account the influence of all subjective action that are required to produce the result. Accuracy (also called validity) describes how closely the segmentation result agrees with the true or most probable value. Efficiency (also called viability) describes the practical viability of the segmentation method. The implemented evaluation framework consists of four elements: a) real life MRA data, b) two expert manual delineations to be used as reference segmentations to denote ground truth, c) a computed generated segmentation algorithm and d) a software tool to incorporate the evaluation metrics (as defined by Udupa) with the proposed segmentation technique.

Metrics of Precision

Let C_{O1} and C_{O2} be manual segmentations of the same object O in two repeated manual segmentation trials. Possible variabilities need to be investigated for possible pairs of C_{O1} and C_{O2} that occurred from one of the following situations:

- T_1: The same operator segments the same object in the same scene twice (assessing the intra-observer variability)
- T_2: Two operators segment the same object in the same scene once (assessing the inter-observer variability)
- T_3: The same operator segments the same object in two corresponding scenes (assessing inter-scan variability)

The third situation is not applicable for this study, but is mentioned for a matter of completeness. Therefore, for the manual segmentation method, all possible pairs of (C_{O1}, C_{O2}) for T_1 will provide for the assessment of intra-operator precision. In the same way, T_2 corresponds to assessment of inter-operator precision [6]. In general, a measure for precision for a given method M, that produces a set of segmentations C_{O1} and C_{O2} for a situation Ti is given by:

$$PR_{T_i}^{M}(\mathrm{O}) = \frac{\left|C_{O1} \cap C_{O2}\right|}{\left|C_{O1} \cup C_{O2}\right|} \qquad (1)$$

$PR_{T_i}^{M}(\mathrm{O})$ represents the amount of tissue that is common to both C_{O1} and C_{O2} as a fraction of tissue in the union of C_{O1} and C_{O2}. The precision of a method M for a given situation (i = 1,2) can be characterised by measuring the coefficient of variation or confidence intervals of the $PR_{T_i}^{M}(\mathrm{O})$ values [6]. The precision of any two segmentation methods M_1 and M_2 for each T_i can be compared by comparing the set of $PR_{T_i}^{M}(\mathrm{O})$ values by using a paired t-test [6], that is if samples are distributed normally, which would allow the use of parametric tests. Alternatively, if the samples are not distributed normally, non-parametric tests should be used.

Metrics of Accuracy

Let C_d^M be the segmentation result of the artery obtained by the automated method M and let C_{td} be the corresponding scene of 'true' delineation, as obtained by the anonymous expert observer. In this case, the False-Negative segmentation is defined as $C_{FN} = C_{td} - C_d^M$; the False-Positive segmentation is defined as $C_{FP} = C_d^M - C_{td}$; the True-Positive segmentation is defined as $C_{TP} = C_{td} \cap C_d^M$; and the True-Negative segmentation is defined as $C_{TN} = U_d - C_{td} - C_d^M$. U_d constitutes a *reference subset* with respect to which all the delineated regions (true or false) in the image can be expressed as a function of, and ideally should be defined such that $U_d = C_{TP} \cup C_{TN} \cup C_{FP} \cup C_{FN}$.
The following measures are describing the delineation accuracy of the automated method M [6].

$$\mathrm{TPVF}_d^M(\mathrm{O}) = \frac{\left|C_{TP}\right|}{\left|C_{td}^M\right|} \qquad (2) \qquad \mathrm{TNVF}_d^M(\mathrm{O}) = \frac{\left|C_{TN}\right|}{\left|U_d - C_{td}\right|} \qquad (3)$$

True positive volume fraction True negative volume fraction

$\mathrm{TPVF}_d^M(\mathrm{O})$ expresses the fraction of total amount of tissue in true delineation C_{td} that was also covered by the delineation with method M. $\mathrm{TNVF}_d^M(\mathrm{O})$ denotes the fraction of total amount of tissue in the reference region U_d that is truly not in the delineation C_{td} and was also correctly not segmented by method M [6].
The false positive and false negative volume fractions, $\mathrm{FPVF}_d^M(\mathrm{O})$ and $\mathrm{FNVF}_d^M(\mathrm{O})$, could similarly be defined [6] and the following desirable relationships for the measures above can be determined from the above equations.

$$\mathrm{FPVF}_d^M(\mathrm{O}) = 1 - \mathrm{TNVF}_d^M(\mathrm{O}) \qquad (4) \qquad \mathrm{FNVF}_d^M(\mathrm{O}) = 1 - \mathrm{TPVF}_d^M(\mathrm{O}) \qquad (5)$$

Since two of the four measures are interdependent, only two measures need to be calculated to assess the delineation accuracy of method M. Either $\mathrm{TPVF}_d^M(\mathrm{O})$ and $\mathrm{TNVF}_d^M(\mathrm{O})$ or $\mathrm{FPVF}_d^M(\mathrm{O})$ and $\mathrm{FNVF}_d^M(\mathrm{O})$ need to be specified [6].

4 Results and Validation

Twelve arteries have been analysed and quantified for the purpose of this study. Six cross-sectional slices were selected along each artery to be segmented and evaluated. To test the performance of the proposed algorithm, the results generated by the segmentation algorithm were compared against the manually segmented images which were used as reference. The size of the cross-section slice is 21 x 21 pixels, while the average arterial area for a non-stenotic part is 40 pixels, and for a stenotic part of the artery is 20 pixels. Note that in extremely narrow cases, (where the accurate quantification of stenosis is critical), cross-sectional areas as low as 4 pixels were found.

Initially the inter-observer agreement was noted. Each clinical expert has created a manual delineation of each artery twice. The mean area value from the two manual delineations is measured for each observer and is used as a comparison against the other observer. Figure 1 presents the difference in mean area measurements between the segmentations from Observer A and Observer B, against the average of the mean areas between the two observers, in the form of a Bland-Altman graph. The difference in the two means is randomly scattered around zero with the majority of the values within ± 5 pixels difference. Note that for average artery areas below 25 pixels, differences up to 10 pixels occur which would result in significant quantification differences. The average percentage difference between the area means is 11%. The difference between the two means for Observers A and B is calculated and

tested for normality using a Kolmogorov-Smirnov test [11]. The distribution of the difference of the means has been found to be normal with P-value > 0.15. A paired t-test has been used to check weather the mean areas for all slices from the two observers are the same. The mean hand-segmented area from Observer A is 43.23 pixels and standard deviation STD = 24.86, while the mean hand-segmented area from Observer B is 44.56 pixels and STD = 26.27, producing a P-value of 0.017, at 90% confidence level ($p > 0.01$). Therefore there is no significant difference between the means of Observer A and Observer B and the inter-observer variability in the area measurements is negligible.

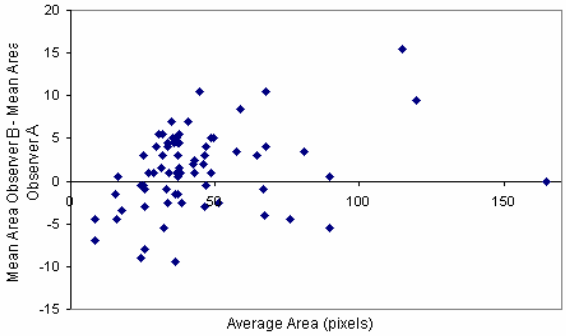


Figure 1: The difference in mean area measurement between the two observers.

	$TPVF_d^M$	$TNVF_d^M$	$PR_{T_1}^M$
Obs. A	0.644 ± 0.10	0.99 ± 0.001	0.639 ± 0.1
Obs. B	0.889 ± 0.15	0.832 ± 0.14	0.83 ± 0.14

(a)

	$TPVF_d^M$	$TNVF_d^M$	$PR_{T_1}^M$
Obs. A	0.758 ± 0.21	0.98 ± 0.02	0.68 ± 0.19
Obs. B	0.722 ± 0.23	0.99 ± 0.01	0.68 ± 0.21

(b)

Table 1: (a) Intra-observer variability of overlap measures, (b) Overlap comparison of observers with algorithm.

While analysing the intra-observer variability, Figures 2 and 3 present the difference in area measurement between the two manual segmentations performed by each of the two observers. From Figure 2 it is concluded that Observer A systematically over-segmented the slices the second time, compared to the first time. The average percentage difference between the areas measured by observer A is 43%. The standard deviation for Observer A is STD = 8. For Observer B, on the other hand, the difference in areas between the two observations is randomly scattered around zero. The average percentage difference between the areas measured by observer B is 11% and the STD = 6. The difference between the two observations is taken for each observer and tested for normality using a Kolmogorov-Smirnov test. The difference between the two observations for Observer A have been found not normally distributed ($p < 0.05$) and a non-parametric test was therefore used to check whether the difference of the two observations is centred around zero. A 1-sample Wilcoxon [10, 11] test was used and found that there is a significant difference in the area measured between the two observations for Observer A (at 95% confidence, $p < 0.05$), denoting therefore a great intra-observer variability in the case of Observer A for the area measurements. The difference between the two observations for Observer B have been found normally distributed with P-value = 0.085 ($p > 0.05$), and a t-test was used to test if there is significant difference between the means of the two observations. The mean hand-segmented area for observation 1 is 43.72 pixels while the mean hand-segmented area for observation 2 is 45.41 pixels, producing a P-value of 0.018 (Sample size: N = 72), at 90% confidence level ($p > 0.01$).Therefore insignificant intra-observer variability was noted for the area measurements of Observer B. After this analysis Observer B was selected to be used as basis for the evaluation of the segmentation algorithm. In Table 1(a) the intra-observer variability of overlap measured described above is displayed. The two numbers in the table represent the mean ± standard deviation, of the metric over all the image slices.

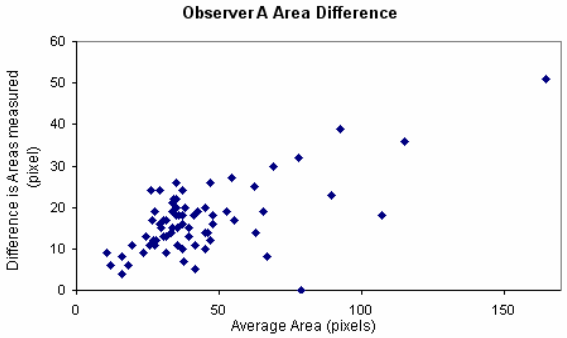


Figure 2: The difference in area measurement between the two manual delineations of Observer A

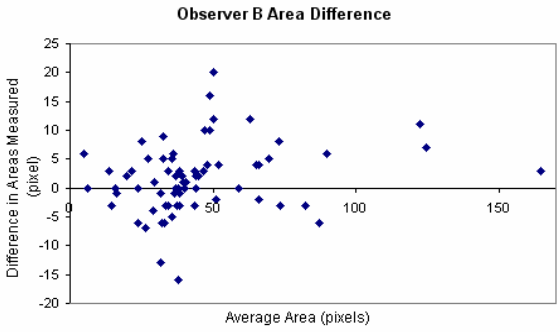


Figure 3: The difference in mean area measurement between the two manual delineations of Observer B

The area measurements obtained from the second observer are compared to the areas obtained from the algorithm. The differences from each of the two second observer manual delineation with the algorithm were examined. The distribution of the two differences proved that a non-parametric test is required; a Kolmogorov-Smirnov test was used and found that P-value < 0.05. A 1-sample Wilcoxon test was used to compare the area measurements from

Observer B's first observation with the area measurements obtained by the algorithm. The resulting P-value = 0.001 denotes that there is a significant difference between the two measurements. The mean difference of the area measurements is 4.81 pixels and STD = 13.54. Another 1-sample Wilcoxon test was used to compare the area measurements from Observer B's second observation with the area measurements obtained by the algorithm. The test produced a P = 0.006 (at 95% confidence, $p < 0.05$) which again denotes a significant difference between the two measurements. The mean difference of the two measured areas is 3.11 pixels, STD = 13.84 (N = 72). Observing the results more closely revealed that for nine from the seventy-two cross-sectional slices, the algorithm overestimated the cross-sectional area measurement due to touching adjacent arteries. Further improvement is required in the algorithm to incorporate the separation of touching arterial structures. This is currently ongoing and will be described in future publications. If those cases were rejected and excluded from the statistical tests, the new mean difference for observation 1 is 3.47 pixels, the STD = 8.76 and the new P-value is P-value = 0.003, whilst for observation 2, the P-value becomes 0.035 with 1.56 pixel mean difference (at 95% confidence) and STD = 7.98 (note that N = 63 if these nine samples are excluded from the analysis). It should be noted that the measured mean differences are small in relation to the artery cross-sectional area (typically 20-30 pixels), and would not lead to clinically significant differences in stenosis quantification. In Table 1(b) the variability of overlap measured between the experts and the algorithm is displayed for both Observers. From these values, it was found that the algorithm's performance was comparable to that of the clinical experts.

5 Concluding Remarks

A computer-assisted method has been presented to quantify the degree of narrowing in the internal carotid artery and evaluated using area and overlap measures. Two clinical experts defined the object boundaries twice, and their results were used as reference to assess the accuracy and precision of the algorithm's performance. The variations between algorithm and observers were similar to the inter- and intra-observer variations, thereby confirming the validity of the proposed algorithm and forming a good basis for further development of the algorithm and its introduction into clinical usage. Future work will include the development of algorithms to correctly segment arterial objects at the presence of touching arteries, the more extended evaluation to include more datasets, a comparison with other segmentation techniques, and the application of the supplementary evaluation metrics suggested in literature [5].

Acknowledgement

The authors would like to thank Dr. Tony Morton-Jones, of the Department of Physics, Astronomy and Mathematics, University of Central Lancashire, for the valuable discussions and advice on the statistical analysis of the results.

References

1. P.J. Yim, R. Mullick, R.M. Summers, H. Marcos, J.R. Cebral, R. Lohner, P.L. Choyke "Measurement of stenosis from magnetic resonance angiography using vessel skeletons", *In Medical Imaging 2000: Physiology and Function from Multidimensional Images, Proceedings of SPIE* 3978, pp 245-254, 2000.
2. C.M. Bemmel, M.A. Viergever, W.J. Niessen "Semi-automatic segmentation and stenosis quantification of 3D contrast-enhanced MR angiograms of the Internal Carotid Artery", *Magnetic Resonance in Medicine*, **51**, pp.753-760, 2004.
3. D.Y. Kim, J.W. Park "Computerized quantification of carotid artery stenosis using MRA axial images", *Magnetic Resonance Imaging*, 22: pp. 353–359, 2004.
4. D.Y. Kim, J.W. Park "Connectivity-based local adaptive thresholding for carotid artery segmentation using MRA images", *Image and Vision Computing, IVC*(23), 14, pp. 1277-1287, 2005
5. W.R. Crum, O. Camara, and D.L.G. Hill "Generalised Overlap Measures for Evaluation and Validation in Medical Image Analysis", IEEE Transactions on Medical Imaging (25), 11, pp 1451-1461, 2006.
6. J.K. Udupa, V.R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L.M. Currie, B.E. Hirsch, J. Woodburn "A Framework for Evaluating Image Segmentation Algorithms", Computerized Medical Imaging and Graphics, 30, pp. 75-87, 2006.
7. D Stampouli, M.R Varley, C.F Walshaw, A.P. Jones, R.W. Bury. and L.-K. Shark "Computer-Assisted identification and modeling of internal carotid arteries from 3D CE-MR angiograms for stenosis quantification", *3rd IEE International Seminar on Medical Applications of Signal Processing*, London, pp 35- 40, 2005.
8. D Stampouli, M.R. Varley, C.F. Walshaw, A.P. Jones, R.W. Bury. and L.-K. Shark "3D Geometry-based Tracking and Segmentation of Carotid Arteries from CE-MR Angiograms Using Locally Adaptive Thresholding", *3rd International Conference Medical Information Visualisation - BioMedical Visualization*, London, UK, 2006.
9. D Stampouli, M.R Varley, C.F Walshaw, A.P. Jones, R.W. Bury. and L.-K. Shark "Evaluation of Computer-Assisted Quantification of Carotid Stenosis", *4th International Conference Medical Information Visualisation - BioMedical Visualization*, Zurich, Switzerland, 2007.
10. R.F. Mould, "Introductory Medical Statistics", Institute of Physics Publishing; 3rd edition, 1998.
11. G. K. Kanji, "100 statistical tests", Sage Publications Ltd, 1999.

Detection of Anatomical Linear Structures in Mammograms and Raw Projection Tomosynthesis Images

Edward M. Hadley[a], Predrag R. Bakic[b], Andrew D. A. Maidment[b] and Reyer Zwiggelaar[a*]

[a]Department of Computer Science, University of Wales, Aberystwyth, UK
[b]Department of Radiology, University of Pennsylvania, Philadelphia, USA

Abstract. Three–dimensional tomosynthesis represents a significant advancement over conventional two–dimensional mammography, effectively eliminating many of the inherent problems caused by representing the three–dimensional anatomy in a two–dimensional image.
Recently, there has been interest in the proportion and distribution of parenchymal linear structures in the breast. The proportion of linear structures has been linked to mammographic risk. This experiment investigates the proportion of linear structures detected in raw tomosynthesis images and compares them to the the proportion detected in corresponding conventional mammograms of the same patient taken at the same time.
Results show a high degree of correlation (0.866) between the proportions of linear structures detected in the two image types.

1 Introduction

Conventional two-dimensional projection mammography plays a very significant role in breast cancer detection, diagnosis and treatment. However, it is well–known that 2D mammography has several inherent limitations, caused by the projection of the 3D breast anatomy on to a 2D plane. These include cancers being obscured by superimposed normal tissue and overlapping normal tissue creating the artificial appearance of densities [1]. These limitations often result in false–negative or false–positive diagnoses, increasing risk to the patient or exposing them to unnecessary, often painful follow–up procedures.

Whereas many of these limitations could be overcome by magnetic resonance imaging (MRI), this is a far more involved procedure and its high cost, inconvenience and low availability prevent the use of MRI from becoming widespread for the detection of breast cancer.

Three–dimensional breast tomosynthesis provides a significant advance over projection mammography. Tomosynthesis effectively eliminates the effects of superimposed tissue on parenchymal structures of interest [2, 3]. This can increase margin visibility, especially in dense breasts and has been shown to improve lesion visibility [4].

Breast tomosynthesis acquires a series of projection x–ray images as the x–ray source moves in an arc around the fixed breast and digital imaging detector. With the exception of their acquisition angle, the ‘raw’ projection images are similar to conventional x–ray mammograms, however they are taken using a significantly lower x–ray dose than that using for conventional mammograms, such that the overall dose received by the patient is similar for the two methods [4].

The raw projection images are subsequently reconstructed in to a three–dimensional volume that can be displayed to a radiologist. Many algorithms have been used in the reconstruction of tomosynthesis images, common examples include filtered back projection and shift–and–add.

Throughout this experiment we use raw tomosynthesis images (the two–dimensional projection images) rather than the three–dimensional reconstructed images since a more direct and straightforward comparison can be made between them and corresponding conventional mammograms.

Figure 1 shows a conventional mammogram and a raw tomosynthesis image of the same breast.

1.1 Linear Structures

Mammographic risk assessment is concerned with estimating the probability of women developing breast cancer and can provide an indication of when to recommend more frequent screening, which has been shown to improve the likelihood of the early detection of breast cancer [5]. Breast density is an important indicator of mammographic

*Edward Hadley: emh05@aber.ac.uk, Reyer Zwiggelaar: rrz@aber.ac.uk.

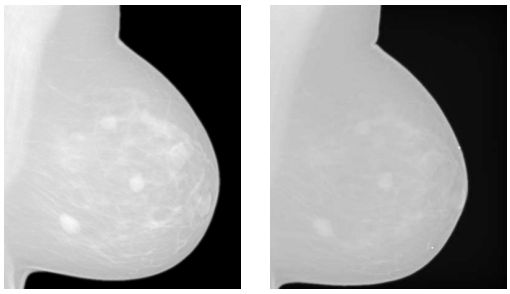


Figure 1. A conventional mammogram (left) and a corresponding raw tomosynthesis image (right).

risk [6] and the best predictor of mammographic sensitivity [7]. However, more recently, it has been suggested that the distribution of linear structures (ducts, blood vessels, etc.) is also correlated with mammographic risk [8, 9]. So far it is not entirely clear if it is just the density of linear structures (either by percentage area or volume) or if the distribution of the linear structures plays a role as well.

Tabár et al. have proposed a mammographic risk assessment model based on four structural components, where the relative proportions of each component is linked to the risk of developing breast cancer [8–10]. One of the four structural components is linear density. It has been shown that the proportion of linear structures in the breast can be used to automatically predict mammographic risk [11, 12].

The purpose of this work is to investigate the sensitivity of the detection of linear structures in tomosynthesis images compared to conventional mammograms.

2 Method

Conventional mammograms and tomosynthesis images were taken for 40 patients. For each patient, two conventional mammograms (left and right breast) and two sets of nine raw tomosynthesis images were taken on the same day. Of the sets of nine raw tomosynthesis images, the central image represented an orthogonal view and corresponded most closely with the MLO conventional mammograms, and as such these central images were selected for use in our experiment. The images were represented as 12–bit greyscale, 1914x2294 pixel DICOM files.

The resulting 80 conventional mammograms and 80 raw tomosynthesis images were processed using the line detector method (described below), which results in a measure of line strength and orientation for each pixel in the breast area of the image.

Finally, for each image the line strength values were thresholded and the proportion of pixels with above–threshold line strength values in the breast area of the image was calculated.

2.1 Line Operator

A study of various methods for detecting linear structures in mammograms [13] showed that Dixon and Taylor's line operator [14] is more accurate than other methods. As such, the line operator was used in our experiments. The method produces a measure of line strength and orientation for each pixel in an image.

The line orientation is determined by calculating the mean pixel brightness of a line of pixels running through the target pixel at a range of orientations. The orientation with the largest mean brightness is taken to be the line orientation. The line strength, S, is then given by

$$S = (L - N), \tag{1}$$

where L is the mean brightness of the line of pixels, and N is the mean brightness of a similarly orientated square of pixels.

Our experiment used a line length of five pixels and twelve orientations as suggested by earlier work [13].

A multi-scale approach was used in order to detect lines of a range of thicknesses and the resultant images were combined to produce line strength values for pixels at the original scale. Scaling of the images was achieved firstly by

blurring the image using a 3x3 Gaussian kernel and subsequently by subsampling to provide a resultant image of half the width and height of the original. Our approach comprised processing with the line operator at three scales.

Figure 2 shows the results of the two images shown in Fig. 1 after processing with the line operator.

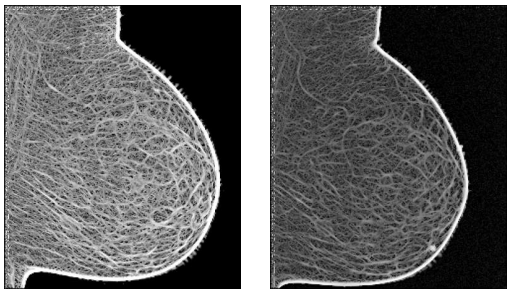


Figure 2. Line operator results showing a conventional mammogram (left) and a corresponding raw tomosynthesis image (right). The linear structures have been enhanced for display.

2.2 Thresholding

Following application of the line operator, the line strength results were thresholded to remove background tissue, aiming to leave only linear structures. The choice of threshold greatly affects the measure of proportion of linear structures. The thresholds were chosen experimentally to select values that remove most background tissue whilst leaving most of the linear structure information intact. Two thresholds were used during this experiment – 0.4% and 0.5%.

The thresholds were applied by normalising the line strength values between 0 and 4096 (the maximum greyscale value for the images used), and then by removing all line strength values below the specified proportion. In the cases used, all line strength values below 0.4% or 0.5% of 4096 were removed. The overwhelming majority of line strength values in the images were considerably below the maximum theoretical value. As such, the threshold values used were very small so as not to remove too much linear structure information.

Figure 3 shows the conventional mammogram and raw tomosynthesis images following application of the line operator and thresholding at 0.4% and 0.5%.

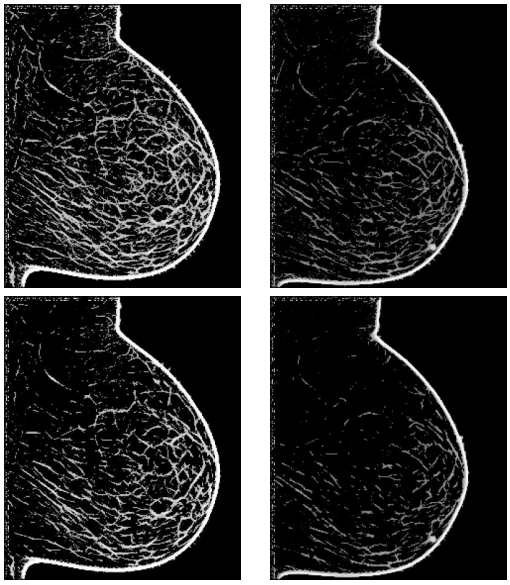


Figure 3. Thresholded line operator results showing a conventional mammogram (left) and a corresponding raw tomosynthesis image (right). The top images have been thresholded at 0.4% and the bottom images have been thresholded at 0.5%. The linear structures have been enhanced for display.

3 Results

The graph in Fig. 4 shows the proportion of above–threshold linearity in raw tomosynthesis images against the corresponding conventional mammograms. A threshold of 0.5% was used in both the conventional mammograms and tomosynthesis images in order to achieve the best correlation. This achieved a Pearson product moment correlation coefficient of 0.866.

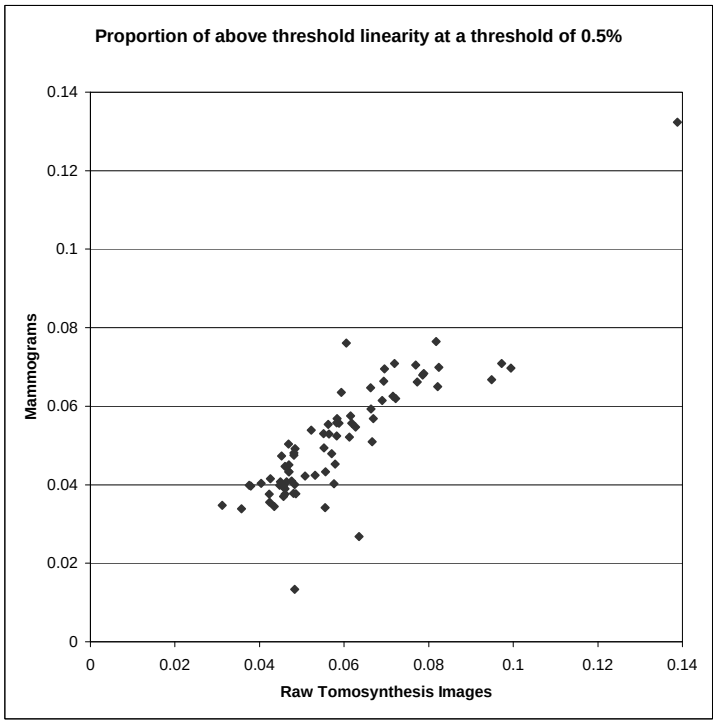


Figure 4. A graph showing the proportion of above–threshold linearity in raw tomosynthesis images against corresponding conventional mammograms at a threshold of 0.5%.

Figure 5 shows the proportions of above threshold linearity in raw tomosynthesis images and mammograms at all combinations of thresholds used.

4 Discussion and Conclusions

Linear structure information in mammograms has been linked to mammographic risk and detected linear structure information has shown promise in improving the specificity of automatic risk assessment [6, 9, 12]. Tomosynthesis offers the potential for significant improvements over mammograms by eliminating the problems associated with the representation of the 3D anatomy in 2D [2–4]. The results shown in Fig. 4 show a high correlation between the detected proportions of linear structures in raw tomosynthesis images compared with conventional mammograms (0.866). This suggests that the linear structure information in tomosynthesis images may be useful in mammographic risk assessment investigations.

The selected threshold affected the correlations. Since the contrast varied between the different image types, the thresholds were selected independently in order to achieve the closest correlation, however the correlations were highest when both were thresholded at 0.5% (0.866) and when both were thresholded at 0.4% (0.813). A larger sample of thresholds would need to be used in order to assess whether this is significant. The correlation remained good when mammograms thresholded at 0.5% were compared with tomosynthesis images thresholded at 0.4% (0.714), however the correlation was less acute when these thresholds were reversed (0.475).

It is suggested that since the contrast was greater in the mammograms (due to the high x–ray dose), the line strength values tended to be higher, and as a result, when the mammogram results were thresholded at a low proportion less background would be removed. Similarly, when tomosynthesis image results were thresholded at a high proportion more linear structure information would be removed (see Fig. 3). When these combinations are compared, the correla-

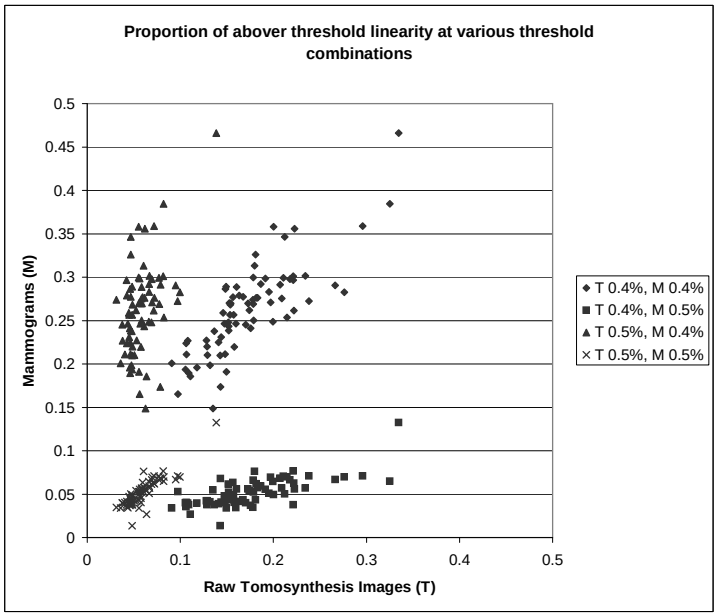


Figure 5. A graph showing the proportion of above–threshold linearity in raw tomosynthesis images against corresponding conventional mammograms at various combinations of thresholds. In the legend, *T* indicates the threshold used for the raw tomosynthesis images and *M* indicates the threshold used for mammograms.

tion would expected to be poorer than for other combinations.

An obvious subsequent investigation would involve attempting to automatically classify images in to BIRADS classes based on the linear structure information in the tomosynthesis images. However, this would require that the images be classified independently by a radiologist, and for clear results, a larger sample size would be beneficial.

References

1. P. T. Huynh, A. M. Jarolimek & S. Daye. "The false-negative mammogram." *Radiographics* **18(5)**, pp. 1137–1154, 1998.
2. J. T. Dobbins & D. J. Godfrey. "Digital x–ray tomosynthesis: current state of the art and clinical potential." *Physics in Medicine and Biology* **48(19)**, pp. R65–106, 2003.
3. M. Varjonen. "Three–dimensional digital breast tomosynthesis in the early diagnosis and detection of breast cancer." *Lecture Notes in Computer Science* **4046**, pp. 152–159, 2006.
4. L. T. Nikalason, B. T. Christian, L. E. Nikalason et al. "Digital tomosynthesis in breast imaging." *Radiology* **205**, pp. 399–406, 1997.
5. C. H. van Gils, J. D. Otten, J. H. Hendriks et al. "High mammographic breast density and its implications for the early detection of breast cancer." *Journal of Medical Screening* **6**, pp. 200–204, 1999.
6. J. N. Wolfe. "Risk for breast cancer development determined by mammographic parenchymal pattern." *Cancer* **37(5)**, pp. 2486–2492, 1976.
7. T. M. Kolb, J. Lichy & J. H. Newhouse. "Comparison of the performance of screening mammography, physical examination, and breast us and evaluation of factors that influence them: An analysis of 27,825 patient evaluations." *Radiology* **225(1)**, pp. 165–175, 2002.
8. I. T. Gram, E. Funkhouser & L. Tabár. "The Tabár classification of mammographic parenchymal patterns." *European Journal of Radiology* **24(2)**, pp. 131–136, 1997.
9. L. Tabár, T. Tot & P. B. Dean. *Breast Cancer - The Art and Science of Early Detection with Mammography.* Georg Thieme Verlag, Stuttgart, 2005.
10. L. Tabár & P. B. Dean. "Mammographic parenchymal patterns. risk indicator for breast cancer?" *Journal of the American Medical Association* **247(2)**, pp. 185–189, 1982.
11. E. M. Hadley, E. R. E. Denton & R. Zwiggelaar. "Mammographic risk assessment based on anatomical linear structures." *Lecture Notes in Computer Science* **4046**, pp. 626–633, 2006.
12. E. M. Hadley, E. R. E. Denton, J. Pont et al. "Risk classification of mammograms using anatomical linear structure and density information." *Lecture Notes in Computer Science* **4478**, pp. 186–193, 2007.
13. R. Zwiggelaar, S. M. Astley, C. R. M. Boggis et al. "Linear structures in mammographic images: Detection and classification." *IEEE Transactions on Medical Imaging* **23(9)**, pp. 1077–1086, 2004.
14. R. N. Dixon & C. J. Taylor. "Automated asbestos fibre counting." *Institute of Physics Conference Series* **44**, pp. 178–185, 1979.

A reduced computational complexity method in non-invasive evaluation of hepatic steatosis and fibrosis by ultrasound image analysis using support vector machines

Nagy Georgiana[a], Gordan Mihaela[b], Vlaicu Aurel[b], Mircea A. Petru[a], Crişan Doinița[c], Vălean Simona[a]

[a] 1st Medical Clinic, 3-5 Clinicilor Str., Cluj-Napoca, Romania, [b] Technical University of Cluj-Napoca, 15 C. Daicoviciu Str., Cluj-Napoca, Romania, [c] "Iuliu Haţieganu" University of Medicine and Pharmacy – Morphopathology Department, 3-5 Clinicilor Str., Cluj-Napoca, Romania

Abstract. Liver biopsy is the gold standard for the quantitative evaluation of steatosis and fibrosis, but is an invasive method. A recent research trend is the investigation of novel non-invasive methods for the quantitative evaluation of the hepatic tissue in respect to steatosis and fibrosis. The goal is to find the best descriptive features and integrate them in reliable classification systems, to accurately determine the steatosis and fibrosis degrees. Several types of features and classifiers are reported. Some of them prove good results, but are computationally expensive. The computational complexity can be important if one considers implementing such systems in the ultrasound machines, which usually are only capable of simple histogram computations, probably due to the hardware limitations. Here we investigate the discriminative power of simple intensity and intensity derivative features extracted from a region of interest to assess the severity degrees in steatosis and fibrosis, using support vector machine classifiers on each feature. Then we build a support vector machine-based classification system to improve the assessment accuracy, integrating the most significant individual classifiers. Due to the simplicity of the feature extraction, the proposed solution is computationally simple. The experimental results are close to the state of the art.

1. Introduction

Chronic hepatic diseases (steatosis, non-alcoholic steatohepatitis, viral and autoimmune chronic hepatitis) represent a priority in hepatological research. The motivation is obvious, concerning the incomplete knowledge about the pathogenesis and the factors influencing the natural evolution of these diseases; also the necessity to identify prophylactic and therapeutic solutions. Additionally, chronic hepatic diseases constitute an important public health problem, by increased incidence, difficulties in prophylaxis and therapy and especially by the evolution in many cases towards cirrhosis and hepatocarcinoma [1], influencing the survival and life quality [2]. Hepatic steatosis is an anatomo-clinical entity, considered as a benign disease with limited evolution potential. Lately, it has been demonstrated that simple steatosis can evolve towards hepatic fibrosis. The identification of chronic hepatic disease evolution stage is essential for diagnosis and for understanding the pathogenesis of disease progression, with immediate consequences regarding therapy. In particular, concerning the evolution of liver fibrosis is essential to discover the evolutionary sequences: from steatosis/steatohepatitis/chronic hepatitis to cirrhosis.

The ultrasonography can identify the diffuse hepatic steatosis by pathological increase of tissue brightness and the emphasis of posterior attenuation [3]. Contrarily, in liver fibrosis, the interpretation of ultrasound images is more difficult and irrelevant, although, probably, liver fibrosis enhances the posterior attenuation phenomenon [4]. In addition, both in steatosis and fibrosis, the image interpretation is subjective (operator dependent), impossible to use for quantitative evaluation. Therefore, the use of ultrasonography is limited as non-invasive diagnosis method and, especially, as time monitoring of hepatic steatosis evolution [5,6,7]. A recent trend is to develop computer-based algorithms to extract suitable quantifiable parameters from ultrasound hepatic images, able to describe and discriminate reliably between the hepatic tissue modifications, the final target being to be able to develop non-invasive hepatic tissue evaluation methods with similar performance as the biopsy [8,9,10]. These parameters must be invariant to the type of ultrasound device used and patient's physiological variability and must ensure a good discrimination between the different histological characteristics of the hepatic tissue [8]. Several such descriptors are reported, starting from simple intensity parameters to more sophisticated ones. These descriptors are extracted from a region of interest (ROI) of the ultrasound image, and afterwards used to classify the hepatic tissue typically with a supervised classifier [8,9]. Some of them, as texture co-occurrence features or fractal dimension, prove good results [9], but are computational expensive. Although not mentioned in literature, the computational complexity can be important as it could impact the possibility of embedding the steatosis and fibrosis evaluation in ultrasound machines. If simple features extracted from the grey level histogram or grey level derivatives histogram can be used as accurate discriminators for different degrees of severity of steatosis and fibrosis, this can ease the diagnosis of these hepatic tissue modifications. Unlike most methods which perform feature extraction in a single ROI, we add discriminative value to our simple features by considering also their *in-depth variation in the ultrasound image*.

2. The proposed methodology for the quantitative characterization of the hepatic tissue modification by ultrasound image analysis

2.1. Overview

Most ultrasound image analysis software provides a graphical user interface allowing the user to select the currently processed image, display and position the ROI from which the features are extracted. In our method, we developed a Windows application in which one can load either the hepatic ultrasound image or the spleen image used as reference as described in Section 3. Afterwards the user specifies the position of the ROI, in cm, considering as reference the hepatic capsule, between 0 cm and 5 cm. As in most ultrasound machines, we considered a square ROI of small size (1cm×1cm = 40×40 pixels). A larger size wouldn't have allowed for a reliable evaluation of the in-depth variation of the extracted features. As standard depths for computing the in-depth variations of the ROI extracted features, we consider the values: 0.5, 2, 3.5cm. Once the ROI is set, the application computes the intensity histogram and the cumulative total gradient in the ROI; from the intensity histogram, the set of features used for discrimination of the degrees of severity of steatosis and fibrosis are computed. Some illustrative samples of ROIs positioned at the three depths in the ultrasound image and their grey level histograms, for the three classes of hepatic tissue examined in the paper: healthy, steatosis and fibrosis, are presented in Figure 1. One can notice the approximate invariance of the histograms for the normal tissue and the most significant change of the histogram shape and median value for the severe steatosis. A slight histogram modification is observed in the case of fibrosis.

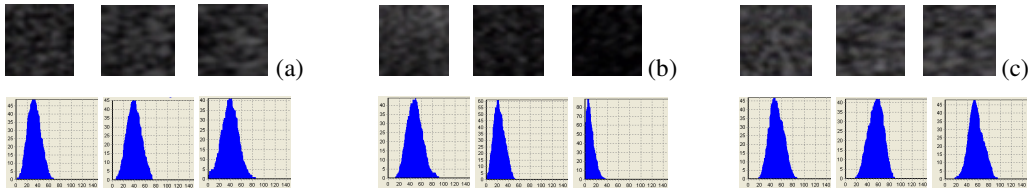


Figure 1. Example ROIs and histograms showing the in-depth variation of the grey levels in ultrasound hepatic images for: (a) normal tissue; (b) severe steatosis; (c) severe fibrosis

2.2. The set of features for ROI description

The features investigated in this paper are a set of simple but widely descriptors of the grey level distribution in a ROI, extracted from the ROI histogram and from the gradient of the grey levels in the ROI (this last parameter is only used for fibrosis). Considering the ROI size denoted by $W \times H$, we define and measure the following statistical parameters used as features: (1) Percentiles of the normalized grey level ROI histogram; we use here the 10^{th} percent percentile, denoted as l_{10}, the 90^{th} percent percentile, denoted as l_{90}, and the 50^{th} percent percentile or the median, denoted as m_l. The median luminance is computed for both hepatic and spleen tissues, the later being used to compute the relative median hepatic luminance feature as the difference between m_l and the spleen median luminance, denoted as $m_{l\text{-}s}$. (2) The standard deviation of the grey levels in the ROI, denoted by s_l, and the coefficient of variation of the luminance, *cvl*, which is a normalization of the standard deviation to the mean luminance in the ROI. (3) The luminance gradient cumulative in the ROI, denoted by *gl*, is computed for fibrosis discrimination only, since the visual examination of the ultrasound image of the hepatic tissue shows an in-depth modification of the tissue sharpness for severe fibrosis. This is the most simple quantitative descriptor able to describe the sharpness of the ROI texture – the sum of the magnitudes of the luminance gradients in each pixel in the ROI, normalized to the total number of pixels in the ROI, *WH*. To compute the luminance gradient magnitude map, we chose to use a simple and rather accurate gradient-based edge detector: the Sobel edge detector [11], although other simple edge detectors could be used as well. Considering these 7 features extracted from ROIs placed at the 3 above mentioned depths for each hepatic US image, denoted as $f_{0.5}$, f_2 and $f_{3.5}$, with f – a generic notation for any of the features, we will form for each feature a 3-dimensional feature vector in the generic form: $\mathbf{f}[3\times 1] = [f_{0.5} \quad f_2 \quad f_{3.5}]^T$. Thus we will have 7 spaces with the feature vectors: $\mathbf{m}_l$, $\mathbf{m}_{l\text{-}s}$, $\mathbf{s}_l$, **cvl**, $\mathbf{l}_{10}$, $\mathbf{l}_{90}$, **gl**.

Each of these spaces is examined in respect to its ability to discriminate the severity degrees of fibrosis and the severity degrees of steatosis, i.e., the ability to accurately separate the cases with mild steatosis from the ones with severe steatosis, and the cases with mild fibrosis as opposed to the ones with severe fibrosis. Considering we do not impose a linear separation of the different degrees of severity of steatosis/fibrosis, we can employ also a classifier with non-linear separation surface. A good choice can be a support vector machine (SVM) non-linear binary classifier [12]. SVMs are powerful binary classifiers from the class of machine learning techniques, based on the optimal separating hyper-plane principle, successfully used in pattern recognition applications. In using them for the

discrimination of the severity degrees of steatosis and fibrosis, we build a training set and a test set of patterns for each case (steatosis/fibrosis), labelling the samples, in each feature space, obtained from the study group as: +1 (positive examples) for the severe modifications of the hepatic tissue, and -1 (negative examples) for the mild modifications of the hepatic tissue. From various kernel functions examined in the non-linear SVMs, the best separation performance in the test set was obtained by the Gaussian RBF kernel, with a penalty factor C=100.

2.3. Building the overall probabilistic classifier through individual SVMs outputs integration

Although some of the feature spaces examined show a satisfactory discrimination performance between the modifications of the hepatic tissue, none of them alone gives a good enough accuracy. We have tried to train an SVM classifier on a 21-dimensional feature space forming a joint vector of all features, but since some of them have different ranges and significances, the results were not better (sometimes even worse). However we noticed on a validation set that the errors of the individual SVM classifiers in the particular feature spaces are unevenly distributed among the test patients. Thus the different "SVM experts" can compensate the diagnostic error between them. This suggests that a combination of the individual classifier decisions can produce a more accurate classification system. The adopted strategy is to first convert the real-valued decision function of the individual SVM classifiers to probabilities according to the method given in Platt [13], then to assign weights to individual classifiers according to their "reliability" in the evaluation of steatosis/fibrosis (assessed on a validation set), and finally to obtain the global probabilistic decision as weighted average of the individual probabilities.

3. Experimental setup and results

3.1. The study group

62 patients (30 women and 32 men) with chronic viral hepatitis B, C and non-alcoholic steatohepatitis (NASH) were included in our study, during 9 months (October 2005 - July 2006). The diagnostic was established by standard methods, including histopathological liver examination. Each patients went through hepatic and spleen ultrasonography in standardized conditions (same machine and examiner) . We used the spleen section as standard reference for the patient himself. Ulterior, the patients went through liver biopsy; the fragment was histopathologically examined with hematoxilin-eosine staining and also with special stainings (trichrom Masson, Gomori for fibrosis, Red oil for steatosis). The severity of necro-inflammatory lesion and liver fibrosis was quantified using the Knodell score and for steatosis using the Brunt classification [7]. Histopathologically, the subjects were classified in: 40 without steatosis or mild steatosis and 22 with steatosis degree 2-3. Regarding fibrosis, 32 patients had mild or no fibrosis (F0-F1) and the remaining 30 had severe fibrosis (F3-F4).

3.2. Results

The experiments performed follow two goals: (A) to examine the discriminative power of the different intensity distribution and in-depth variation features for the identification of steatosis and fibrosis severity degrees; (B) using the best individual feature-based SVM classifiers, combined with confidence ranking in an optimal classification scheme for steatosis and fibrosis evaluation, to evaluate the accuracy of their decisions. To do so, we divided the study group into a training set and a test set, both relatively balanced in respect of the number of positive and negative examples. 30 patients formed the training set (15 representing positive examples and 15 – negative examples, for both fibrosis and steatosis – although not necessarily the same patients in the group present simultaneously fibrosis and steatosis, thus the training patterns labels will differ in the two hepatic tissue modifications examined). The remaining 32 patients were included in the test set. 25 of the patients in the test set present rather severe steatosis and 7 of them have almost no steatosis. In respect to fibrosis, 17 of the 32 test patients have severe fibrosis, whereas the remaining 15 have almost no fibrosis. Using these data we have trained, in each individual feature space, a set of SVM classifiers with Gaussian RBF kernel [12], with different values of the width of the Gaussian kernel (between 0.0001 and 10) and different values of the penalty factor C (between 1 and 1000), and selected for further use in classification and evaluation of the discriminative performance of that particular feature, the one that minimizes the training error at the minimal number of resulting support vectors and smallest C. To asses the classification performance of each best SVM classifier and also of the overall classification system (resulting by the individual classifiers combination), we compute the following standard parameters [9]: accuracy, sensitivity, specificity, the predicted positive value (PPV) and the predicted negative value (PNV) in the test set. Let us denote the total number of test patterns by N_t, the number of positive test patterns by N^+_t, the number of negative test patterns by N^-_t. Similarly we denote the number of positive test patterns falsely classified as negative by FN and the number of negative test patterns falsely classified as positive by FP. Then:

$$Accuracy\,[\%] = (FP + FN)/N_t \cdot 100\,;\; Sensitivity\,[\%] = \left(1 - FN/N^+{}_t\right)\cdot 100\,;\; Specificity\,[\%] = \left(1 - FP/N^-{}_t\right)\cdot 100\,; \quad (1)$$

$$PPV\,[\%] = \left(\left(N^+{}_t - FN\right)/\left(N^+{}_t - FN + FP\right)\right)\cdot 100\,;\; PNV\,[\%] = \left(\left(N^-{}_t - FP\right)/\left(N^-{}_t - FP + FN\right)\right)\cdot 100 \quad (2)$$

A. The discriminative power of each individual feature to the severity degrees of steatosis and fibrosis, according to the best selected SVM classifier in that feature's space, can be evaluated examining the 5 classification performances in equations (1) and (2), given in Table 1 for steatosis and in Table 2 for fibrosis.

Feature	Accuracy[%]	Sensitivity[%]	Specificity[%]	PPV[%]	PNV[%]
$\mathbf{m}_l$	68.75	84	14.29	77.78	20
$\mathbf{m}_{l\text{-}s}$	**75**	**92**	14.29	79.31	33.33
$\mathbf{s}_l$	65.63	68	57.14	85	33.33
cvl	65.63	60	**85.71**	93.75	37.5
$\mathbf{l}_{10}$	**75**	**92**	14.29	79.31	33.33
$\mathbf{l}_{90}$	71.88	68	**85.71**	**94.44**	**42.86**

Table 1. The discrimination performance of the various features extracted from the ROI and their in-depth variation, for steatosis classification/evaluation

Feature	Accuracy[%]	Sensitivity[%]	Specificity[%]	PPV[%]	PNV[%]
$\mathbf{m}_l$	59.38	35.29	**86.67**	75	54.17
$\mathbf{m}_{l\text{-}s}$	46.88	76.47	13.33	50	33.33
$\mathbf{s}_l$	34.38	47.06	20	40	25
cvl	50	29.41	73.33	55.56	47.83
$\mathbf{l}_{10}$	46.88	70.6	20	50	37.5
$\mathbf{l}_{90}$	56.25	**94.12**	13.33	55.17	66.67
gl	**75**	76.47	73.33	**76.47**	**73.3**

Table 2. The discrimination performance of the various features extracted from the ROI and their in-depth variation, for fibrosis classification/evaluation

Examining the results in Table 1, we can see that, for steatosis, there is no feature giving maximum performance in respect to all performance parameters computed. As best among them we should mention the in-depth variation vectors of: the 10^{th} percentile of the luminance ($\mathbf{l}_{10}$); the relative median hepatic luminance referenced to the spleen ($\mathbf{m}_{l\text{-}s}$); the 90^{th} percentile of the luminance ($\mathbf{l}_{90}$), followed by the remaining features. Therefore a weighted combination of the classifiers in the individual feature spaces can improve the steatosis degree evaluation performances, as shown in the following set of results. A similar evaluation of the results presented in Table 2 shows that in the case of fibrosis, the discriminative power of all simple ROI extracted features examined is poorer than in the case of steatosis. However in this case, the best performance given by the vector of in-depth variation of the luminance gradient cumulative in the ROI ($\mathbf{g}_l$), followed by the median of the hepatic luminance vector ($\mathbf{m}_l$). A weighted combination of the classifiers in the individual feature spaces can improve the fibrosis degree evaluation.

B. The decision performance of the proposed classification scheme for steatosis and fibrosis, obtain through the weighted fusion of the best individual feature-based SVM classifiers probabilistic decisions, is presented in Table 3 – evaluated again through the parameters in the equations (1) and (2). The different classifier decisions weights were obtained on a validation sub-set of patients.

Evaluated histopathological change	Accuracy[%]	Sensitivity[%]	Specificity[%]	PPV[%]	PNV[%]
Steatosis	84.38	88	71.43	91.67	62.5
Fibrosis	78.13	76.47	80	81.25	75

Table 3. The decision performances of the mixed classification schemes for steatosis and fibrosis

Comparing these results to the state of the art, we can see that they outperform the same performances obtained with comparable simple features. In the case of steatosis, the reported accuracies are [10]: with fractal measures– 69%; with Fourier measures from the Fourier power spectrum – 82% (inferior to our method). The grey level co-occurrence matrix-based features slightly outperform our accuracy (87%) and so do the non-separable quincunx wavelet transform decomposition features (90% accuracy), but the computation of co-occurrence matrix and of the wavelet transform is more expensive than computing 1-D histogram features as we did. In the case of fibrosis, the results reported in [9] using the fractal dimension of the texture edge co-occurrence matrix (TECM) are below our performances: 54.3% accuracy, 62.6% sensitivity and 51.3% specificity. When the features used are the entropies of the TECM, the performances are better than those provided by our method: 82% accuracy, 93.9% sensitivity and 78% specificity. However the mathematics behind TECM and thus also its computational complexity is higher.

4. Conclusion

In this work we investigated a set of simple features, extracted from the in-depth variation of various parameters of the intensity histogram and intensity gradient in a hepatic ROI, in respect to their ability to discriminate the severity degree of steatosis and fibrosis. Furthermore, based on the results of this examination, we developed a classification system based on the weighted integration of several individual SVM classifiers probabilistic decisions, to maximize the discrimination accuracy for steatosis and for fibrosis, using these simple features. The experimental results are close to the state of the art in the field, even with such very simple tissue descriptors, which can be explained by the fact that we take into account also their in-depth variation in the hepatic tissue. In our future work we will try to increase the decision performance especially for fibrosis, by considering the use of other parameters, while trying to keep the computational complexity of their extraction at a low level.

References

1. E.R.Schiff & M.F. Sorrell. *Schiffs disease of the liver, 10th* Ed. Lippincot, Williams & Wilkins 2007.
2. L.M. Martin, M.J. Sheridan & Z.M. Younossi. "The impact of liver disease on health-related quality of life: a review of the literature". In *Curr Gastroenterol Rep*, pp. 79-83, 2002.
3. E.M. Brunt. "Nonalcoholic steatohepatitis: definition and pathology". In *Semin Liver Dis.* , pp.3-16, 2001.
4. R.I. Badea. Ficatul. In: R.I. Badea, S.M. Dudea, P.A. Mircea & F. Stamatian (eds). *Tratat de ultrasonografie clinica*, vol.1. Ed. Medicala, Bucuresti 2000.
5. U.L. Mathiese, LE. Franzen, H. Aselnis & M. Resjo. "Increased liver echogenicity at ultrasound examination reflects degree of steatosis but not of fibrosis in asymptomatic patients with abnormalities of liver transaminases". In *Dig Liver Dis, pp. 516-22,* 2002.
6. S. Saadeh, Z.M. Younossi, & E.M. Remer. "The utility of radiological imaging in non-alcoholic fatty liver disease". In *Gastroenterology*, pp. 745-50, 2002.
7. E.M. Brunt, C.G. Janney, A.M. Di Bisceglie, B.A. Neuschwander-Tetri & B.R. Bacon. "Non-alcoholic steatohepatitis: a proposal for grading and staging the histological lesions". In Am J Gastroenterol, pp. 2477-74, 1999.
8. Y. M. Kadah, A. A. Farag, J. M. Zurada, A. Badawi & A. Youssef. "Classification Algorithms for Quantitative Tissue Characterization of Diffuse Liver Disease from Ultrasound Images". In: *IEEE Transactions on Medical Imaging*, pp. 466-478, Vol. 15, No. 4, , August 1996.
9. G.T. Cao, P.F. Shi & B. Hu. "Liver fibrosis identification based on ultrasound images captured under varied imaging protocols". In: J Zhejiang Univ Sci B. , pp. 1107-14, Nov 6(11), 2005.
10. A. Mojsilovic, S. Markovic & M. Popovic. "Characterization of visually similar diffuse diseases from B-scan liver images with the nonseparable wavelet transform". In: ICIP, pp. 547-550, 1997.
11. W. K. Pratt, *Digital Image Processing: PIKS Inside, 3rd Edition*, John Wiley & Sons, 2001
12. V.N. Vapnik, *Statistical Learning Theory*, J. Wiley, N.Y., 1998
13. J. Platt. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in Large Margin Classifiers* (A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, Eds.), MIT Press, Cambridge, MA, 2000.

Dealing with false positive reduction in mammographic mass detection

Xavier Lladó*, Arnau Oliver, Joan Martí, and Jordi Freixenet

Institute of Informatics and Applications, University of Girona, 17071, Girona, Spain

Abstract. In this paper we analyze a set of false positive reduction methods in the field of mammographic mass detection. The main goal of this false positive reduction process is the discrimination between the true recognized masses and the ones which actually are normal parenchyma. We describe three different approaches to extract breast mass image features. The first approach is based on modeling the tissue variation of both kinds of regions of interest (RoIs) by extracting the principal components (PCA) of a set of already classified RoIs. The second approach is based on an extension of the PCA approach by using the recently proposed 2DPCA algorithm. Finally, the third approach is based on Local Binary Patterns (LBP) for representing texture information and preserving at the same time the spatial structure of the masses. Once those image descriptors are extracted, the system is trained and used for classifying the unknown RoIs. We evaluate our false positive reduction approaches using a set of 1792 suspicious RoIs extracted from the DDSM database, providing a comparison when using different ratios of number of RoIs depicting masses and number of RoIs depicting normal tissue, and also when using different mass sizes, a critical aspect in mass detection systems.

1 Introduction

Breast cancer is one of the most devastating and deadly diseases for women in their 40s in the European Union as well in the United States. It is estimated that between one in eight and one in twelve women will develop breast cancer during their lifetime [1]. The most used method to detect breast cancer is mammography, because it allows the detection of the cancer at its early stages, a crucial issue for a high survival rate. During the last decade several algorithms have been proposed for the automatic detection of masses in mammograms [2, 3]. However, the main drawback of these methods is the high number of obtained false positives [4]. A false positive is a Region of Interest (RoI) – a sub-image containing the suspicious region – being normal tissue but interpreted by the automatic algorithm as a real mass. Therefore, almost all works trying to detect masses in mammography need a final step in order to reduce the number of false positives. This is due to the complexity of the internal breast tissue, which induces the detection of regions which are not masses, but normal variations in tissue characteristics.

In this paper we present three different approaches to perform the false positive reduction in mass detection. The first one is based on modeling the tissue variation by extracting the principal components of a set of already classified RoIs [5]. The second approach is based on an extension of the PCA approach by using the recently proposed 2DPCA algorithm [6]. Finally, our third approach is based on Local Binary Patterns (LBP) for representing texture information and preserving at the same time the spatial structure of the masses. The idea of this proposal is inspired on the recent work of Ahonen et al. [7] in which Local Binary Patterns (LBP) are successfully applied to the face recognition problem. To our knowledge this is the first attempt to use LBP in the field of mammographic mass detection. In order to analyze these approaches, we perform experiments on a complete set of 1792 RoIs extracted from the DDSM database, evaluating the results when using different RoI image sizes, and when using different ratios of number of RoIs depicting masses and RoIs depicting normal tissue in the database.

The rest of the paper is organized as follows. Section 2 presents a briefly state of the art on mass false positive reduction. Section 3 describes our three approaches for mass false positive reduction. Experimental results are presented in Section 4. Finally, the paper ends with conclusions.

2 Related work

Different algorithms have been proposed so far for the mass false positive reduction task. Analyzing these algorithms, we can distinguish between two main strategies. The first one includes the set of algorithms which firstly extracts image features from the RoIs, usually related to their texture, and subsequently trains the classifier. On the other hand, a second strategy handles this problem as a template matching algorithm. Each new image is compared to all the RoIs of the database and then it is classified as an image containing a mass or not. Table 1 summarizes some works belonging to both strategies.

*Corresponding author, e-mail: llado@eia.udg.es

Classifier-Based					
Author	**Year**	**Features**	**Classifier**	**RoIs**	**Results**
Sahiner [8]	1996	Texture, Morphologic	LDA, NN	$168/504$	$Az = 0.90$
Christoyianni [9]	2002	Grey-level, Texture, ICA	NN	$119/119$	88.23%
Qian [10]	2001	Texture, Shape	NN	$200/600$	$Az = 0.86$
Tourassi [11]	2005	Grey-level	NN	$681/984$	$Az = 0.84$

Template-Based					
Author	**Year**	**Features**	**Similarity**	**RoIs**	**Results**
Chang [12]	2001	Grey-level, shape	Likelihood function	$300/300$	$Az = 0.83$
Tourassi [13]	2003	Grey-level	Mutual Information	$809/656$	$Az = 0.87$

Table 1. Summary of the reviewed works on false positive reduction, with the features used, the classifier/similarity used (where LDA means linear discriminant analysis, NN neural network analysis, and ICA independent component analysis), the number of RoIs depicting masses vs the number of normal RoIs, and the results obtained. Note that for all works accuracy is given in terms of Az (the area under the ROC curve) except for the work of Christoyianni et al. [9] which just gives the correct classification percentage.

For instance, Sahiner et al. [8] extracted a huge set of features, and subsequently used genetic algorithms to select the most discriminative ones. With this subset of features, a neural net (NN) and a linear classifier (LDA) are trained and used to classify a new RoI. A similar strategy is used by Christoyianni et al. [9], who extracted grey-level, texture, and features related to independent component analysis (ICA), and use them to train a neural net. Note also, that they apply a principal component analysis (PCA) pre-processing step to reduce the complexity of the problem. On the other hand, Qian et al. [10] analyzed the implementation of an adaptive module to improve the performance of an automatic procedure which consists of training a Kalman-filter based neural net using features obtained from a wavelet decomposition.

The works of Chang et al. [12] and Tourassi et al. [13] are based on comparing a new RoI with all the RoIs in the database (template-based approaches). The two most clear differences between them arise from the similarity measure and the database used. More specifically, the former developed a likelihood measure which depends on the grey-level and the shape of the RoIs. Both parameters were compared with the new RoI and the set of RoIs present in the database, which was only composed by RoIs depicting masses. From this comparison a likelihood measure was computed. On the other hand, the work of Tourassi et al. [13] consists of comparing all the RoIs of the database (including RoIs with and without masses) with the new one using a mutual information based similarity measure. Thus, the new RoI will be labeled as belonging to the closest class.

Note that with the last strategy, the similarity used for classifying has to be re-computed for each new element, as it measures the difference between the new RoI and all the RoIs in the database. Observe also that among all those works showed in Table 1, one of the main differences in terms of producing their results is the ratio between the number of RoIs depicting masses and the total number of RoIs. This is an important issue because the number of wrong classified RoIs will increase as the number of normal RoIs increases. One should remember that the aim of this step is to reduce the number of false positives, which is usually higher than the number of true positives. We will analyze this issue in the experimental results.

3 Approaches to mass false positive reduction

In the following sections, we will present three different approaches to perform the false positive reduction: the first one based on the use of PCA like in the eigenfaces algorithm; the second one based on the recent developed 2DPCA decomposition instead of using the typical PCA; and finally, the third one based on the textural features extracted from LBP. Note that these methods should be grouped into the classifier-based strategy.

3.1 PCA-based false positive reduction

Our first approach to deal with the mass false positive reduction is based on the idea of the eigenfaces approach [14]. We assume that a database of already classified RoIs (RoIs containing masses and RoIs of normal tissue[1]) is available. Different instances for each class are included in the database. Their intra class variability is mainly due to grey-level and texture differences and to the shape and size of the mass or other structures present in the RoI.

[1]Note that although in this work the database contains only two types of RoIs (masses or normal tissue) this can be extended to include other RoIs containing microcalcifications or architectural distortions.

According to the Karhunen-Loeve transform, the usefulness of the different eigenvectors to characterize the variation among the images is ranked by the value of the corresponding eigenvalue. Note that in this case, instead of eigenfaces we should refer them as eigenrois. Thus, the eigenrois span the RoI subspace of the original image space, and each RoI can be transformed into this space by using them. The result of this transformation is a vector of weights describing the contribution of each eigenroi in representing the corresponding input image. Therefore, we propose to use this vectors to construct the models for the training step (see [5] for more details). In our experiments we retained 95% of the information when computing the eigenrois.

When a new RoI has to be tested, it will be classified as belonging to the most similar class. Although in this paper we use the k-Nearest Neighbour classifier to perform the classification, in our previous work [5] we also used a Bayesian combination of this algorithm with the $\mathrm{C}4.5$ decision tree to compare the results.

3.2 2DPCA-based false positive reduction

The 2DPCA approach [15] is a recent improvement of the typical eigenfaces approach. As the authors argue 2DPCA has important advantages over PCA in two main aspects: firstly, it is simpler and more straightforward to use for image feature extraction since 2DPCA is directly based on the image matrix, and secondly, it is easier to accurately evaluate the covariance matrix. In the original eigenfaces approach, each input image of size $m \times n$ is transformed into a vector of size $m \cdot n$, in contrast to the natural way to deal with two dimensional data, which would be treating it as a matrix.

In [6] we presented an extension of the PCA approach to mass false positive reduction by using the 2DPCA algorithm. One of the main differences with respect the standard PCA approach is that while for PCA each principal component is a scalar, for 2DPCA each principal component is a vector. Therefore, it is this set of vectors for image that is used to construct the feature image. In a similar way to the eigenfaces approach, comparing images means to compare the constructed features. As the dimension of the feature space has increased in one dimension respect to the PCA approach, now the comparison of images is done by comparing matrices (see [6] for the detailed mathematical background).

3.3 LBP-based false positive reduction(LBP)

The original LBP operator was introduced by Ojala et al. [16] with the idea to perform grey scale invariant two-dimensional texture analysis. The LBP operator labels the pixels of an image by thresholding the neighbourhood (i.e. 3×3) of each pixel with the centre value and considering the result of this thresholding as a binary number. When all the pixels have been labeled with the corresponding LBP codes, the histogram of the labels is computed and used as a texture descriptor. The authors proposed to derive different LBP operators at circular neighbourhoods of any quantization of the angular space and at any spatial resolution. Therefore, it is possible to derive the operator for a general case based on a circularly symmetric neighbourhood of P members on a circle of radius R. Another extension of LBP was the use of the so called uniform patterns [17]. A LBP is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular. In this paper, we shall refer the uniform LBP operator as $LBP^{u2}_{P,R}$, where (P, R) indicates the neighbourhood and $u2$ uniform patterns.

Inherently to the mass false positive reduction problem, and in particular with the images we are dealing with, texture and the image spatial information play a key role in correctly detecting the masses. Due to this fact, we base this approach on the use of LBP for representing salient micro-patterns, and an adaptation of these descriptors for preserving also the spatial structure of the masses. The idea of our proposal has been inspired on the recent work of Ahonen et al. [7] in which LBP is used to perform face recognition. Our general procedure consists in using the LBP texture descriptor to build several local descriptions of the RoI and combining them into a global description. Afterwards, this global LBP descriptor is used to reduce the false positives, classifying the RoIs between true masses and normal tissue.

Initially, the RoI image – which contains the suspicious mass – is divided into several local regions (for instance, 5×5 rectangular regions). From these regions, texture descriptors are independently extracted using LBP and then concatenated to form a global description of the RoI based on the textural information of each region and its spatial distribution. Note that following this methodology, the basic LBP histogram is extended into a spatially enhanced histogram which encodes both the local region appearance and the spatial relations (global geometry) of the mass. In the final histogram, the RoI is described on three different levels of locality: the labels for the histogram contain the pixel-level texture patterns, the labels are summed over a small region to produce information on a regional level and finally the regional histograms are concatenated to build a global description of the mass. The final step is the mass classification. For this purpose we use the well-known Support Vector Machines (SVM) [18] with polynomial kernel to perform a binary decision between RoIs depicting a true mass and RoIs depicting normal parenchyma.

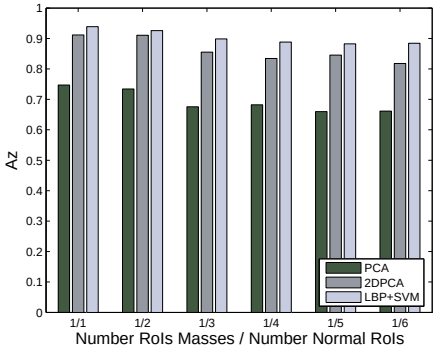


Figure 1. Az values obtained with the methods PCA, 2DPCA and LBP+SVM. Each Az value is the mean computed from the results of different RoI image sizes.

4 Experimental results

The three described approaches have been evaluated using a database of 1792 RoIs extracted from the DDSM mammographic database [19]. From this set, 256 depicted a true mass, while the rest 1536 were normal, but suspicious tissue. According to the size of the lesion, we use six different groups of RoI images, corresponding to the following mass sizes intervals: $< 0.10\,cm^2$, $(0.10-0.60)\,cm^2$, $(0.60-1.20)\,cm^2$, $(1.20-1.90)\,cm^2$, $(1.90-2.70)\,cm^2$, $> 2.70\,cm^2$. The number of masses in each interval were 28, 32, 37, 57, 69, and 33, respectively. The evaluation of our experiments is done by using a leave-one-out scheme and Receiver Operating Characteristics (ROC) analysis [20]. In such analysis, widely used in the medical field, a graphical curve represents the true positive rate as a function of the false positives rate. The percentage value under the curve (known as Az) is an indication for the overall performance of the observer, and is typically used to analyze the performance of the algorithms.

In order to perform a more global evaluation we compute the Az value for different ratios of number of RoIs depicting masses and number of RoIs depicting normal tissue (from ratio $1/1$ to ratio $1/6$). The idea of analyzing these different ratios is twofold: firstly, to evaluate the performance of our method on different levels of difficulty, and secondly, to allow the comparison of our proposal with existing methods (see the methods in Table 1). Notice that previous works only provide results for specific ratios. Hence, analyzing all these ratios will enable the comparison with them.

Figure 1 shows the obtained mean Az values for each specific ratio when testing our proposal at different RoI image sizes. We include a quantitative comparison with the three described approaches: PCA, 2DPCA and LBP. For the LBP approach we empirically decided to use the basic $LBP^{u2}_{8,1}$ operator with the RoI image divided into 5×5 regions, obtaining an overall mean Az value of 0.88 ± 0.05. With the aim of improving these results and obtaining a more accurate description of the central area, we added a new set of LBP operators for the 3×3 central regions. In particular, we computed two new LBP responses varying the radius R according to each RoI image size (i.e. $R = 1, 4$ and 6 were used for the first RoI image size). The global descriptor was then obtained concatenating the 43 $(25+9+9)$ histograms of all the regions and LBP operators. This was the final descriptor we used for the experiments presented in Figure 1 since provided a good trade-off between performance and feature vector length. Observe that the performance of LBP was clearly better than the PCA method. The results were also better than the 2DPCA approach, specially for the cases in which we had smaller ratios $1/4$, $1/5$ and $1/6$. Note also that the use of 2DPCA itself allowed to obtain better results than the original PCA approach. Using PCA we obtained an overall mean Az value of 0.67 ± 0.09 for different RoI image sizes and ratios, using 2DPCA the Az value was 0.87 ± 0.08, while using the LBP approach the Az value was 0.91 ± 0.04. Note that these results could be used to perform a qualitative comparison – in terms of Az value – with the rest of the approaches described in Section 2. However, we want to clarify that not all the methods used the same databases and therefore our aim is only to provide a general view of the performance of our approaches. Note that using the LBP approach we obtained better results than those reported by the works which used ratios of $1/1$ and $1/2$, and similar results than those obtained by Sahiner et al. for the ratio $1/3$.

Regarding the behaviour when varying the RoI image sizes, we observed that LBP provided better and more constant overall results than PCA and 2DPCA methods (see the standard deviations). Finally, Table 2 illustrates the Az values for the specific ratio $1/3$. Note that the three approaches are more suitable for false positive reduction of larger masses than smaller ones. This is due to the fact that larger masses have a larger variation in grey-level contrast with respect

Alg.	Lesion Size (in cm^2)					
	<0.10	**0.10-0.60**	**0.60-1.20**	**1.20-1.90**	**1.90-2.70**	**>2.70**
PCA	0.53	0.70	0.70	0.68	0.72	0.83
2DPCA	0.81	0.83	0.87	0.84	0.89	0.93
LBP	0.86	0.90	0.96	0.91	0.86	0.92

Table 2. Az results (ratio $1/3$) for the classification of masses using RoIs extracted from the DDSM database, detailed per size (in cm^2).

to their surrounding tissue than smaller masses, which are usually more subtle, even for an expert.

5 Conclusions

In this paper we have presented three different approaches for the mass false positive reduction problem: one based on PCA, another based on 2DPCA, and the last one based on LBP. Our experiments have shown that LBP and 2DPCA features are effective and efficient for false positive reduction at different RoI image sizes, outperforming the traditional PCA approach. Moreover, LBP provides the best results when using different ratios of number of RoIs with masses and number of RoIs with normal tissue. Regarding the behaviour when varying the RoI image sizes, LBP also provides better and more constant results than PCA and 2DPCA methods.

Acknowledgments

This work was partially supported by MEC grant TIN2006-08035 and grant IdIBGi-UdG 91060080.

References

1. Eurostat. “Health statistics atlas on mortality in the European Union.” *Official Journal of the European Union* 2002.
2. R. Zwiggelaar, T. C. Parr, J. E. Schumm et al. “Model-based detection of spiculated lesions in mammograms.” *Med. Image Anal.* **3(1)**, pp. 39–62, 1999.
3. B. Sahiner, N. Petrick, H. P. Chan et al. “Computer-aided characterization of mammographic masses: Accuracy of mass segmentation and its effects on characterization.” *IEEE Trans. Med. Imag.* **20(12)**, pp. 1275–1284, 2001.
4. R. M. Nishikawa & M. Kallergi. “Computer-aided detection, in its present form, is not an effective aid for screening mammography.” *Med. Phys.* **33**, pp. 811–814, 2006.
5. A. Oliver, J. Martí, R. Martí et al. “A new approach to the classification of mammographic masses and normal breast tissue.” In *IAPR Int. Conf. Pattern Recogn.*, volume 4, pp. 707–710. 2006.
6. A. Oliver, X. Lladó, J. Martí et al. “False positive reduction in breast mass detection using two-dimensional PCA.” In *Lect. Not. Comp. Sc.*, volume 4478, pp. 154–161. 2007.
7. T. Ahonen, A. Hadid & M. Pietikäinen. “Face detection with local binary patterns: Application to face recognition.” *IEEE Trans. Pattern Anal. Machine Intell.* **28(12)**, pp. 2037–2041, 2006.
8. B. Sahiner, H. P. Chan, D. Wei et al. “Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue.” *Med. Phys.* **23**, pp. 1671–1684, 1996.
9. I. Christoyianni, A. Koutras, E. Dermatas et al. “Computer aided of breast cancer in digitized mammograms.” *Comp. Med. Imag. Grap.* **26**, pp. 309–319, 2002.
10. W. Qian, X. Sun, D. Song et al. “Digital mammography - wavelet transform and kalman-filtering neural network in mass segmentation and detection.” *Acad. Radiol.* **8(11)**, pp. 1074–1082, 2001.
11. G. D. Tourassi, N. H. Eltonsy, J. H. Graham et al. “Feature and knowledge based analysis for reduction of false positives in the computerized detection of masses in screening mammography.” In *IEEE Conf. Eng. Med. Biol. Soc.*, pp. 6524–6527. 2005.
12. Y. H. Chang, L. A. Hardesty, C. M. Hakim et al. “Knowledge-based computer-aided detection of masses on digitized mammograms: A preliminary assessment.” *Med. Phys.* **28(4)**, pp. 455–461, 2001.
13. G. D. Tourassi, R. Vargas-Vorecek, D. M. Catarious et al. “Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information.” *Med. Phys.* **30(8)**, pp. 2123–2130, 2003.
14. M. A. Turk & A. P. Pentland. “Eigenfaces for recognition.” *J. Cognitive Neurosci.* **3(1)**, pp. 71–86, 1991.
15. J. Yang, D. Zhang, A. F. Frangi et al. “Two-dimensional PCA: a new approach to appearance-based face representation and recognition.” *IEEE Trans. Pattern Anal. Machine Intell.* **26(1)**, pp. 131–137, 2004.
16. T. Ojala, M. Pietikäinen & D. Harwood. “A comparative-study of texture measures with classification based on feature distributions.” *Pattern Recogn.* **29(1)**, pp. 51–59, 1996.
17. T. Ojala, M. Pietikäinen & T. Mäenpää. “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns.” *IEEE Trans. Pattern Anal. Machine Intell.* **24(7)**, pp. 971–987, 2002.
18. V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, New York, 1998.
19. M. Heath, K. Bowyer, D. Kopans et al. “The Digital Database for Screening Mammography.” In *Int. Work. Dig. Mammography*, pp. 212–218. 2000.
20. C. E. Metz. “Evaluation of digital mammography by ROC analysis.” In *Int. Work. Dig. Mammography*, pp. 61–68. 1996.

Efficient Construction of MDL Statistical Shape Models using Non-Parametric Surface-Based Regularisation

Carole Twining*, Rhodri Davies, and Chris Taylor

Imaging Science and Biomedical Engineering, University of Manchester, Manchester, UK.

Abstract. The method of establishing automatic groupwise correspondence for groups of shapes or images by the optimisation of a Minimum Description Length (MDL) objective function is a well-established technique. It has been shown to produce qualitatively better models than other approaches. However, the computational cost of optimisation is high, leading to long convergence times. In this paper, we discuss two techniques that lead to substantial decreases in convergence times (up to two orders of magnitude). Firstly, we show how topologically non-trivial shapes can be mapped to regular grids (called *shape images*), which simplifies the computational cost of interpolation. Secondly, we consider regularisation, and show that a non-parametric fluid regularizer can be applied in a principled manner, acting on the surface of the shape itself, without losing the computational gain made by the use of shape images. Quantitative evaluation is performed on several biological datasets.

1 Introduction

Statistical models of shape and appearance have proven to be powerful tools for image interpretation and morphological analysis. A key step in the construction of such a model is establishing a dense correspondence across a training set of examples of the object to be modeled. It is important that the correct correspondence is chosen, otherwise the model built from this correspondence will not truly represent the actual variation seen in the training set, for example, by not being specific enough, or by having poor generalisation ability.

It has been shown previously that for either shapes or images, this fundamental problem of groupwise correspondence can be solved by treating it as an integral part of the learning process (e.g, [1,2]). The key idea is that the quality of the correspondence can be evaluated by measuring the quality of the model build using that correspondence. This then casts model-building into the form of an optimisation problem. The required framework consists of: an *objective function* (which provides a quantitative measure of model quality), a method of *manipulating* the groupwise correspondence (which enables movement in the search space of our optimisation problem), and finally an *optimisation algorithm*.

It has been shown that an objective function based on the information theoretic concept of Minimum Description Length (MDL, [3]) produces models which are quantitatively better than those produced using other methods, including manual annotation [1, 2]. Such an MDL approach can be applied to the groupwise correspondence problem for either shapes or images, but in this paper we will focus on the case of MDL groupwise correspondence for shapes.

In previous approaches, each shape in the training set was first parameterised by mapping it to the corresponding topological primitive (e.g., the unit sphere). The groupwise correspondence was then manipulated by re-parameterising each shape in turn, by means of re-parameterisations of the sphere. Although this allowed the topology of the shape to be maintained, it entailed the computationally-costly step of interpolating triangulated shape meshes on the surface of a sphere. In this paper, we adapt ideas from the computer graphics literature, which show how triangulated shape surfaces can be mapped onto regular grids, which we call *shape images*. Interpolation and numerical differentiation are much simpler and faster on regular grids, which leads to an order of magnitude decrease in convergence times.

The final issue is the optimisation algorithm used. The groupwise correspondence problem is in general ill-posed, and insoluble without some form of regularisation. Previous approaches to shape correspondence used parametric regularisation, confining the allowed re-parameterisation transformations to some set of parametric transformations (e.g., Cauchy functions [1], B-splines [4], or polynomials [5]). But the limited nature of the allowed transformations can also make the optimum correspondence difficult to find and represent. By analogy with techniques in image registration [6], we show that it is possible to use a non-parametric, fluid regularisation method. We also show that it is possible to introduce this regulariser in a principled way, so that the fluid regulariser acts on the shape surface itself, whilst retaining all the computational advantages of the shape image representation. We perform a quantitative evaluation on biological datasets, and show that it converges several orders of magnitude faster than previous approaches.

2 Statistical Shape Models

We start with a training set of shapes $\{S_i : i = 1, \ldots n_S\}$, where S_i represents the entire shape in $\mathbb{R}^3$. We suppose we have an initial *parameterisation* or *chart* $\vec{\mathfrak{X}} \subset \mathbb{R}^2$, so that $\vec{S}_i(\vec{x})$ is a single point of the i^{th} shape, with parameter value

*Email: carole.twining@manchester.ac.uk

$\vec{x} \in \mathbb{R}^2$ and shape function $\vec{S}_i(\cdot)$. The dense pointwise correspondence between shapes is then given by $\vec{S}_i(\vec{x}) \sim \vec{S}_j(\vec{x})$. The statistical properties of the collection of shapes is then captured by the mean shape $\widehat{\vec{S}}$ and the shape covariance matrix$\vec{D}$ (with elements D_{ij}):

$$\widehat{\vec{S}}(\vec{x}) \doteq \frac{1}{n_S}\sum_{i=1}^{n_S}\vec{S}_i(\vec{x}),\ D_{ij} \doteq \int dA(\vec{x})(\vec{S}_i(\vec{x}) - \widehat{\vec{S}}(\vec{x})) \bullet (\vec{S}_j(\vec{x}) - \widehat{\vec{S}}(\vec{x})), \tag{1}$$

where $\bullet$ is the usual vector dot product in $\mathbb{R}^3$, and $dA(\vec{x})$ is the area measure defined on the mean shape (see [2] for further details). For a linear model, PCA is performed on the covariance matrix, which yields the eigenvectors (representing the modes of shape variation), and the corresponding eigenvalues $\{\lambda_\alpha\}$.

2.1 Manipulating Correspondence

The groupwise correspondence across the set of shapes is then manipulated by re-parameterising each shape. For re-parameterising just the i^{th} shape, we have the re-parameterisation function $\phi_i(\vec{x})$, where:

$$\vec{x} \overset{\phi_i}{\longmapsto} \vec{x}' \doteq \phi_i(\vec{x})\ \&\ \vec{S}_i \overset{\phi_i}{\longmapsto} \vec{S}'_i,\ \text{where } \vec{S}'_i(\vec{x}') \doteq \vec{S}_i(\vec{x})\ \Rightarrow\ \vec{S}_j(\vec{x}) \sim \vec{S}_i(\vec{x}) \overset{\phi_i}{\longmapsto} \vec{S}_j(\vec{x}) \sim \vec{S}'_i(\vec{x}) \doteq \vec{S}_i\left(\phi_i^{-1}(\vec{x})\right). \tag{2}$$

Hence both the shape-functions $\{\vec{S}_i(\cdot)\}$ and the groupwise correspondence change under re-parameterisation. The re-parameterisation is hence valid provided that ϕ_i is a homeomorphic mapping. Manipulating the groupwise correspondence changes the mean shape and the shape covariance matrix (1), hence leads to a different model.

Note that the re-parameterisation function ϕ_i acts on the space of the original chart $\vec{\mathfrak{X}}$, *not* on the shape surface itself. Hence the problem of constructing valid re-parameterisation functions depends on the initial chart construction (See §3).

2.2 The Objective Function

A general objective function will be denoted by $\mathcal{L}_S(\{\vec{S}_i(\cdot)\})$, which depends *explicitly* on the set of shape functions $\{\vec{S}_i(\cdot)\}$, hence only *implicitly* on the set of re-parameterisation functions $\{\phi_i(\cdot)\}$, as described above. In this paper, we will focus on objective functions based on the MDL principle (e.g., [1, 2]), where the explicit dependance is on the set of eigenvalues $\{\lambda_\alpha\}$ of $\vec{D}$ (1), which we will denote by $\mathcal{L}_S = \mathcal{L}_{\text{MDL}}(\{\lambda_\alpha\})$.

3 Coordinate Charts for Shapes: Shape Images

A typical shape will be represented by some finite set of points, where the continuous shape surface is obtained from these by some sort of interpolation. So, these points could be spline control points. For the simplest case of bilinear interpolation, the shape is represented by a triangulated mesh. In previous work [1], the initial chart $\vec{\mathfrak{X}}$ was constructed by mapping the shapes onto the topological primitive (e.g., a sphere). Re-parameterisation functions then act on the surface of the unit sphere, which maintains the topology of the shape. However, re-constructing the positions of manipulated points on the shape involves interpolation on the triangulated mesh. Such a procedure is computationally intensive.

The computational cost could be considerably reduced if instead interpolation could be performed on some regular grid, which means that our chart would be some region of the plane. It is a well-known result that a sphere cannot be totally covered by a single such chart, hence using just a single chart will mean that there will be points where derivatives cannot be defined. Such a chart can be constructed following the work of Praun & Hoppe [8]. As is shown in Figure 1, the shape is first mapped to the unit sphere (for example, by using the method of method of Brechbühler et al. [9]). Then the sphere is mapped to the inscribed octahedron using a gnomonic projection [8]. The octahedron is then unfolded to give the final square chart. Whichever unfolding is used, there are points where derivatives cannot be defined (the points where an edge of the chart folds back on itself, marked by plus symbols in the Figure), but these are just a few points. A regular pixel grid is then constructed on this chart. Praun and Hoppe stored information about texture, colour and geometry at each pixel, and hence christened them *geometry images*. We instead store the coordinates of the corresponding point on the physical shape for each pixel, hence call these *shape images*. For the i^{th} shape image, we have the common grid of all pixel positions $\vec{\mathbf{X}} = \{\vec{x}\}$. The shape is sampled at each point on this grid to construct the shape image $\mathbf{S}_i = \{\vec{S}_i(\vec{x}) : \vec{x} \in \vec{\mathbf{X}}\}$.

4 Non-Parametric Fluid Regularisation

As mentioned previously, the groupwise correspondence problem is in general ill-posed, and can only be solved if some regularisation is involved. One approach (the parametric approach), is to consider only re-parameterisation transformations from some restricted set of parametric transformations (e.g., transformations of the sphere based on Cauchy kernels [1], B-splines [4], and polynomials [5]). The optimisation then proceeds on the finite-dimensional space of parameters. This can hence be considered as *hard* regularisation, in that the subspace, hence allowed transformations,

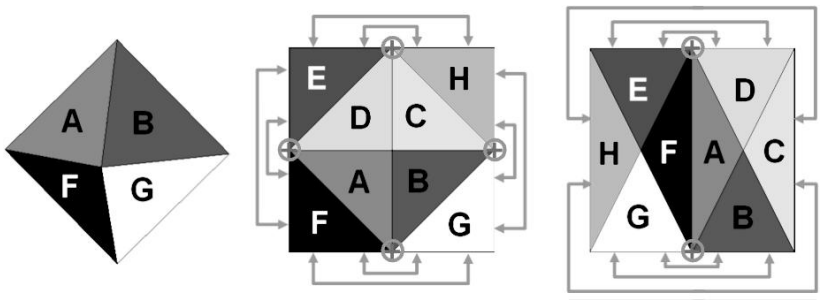


Figure 1: **Left:** The octahedron, **Centre & Right:** Two unfoldings. The labels show the faces, the gray lines the boundary conditions, and the gray circles the singular points.

is strictly bounded. But as mentioned previously, the required correspondence may be difficult to find or represent within the limited space of allowed transformations.

An alternative approach can be thought of as *soft* regularisation, where the subspace is not held fixed, but can be expanded provided the evidence from the data warrants it. We hence adopt the fluid-based approach as introduced by Christensen et al. [6] for the case of image registration, which is a particular form of soft regularisation.

4.1 Fluid-Based Regularisation

Rather than working with the re-parameterisation function directly, we consider the related displacement field $\vec{u}(\vec{x})$:

$$\vec{y} + \vec{u}(\phi(\vec{y})) \doteq \phi(\vec{y}), \;\; \vec{x} = \phi(\vec{y}) \;\; \Rightarrow \;\; \vec{y} = \vec{x} - \vec{u}(\vec{x}). \tag{3}$$

The shape part of the objective function then gives a driving force:

$$\vec{F}^S(\vec{x}) \propto \frac{\delta \mathcal{L}_S(\{\vec{S}_i\})}{\delta \vec{u}(\vec{x})}. \tag{4}$$

Hence the optimum value of the shape part alone corresponds to zero driving force. The displacement field (hence the re-parameterisation function) is then made a function of computational time t. The driving force from the shape part of the objective function is then balanced by a force from the regulariser, based on the viscous forces for a fluid with displacement field $\vec{u}(\vec{x}, t)$:

$$\vec{v}(\vec{x}, t) \doteq \left.\frac{\partial}{\partial t}\right|_{\vec{y}} \phi(\vec{y}, t) = \left.\frac{\partial}{\partial t}\right|_{\vec{y}} \vec{u}(\phi(\vec{y}, t), t) = \frac{\partial \vec{u}(\vec{x}, t)}{\partial t} + (\vec{v}(\vec{x}, t) \bullet \vec{\nabla})\vec{u}(\vec{x}, t). \tag{5}$$

$$\vec{F}^{visc}(\vec{x}, t) = \mu \vec{\nabla}^2 \vec{v}(\vec{x}, t) + (\lambda + \mu)\vec{\nabla}(\vec{\nabla} \bullet \vec{v}(\vec{x}, t)), \tag{6}$$

$$\vec{F}^{visc} + \vec{F}^S = 0. \tag{7}$$

$\vec{v}(\vec{x}, t)$ is the Eulerian velocity of the fluid. The fluid parameters are the shear viscosity μ and second viscosity coefficient λ (related to bulk viscosity).[1]

We can calculate the viscous force (6) on the shape image as in [7]. However, it could be argued that given the mapping between the shape and the shape image, this means that the regulariser does not act uniformly on the shape, but instead depends on the details of the mapping. A more principled approach would be to calculate the fluid forces directly on the shape itself, for a fluid flowing on the curved surface of the shape itself. In mathematical terms, this would mean replacing the ordinary derivatives in (6) by covariant derivatives, which then automatically corrects for the details of the mapping, and produces an exact derivative on the shape itself.

Let P be a particular point on the shape, with position $\vec{S}_P$. The corresponding point on the shape image can, without loss of generality, be assigned cartesian coordinates $(0, 0)$, with shape image coordinates about this point being (i, j). We then have the shape difference function about P: $\vec{S}_{diff} \doteq \vec{S} - \vec{S}_P$. It is then a standard result (see [10] for a simple derivation) that the induced Riemannian metric on the shape at P can be written as:

$$M_P \doteq \begin{pmatrix} (\nabla_i \vec{S}_{diff})^2 & (\nabla_i \vec{S}_{diff}) \bullet (\nabla_j \vec{S}_{diff}) \\ (\nabla_i \vec{S}_{diff}) \bullet (\nabla_j \vec{S}_{diff}) & (\nabla_j \vec{S}_{diff})^2 \end{pmatrix}. \tag{8}$$

[1] As regards a physical model, this corresponds to a fluid where inertial terms have been neglected, as well as pressure terms, acting under a driving force. Shear viscosity acts so that if we apply a driving force to one part of a fluid, neighbouring parts tend to get dragged along as well. Whereas bulk viscosity acts to resist non-uniform compression or expansion.

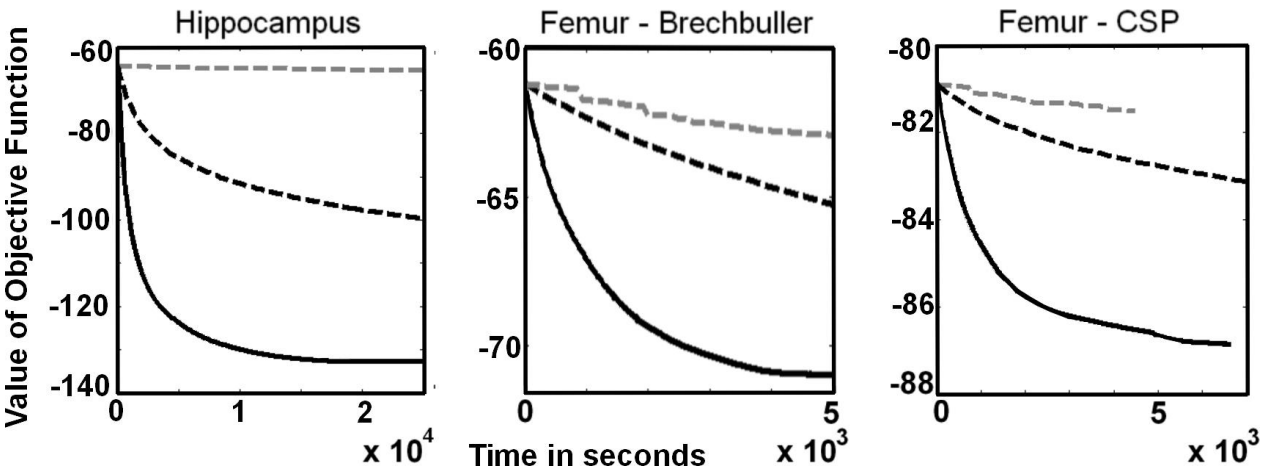


Figure 2: Values of the objective function plotted as a function of time.**Left:** Hippocampus dataset, **Centre & Right:** Femur dataset with varying initial mappings to the sphere.
Gray dashed line: Parametric regularisation on the sphere.
Black dashed line: Parametric regularisation on the shape image.
Black solid line: Fluid regularisation on the shape.

This includes only first-derivatives of the shape wrt the shape image coordinates, which are evaluated directly on the shape image itself. In general, covariant derivatives of vectors or tensors include terms containing christoffel symbols, which are constructed from first-derivatives of the metric. The christoffel symbols hence include *second* derivatives of the shape, which can be numerically unstable to compute. We hence make the approximation that the shape is piecewise linear, and hence neglect the terms involving the christoffel symbols. The viscous force calculated on the shape can then be approximated by:

$$\left(\vec{F}^{visc}\right)_c = \mu \left[(M_P^{-1})_{ab}\nabla_a\nabla_b\right]\vec{v}_c + (\lambda+\mu)\left[(M_P^{-1})_{ca}\nabla_a\nabla_b\right]\vec{v}_b, \tag{9}$$

where the sub-scripts indicate components in the shape image coordinates. We hence see that the effect of calculating the viscous forces on the shape itself is that the effective viscous forces on the shape image become a bit more complicated, where we have a position-dependant mixture of spatial derivatives. Although this mixture depends on position, M_P^{-1} has only to be calculated once for each sample point in the shape image, since the shape image (and hence M_P^{-1}), remains fixed as the displacement field evolves. For exact details of the fluid implementation and the details of calculating the shape term driving force (4), see [10].

5 Experiments

The aim of these experiments was to perform a direct comparison of the charting method of shape images as compared to other techniques, and to compare parametric with fluid regularisation, using the same MDL objective function. The datasets used were a set of 82 surfaces of the left hippocampus, and a set of 15 examples of the distal third of the human femur. Both were segmented from 3D MR images. We used several methods to map shapes to the unit sphere, the method of Brechbühler et al. [9], as well as the Consistent Spherical Parameterisation (CSP) method of Davies et al. [11]. For parametric regulariser, we used a set of parametric transformations based on clamped-plate splines, applied either on the surface of the sphere (which is very similar to the method used by Heimann et al. [12]) or on the shape image, with gradient descent optimisation. The final comparison is then with the non-parametric fluid regularisation on the shape. In Figure 2, we show the values of the objective function plotted as a function of time in seconds. In Table 1, we show the results for a quantitative comparison of the convergence rates for the graphs shown in Figure 2, with higher values indicating faster convergence. Starting with the hippocampus dataset (Figure 2, **Left**), we can see that for parametric regularisation (dashed lines), the chart mapping to the shape image improves the convergence rate considerably when compared to using the sphere. This result is repeated for the femur data, for both methods of mapping to the sphere.

Table 1: A table showing the convergence rate of the methods for the various datasets, compared to the parametric case on the sphere. Higher values indicate faster convergence.

	Hippocampus	Femur	
		Brechbuller	CSP
Parametric on sphere	1	1	1
Parametric on shape image	24	4	9
Fluid on the shape	**384**	**56**	**153**

When we compare fluid with parametric regularisation (solid line compared to dashed lines), we see that fluid regularisation gives an appreciable advantage. If we compare the two methods for mapping the shape to the sphere (**centre & right** Figure 2 & Table 1), it can be seen that for this particular dataset, the exact result for fluid regularisation and rate of convergence does depend on the choice of chart. Note that for fluids, we would expect a possible difference in convergence rates for different parameterisations, since the scaling criterion depends on pixel size. Hence a region of the shape where the mapped pixels are small in terms of shape units will give a slower flow that a mapping where the pixels are larger. But we also might expect an additional effect, given that different initial sphere mappings lead to a different initial correspondence, hence we have to follow a different path in the space of groupwise correspondence. We also have to allow for the probability that the algorithm may get stuck in a local minimum. Note however that the degradation in performance for the Brechbühler et al. mapping as opposed to CSP is seen for both parametric regularisation on the shape image, and non-parametric regularisation. We cannot however make a direct comparison between the two methods of mapping to the sphere, since the values of the objective function cannot be directly compared, as the differing scales on the y-axes show.

6 Discussion & Conclusions

The results clearly show that for parametric regularisation, the use of shape images improves the convergence rate by about an order of magnitude. It is clear that using fluid regularisation can give a further improvement, leading to an overall improvement by up to two orders of magnitude.

What is not clear from this data is the exact effect of choice of initial sphere mapping. It should also be noted that the femur dataset contains only 15 examples, whereas the hippocampus set contained 82. It could be that some of the results we are seeing on the femur dataset are the result of this small sample size, and not indicative of the behaviour for larger datasets.

The possibility of getting stuck in a local minimum should also be considered, and we intend to investigate techniques for introducing a stochastic element, so that the fluid can be moved out of such a minimum.

Finally, it would also be desirable to perform a full quantitative evaluation of the final models obtained, and analyse the differences (if any) in groupwise correspondence that these methods give, for a range of biological datasets.

Investigating the detailed behaviour of the fluid regularisation approach, and further refinement of the algorithm forms the basis of our future research.

Acknowledgements

Our thanks to Tomos Williams and Chris Wolstenholme for processing the femur dataset. . Also thanks to Anna Mills for helping with the exact details of the implementation.

References

1. R. H. Davies, C. J. Twining, T. F. Cootes et al. "3D statistical shape models using direct optimisation of description length." In *Proceedings of ECCV 2002*, pp. 3–21. 2002.
2. R. H. Davies, C. J. Twining, T. F. Cootes et al. "A minimum description length approach to statistical shape modelling." *IEEE Transactions on Medical Imaging* **21**, pp. 525–537, 2002.
3. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific Press, 1989.
4. P. Horkaew & G. Yang. "Optimal deformable surface models for 3D medical image analysis." In *Proceedings of IPMI 2003*, pp. 12–24. 2003.
5. J. Hladůvka & K. Bühler. "MDL spline models: gradient and polynomial reparameterisations." In *Proceedings of* BMVC *2005*, pp. 869–878. 2005.
6. G. Christensen, R. Rabbitt & M. Miller. "Deformable templates using large deformation kinematics." *IEEE Transactions on Image Processing* **5**, pp. 1435–1447, 1996.
7. R. H. Davies, C. J. Twining, T. G. Williams et al. "Building MDL statistical shape models using shape images and non-parametric regularisation." In *Proceedings of ISBI 2007*. 2007.
8. E. Praun & H. Hoppe. "Spherical reparameterization and remeshing." In *Proceedings of SIGGRAPH 2003*, pp. 340–349.
9. C. Brechbühler, G. Gerig & O. Kübler. "Parameterisation of closed surfaces for 3D shape description." *Computer Vision, Graphics and Image Processing* **61**, pp. 154–170, 1995.
10. C. Twining, R. Davies & C. Taylor. "Non-parametric surface-based regularisation for building statistical shape models." In *Proceedings of IPMI 2007*. 2007.
11. R. H. Davies, C. J. Twining & C. J. Taylor. "Consistent spherical parameterisation for statistical shape modelling." In *Proceedings of ISBI 2006*, pp. 1388–1391.
12. T. Heimann, I. Wolf, T. Williams et al. "3D active shape models using gradient descent optimization of description length." In *Proceedings of IPMI 2005*, pp. 566–577.

Finding indications of Osteoporosis in Panoramic Dental Radiographs using Active Shape Models: The OSTEODENT project.

P. D. Allen[a*] J. Graham[a] E. J. Marjanovic[a] R. Jacobs [b] K. Karayianni[c] C. Lindh[d]
P. F. van der Stelt[e] K. Horner[f]
H. Devlin [f]

[a] ISBE, University of Manchester, UK.[b] Oral Imaging Centre, Katholieke Universiteit Leuven, Belgium,
[c] Dental School, University of Athens, Greece, [d] Faculty of Odontology, Malmo University, Sweden,
[e] Academic Centre for Dentistry, Amsterdam, The Netherlands, [f] School of Dentistry, University of Manchester, UK.

Abstract. Dental panoramic tomograms (DPTs) are one of the most widely used radiological investigations, usually fior dental purposes. Here we describe a novel application of these images for estimating reduced bone mineral density (BMD). It has been shown that measurement of the width of the inferior mandibular cortex (IMC) on DPTs is predictive of BMD in hip and spine osteopenia and osteoporosis. We have trained an active shape model on 132 DPTs of female subjects and used it to locate and measure the width of the IMC. The method has been tested on a separate set of DPTs from 664 female subjects whose BMD has been established by dual-energy X-ray absorptiometry. We demonstrate that widths measured after fully automatic search are significantly correlated with BMD. ROC analysis for the detection of osteoporosis gives area-under-curve (AUC) values of 0.782 using fully automatic analysis, which is equivalent to human expert measurement, but with greater reliability. A straightforward user interaction improves the anatomical precision of the analysis and increases the AUC value to 0.82.

1 Introduction

Osteoporosis is a serious health problem among the elderly. It results in a general loss of bone mineral density (BMD) leading to an increased risk of fracture. Hip fracture in particular is associated with increased rates of morbidity and mortality. Definitive assessment of BMD uses dual-energy X-ray absorptiometry (DXA). Referral for DXA is often based on factors such as previous fracture, family history, and height loss. However, as mandibular BMD is related to systemic BMD [1] there has recently been interest among dental researchers in the possibility of identifying those at risk of reduced BMD from dental radiographs.

An example of a dental panoramic tomogram (DPT) of a normal patient is shown in figure 1. The schematic close up of the right mandible in figure 2 indicates the inferior mandibular cortex (IMC), a region of cortical bone. It has been shown that careful manual measurements by experts of the thickness of this cortex is correlated with systemic BMD and hence osteoporosis [2]. This measurement is conventionally made at a point closest to the mental foramen, and is referred to as the Mental Index (MI) (figure 2); measurement at this site is a better indicator of low BMD than the equivalent indices at the gonion (GI) and the ante-gonion (AI) (figure 1) [2]. Since the mental foramen is a very indistinct feature and the endosteal border can become very diffuse in cases of osteoporosis, there is considerable room for subjectivity in the precise placement of the MI measurement. It is therefore impractical for the measurement to be carried out by general dental practitioners, and expert radiological attention is required [3].

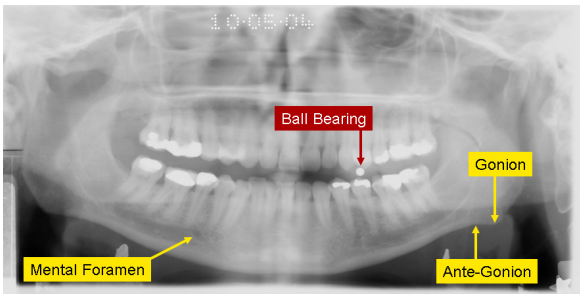


Figure 1. *An example of a panoramic dental tomogram of a normal (non-osteoporotic) patient. The positions of anatomical points key to manual annotation and the ball bearing calibration object are shown.*

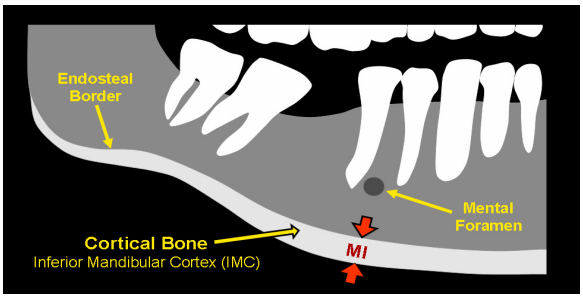


Figure 2. *Schematic diagram of the right mandible, showing the point at which the inferior mandibular cortex thickness is measured by dentists (mandibular index MI).*

Dental radiographs account for 32% of all medical radiological examinations in the UK [4]. This widespread investigation opens the possibility of obtaining valuable medical information about patients' osteoporotic status from a routine

*danny.allen@manchester.ac.uk

radiological examination. To be widely applicable, the measurement of cortical width would need to be made routinely and easily. This paper describes an automatic method of measuring radiographic indices using the Active Shape Model (ASM) method [5] to locate the upper and lower borders of the inferior mandibular cortex, and hence measure its thickness.

2 Image Data

In this study we have used separate training and test data sets.

2.1 Training Images:

The training images had been collected for a previous study [6] and consisted of 132 female patients aged 45-55 who attended the University Dental Hospital, Manchester for routine dental treatment. DPTs were obtained using either a Cranex DC-3 unit (Soredex Orion Corporation, Finland) or a Planmeca PM 2002C unit (Planmeca, Finland) using the same film/cassette combination. The films were digitised using a Kodak LS85 digitiser (Eastman Kodak, Rochester, NY) at a resolution of 25.64 pixels/mm. Manual annotation of these digitised images is described below (section 3.1).

2.2 Test Images:

The test Images consisted of 664 DPTs from female patients recruited from the area surrounding four European centres (Manchester, Leuven, Athens and Malmö). In three centres images were obtained using a conventional film/cassette combination and then digitised in the same way as the training data; at the remaining centre the images were captured digitally. A 3mm diameter ball bearing set in a plastic cube gripped between the teeth was used to normalise the images for variations in calibration between centres (see figure 1). BMD ground truth was established for each subject using DXA at the four recruitment centres. The European spine phantom (ESP) was used to standardise measurements between equipment at different sites using the method described by Pearson [7]. T and Z scores were calculated from the DXA using Hologic reference data for the lumbar spine and NHANES reference data for the proximal femur [8] Patients were diagnosed as osteoporotic according to the World Health Organization (WHO) criteria: T-score value ≤ -2.5, i.e. those havng bone mineral density 2.5 SD or more below the mean value of the young sex matched reference population.

3 Active Shape Model Search

Active Shape Models (ASMs) have been thoroughly described elsewhere [5, 9, 10] and only a brief description will be given here. The key point is that a Point Distribution Model (PDM) model is 'trained' using points placed on a training set of example shapes, usually manually (see section 3.1), at anatomically consistent locations around the border of the object. The PDM provides shape constraints on a flexible template search [5]. While a number of search strategies may be employed, in this application we use a process of iterative local refinement in which the position of each landmark point is updated to the nearest best grey-level match within the constraints of the PDM, the best match being based on edge strength and edge direction. The search is divided into two phases: firstly locating the inferior border only, and using this result to initiate a search using the complete model of the IMC. This two-stage search allowed the rather faint endosteal (upper) border to be defined robustly by first locating the much better defined inferior border.

3.1 Manual Annotation

To build a PDM a set of consistently-located points need to be placed on the inferior mandibular cortex in each of the training images. This was done by two experts using a custom-written graphical user interface [11]. Two key anatomical landmarks, the points closest to the mental foramen, and the points at the ante-gonion (see figure 1), henceforth referred to as the MF and AG points, were used to define the region to be modelled. Between these a number of equally spaced points were placed to define the upper and lower borders of the IMC.

The nature of the image formation results in shadows of the spine and other anatomical features being superimposed on the images, particularly in the centre of the image, between the MF points and laterally beyond the AG. These shadows make the endosteal border poorly defined in these regions. ASM search was therefore conducted using a model built from 200 points between the AG and MF points (50 equally spaced points on the upper and lower boundaries on each side).

3.2 Experimental Procedure

To test fit accuracy, leave-one-out cross-validation was used on the training set, using the ground-truth provided by the manual annotations. To evaluate sensitivity to BMD, search was applied to the unseen test set followed by ROC analysis, using osteoporosis ($T < -2.5$) as the classification. Two versions of the ASM search described in section 3 were tested experimentally on both data sets:

The first was an unconstrained search, without any manual interaction, employing a coarse-to fine multi-resolution search strategy [5] initialised at the mean position and pose calculated from the training data. We refer to fits determined this way as 'unconstrained fits' or 'UFits'.

The second version used four manually defined reference points on the lower mandible edge at the left and right AG and MF as reference points. The initial configuration for the search was obtained by stretching and positioning the mean shape so that its AG and MF points matched the reference points. An edge based ASM search was then initiated, making no further reference to the manual points during the search. The results of this search were then used to initialise a full endosteal border ASM search by warping the mean example of the full endosteal PDM so that its lower edge matched the results of the lower edge ASM fit.

The use of this straightforward interaction allowed us to largely decouple the effects of location and shape in ASM search. Starting the search so close to the true position effectively guarantees that the search will finish up with the correct pose. The quality of ASM fit is determined solely by the ability of the PDM to represent the variation in shape that occurs among the images. We refer to fits determined this way as 'constrained fits' or '4PFits'.

4 Results

To assess accuracy we use the mean point-to-point, and the mean point-to-curve difference between the final point positions and the manually placed points. Since our goal is to measure mandibular cortical thickness, we also compare the measurements of thickness derived from model fits with those from manual annotation. The thickness is measured as the distance between corresponding points on the lower and upper border of the mandibular cortex. The comparison is done using a Bland and Altman analysis [12] in which the difference between two sets of measurements are plotted against their mean. From this analysis the *bias* is measured as the mean difference between the two sets of measurements, and the limits of agreement are the mean difference +/- 1.96 standard deviations.

The sensitivity of the derived measurements to osteoporosis is assessed by calculating the correlation coefficient between the measured width of the IMC and BMD. Receiver Operator Characteristic (ROC) analysis [13] provides a method of assessing the sensitivity and specificity of the measures. The area under the curve (AUC) can be used to quantify the overall diagnostic efficacy of the parameter in question - ranging from 0.5 (no better than chance) to 1.0 (perfect discrimination).

4.1 Fit Accuracy (Training Images)

These tests use the training set of images (section 2.1) to conduct leave-one-out trials of fit accuracy. The accuracy results are summarised in table 1, showing means and variances of point-to-point and point-to-curve errors between manual delineation of the IMC and different ASM fits. Four comparisons are made:

1. **Manual 1-2** - Comparison of the manual annotation of the two observers.
2. **Manual-UFit** - Comparison of the mean of the manual points with that of the unconstrained model fit.
3. **Manual-4PFit** - Comparison of the mean of the two sets of manual points against the results of the four-point initialised model fit (4PFit).
4. **Fit1-Fit2** - This provides an indication of the effect of variation in interactively placing the four initialisation points for the 4PFit. To estimate the magnitude of this effect we perform two 4PFits, each initialised by a different observer.

The orthogonal distance of any given point in the model from the border of the IMC is clearly defined by the local grey level gradient. However there are no features such as edges to define its position *along* the edge of the mandible.

Table 1. Model Fit Accuracy Results (See Text).

Comparison	Point Difference (mm - $\mu(\sigma)$)		Cortical Thickness (mm)		
	point-to-point	point-to-curve	bias	lower lim	upper lim
Manual 1-2	2.45 (2.45)	0.31 (0.33)	-0.02	-0.77	0.72
Manual-UFit	5.73 (4.57)	0.49 (1.58)	-0.30	-1.01	0.40
Manual-4PFit	0.59 (0.54)	0.31 (0.40)	-0.25	-0.79	0.28
Fit1-Fit2	2.31 (2.44)	0.14 (0.24)	-0.04	-0.21	0.29

The unconstrained fit can only be expected to locate the transverse position of the border, while the correct anatomical position of the resulting points along the border with respect to the AG and MF points cannot be guaranteed. As a result the fully automatic method (UFit) results in much larger point-to-point error than the difference between two manual observers. The relatively low corresponding point-to-curve error shows the correct transverse placement of the model. The initialisation constraint introduced in the 4PFit results in a much lower point-to-point error than the difference between two manual observers. The corresponding point-to-curve errors are very similar to those between the two manual observers. The Bland-Altman results also appear in table 1 and show that the unconstrained fit has similar limits of agreement to the comparison of manual observers. The constrained fit has rather tighter limits of agreement than either. Both the ASM methods show a slight bias in tending to measure the thickness as less than the average human measure. The Fit1-Fit2 measure shows similar point-to-point consistency with human observers, though rather better point-to-curve consistency. The Bland-Altman results show very little difference in width measurement.

4.2 Sensitivity to Reduced BMD (Test Images)

Once the model is fitted, cortical thickness may be measured at any site between the MF and the AG. The correlation between measured thickness and BMD varies with measurement site. Figure 3 shows the variation with position along the mandible of the correlation with BMD measured at the hip for the 4PFit results (for each point along the mandible, each thickness measurement is the average of the thickness measured at 10 consecutive model points). The cortical thickness measurements show greatest sensitivity to BMD in the lateral halves of the mandible, and least around the MF points. This would suggest that the MF points are not the optimal place to measure cortical thickness in the detection of osteoporosis. The BMD values from the femoral neck and spine show a similar pattern and the minimum T-Score of the three sites is also shown in the figure.

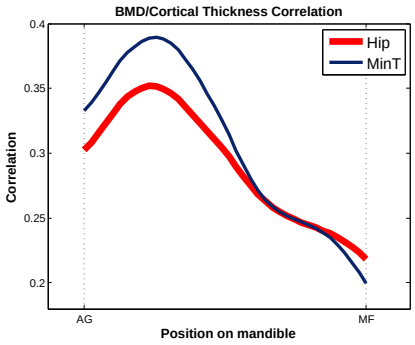


Figure 3. *Correlation coefficient between cortical thickness and Hip BMD plotted as a function of position along mandible.*

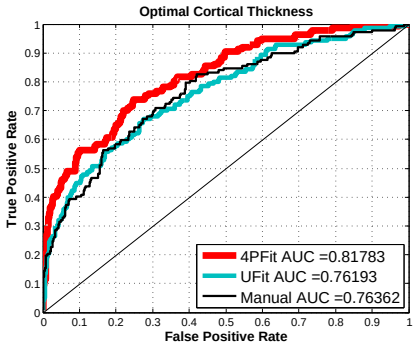


Figure 4. *ROC curve for reduced BMD using the cortical thickness from the optimal region suggested by figure 3 as the discriminating parameter.*

Using the average cortical thickness over this region as the predictor of osteoporosis results in the ROC curve shown in Figure 4. The curves for both 4PFit and UFit searches are shown. The figure also shows the curve generated by measuring the thickness manually from the radiographs - the average of five measurements by expert radiologists at two of the collaborating sites. AUC measurements for manual and Ufit measurements are 0.76. 4PFit search achieves a significantly better AUC of 0.82 (p=0.007).

5 Discussion

We have shown that we can measure the width of the inferior mandibular cortex in panoramic dental tomograms accurately and reliably using an edge-based ASM. The calibration using the ball-bearing was used to correct minor differences in image scale-factor between centres. The optimal position for making the measurement is different from that conventionally used for manual measurement. This is also true if the thickness measurements are made from the manually-placed annotation points (data not shown). Fully automatic measurement is in specificity and sensitivity equivalent to expert manual measurement. The greater anatomical precision that results from the four-point initialisation results in significantly improved prediction of osteoporosis. This initialisation requires only the lower mandible edge to be identified - a clearly visible feature in all patients - and the points do not require to be placed with great accuracy. While fully automatic operation would be desirable, the use of this straightforward interaction delivers benefits.

Osteoporosis diagnosis is not part of everyday dental practice and so any involvement in this task should be facilitated for the dentist. Measurement of cortical width will only be practical if it is fully automatic or very nearly so. The limited and straightforward interaction described here may be sufficiently unobtrusive to be practical. However the results of automatic search indicate that useful measurement can be made without dentist involvement at all. The improved specificity and sensitivity arising from being able to make measurements at anatomically precise locations holds out the possibility of improved diagnostic performance.

Acknowledgment

This work was supported by a research and technological development project grant from the European Commission Fifth Framework Programme 'Quality of Life and Management of Living Resources' (QLK6-2002-02243).

References

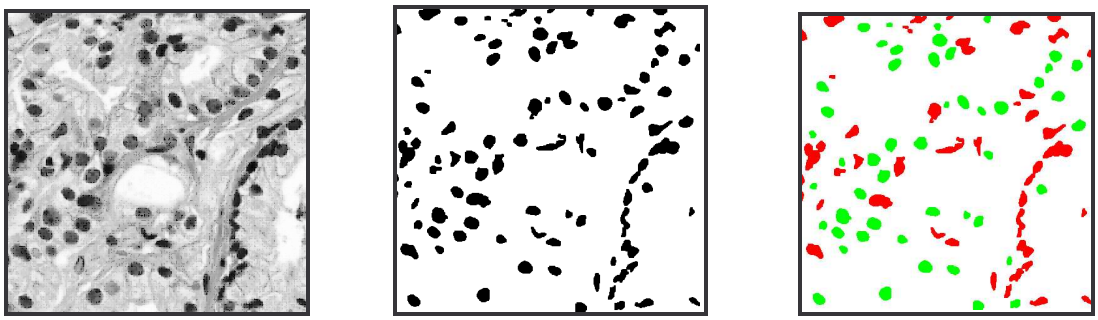
1. K. Horner, H. Devlin, C. W. Alsop et al. "Mandibular Bone Mineral Density as a Predictor of Skeletal Osteoporosis." *The British Journal of Radiology* **69(827)**, 1996.
2. H. Devlin & K. Horner. "Mandibular radiomorphometric indices in the diagnosis of reduced skeletal bone mineral density." *Osteoporos International* **13(5)**, 2002.
3. C. V. Devlin, K. Horner & H. Devlin. "Variability in measurement of radiomorphometric indices by general dental practitioners." *Dentomaxillofacial Radiology* **30(2)**, 2001.
4. R. J. Tanner, B. F. Wall, P. C. Shrimpton et al. "Frequency of Medical and Dental X-Ray examinations in the UK." *1997/98/ NRPB-R320. Chilton: National Radiological Protection Board* 2001.
5. T. F. Cootes, C. J. Taylor, D. H. Cooper et al. "Active Shape Models - Their Training and Application." *Computer Vision and Image Understanding* **61(1)**, 1995.
6. K. Horner, H. Devlin & L. Harvey. "Detecting patients with low skeletal bone mass." *Journal of Dentistry* **30(4)**, 2002.
7. J. Pearson, J. Dequeker & M. H. et al. "European Semi-Anthropomorphic Spine Phantom for the Calibration of Bone Densitometers - Assessment of Precision, Stability and Accuracy - the European Quantitation of Osteoporosis Study- Group." *Osteoporosis International* **5**, 1995.
8. A. C. Looker, H. W. Wahner, W. L. Dunn et al. "Updated data on proximal femur bone mineral levels of US adults." *Osteoporosis International* **8**, 1998.
9. G. Behiels, F. Maes, D. Vandermeulen et al. "Evaluation of Image Features and Search Strategies for Segmentation of Bone Structures in Radiographs Using Active Shape Models." *Medical Image Analysis* **6(1)**, 2002.
10. T. F. Cootes, A. Hill, C. J. Taylor et al. "Use of Active Shape Models for Locating Structures in Medical Images." *Image and Vision Computing* **12(6)**, 1994.
11. H. D. D.J.J Farnell1, J. Graham1 & K. Horner. "Measuring Indicators of Osteoporosis from Dental Panoramic Radiographs: Preliminary Results." *Medical Image Unserstanding and Analysis* pp. 111–114, 2005.
12. J. M. Bland & D. G. Altman. "Statistical methods for assessing agreement between two methods of clinical measurement." *Lancet* **8**, 1986.
13. J. A. Hanley & J. M. McNeil. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology* **143**, 1982.

Detection of Nuclei by Unsupervised Manifold Learning

Muhammad Arif[a] and Nasir Rajpoot[b]

[a] Pakistan Institute of Engineering and Applied Sciences, Nilore, Islamabad, Pakistan
[b] Department of Computer Science, University of Warwick, Coventry, UK

Abstract: Shape related signatures of nuclei in a tissue section are important for diagnosis and prognosis of cancer. Understandably, the process of demarcation of nuclei for cytometry with high degree of confidence is the most difficult part as the tissue section is fraught with staining artifacts and frequently contains other objects such as overlapping nuclei, nuclear debris, and extracellular structures. In this paper, we address this problem using a novel clustering algorithm for various shapes in prostate histopathology images using an unsupervised manifold learning paradigm. Experimental results with two-dimensional embedding of the shapes using diffusion maps demonstrate that various shapes in the tissue section are organized in accordance to the degree of complexity of their boundaries. This important observation can be exploited in the development of computerized techniques for image based cytometry.



(a) (b) (c)

Figure 1: (a) A typical prostate tissue specimen (b) Binary image showing potential nuclei, overlapping nuclei, nuclear debris, and extracelluar structures (c) Nuclei detected (shown in green) by our detection algorithm; the objects shown in red are the ones rejected by the algorithm.

1 Introduction

Today cancer is one of the major causes of deaths throughout the world. Mortality rate can be reduced if it is diagnosed at an early stage. Traditionally, cancer and its level of malignancy (i.e., grading) are detected by pathologists who study microscopic images of tissues removed from the patients. For over a decade now, there has been a considerable amount of interest in research and development of computer-assisted systems for diagnosis and prognosis of cancer, due to the following reasons:

1. The judgment of a pathologist is subjective and significant inter- and intra-observer variability and poor reproducibility has been reported [1, 2]. Analyzing the tissues quantitatively using machine vision techniques can reduce this variability.
2. In the very early stages of the malignancy, the pathologist may be unable to detect the changes in the tissues. By introducing objectivity in the analysis, these otherwise undetectable changes can be detected [3].
3. A significant number of imaging modalities have been developed in recent years. To increase the reliability of judgment, it would be more judicious to fuse the information extracted from the huge amount of data generated by these imaging modalities as well as from the computer based analysis of tissue specimens.
4. The task of analyzing histopathological data by human experts is dull, tedious and laborious. Machines are well suited for such chores with high throughput. Hence nature of the job also demands automation of the process or at least development of ancillary systems to decrease the workload for pathologists.

One of the main areas of research into quantitative histopathology for computer-assisted diagnosis systems has been to quantify changes at one of the two levels: tissue level and cellular level [4]. Computerized methods at both levels extract diagnostically important features from stained tissues and employ statistical and machine learning methods to detect any abnormality. Most of the techniques at tissue level rely on the texturedness of the tissue and on the spatial arrangement of cells in histological sections [5, 6, 7, 8, 9, 10]. The second class of methods exploits the morphological features of cells and nuclei [11,14,15], such as quantitative descriptors of nuclei including area,

E-mail addresses: {arif,nasir}@dcs.warwick.ac.uk

radius, major/minor axis, compactness, perimeter, fractal dimension and texture of the chromatin inside the nucleus, to name a few [13].

Nuclei are particularly very important objects in the tissue to be analyzed for the development of computer-assisted systems as they can be easily differentiated from rest of the structural components by selective staining methods. The success of nuclear morphometry based techniques depends on how reliably the nuclei have been recognized in a histological section fraught with overlapping nuclei, nuclear debris, cellular and extra cellular structures, staining and image acquisition related artifacts.

Inspired by the seminal work of Coifman et al. on diffusion maps [16], we address the problem of separating various parts of the tissue in a stained specimen by posing it as an unsupervised manifold learning problem in a low-dimensional space and by using diffusion maps for manifold learning. Preliminary results are very encouraging and demonstrate that various objects in the tissue section are organized in the new representation (using only 2 dimensions) in accordance with the degree of complexity of their boundaries. The proposed detection algorithm is fast and only requires a thresholding operation, with the threshold being the same for all images that we experimented with, after low-dimensional embedding to separate the nuclei from a given histopathological image.

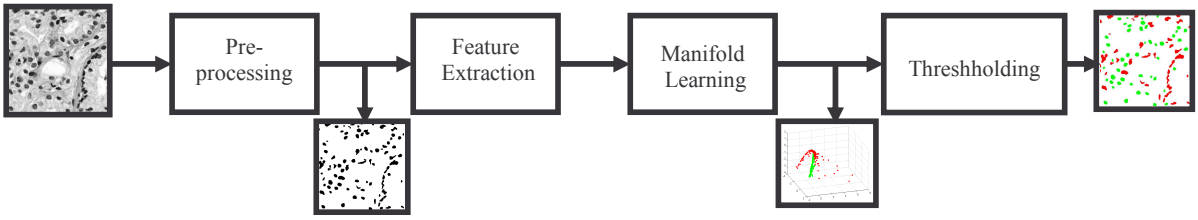


Figure 2. Block diagram of the proposed algorithm

2 Materials and Methods

We used Haematoxylin and Eosin (H&E) stained color images of prostate tissue samples for experimentation. All of the images were captured under similar conditions. The images were converted into gray scale and filtered using edged preserving smoothing filter to remove the noise. A typical specimen of the stained tissues is shown in the Figure 1 (a).

2.1.1 Preprocessing and Feature Extraction

The general block diagram of the proposed algorithm is shown in Figure 2. The gray images were processed by an edge-preserving smoothing filter, namely *anisotropic diffusion*, to remove the noise. In a typical H&E stained image of a tissue section, cell nuclei are darker than the surrounding cytoplasm and extracelular matter. Exploiting this fact, the class of potential nuclei are segmented from the specimen using *k*-means clustering on the intensity values. The resultant image with expected nuclei is then converted into a binary image. Morphological operations are performed on the dichotomized image to smooth it and to remove segmentation artifacts and objects with very small areas. The resultant binary image is shown in Figure 1 (b).

As can be seen in Figure 1(b), the processes of segmentation and binarization do not isolate the nuclei completely from other objects such as custers of overlapping nuclei, nuclear debris, cellular and extracellular structures and staining artifacts. In order to learn the shape manifolds of nuclei and other objects in an unsupervised manner, we extract the boundaries of all objects. The extracted boundary points of each object are then resampled into an equal number of points N using cubic spline interpolation. A centroidal distance function r_i, for $i=1,2,...,N$, is computed as follows [12],

$$r_i = \sqrt{([x_i - x_c]^2 + [y_i - y_c]^2)} \tag{1}$$

where (x_i,y_i) denote the coordinates of the ith boundary point of the object. The distance vector $\boldsymbol{r}=\{r_1,r_2,...r_N\}$ is transformed into frequency domain using FFT. Our feature vector $\boldsymbol{f}$ for the detection algorithm is derived as follows,

$$\boldsymbol{f} = \left[\frac{|F_1|}{|F_0|}, \frac{|F_2|}{|F_0|}, \ldots, \frac{|F_{N/2}|}{|F_0|}\right]^T \tag{2}$$

where $|F_i|$ denote the ith Fourier coefficient, with F_0 being the DC component. In the above equation, taking the magnitude of the coefficients yields rotation invariance and their division by $|F_0|$ results in scale invariance.

2.1.2 Unsupervised Manifold Learning

To improve the performance of a classifier or a clustering algorithm, it is desirable to extract relevant features concealed in high dimensional data. Generally, the relevant features lie on a low-dimensional sub-manifold embedding of the high dimensional space. Many algorithms have been proposed to learn a suitable low-dimensional embedding from the given data. The classical techniques such as PCA, ICA, MDS etc. assume that the low-dimensional manifold is linear. In recent years, locality preserving dimensionality reduction techniques such as Locally Linear Embeddings [17], Laplacian Eigen Maps [18], Diffusion Maps [16] and Hessian Eigen Maps [19] have been proposed when the features are lying in a non-linear low-dimensional manifold. These methods exploit the spectral properties (eigenvectors and eigenvalues) of some kind of similarity matrices.

In this work, we employ a diffusion map based framework for unsupervised manifold learning as it clearly defines a robust (with respect to noise) distance metric on the data set reflecting its connectivity [20]. As a consequence, the number of Eigenvectors in low-dimensional embedding can be determined by time parameter in Markov chain. In other methods, it is not clear how to choose the number of eigenvectors in the embedding [20].

For a given set of feature vectors $\Omega=\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n\}$ corresponding to boundaries of various objects in the tissue section, the first step of diffusion maps based framework is to consider feature vectors as nodes of a symmetric graph and compute pairwise similarity matrix between points $\boldsymbol{f}_i$ and $\boldsymbol{f}_j$ using Gaussian Kernel with width ε,

$$w(\boldsymbol{f}_i, \boldsymbol{f}_j) = \exp\left(-\frac{\left\|\boldsymbol{f}_i - \boldsymbol{f}_j\right\|^2}{2\varepsilon}\right) \tag{3}$$

The next step is to define Markov Random Walk on the graph by defining the Markov matrix $\boldsymbol{P}$, with its (i,j)th element p_{ij} given as follows

$$p_{ij} = \frac{w(\boldsymbol{f}_i, \boldsymbol{f}_j)}{d(\boldsymbol{f}_i)} \tag{4}$$

where $d(\boldsymbol{f}_i)$ denotes the degree of node $\boldsymbol{f}_i$ in the graph and is defined as

$$d(\boldsymbol{f}_i) = \sum_{z \in \Omega} w(\boldsymbol{f}_i, \boldsymbol{z}). \tag{5}$$

The (i,j)th element of $\boldsymbol{P}^t$ (i.e., the tth iterate of $\boldsymbol{P}$) represent probability of going from node $\boldsymbol{f}_i$ to node $\boldsymbol{f}_j$ in t steps. If $\{\lambda_l\}$ is the sequence of eigenvalues of $\boldsymbol{P}$ such that $|\lambda_0| \geq |\lambda_1| \geq \ldots$, and $\{\psi_l\}$ are the corresponding eigen functions, then a mapping from the data set Ω to a low-dimensional Euclidean space $\mathfrak{R}^m$,where m is the dimensionality of the lower-dimensional space and can be a function of t, is given by (see [16, 19] for details),

$$\Psi^t : \boldsymbol{f} \mapsto (\lambda_1^t \psi_1(\boldsymbol{f}), \lambda_2^t \psi_2(\boldsymbol{f}), \ldots . \lambda_m^t \psi_m(\boldsymbol{f}))^T \tag{6}$$

Spectral fall-off and the time t of the random walk are the main factors contributing to dimensionality reduction. Consequently, for large value of t, we will be able to capture large-scale structures in the data with fewer diffusion coordinates [21]. A natural consequence of this embedding is that it captures the intrinsic dimensionality of the high dimensional data which could be depicted by the way the objects are organized in the new, lower dimensional space.

The two-dimensional diffusion map for $(N/2\text{-}1)$-dimensional feature vectors (with $N = 100$ boundary points) for the tissue sample shown in Figure 1 (a) is shown in Figure 3 (a). In both figures, nuclei are shown in green and rest of the entities in red. These figures show that the low-dimensional embeddings have successfully captured in high dimensional data characterizing nuclei-like objects.

3 Results and Discussion

Diffusion coordinates were computed for high dimensional feature vector data corresponding to objects in the segmented tissue of Figure 1 (a) as outlined in the previous section. Figure 3 (a) shows the low-dimensional embedding of high dimensional original feature vectors extracted from the Figure 1 (b) in new space. To investigate for some interesting patterns emerging from the embedding, two non-overlapping sections of the Figure 3 (a) have been zoomed in (b) and (c) and the objects in the binarized image corresponding to Figure 1 (a) have been overlaid approximately in their respective coordinates in the embedding.

Visual examination of these plots demonstrates that as we move roughly from left to right and from bottom to top, the degree of complexity in the shape of objects increases. To put it another way, we may say that regular shaped nuclei have been clustered on the bottom right of the plot while non-regular shapes such as bunches of overlapping nuclei, possibly irregular nuclei, debris of nuclei, nuclei in different stages of mitosis, staining artifacts etc. have

distinct regions in sub-manifold. The same pattern was observed when the experiment was repeated for several specimens.

Interestingly almost all the regular shaped nuclei can easily be isolated from rest of the objects by thresholding the values of the first two eigen vectors at zero crossings. The threshold values of zero are chosen due to the nature of the problem (binary clustering) using the spectral clustering theory [22]. Figure 1 (c) shows the tissue section with potential nuclei with green color by this technique. Figure 4 (c) also shows the results of successful classification of objects in the binary image by thresholding for another specimen shown in Figure 4 (a). Note that in Figure 3 (a) the coordinates of potential nuclei in diffusion maps recovered by mere thresholding have been coded in green and other objects in red.

Intuitively, these observations seem to be very promising in the field of quantitative cytometry. However, it has yet to be established, by extensive experimentation on ground-truth, how significant these findings are for diagnosis and prognosis of cancer.

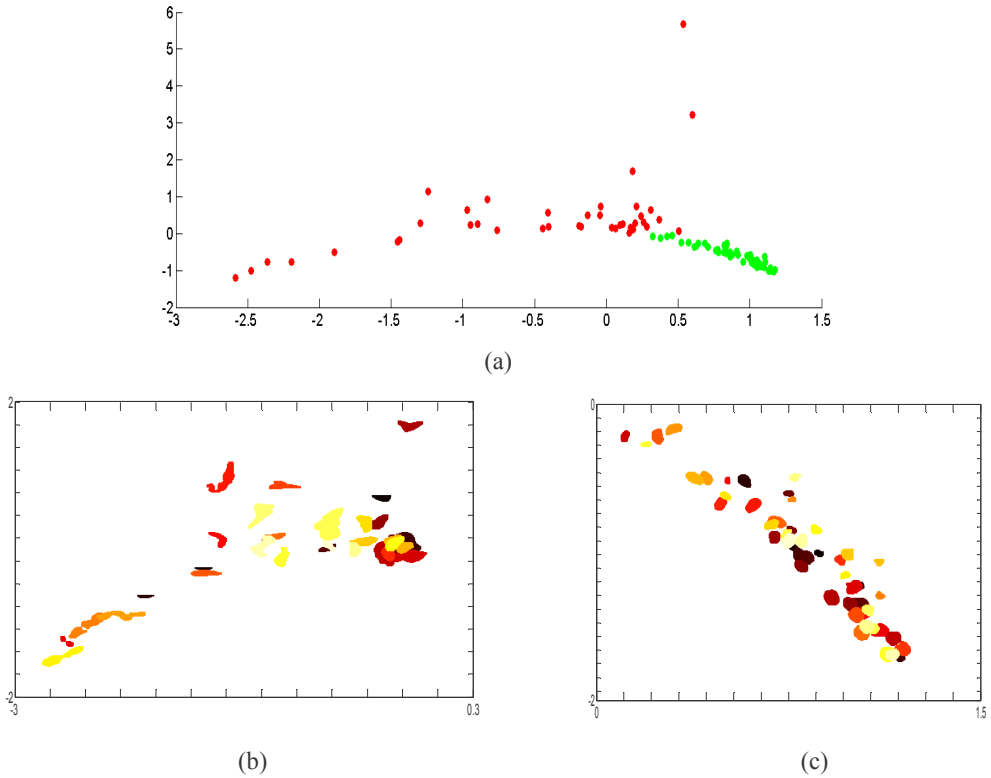


Figure 3. (a) Diffusion maps for the binary image of Figure 1(b). (b) & (c) The objects in the binary image corresponding to the their embedded coordinates overlaid approximately on their respective coordinates in (a).

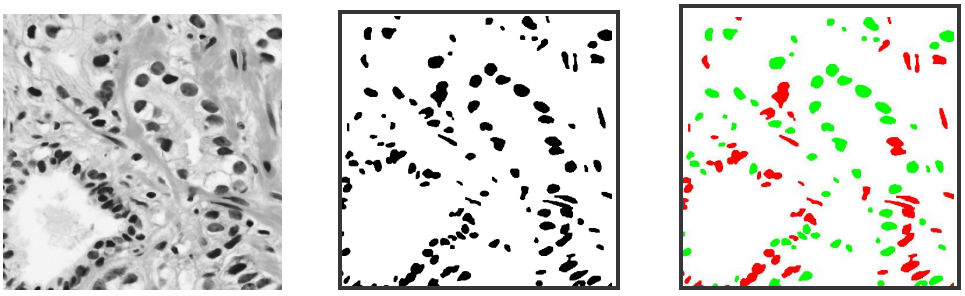


Figure 4: Another prostate tissue specimen (b) Binary image containing potential nuclei, overlapping nuclei, nuclear debris, and extracelluar structures (c) Potential Nuclei (shown in green) detected via shape manifold learning with diffusion maps.

Conclusions

The problem of detecting nuclei in a tissue section for image-based cytometry has been addressed in this paper using an unsupervised manifold learning paradigm. Preliminary results demonstrate that this technique successfully reorganizes various objects of the tissue section in accordance with the complexity of the objects' boundaries. The pattern that emerges in the diffusion space can be incorporated in the development of efficient and reliable systems for quantitative histopathology.

Acknowledgements

The first author would like to gratefully acknowledge the Higher Education Commission, Government of Pakistan, for financial support and the Department of Computer Science, University of Warwick, UK for hosting his visit and for providing facilities to carry out this research.

References

1. S.M. Ismail, A.B. Colclough, J.S. Dinnen et al., "Observer Variations in Histopathological Diagnosis and Grading of Cervical Intraepithelial Neoplasia," *Br. Med. J.*, 298(6675): 707-710, April 1989.
2. P.J Klinkhamer, G.P. Vooijs, & A.F. de Haan, "Intraobserver and Inerobserver Variability in the Diagnosis of Epithelial Abnormalities in Cervical Smears, *" Acta Cytology*, 32(6):794-800, Nov-Dec 1998.
3. P. Bartlets, "Computer-Generated Diagnosis and Image Analysis: An Overview" *Cancer Supplement*, Vol. 69, No. 6, March 15, 1992.
4. C. Demir & B. Yener, "Automated Cancer Diagnosis Based on Histopathological Images: A Systematic Survey," *Rensselaer Polytechnic Institute Technical Repor,t* TR-05-09, 2005.
5. C. Demir, S. Humayun Gultekin, & B. Yener, " Learning the Topological Properties of Brain Tumors," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* Vol. 2, No. 3, July-September 2005.
6. Qing Ji, John Engel, and eric Craine, " Texture Analysis for Classification of Cervix Lesions," *IEEE Transactions on Medical Imaging,* Vol. 19, No. 1, November 2000.
7. K. Masood, N. Rajpoot, K. Rajpoot et al., "Hyperspectral Texture Analysis for Colon Tissue Biopsy Classification," *International Symposium on Health Informatics and Bioinformatics*, 2007.
8. W. Christen-Barry & A. Partin, "Quantitative Grading of Tissue and Nuclei in Prostate Cancer for Prognosis Prediction," *John-Hopkins APL Technical Digest*, Vol. 18, Number 2, 1997.
9. H. Qureshi, N. Rajpoot, K. Masood et al., "Classification of Meningiomas using Discriminant Wavelet Packets and Learning Vector Quantization," *In Proceedings of Medical Image Understanding and Analysis*, 2006.
10. B. B. Chaudhuri, K. Rodenacker, & G. Burger, " Charactrization and Featuring of Histological Section Images," Pattern Recognition Letters, 7, pp. 245-262, 1988.
11. J. P. Thiran & B. Macq, " Morphological Feature Extraction for the Classification of Digital Images of Cancerous Tissue," *IEEE Transactions on Biomedical Engineering*, Vol. 43, No. 10, October 1996.
12. H. Kauppinen, T. seppanen, & M. Pietikainen, "An experimental Comparison of Autoregressive and Fourier-based Descriptors in 2-D Shape Classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 17, pp. 201-207, 1995.
13. K. Rodenacker, & E. Bengtsson, "A Feature Set for Cytometry on Digitized Microscopic Images," *Analytical Cellular Pathology*, pp. 1-36, 25(2003), ISO Press.
14. W. H. Wolberg, W. N. Street, D. M. Heisey, & O. L. Mangasarian, "Computer-derived Nuclear Features Distinguish Malignant from Benign Breast Cytology," *Hum. Pathology*, 26(7):792-796, July 1995.
15. K. Yogesan, T. Jorgensen, F. albregtsen et al., "Entropy-Based Texture Analysis of chromatin Structure in Advanced Prostate Cancer," *Cytometry,* 24:268-276, 1996.
16. R. R. Coifman & S. Lafon, "Diffusion Maps," *Applied and Computational Harmonic Analysis*, Special Issue on diffusion maps and wavelets, Vol. 21, pp. 5-30, July 2006.
17. S. Roweis & L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, pp. 2323-2326, 2000.
18. M. Bilkin & P.Niyogy, "Laplacian Eigen Maps for Dimensionality Reduction and Data Representation," *Neural Computation*, Vol. 6. No. 15, pp. 1373-1396, June 2003.
19. D. Donoho & C. Grimes, " Hessian Eigenmaps: New Locally Linear Embedding Techniques for High Dimensional Data," *Proc. Nat'l Academy of Sciences,* vol. 100, no. 10, pp. 5591-5596, May 2003.
20. S. Lafon & A. B. Lee, "Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parmeterization," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 28, No. 9, September 2006.
21. S. Lafon, Y. Keller & R. R. Coifman, "Data Fusion and Multicue Data Matching by Diffusion Maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 28, No. 11, November 2006.
22. J. Shi & J. Malik, "Normalized cuts and image segmentation," *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 22:888-905, 2000.

Excitation/emission resolved fluorescence imaging with FR-IsoMap

Kevin R. Koh[1,2], Karim Lekadir[1], Daniel S. Elson[2,3], and Guang-Zhong Yang[1,2]

[1]Visual Information Processing Group, Department of Computing, [2]Institute of Biomedical Engineering and
[3]Department of Biosurgery and Surgical Technology
Imperial College London, UK

Abstract: Fluorescence imaging has been demonstrated to be able to detect and diagnose different tumour types in a range of biological tissues including cervix, lung, and along the gastrointestinal tract. An approach based on laser induced tissue autofluorescence may be applied minimally invasively using endoscopes, although there is a high degree of overlap in endogenous fluorophore excitation/emission spectra. This paper presents a tuneable laser induced fluorescence imaging system that is able to increase specificity through simultaneous variation of excitation and emission wavelengths. The method is validated using fluorescent laser dyes with data analysis being performed with fixed reference isomap (FR-IsoMap). The method is compared to principle component analysis (PCA) for dimensionality reduction to demonstrate the improved specificity and sensitivity achieved based on one-way Multivariate Analysis of Variance (MANOVA).

1. Introduction

Fluorescence spectroscopy is widely used for the characterization of pre- and early malignant tissue transformations. Although the application of light illumination for the study of tumour fluorescence was first pioneered by Policard in the early 1920s, there has been a recent increase in the use fluorescence spectroscopy as a diagnostic technique in modern biology and medicine [1, 2]. Tissue autofluoresce when excited by light in the ultraviolet and visible regions of the electromagnetic spectrum due to the presence of naturally occurring biomolecules known as endogenous fluorophores. Different fluorophores are excited by different ranges of excitation wavelengths [3]. Fluorescence contrast arises when endogenous fluorophores vary in concentration and distribution between different tissue types as well as between different regions and layers of the same tissue [4]. Studies have shown that many endogenous fluorophores are either closely related to the structural matrix of tissues or involved in cellular metabolism [5]. Examples of the former comprise of structural proteins like collagen and elastin, whose fluorescent properties are the result of cross-links between amino acid chains that confer high tensile strength to the protein [6]. Fluorophores that have a role in cellular metabolism include co-enzymes like the pyridine nucleotide NADH and the flavoprotein FAD, both of which play an important role in the electron transport chain during ATP synthesis. Other endogenous fluorophores include the aromatic amino acids tryptophan, tyrosine and phenylalanine, certain groups of vitamins (e.g. vitamins A, K, D and some B6 compounds), lipopigments (e.g. lipofuscin, ceroids and phospholipids) and various porphyrins [7]. Endoscopic fluorescence imaging approaches have been employed by many groups for the detection and diagnosis of carcinomas of the colon [7, 8, 9] and oesophagus [10, 11], which are known to develop from premalignant lesions. As these are along regions of the gastrointestinal (GI) tract where non-invasive techniques can be employed for the early detection of such lesions, an endoscopic approach may be used.

The purpose of this paper is to increase the specificity of the fluorescence signal through excitation and emission resolution. A tuneable source based on a supercontinuum fibre laser system is incorporated into an endoscopic device for the characterization and screening of cancerous tissues. As an initial system validation, the analysis of fluorescent dyes of known excitation and emission wavelengths at different concentrations and in different mixtures by using FR-IsoMap [12] is presented. Results obtained were then compared to that of PCA [13, 14] to demonstrate the improved specificity and sensitivity achieved, based on one-way Multivariate Analysis of Variance (MANOVA).

2. Materials and Methods

The fluorescent dyes (and absorption maxima) consisted of Coumarin 334 (C334, 450 nm), Fluorescein (Fl, 512 nm), Rhodamine 6G (R6G, 530 nm), Rhodamine B (RB, 552 nm) and Cresyl Violet (CV, 601 nm) dissolved in 99% ethanol at various concentrations. These were then pipetted into 100-μl multi-well glass bottom PCR plates for excitation/emission spectroscopic imaging. Two sets of validation tests were carried out: the first consisting of single dyes of varying concentrations (0.3 μM, 1.0 μM, 3.0 μM and 10.0 μM) and the second consisting of mixtures of 2 dyes in varying proportions to a total of 3.0 μM concentration. Fluorescence was induced by a supercontinuum fibre laser, which had been previously demonstrated for fluorescence and fluorescence lifetime imaging applications [15, 16], and emitted broadband (400 – 1800 nm) pulsed light generated in highly nonlinear holey fibre. The output from the supercontinuum fibre laser source (Fianium SC450-2) was passed through a computer controlled spectral filter

consisting of a dispersive prism (SF2) and motor controlled slit. The spectrally filtered light was then coupled into 2 m of high NA (numerical aperture) polymer clad 200 mm core diameter optical fibre (Thorlabs BFL37-200) using a ×20 achromatic microscope objective. The fibre was chosen to efficiently deliver the light to the sample and illuminate a large field of view. It was positioned 120 mm from the horizontal sample, pointing upwards at the underside of the wellplate. Fluorescence emitted from the wellplate was imaged onto a CCD camera (QImaging Retiga Exi) by two achromatic lenses with focal lengths of 100 and 50 mm, respectively. For emission wavelength selection, an electronically controllable filter wheel was placed between these two lenses so that the 5 mm thick interference filters caused only small lateral translations of the image at the CCD chip. The 5 different excitation wavelengths used are shown in Fig. 1 and these were paired up with longpass interferometric filters with cut-on wavelengths of 500, 550 and 600 nm as shown in Table 1.

Excitation wavelength (nm)	Excitation bandwidth (nm)	Emission filter
470	10	Longpass 500 nm
490	20	Longpass 500 nm
490	20	Longpass 550 nm
510	20	Longpass 550 nm
530	25	Longpass 600 nm
560	35	Longpass 600 nm

Table 1. Combinations of excitation/emission wavelengths used

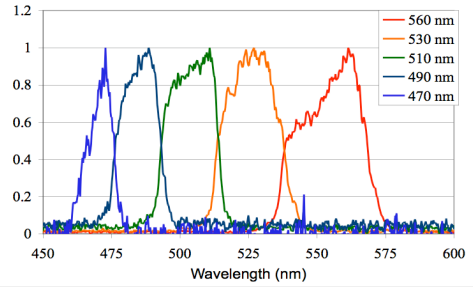


Figure 1. Normalised excitation spectra

The main advantage of the proposed fluorescence laser system is its ability to use several combinations of acquisition parameters (e.g. excitation and emission wavelengths), which increases its discrimination capability. The analysis of the acquired data, however, is not straightforward due to the high dimensionality of the problem. To overcome this difficulty, a solution is to reduce the number of dimensions to a few underlying axes such the variability induced by the data is preserved, with the aim to produce a new intuitive format for tissue characterisation. For example, by reducing the dimensionality of the problem to three factors, a colour map which exhibits the different constituents can be constructed by using the new 3D coordinates as RGB channels. For this purpose, one of the most popular methods for dimensionality reduction is PCA [13, 14]. By considering the cloud of points generated by the pixel data, the technique aims to find directions which exhibit the maximum of variations. The technique uses eigen-decomposition to extract a set of orthogonal axis from the covariance matrix. A linear model which incorporates the axis with the main variation can be generated to describe the data. This linear model, however, may not be optimal when the cloud of points in question lies on a more complex manifold with non-linear characteristics. In this case, non-linear dimensionality reduction is required and a set of techniques have been developed in the recent years. Amongst them, isometric mapping (Isomap) [17] is a technique capable of producing a low-dimensional mapping for the data such that the inter-point dissimilarity relationships in the manifold are preserved in a way similar to multidimensional scaling (MDS), but by using the geodesic distances to describe the shape of the manifold and thus the non-linear characteristics of the manifold. The method, however, when applied to different datasets of the same tissue type, can produce results which are not comparable. This is because Isomap produces an optimal embedding up to a geometrical transformation of the axes. Therefore, a fixed-reference Isomap (FR-IsoMap) is developed which is capable of generating consistent embedding between samples. The method is based on the construction of a fixed coordinate system using Isomap from a training sample selected such that the main variability and tissue characteristics are well represented. For new test data, instead of reapplying Isomap, a mapping is computed for each pixel such that the inter-point distances to the nearest neighbours in the training samples are well preserved. This allows the generation of consistent mappings across samples and therefore meaningful and comparable colour maps for tissue characterisation. More technical details about the method can be found in this paper [12].

3. Results

Figs. 2(a) and (b) show all 5 fluorescent dyes at concentrations of 0.3 μM, 1.0 μM, 3.0 μM and 10.0 μM. Figs. 2(c) and (d) show wells where dye solutions of 3.0 μM concentration were mixed in different proportions. These images represent the excitation spectroscopy images of the dye solutions obtained after processing with PCA (Figs. 2(a) and (c)) and FR-IsoMap (Figs. 2(b) and (d)) respectively.

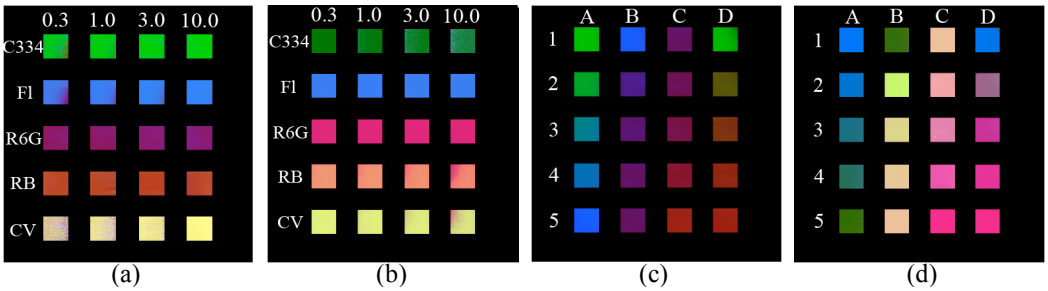


Figure 2. Processed images of single dye solutions at different concentrations (marked in μM) (Figs. 2(a) and 2(b)) and 2-dye mixtures (Figs. 2(c) and 2(d)) obtained after processing with PCA and FR-IsoMap.

For detailed quantitative analysis, MANOVA was applied to images obtained through PCA and FR-IsoMap. MANOVA evaluates the null hypothesis that the means of colour pixels of laser dyes in each well are the same n-dimensional multivariate vector, and that any difference observed within each well is due to random chance. For this test, a region of approximately 1900-2000 pixels was chosen manually from each well as the test sample. Information obtained after testing included d, the number of dimensions being measured, the p values for each dimension, a vector of Mahalanobis Distances (MDs) from each pixel to the mean of its region as well as a matrix consisting of the average MDs between pairs of pixel region means. The MD is a measure of similarity between variables by which different patterns can be identified and analysed.

In this study, MANOVA results for images (Figs. 2(a) and 2(b)) gave values of d=3 and p<0.005 signifying the accurate identification of the 3 RGB colour channels by both PCA and FR-IsoMap. Table 2 shows the average MDs obtained between the dye pair combinations of 0.3 μM concentration as well as the average MD from each pixel to the mean of its region within each dye. This low concentration was chosen for comparison so as to determine if PCA and FR-IsoMap could accurately separate the various dyes in this 'worst-case scenario' of least fluorescence.

	Coumarin 334	Fluorescein	Rhodamine 6G	Rhodamine B	Cresyl Violet
Coumarin 334	7.56/ **0.17**	155.58/ **534.78**	342.34/ **1534.00**	445.75/ **1476.04**	610.43/ **1317.62**
Fluorescein	155.58/ **534.78**	7.38/ **0.19**	104.15/ **1337.89**	228.42/ **1311.31**	382.92/ **1128.52**
Rhodamine 6G	342.34/ **1534.00**	104.15/ **1337.89**	0.23/ **0.24**	63.19/ **161.59**	308.05/ **498.41**
Rhodamine B	445.75/ **1476.04**	228.42/ **1311.31**	63.19/ **161.59**	0.48/ **0.28**	131.14/ **109.31**
Cresyl Violet	610.43/ **1317.62**	382.92/ **1128.52**	308.05/ **498.41**	131.14/ **109.31**	19.99/ **0.08**

Table 2. Average interpixel Mahalanobis distances within and between dyes of 0.3 μM concentration after analysis by PCA (normal text) and FR-IsoMap (bold text).

It can be observed from the results in Table 2 for the 0.3 μM dye solutions that a larger average interpixel distance between almost all dyes is obtained when FR-IsoMap analysis is applied, signifying that dye types were better separated when using this technique of analysis compared to PCA. Interpixel distances within the same dye for FR-IsoMap were also much smaller compared to PCA for almost all dyes. MANOVA was subsequently applied to PCA and FR-IsoMap analysis results of 1.0 μM, 3.0 μM and 10.0 μM concentrations from the first validation test set. Results showed that the FR-IsoMap performance decreased as dye concentration increased from 1.0 μM to 10.0 μM, with decreased average interpixel distances between almost all dyes and increased interpixel distances within the same dyes. At 10.0 μM dye concentrations FR-IsoMap performed significantly poorer than PCA, implying that FR-IsoMap performs best at low dye concentrations up to a threshold of about 3.0 μM i.e. at low levels of fluorescence.

MANOVA was next applied to PCA and FR-IsoMap images obtained from the second validation set which consisted of 2-dye mixtures (Figs. 2(c) and 2(d)). Values of d=3 and p<0.005 from MANOVA analysis were obtained across all images with MDs having much larger values than the first validation set. To determine which analysis method best distinguished the dye mixtures, MD values obtained from analysis of PCA and FR-IsoMap images of selected dye mixture transitions were plotted and compared as seen in Fig. 3. This gives a clearer indication of the spread in MD values for both analytical methods as the analysis moves from one dye-pair to the next. From Fig. 3, it is evident that FR-IsoMap (red line) gives a wider spread of MD values. In some cases, these values are also more than 1.5 times that obtained from PCA.

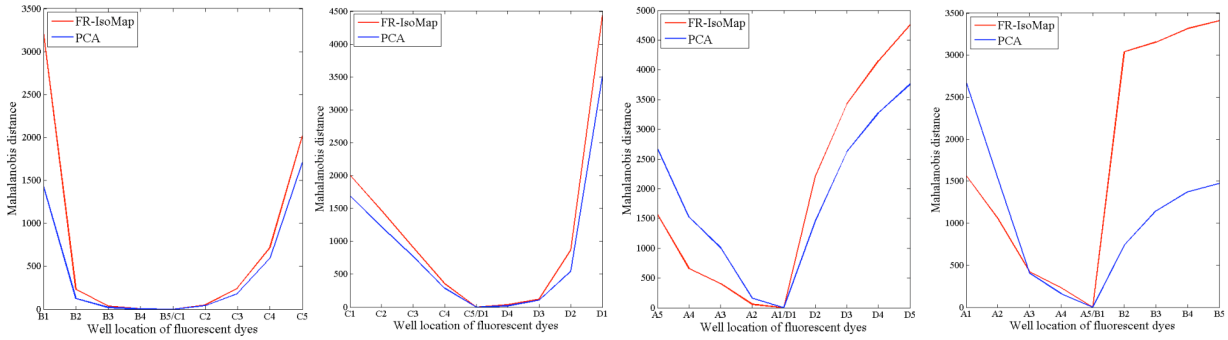


Figure 3. Graphical representations of the Mahalanobis distances from 4 different groups of dye mixture transitions, where red and blue curves represent FR-IsoMap and PCA respectively. Transitions are shown for (a) Fl to R6G to RB; (b) R6G to RB to C334; (c) Fl to C334 to RB; (d) C334 to Fl to R6G. For each transition group, the central dye is taken as the origin that is used to measure the Mahalanobis distances to the means of other dye mixtures.

To establish if acquisition parameters such as camera exposure timings had an effect on the outcome of PCA and FR-IsoMap analysis, individual well regions of PCA and FR-IsoMap images captured at 25 ms and 80 ms were manually selected for comparison of their mean RGB values and standard deviations. It has been shown that increasing the exposure timing from 25 ms to 80 ms does not appear to have an effect on the PCA results, with similar RGB values and standard deviations obtained under both conditions. On the other hand, RGB values obtained from FR-IsoMap analysis appear slightly different. For example, analysis of Fl by FR-IsoMap gives similar R and G values under both conditions. The difference here is the B value, which greatly decreases by almost six-fold with this exposure time increase. A similar observation can be made for the FR-IsoMap analysis of R6G: In this case the R and B values remain similar under both conditions whereas the G value increases by almost 40-fold when exposure time increases from 25 ms to 80 ms.

To investigate if changes in dye concentration affected PCA and FR-IsoMap analysis methods, dye images of 0.3 μM, 1.0 μM, 3.0 μM and 10.0 μM concentrations from Figs. 2(a) and (b) were subjected to MANOVA tests. Results in Fig. 4 indicate that both methods demonstrate a slight change in MDs as concentrations increase from 0.3 μM to 3.0 μM for most dyes. Larger increases in MDs occurred at 10.0 μM concentration levels for both analysis methods, with FR-IsoMap showing a larger change than PCA. This implies that both analysis methods are relatively independent of concentration changes within the range of 0.3 μM to 3.0 μM.

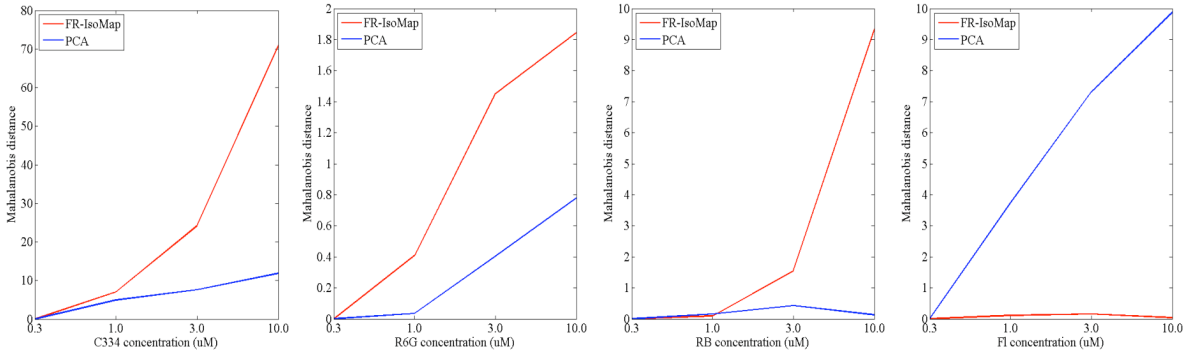


Figure 4. Change in Mahalanobis distances measured in PCA (blue curve) and FR-IsoMap (red curve) images as individual dye concentrations of C334, R6G, RB and Fl increase from 0.3 μM to 10.0 μM.

4. Conclusions

We have presented in this paper the results of excitation/emission resolved fluorescence imaging with FR-IsoMap. Data analysis results were compared to that of PCA based on MANOVA tests. The results demonstrate that both techniques provide accurate separation of dyes. FR-IsoMap analysis produced larger separation values between dyes and smaller interpixel distances within dyes when compared to PCA at lower dye concentration levels, indicating that dye types were better distinguished with FR-IsoMap under these conditions. This implies that FR-IsoMap would

be able to better discriminate the different endogenous fluorophores in biological tissue at low fluorescence intensities. In addition to being able to discriminate between dye mixtures, FR-IsoMap was also shown to be relatively independent of concentration at lower concentration levels, hence demonstrating the robustness of this technique under normal physiological conditions. Further investigation is required on the effect of system acquisition parameters on the results obtained. Future work will include further analytical work with FR-IsoMap for biological tissue where the real test will be discriminating between the myriad of endogenous fluorophores present in the tissue.

5. References

1. A. Policard. "Etudes sur les aspects offerts par des tumeurs expérimentales examinées à la lumière de Woods". *CR Soc Biol* **91,** pp 1423–1425, 1924.
2. G.A. Wagnieres, W.M. Star and B.C. Wilson. "In vivo fluorescence spectroscopy and imaging for oncological applications". *Photochemistry and Photobiology* **68(5)**, pp 603-632, 1998.
3. R.S. DaCosta, H. Andersson and B.C. Wilson. "Molecular fluorescence excitation-emission matrices relevant to tissue spectroscopy". *Photochemistry and Photobiology* **78(4)**, pp 384-392, 2003.
4. R.S. Dacosta, B.C. Wilson and N.E. Marcon. "Optical techniques for the endoscopic detection of dysplastic colonic lesions". *Current Opinion in Gastroenterology* **21(1)**, pp 70-79, 2005.
5. R. Richards-Kortum and E. Sevick-Muraca. "Quantitative optical spectroscopy for tissue diagnosis". *Annual Review of Physical Chemistry* **47**, pp 555-606, 1996.
6. N. Ramanujam. "Fluorescence Spectroscopy In Vivo". In *Encyclopedia of Analytical Chemistry*, pp 20-56, John Wiley & Sons Ltd, 2000.
7. T.D. Wang, J.M. Crawford, M.S. Feld, et al. "In vivo identification of colonic dysplasia using fluorescence endoscopic imaging". *Gastrointestinal Endoscopy* **49(4)**, pp 447-455, 1999.
8. K.T. Schomacker, J.K. Frisoli, C.C. Compton, et al. "Ultraviolet Laser-Induced Fluorescence of Colonic Tissue - Basic Biology and Diagnostic Potential". *Lasers in Surgery and Medicine* **12(1)**, pp 63-78, 1992.
9. R.M. Cothren, M.V. Sivak, J. Van Dam, et al. "Detection of dysplasia at colonoscopy using laser-induced fluorescence: A blinded study". *Gastrointestinal Endoscopy* **44(2)**, pp 168-176, 1996.
10. M. Panjehpour, B.F. Overholt, J.L. Schmidhammer, et al. "Spectroscopic Diagnosis of Esophageal Cancer - New Classification Model, Improved Measurement System". *Gastrointestinal Endoscopy* **41(6)**, pp 577-581, 1995.
11. M. Panjehpour, B.F. Overholt, T. VoDinh, et al. "Endoscopic fluorescence detection of high-grade dysplasia in Barrett's esophagus". *Gastroenterology* **111(1)**, pp 93-101, 1996.
12. K. Lekadir, D.S. Elson, J. Requejo-Isidro, et al. "Tissue characterization using dimensionality reduction and fluorescence imaging". *Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv.* **9(Pt 2)**: pp 586-93, 2006.
13. J.E. Jackson. *A User's Guide to Principal Components*. John Wiley & Sons, Inc., 2003.
14. I.T. Jolliffe. *Principal Component Analysis*. 2nd ed. Springer, 2002.
15. G. McConnell. "Confocal laser scanning fluorescence microscopy with a visible continuum source". *Optics Express* **12(13)**, pp 2844-2850, 2004
16. C. Dunsby, P.M.P. Lanigan, J. McGinty, et al. "An electronically tuneable ultrafast laser source applied to fluorescence imaging and fluorescence lifetime imaging microscopy". *Journal of Physics D - Applied Physics* **37(23**), pp 3296-3303, 2004.
17. J.B. Tenenbaum, V. de Silva and J.C. Langford. "A global geometric framework for nonlinear dimensionality reduction". *Science* **290(5500)**, pp 2319+, 2000.

Medical Image Computing Algorithms for Understanding Early Brain Development

Simon K. Warfield*

Harvard Medical School and Children's Hospital,
Dept. Radiology, Wolbach 215, 300 Longwood Avenue, Boston MA 02115

Abstract. In the last trimester before birth the developing brain undergoes tremendous changes as it grows. Premature birth during this period is associated with an increased risk of adverse outcomes, with up to 50% of very low birth weight infants going on to develop cognitive and motor deficits.
The analysis of magnetic resonance images has a crucial role to play in characterizing normal brain development, and in understanding the impact of early brain injury upon the path of later brain maturation. However, there are a number of unique challenges in quantitatively assessing early brain MRI due to the limited contrast between different types of tissue, the rapid progression of brain maturation, and the logistical challenges of imaging newborn infants. Particular challenges include overcoming the effects of image acquisition artifacts, imaging system noise, and patient-specific normal and pathological variability.
Advances in image acquisition and medical image computing algorithms now enable sophisticated characterization of early brain development. We describe recent work developing and evaluating image analysis algorithms that improve our capacity to characterize early brain maturation and early brain injury and enable us to assess potential therapeutic interventions.

1 Introduction

A full-term baby is born after approximately 40 weeks (38–42 weeks) of gestation. In the United States, 12% of neonates are born prematurely (< 38 weeks gestation) each year, and the incidence of children born with very low birth weight was 7.9% in 2003 [1, pp. 17]. Preterm birth is a leading cause of infant death and is associated with nearly one-half of all congenital neurological deficits, such as cerebral palsy. The most rapidly growing segment of the preterm population is the 47% of premature infants born between two and three months too early and weighing less than 1500g (3lbs 3oz). Over 50% of these children go on to have significant learning disabilities. In addition, around 10% will have cerebral palsy and a further 15-40% have more moderate motor control disability [2]. Prevention of these consequences of perinatal brain injury will require a better understanding of the pathogenesis and subsequent alteration of the brain development process and brain plasticity. The two major forms of perinatal brain injury in the premature neonate are 1.) forms of intraventricular hemorrhage and 2.) focal and diffuse injury to the white matter [2, 3, 4, 5].

The motor control and learning difficulties in these premature infants are associated with brain changes. The brain undergoes an incredible structural transformation during the 16 weeks or so preceding a full-term birth, and continues to rapidly develop for the next several years and on into adulthood. Brain dysfunction is thought to result from increased vulnerability of unmyelinated and myelinated white matter, cortical and deep gray matter, and specific cerebellar changes during the last trimester of pregnancy, the time of most rapid glial cell migration and brain connective tissue development.

Outcomes for premature infants have been associated with brain changes. We have previously developed automatic MRI segmentation algorithms which carry out tissue classification [6, 7]. Utilizing tissue classification, we have been able to identify risk factors for later adverse developmental outcomes [8]. Similarly, a recent study using MRI acquisition and tissue classification of originally premature and control subjects at eight years of age found that preterm birth is associated with regionally specific, long-term reductions in brain volume and that morphological abnormalities are, in turn, associated with poorer cognitive outcome [9]. More work needs to be done to provide effective tools that can improve our understanding of what brain changes occur in premature infants, how these changes arise, and what interventions and treatments can ameliorate the effects of these changes.

Tissue classification is a critical procedure in quantitative assessment of brain maturation [8, 10]. Tissue classification may be an effective MRI biomarker of later neurodevelopmental outcome, and a means of assessing potential interventions. We can expect that more accurate predictions of later outcome will be derived from more accurate and robust tissue classification procedures. In particular, the sensitivity, specificity and predictive value of a tissue classification algorithm are key indicators of quality.

*Email: simon.warfield@childrens.harvard.edu

Unfortunately, in clinical imaging studies such as MRI of newborns, there is no widely accepted reference standard against which different algorithms may be compared. We have developed and utilized an algorithm called STAPLE, for simultaneous truth and performance level estimation, which allows us to estimate a reference standard from a collection of segmentations of a particular image. We simultaneously compute performance parameters which in the case of binary segmentations are estimates of sensitivity and specificity, and the natural generalizations of these in the case of multiple unordered labels (for example gray matter, white matter and CSF). From these we can also estimate predictive values and these enable us to assess different methodologies for tissue classification.

One common approach used with clinical imaging data is manual tracing of key structures by neuroanatomical experts. Manual tracing is often time consuming, and therefore expensive, while also being tedious for the expert, and exhibits intra-expert and inter-expert variability. Automated and semi-automated algorithms can capture expertise through expert initialization or parameter setting, and then operate fully automatically to achieve deterministic segmentations with perfect reproducibility. Different algorithms may achieve the same perfect reproducibility while operating at a different level of accuracy — the algorithms may have different levels of bias or systematic error. We seek to identify algorithms that generate results with an accuracy indistinguishable from that of human experts while maintaining perfect reproducibility (or very high reproducibility compared to human experts in the case of stochastic algorithms).

2 Materials and Methods

2.1 Data Acquisition

Five data sets were selected at random from over 200 MRI scans acquired in previous studies. All scans were acquired on 1.5T clinical MRI scanners, and typically included a T1-weighted SPGR (spoiled gradient recalled acquisition in the steady state) and a T2-weighted FSE (fast spin echo) sequence. The scans were acquired in the coronal plane, consisting of 256x256 voxels in-plane with slices covering the entire brain, and having a resolution of approximately 0.7mm $\times$ 0.7mm with a slice thickness of 1.5mm for the SPGR and 3.0mm for the FSE.

2.2 Interactive MRI Segmentation

Each T2w scan was aligned with the SPGR by rigid registration, computing a transformation by maximization of mutual information. A mask of the intracranial cavity was created using signal intensity and connectivity information as described in [11]. A single coronal slice near the center of the brain was selected, and utilizing the T1w and T2w scans, an expert traced the boundaries of the following structures: cortical gray matter, sub-cortical gray matter, unmyelinated white matter, myelinated white matter and cerebrospinal fluid. This was repeated approximately once a week for five weeks, in order to obtain five repeated segmentations of each of the five subjects.

2.3 Computation of Reference Standard

We computed a probabilistic reference standard from each set of five segmentations, of each subject using STAPLE [12]. This models the reference standard as an unknown random variable, and considers each segmentation as an attempt to generate the true segmentation which is different from the true segmentation in a manner that depends on the inherent performance of the segmentation generator responsible for each segmentation. The algorithm is formulated as an instance of the Expectation-Maximization algorithm [13].

Detailed descriptions of the EM algorithm and generalizations are available [13, 14]. The essence of the EM algorithm is the observation that certain maximum likelihood estimation problems would be considerably simplified if some missing data were available, and this is the case for our problem. The observable data, the segmentation decisions at each voxel, is regarded as being incomplete and is regarded as an observable function of the complete data.

The *complete* data consists of the segmentation decisions at each voxel, which are known, and the true segmentation, which is not known. If we did know the true segmentation, it would be straightforward to estimate the performance parameters by maximum likelihood estimation. Since the complete data is not available, the complete data log likelihood cannot be constructed and instead must be estimated. Doing so requires evaluating the conditional probability density of the *hidden* true segmentation given the segmentation decisions and a previous estimate of the performance level of each segmentation generator. The expectation of the complete data log likelihood with respect to this density is then calculated, and from this estimate of the complete data log likehood, the performance parameters are found by maximum likelihood estimation. We iterate this sequence of estimation of the conditional probability of the true segmentation and performance parameters until convergence is reached. Convergence to a local maximum is guaranteed.

Consider a brain MRI tissue segmentation application where we are interested to identify the true label as one of white matter, gray matter or cerebrospinal fluid. We allow $\mathbf{T}$ to take on one of $\mathbf{L}$ labels where $\mathbf{T_i} \in \{0, 1, \ldots, L-1\}$ and similarly the segmentation decisions may also be one of these L labels, $D_{ij} \in \{0, 1, \ldots, L-1\}$, and $\mathbf{D}$ is an $N \times R$ matrix describing the decisions made for each segmentation at each voxel of the image. In this case, we characterize each rater performance by θ_j, an $L \times L$ matrix of parameters where the element of the matrix characterizes the probability rater j will decide label s' is present when the true label is s. The perfect rater would have a performance matrix with 1 on the diagonal and 0 in each off-diagonal element.

In order to evaluate the conditional expectation of the complete data log likelihood for this case, we need to compute the conditional probability that the true label at voxel i is s given the set of segmentations and the previous estimate of the performance characteristics. We find [12] that

$$\begin{aligned} W_{si}^{(k)} &\equiv f(T_i = s | \mathbf{D_i}, \boldsymbol{\theta}^{(k)}) \\ &= \frac{f(T_i = s) \prod_j f(D_{ij} | T_i = s, \theta_j^{(k)})}{\sum_{s'} f(T_i = s') \prod_j f(D_{ij} | T_i = s', \theta_j^{(k)})} \end{aligned}$$

With this conditional probability, we can evaluate the expected value of the complete data log likelihood function, with the goal of identifying the performance parameters by maximum likelihood estimation. Considering each rater separately, we find the new parameter estimates $\theta_j^{(k+1)}$ by:

$$\begin{aligned} \theta_j^{(k+1)} &= \arg\max_{\theta_j} \sum_i E\left[\ln f(D_{ij} | T_i, \theta_j) | \mathbf{D}, \theta_j^{(k)}\right] \\ &= \arg\max_{\theta_j} \sum_i \sum_s W_{si}^{(k)} \ln f(D_{ij} | T_i = s, \theta_j) \\ &= \arg\max_{\theta_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_s W_{si}^{(k)} \ln f(D_{ij} = s' | T_i = s, \theta_j) \\ &= \arg\max_{\theta_j} \sum_{s'} \sum_{i:D_{ij}=s'} \sum_s W_{si}^{(k)} \ln \theta_{js's} \qquad (1) \end{aligned}$$

which after some substitution and manipulation leads to

$$= \frac{\sum_{i:D_{ij}=s'} W_{si}^{(k)}}{\sum_i W_{si}^{(k)}} \qquad (2)$$

3 Results

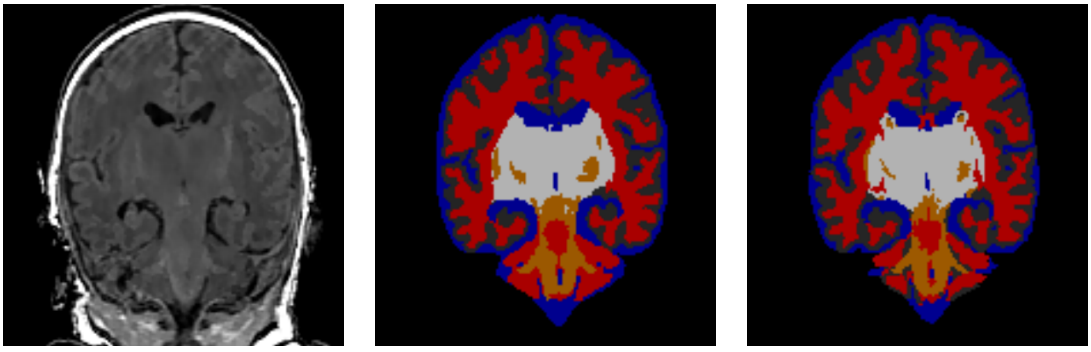
Figure 1 illustrates one slice from one T1w newborn MRI, and shows the STAPLE estimate of the optimal tissue classification together with four of the five interactive segmentations carried out on this subject.

Table 1 shows the values of the posterior probabilities for each repeated segmentation of each tissue class. This table illustrates how we can examine the variability intrinsic to manual segmentation with respect to the reference standard computed by STAPLE.

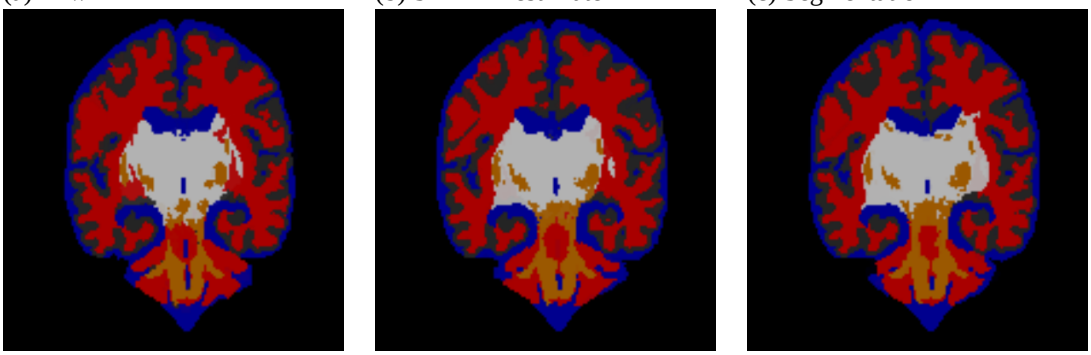
Table 2 reports the mean posterior probability for each tissue class over five segmentations of all five subjects. The table illustrates the variability of the mean posterior probability across different subjects and across different types of tissue, indicating that some subjects and some tissue classes are more challenging to segment than others.

4 Discussion

We have estimated a reference standard segmentation for five subjects, from five repeated segmentations carried out by an expert. The reference standard was estimated using STAPLE in each case. The visual appearance of the estimated reference standard appears more regular than the individual segmentations, with fewer isolated voxels having a tissue



(a) T1w MRI (b) STAPLE estimate (c) Segmentation 1



(d) Segmentation 2 (e) Segmentation 3 (f) Segmentation 4

Figure 1. Illustration of (a) newborn T1w MRI, (b) STAPLE estimate of the true segmentation and (c-f) manual segmentations.

Table 1. Evaluation of five manual segmentations of a newborn MRI with STAPLE. The performance of the segmentation of each tissue class is summarized with the posterior probability, which is the probability that the true label is $T = t$ when the manual segmentation was $D = t$. The tissue classes considered are background, cortical gray matter, cerebrospinal fluid, myelinated white matter, unmyelinated white matter and subcortical gray matter in turn.

Segmentation	BG	CGM	CSF	MWM	UNWM	SCG
1	0.997	0.763	0.945	0.889	0.914	0.904
2	0.998	0.937	0.926	0.961	0.899	0.920
3	0.999	0.939	0.903	0.911	0.922	0.985
4	0.999	0.958	0.923	0.900	0.973	0.945
5	0.998	0.871	0.911	0.934	0.913	0.954

label different from their neighboring voxels, and labels that more closely follow the signal intensity boundaries in the MRI than the individual segmentations.

STAPLE estimates performance parameters for each segmentation in terms of the probability of a correct segmentation decision $D = d$ when the true segmentation is $T = t$. In the case of binary labels, we can recognize Pr($D = 1|T = 1$) and Pr($D = 0|T = 0$) as sensitivity and specificity. These measures, and the generalization to multi-category labels Pr($D = t|T = t$) are valuable indicators of performance. These probabilities vary with the amount, or relative proportion or prior probability, of each label $T = t$ in a given image. Furthermore, it is sometimes of interest to know the posterior probability Pr($T = t|D = d$), the probability that the true label matches the segmentation, and these can be computed from the performance parameters of each segmentation generator and an empirical estimate of the prior probability of each tissue class. Since the relative proportion of each tissue class is accounted for, such posterior probabilities can readily be combined and compared.

An excellent indicator of overall performance is the mean posterior probability. Table 2 above illustrates this measure, and indicates that cortical gray matter and myelinated white matter (both structures with complex boundaries) are more difficult to segment than other structures, such as subcortical gray matter or unmyelinated white matter. We can use this data to estimate the reliability of predictions of later outcomes from newborn MRI, and to judge the quality of

Table 2. Evaluation of 25 manual segmentations of five newborns with STAPLE. The performance of the segmentation of each subject and tissue class is summarized with the mean posterior probability over the repeated segmentations. The tissue classes considered are background, cortical gray matter, cerebrospinal fluid, myelinated white matter, unmyelinated white matter and subcortical gray matter in turn.

Subject	BG	CGM	CSF	MWM	UNWM	SCG
1	0.998	0.894	0.922	0.919	0.924	0.941
2	0.999	0.881	0.935	0.931	0.921	0.949
3	0.998	0.918	0.947	0.982	0.988	0.959
4	0.999	0.912	0.927	0.971	0.946	0.951
5	0.998	0.916	0.939	0.897	0.930	0.960

segmentation algorithms designed for this task.

Acknowledgements

This investigation was supported in part by a research grant from CIMIT, grant RG 3478A2/2 from the NMSS, by NSF ITR 0426558, and by NIH grants R03 CA126466, R01 HL074942, P30 HD018655 and R01 RR021885. We acknowledge with thanks the substantial aid in preparing this paper provided by Andrea U. J. Mewes and by Neil I. Weisenfeld.

References

1. J. Martin, B. Hamilton, P. Sutton et al. "Births: Final data for 2003." *National Vital Statistics Reports* **54(2)**, 2005.
2. J. J. Volpe. "Brain injury in the premature infant - from pathogenesis to prevention." *Brain and Development* **19**, pp. 519–534, 1997.
3. A. Hill. "Current concepts of hypoxic-ischemic cerebral injury in the term newborn." *Pediatr Neurol* **7(5)**, pp. 317–25, 1991.
4. Y. Nakamura, O. T. & H. T. "Vascular architecture in white matter of neonates: its relationship to periventricular leukomalacia." *J Neuropathol Exp Neurol* **53(6)**, pp. 582–9, 1994.
5. J. Perlman, R. Risser & R. Broyles. "Bilateral cystic periventricular leukomalacia in the premature infant: Associated risk factors." *Pediatrics* **97**, pp. 822–827, 1996.
6. S. K. Warfield, M. Kaus, F. A. Jolesz et al. "Adaptive, Template Moderated, Spatially Varying Statistical Classification." *Med Image Anal* **4(1)**, pp. 43–55, Mar 2000.
7. N. I. Weisenfeld, A. J. U. Mewes & S. K. Warfield. "Highly Accurate Segmentation of Brain Tissue and Subcortical Gray Matter from Newborn MRI." In R. Larsen, M. Nielsen & J. Sporring (editors), *MICCAI (1)*, volume 4190 of *Lecture Notes in Computer Science*, pp. 199–206. Springer, 2006.
8. D. Thompson, S. Warfield, J. Carlin et al. "Perinatal risk factors altering regional brain structure in the preterm infant." *Brain* **130("Pt 3")**, pp. 667–77, 2007.
9. B. S. Peterson, B. Vohr, L. H. Staib et al. "Regional Brain Volume Abnormalities and Long-term Cognitive Outcome in Preterm Infants." *JAMA* **284(15)**, pp. 1939–1947, October 2000.
10. P. Hüppi, S. Warfield, R. Kikinis et al. "Quantitative Magnetic Resonance Imaging of Brain Development in Premature and Mature Newborns." *Ann. Neurol.* **43(2)**, pp. 224–235, Feb 1998.
11. V. Grau, A. U. J. Mewes, M. Alcaniz et al. "Improved Watershed Transform for Medical Image Segmentation Using Prior Information." *IEEE Trans Med Imag* **23(4)**, pp. 447–458, 2004.
12. S. K. Warfield, K. H. Zou & W. M. Wells. "Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation." *IEEE Trans Med Imag* **23**, pp. 903–921, 2004.
13. A. Dempster, N. Laird & D. Rubin. "Maximum-likelihood from incomplete data via the EM algorithm." *J. Royal Statist. Soc. Ser. B.* **39**, pp. 34–37, 1977.
14. G. J. McLachlan & T. Krishnan. *The EM Algorithm and Extensions*. Wiley-Interscience, New York, New York, 1996.

A Model to Investigate the Feasibility of FDG as a Surrogate Marker of Hypoxia.

Catherine J Kelly[a*], Kieran Smallbone[b], Tiina Roose[c] and Michael Brady[a]

[a]Wolfson Medical Vision Laboratory, University of Oxford, [b]Manchester Centre for Integrative Systems Biology, University of Manchester, [c]Centre for Mathematical Biology and Oxford Centre for Industrial and Applied Mathematics, University of Oxford

Abstract. Fmiso-PET is a non-invasive modality used for the assessment of tumour hypoxia, and increasingly for planning radiotherapy. However, the availability and contrast properties of Fmiso are not ideal. Recent efforts to compare FDG binding with that of Fmiso, in order to ascertain FDG's potential as a marker of hypoxia, have met with mixed results. The potential reasons for correlated and disparate binding patterns between the two tracers have been postulated, but not formally outlined as yet. We present a model of a key component of the image formation process - tracer pharmacokinetics. This involves a series of coupled PDEs, describing the interplay between concentrations of oxygen, glucose, HIF, Fmiso and FDG. We use this model to assess the general feasibility of FDG as a surrogate marker of hypoxia and find that its utility is dependent on activity of oncogenic pathways.

1 Introduction

One of the current challenges in molecular imaging is the non-invasive imaging of tumour hypoxia. Determining the oxygenation status of a tumour is crucial as hypoxia both affects the efficacy of radiotherapy and is also indicative of a malignant phenotype.

Positron Emission Tomography (PET) offers a non-invasive means of assessing tumour hypoxia. Its utility is dependent on the specificity and sensitivity of the tracer used. The standard hypoxia tracer [18F]-fluoromisonidazole (Fmiso), is specifically reduced in hypoxic tissues, becoming covalently bound to intracellular molecules. However, this specificity is compromised by poor distribution kinetics. The lipophillic nature of the compound means that it distributes into all tissues, with typical tumour-to-blood ratios of approximately 1.2.

Far better distribution characteristics are exhibited by the metabolic tracer, [18F]-fluorodeoxyglucose (FDG), a hydrophillic molecule, whose entry to cells is transporter-mediated. Conveniently, FDG has a direct link to tumour hypoxia, as the transcription factor hypoxia-inducible factor (HIF) controls regulation of genes encoding glycolytic enzymes [1]. Recently, there have been several attempts in the clinic to assess the feasibility of FDG as a surrogate marker for hypoxia [2] [3]. These have demonstrated mixed results, with some correlation between Fmiso and FDG reported within cancer types, but less correlation within individual cancers.

Elevated levels of HIF are found in the nucleus within minutes of cellular exposure to hypoxia. Glycolytic enzymes are upregulated and glycolysis increases approximately 2 and a half fold.

The regulation of glycolysis in tumours is complex and extends beyond HIF. The Warburg effect, where cells utilise glycolysis even under aerobic conditions, can be initiated both via and in the absence of HIF by oncogenic pathways such as the Akt pathway [4]. Akt upregulates glycolysis by both stabilising HIF and, independently, activating hexokinase-2 and glucose transporters.

In this paper, we describe a key component of the imaging process - tracer pharmacokinetics. We present a formal model of both HIF- and oncogenic pathway-mediated glycolysis and use this as a basis to describe FDG and Fmiso kinetics in response to hypoxia.

We use our model to (i) investigate the relationship between the microenvironmental oxygenation and tracer binding, and (ii) to compare the different effects of oncogenic pathway activation on FDG binding. Finally, based on the results of these studies, we discuss the feasibility of FDG as a surrogate marker of hypoxia.

*Email: ckelly@robots.ox.ac.uk

2 MODEL

A simplified model of HIF-mediated metabolic response to hypoxia is presented in Figure 1. The two main processes mediated by HIF are (i) the upregulation of glycolysis and (ii) the downregulation of oxidative phosphorylation, the former clearly leading to increased FDG avidity. The activation of oncogenic pathways (e.g. Akt) has two potential effects. The first is stabilisation of HIF levels. The second is the HIF-independent upregulation of glycoysis. The latter suggests a potential divergence between hypoxia- and oncogene-mediated FDG binding.

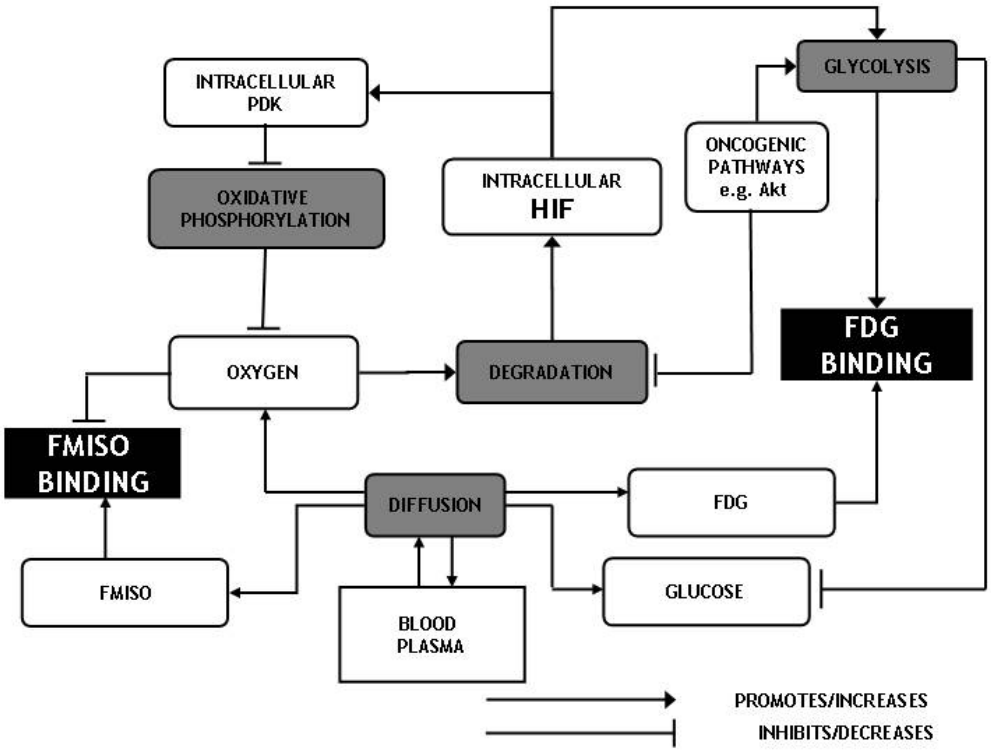


Figure 1. Proposed model of HIF- and oncogenic pathway-mediated glycolysis and the relationship with FDG and Fmiso binding. Molecular components (e.g. tracers) are shown in white boxes. Regulatory and diffusive processes are shown in black boxes. Oncogenic pathways can affect glycolysis (and thus FDG binding) directly through upregulation of enzymes, or indirectly through HIF stabilisation.

$$\frac{\partial H}{\partial t} = r_H\,(1-H) - d_H^{\star} H\left(\frac{O}{k_H+O}\right) \quad (1)$$

$$\frac{\partial O}{\partial t} = -\,d_O\left(\frac{O}{k_O+O}\right)(1-\gamma H(t-\tau)) + D_O\nabla^2 O \quad (2)$$

$$\frac{\partial G}{\partial t} = -\,d_G^{\star}\left(\frac{G}{k_G+G}\right)(1+k_A H) + D_G\nabla^2 G \quad (3)$$

$$\frac{\partial F}{\partial t} = -\,d_G^{\star}\left(\frac{F}{k_G+G}\right)(1+k_A H) + D_G\nabla^2 F \quad (4)$$

$$\frac{\partial M}{\partial t} = -\,d_M M\left(\frac{k_{mM}}{k_{mM}+O}\right) + D_M\nabla^2 M \quad (5)$$

The model dynamics are described using the system of non-dimensionalised PDEs outlined in Equations 1-5. The main variables are HIF (H), Oxygen (O), Glucose (G), FDG (F) and FMISO (M). H is produced and retained intracellularly in response to local oxygen concentration. Nutrients (O and G) and tracers (F and M) are supplied by the blood and then diffuse through a network of 'cells', being metabolised or bound, depending on local environmental

conditions and the oncogenic state of each cell.

Activation of the oncogenic pathway is represented as a random occurrence. Once activated, two processes are affected, (1) Degradation of HIF is inhibited (stabilisation), (2) Glycolysis is upregulated (Warburg). These are represented as a decrease in the maximal rate of HIF degradation ($d^{\star}_H$) and an increase in the basal rate of glucose consumption ($d^{\star}_G$). Under normal conditions these are d^n_H and d^n_G (*n*ormal), and under oncogenic activation these are d^a_H and d^a_G (*a*kt). As little comparative quantitative data yet exists for HIF-dependent and independent Akt-mediated glycolysis, values for d^a_H and d^a_G are estimated as multiples of their normal counterparts, i.e. $d^a_H = \frac{1}{\alpha} d^n_H$ and $d^a_G = \beta d^n_G$. To investigate the relative effects of each pathway on tracer uptake, we compare profiles of FDG and Fmiso for $\alpha = 1$, $\beta = 1$ (no oncogenic regulation), $\alpha = 2$, $\beta = 1$ (HIF stabilisation) and $\alpha = 1$, $\beta = 2$ (HIF-independent regulation of glycolysis).

The concentrations presented here are normalized by the maximal plasma or intracellular level of each component, to give relative quantities e.g. F is the concentration of FDG *relative to its maximum plasma level*. By looking at relative quantities, trends and relationships are often clearer. Equations were scaled to a time scale of 1 minute to best represent changes during a conventional PET scan, and a spatial scale of 1 cell width ($20\mu m$). Full parameterisation is available in the appendix (Section 6). As tracer pharmacokinetics are conventionally analysed using compartment modelling techniques, we parameterise a 2-compartment model using FDG curves for each condition (Akt-active/inactive, HIF-dependent/independent, Hypoxia/Normoxic) and, assuming irreversible kinetics, calculate the metabolic rate of glucose ($\text{MR}_{gluc} = \frac{K_1 k_3}{k_2 + k_3}$).

Our model equations were integrated in Matlab using the finite difference method. We model our physical system as the simple one-dimensional case in which a single vessel supplies tissue. Boundary conditions are Dirichlet at the vessel and Neumann at the distal end of the grid.

3 RESULTS

3.1 Molecular Profiles

We initially present profiles of the non-dimensionalised concentrations described above, as a function of increasing distance from a blood vessel, *in the absence of oncogenic pathway activation*. Profiles for total concentration are shown in Figure 2 (tracers will exist in free and bound states, but imaging will only show the combined concentration). Oxygen and glucose are both depleted as distance increases. Oxygen is almost fully metabolised after only a few cells widths, in agreement with previous calculations [5]. HIF increases with distance, although levels plateau. Both Fmiso and FDG show an increase in total concentration.

Inspection of the temporal profiles of total Fmiso and FDG for normoxic (cell adjacent to blood vessel) and hypoxic (8 cells away from blood vessel) tissue is shown in Figure 3. Contrast between activity in hypoxic and normoxic tissues develops at later time points. Due to the slower diffusive properties of Fmiso, contrast is obtained later than using FDG. Maximum contrast (at 120 minutes) during the scan period is higher for FDG than for Fmiso (1.65 and 1.19 respectively).

3.2 Effect of Oncogenic Pathways

Profiles in the previous section are generated under the assumption that increased glycolysis is entirely hypoxia mediated. In this section we present profiles of FDG *in the presence of oncogenic pathway activation.*

Figure 4 shows the normoxia/hypoxia temporal profile of FDG where $\alpha = 2$ and $\beta = 1$, i.e. when the oncogenic pathway acts through HIF stabilisation alone. Figure 5 shows the normoxia/hypoxia temporal profile of FDG where $\alpha = 1$ and $\beta = 2$, i.e. when the oncogenic pathway acts through HIF-*independent* upregulation of glycolysis. These figures show that hypoxia:normoxia contrast is reduced in both instances (from 1.65 to 1.52 for HIF-dependent Akt activity, and 1.32 for HIF-independent Akt activity), but that the action of HIF stabilisation, as it is represented in our model, is relatively small compared to the dominant effect of HIF-independent upregulation of glycolysis.

Table 1 shows the MR_{gluc} for each condition, calculated using a standard 2-compartment model. In all cases, MR_{gluc} increases in hypoxia, but the hypoxia:normoxic ratio decreases under Akt-active conditions, particularly when upregulation of glycolysis is HIF-independent. Also evident from Figure 5 and Table 1 is that FDG levels in Akt-regulated normoxic tissue could potentially match or even supersede those in HIF-regulated hypoxic tissue, depending on the value of β.

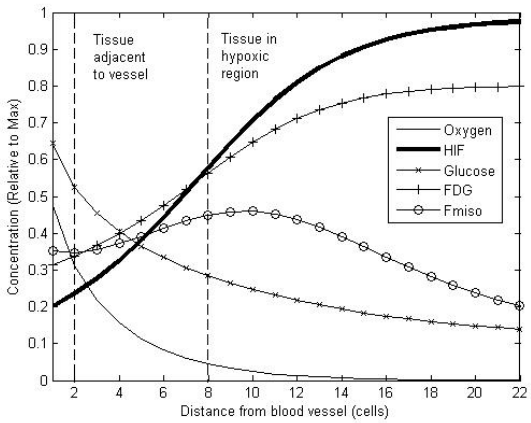


Figure 2. Concentration profiles with increasing distance from blood vessel at 120 minutes post-injection. Dashed lines indicate normoxic (adjacent to vessel) and hypoxic (8 cells from vessel) regions.

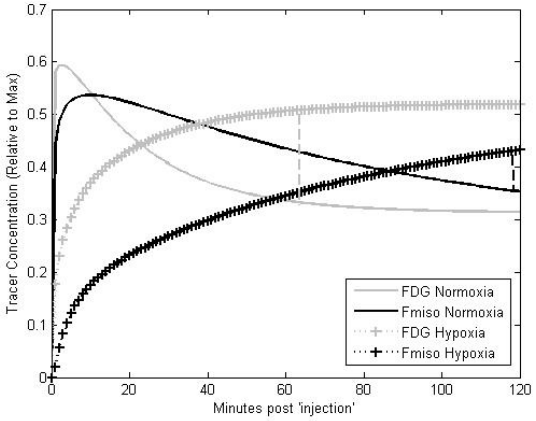


Figure 3. Temporal Profiles of FDG and Fmiso for normoxic and hypoxic tissue. Dashed lines indicate hypoxia:normoxia contrast at conventional scan times (60 and 120 minutes respectively). Contrast FDG: 1.6, Fmiso: 1.19. (Max contrast FDG=1.65 at 120 mins)

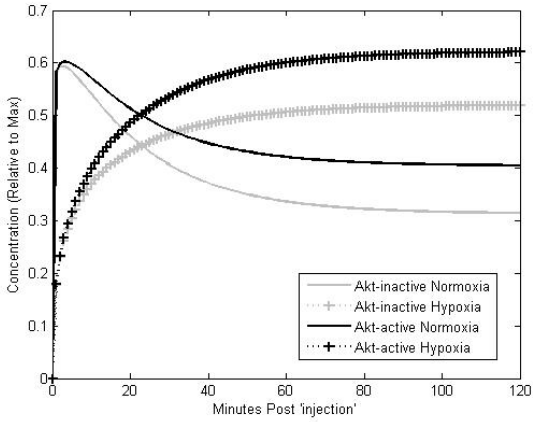


Figure 4. Temporal Profiles of FDG for normoxic and hypoxic tissue under Akt-mediated HIF stabilisation ($\alpha = 2$, $\beta = 1$), compared to Akt-inactive tissue ($\alpha = 1$, $\beta = 1$). Max contrast Akt-inactive: 1.65, Akt-active: 1.52.

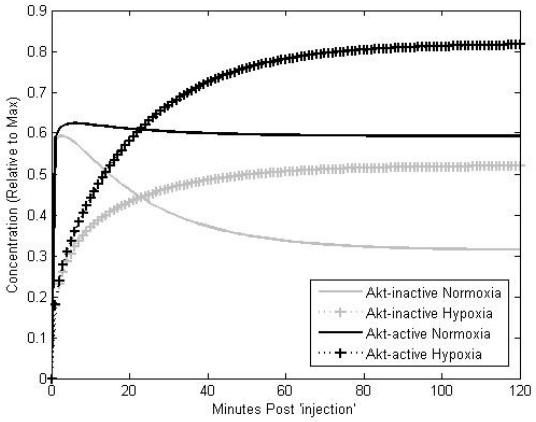


Figure 5. Temporal Profiles of FDG for normoxic and hypoxic tissue under Akt-mediated upregulation of glycolysis ($\alpha = 1$, $\beta = 2$) compared to Akt-inactive tissue ($\alpha = 1$, $\beta = 1$). Max contrast Akt-inactive: 1.65, Akt-active: 1.32.

4 DISCUSSION AND SUMMARY

This model highlights several significant differences between what can be inferred, regarding hypoxia, from FDG and Fmiso images. Figure 3 would suggest that in the absence of other glycolytic regulators, FDG would be a sound candidate for a surrogate marker of hypoxia, as proposed previously [6]. Its uptake is increased in the presence of hypoxia and less time is required to achieve contrast than when using Fmiso. However, regulation of glycolysis by other factors is an issue. Akt acts via both stabilisation of HIF levels and direct upregulation of glycolytic enzymes, and has been shown to enhance FDG uptake in mice transfected with Akt-active cells [4]. Whilst HIF stabilisation has a minor effect on FDG uptake, Figure 5 demonstrates that HIF-independent regulation of glycolysis (Warburg Effect) could be a key factor in lack of correlation between FDG and Fmiso binding in clinical images. Any non-hypoxia mediated regulation of FDG may potentially lead to a reduction in hypoxia:normoxia contrast.

Although FDG is only an independent marker of hypoxia for hypoxia-regulated tumours, this in itself is potentially useful. If FDG and Fmiso are used in conjunction with one another, the pattern of correlation may provide further tumour characterisation. High correlation between tracers may indicate a lack of oncogenic pathway activation and thus a better prognosis. Low correlation, where FDG binds and Fmiso does not, may indicate enhanced oncogenic activity. These are presently just hypotheses and require further investigation.

Table 1. MR_{gluc} for each condition

Akt Status	Oxygen Status	MR_{gluc}
Akt-inactive	Normoxia	0.004
Akt-inactive	Hypoxia	0.008
Akt-active (HIF)	Normoxia	0.007
Akt-active (HIF)	Hypoxia	0.012
Akt-active (Gly)	Normoxia	0.015
Akt-active (Gly)	Hypoxia	0.025

To summarise, in the absence of alternative regulation, our model would suggest that FDG is a good candidate for a surrogate marker of hypoxia. However, when other regulating factors are considered, it becomes clear that FDG can no longer act as an independent hypoxic marker. The potential exists for using Fmiso and FDG as complementary imaging agents, to provide a more complete tumour characterisation. This complementarity presents a strong argument for image fusion.

5 ACKNOWLEDGEMENTS

The authors wish to acknowledge the advice of Professor Adrian Harris and his research group at the Weatherall Institute of Molecular Medicine, University of Oxford.

6 APPENDIX

Symbol	Value	Details
r_H	$6.9 \cdot 10^{-3}$	Maximum HIF production rate constant
d_H^n	$5.1 \cdot 10^{-2}$	Maximum HIF degradation rate constant
k_H	0.4	Oxygen concentration at half maximum degradation rate
d_O	2.8	Oxygen consumption $\mathrm{V_{max}}$
k_O	0.15	Oxygen consumption $\mathrm{k_m}$
γ	0.99	Proportion of reduction
D_O	$2.2 \cdot 10^2$	Oxygen diffusion coefficient
d_G^n	$9.4 \cdot 10^{-3}$	Glucose consumption $\mathrm{V_{max}}$ (normal tissue)
k_G	$1 \cdot 10{-}2$	Glucose consumption $\mathrm{k_m}$
k_A	4	Effect of HIF on Glycolysis rate constant
D_G	84	Glucose diffusion coefficient
d_M	0.09	Maximal Fmiso binding $\mathrm{V_{max}}$
k_m	$3 \cdot 10{-}2$	Fmiso consumption $\mathrm{k_m}$
D_M	8.4	Fmiso diffusion coefficient
τ	$9.6 \cdot 10^2$	Transcriptional delay

References

1. G. Semenza, P. Roth, H. Fang et al. "Transcriptional regulation of genes encoding glycolytic enzymes by hypoxia-inducible factor 1." *J Biol Chem* **269**, pp. 23757–63, 1994.
2. L. Cher, C. Murone, N. Lawrentschuk et al. "Correlation of hypoxic cell fraction and angiogenesis with glucose metabolic rate in gliomas using 18f-fluoromisonidazole, 18f-fdg pet, and immunohistochemical studies." *J Nucl Med* **47(3)**, pp. 410–8, 2006.
3. J. Rajendran, D. Mankoff & F. O. Sullivan. "Hypoxia and glucose metabolism in malignant tumors: evaluation by [18f]fluoromisonidazole and [18f]fluorodeoxyglucose positron emission tomography imaging." *Clin Cancer Res* **10**, pp. 2245–52, 2004.
4. R. Elstrom, D. Bauer, M. Buzzai et al. "Akt stimulates aerobic glycolysis in cancer cells." *Cancer Res* **64**, pp. 3892–3899, 2004.
5. R. Thomlinson & L. Gray. "The histological structure of some human lung cancers and the possible implications for radiotherapy." *Br J Cancer* **9**, pp. 539–49., 1955.
6. A. Pugachev, S. Ruan, S. Carlin et al. "Dependence of FDG uptake on tumor microenvironment." *Int J Radiat Oncol Biol Phys* **62(2)**, pp. 545–53, 2005.

Improving the Contrast of Ultrasound Images using an Autoregressive Model Based Filter

E von Lavante[a*] and J A Noble[a]

[a]Wofson Medical Vision Lab, University of Oxford

Abstract. An autoregressive model has been employed to recover from the backscattered RF-signal a tissue harmonic image which is not degraded by harmonic leakage.A significant overall improvement of the image quality has been achieved with the filtered image, where measurements have shown a significant (up to 45 %) increase in contrast between cancer masses and benign tissue.

1 Introduction

Ultrasonic imaging has become an indispensable tool for medical diagnosis due to its noninvasive nature, low cost, portability and real time image formation. However, despite its widespread use, even for skilled radiologists it is very challenging to correctly interpret ultrasound images. This is due to strong image artifacts characteristic to ultrasound imagery, such as attenuation, speckle, shadowing and general low contrast.

Of these effects, the speckle texture reduces the perceived spatial resolution, reducing the effectiveness of ultrasound to identify subtle tissue features. Speckle is a form of locally correlated multiplicative noise [1], which both corrupts the image and conveys important information. Speckle is furthermore the main contributor to the low SNR (signal to noise ratio) exhibited by ultrasound images.

Recent advances in signal processing have reduced the effect of speckle on the overall image appearance with the introduction of new techniques such as dynamic beam focussing and tissue harmonic imaging. However, speckle still remains the predominant texture feature in any ultrasound B-mode image. A lot of research has gone into speckle reducing filters, with one of the most recent ones being the Speckle Reducing Anisotropic Filter [2] derived from the Frost and Lee filters [3, 4]. Despite their general effectiveness in reducing speckle, they tend to oversmooth the image, and are consequently rarely used for diagnostic breast ultrasound.

Other attempts to improve the ultrasound image are based on deconvolution methods which try to estimate the point-spread function of the underlying tissue [5–7]. However, results of these methods have so far been unconvincing for improving the general image quality of ultrasound images, and their application for clinical use remains very limited.

Of all methods mentioned above, the only widely used clinical method appears to be harmonic imaging [8]. Consequently, our approach is motivated by this idea.

1.1 Tissue Harmonic Imaging

Due to the nonlinear nature of tissue, the back-scattered signal from an ultrasound pulse interacting with its target will contain, besides the fundamental frequency of the emitted pulse, also higher harmonics. These higher harmonics are mostly generated due to the peaks of transmitted pulse travelling faster than the troughs due to tissue having different velocities for sound wave propagation in compressed tissue as opposed to relaxed tissue. This effect causes very weak harmonics which are accumulated as the emitted ultrasound pulse propagates through tissue [9]. Hence, the signal received at the ultrasound transducer is made up of the fundamental frequency generated by direct reflections of the ultrasound pulse at tissue interfaces and inhomogeneities, and the higher harmonics generated by the tissue itself, called here the tissue harmonic signal.

The signal arising from the tissue harmonics has several advantages over the signal from the fundamental frequency. The distortion of the emitted ultrasound pulse is caused to a large degree by the skin and body wall of the patient. In theory at least, the tissue harmonic signal is not affected by this distorting first pass through the skin layers, as the signal is generated directly in the tissue itself, and not by the transducer. Correspondingly, the tissue harmonic signal will contain less noise and especially less lateral blurring than the signal from the emitted pulse. Additionally, the tissue harmonic signal is much more focused as the higher harmonics are generated mainly in the centre of the emitted beam where the acoustical pressure is maximal, again causing an improved lateral resolution.

*etienne@robots.ox.ac.uk; Wofson Medical Vision Lab, University of Oxford, Parks Road, Oxford OX1 3PJ

There are two established methods for harmonic imaging, the first one is based on using the pulse inversion technique [10], and the other one is by the application of finite impulse response (FIR) filters to filter out the fundamental harmonic and second harmonic of the received signal. While the pulse inversion technique in very effective in reducing the effects of harmonic leakage from the fundamental harmonic, it has the major disadvantage of halving the effective frame rate and being more susceptible to motion artefacts. With the other method one has the problem that harmonic leakage of the fundamental harmonic effects the higher harmonics [11], adding additional noise to these harmonics. In theory one can overcome this problem by emitting a very narrow band pulse. However in practise one cannot avoid side lobes to the emitted pulse and any pulse will become more broad band as it propagates through tissue due to the above mentioned non-linearity of tissue.

2 Method

One way to overcome the effects of harmonic leakage is to develop a linear predictive model of the emitted pulse and then consequently using parametric spectral estimation to both estimate the spectral contents of the emitted pulse and the remaining tissue harmonic signal. In the case one has found the proper parametric model, the advantage of using such a model compared to Fourier transform based methods is its ability to detect peaks at superresolution, greater resistance to spectral ringing and reduced computational complexity. In several pieces of work [12, 13] an autoregressive (AR) model of the received RF-signal has been used to estimate various tissue parameters from ultrasound data. The output $y(n)$ of such an AR-model of order M is defined as the output of the following linear filter driven by white Gaussian noise $\eta(n)$:

$$y(n) = \eta(n) - \sum_{k=1}^{M} a(k)y(n-k) \tag{1}$$

where a are the AR coefficients of an order M AR-model. Intuitively, most users of ultrasound devices regard the B-mode image as a two dimensional representation of the power of the reflected signal of an insonified target. With this observation and using the model in (1), one can directly compute the power spectral density (PSD) $P_{yy}(f)$, and consequently the power of the signal, as:

$$P_{yy}(f) = \frac{r_p}{\|1 + \sum_{k=1}^{M} a^{(M)}(k)e^{-2j\pi fk}\|^2} \tag{2}$$

In linear predictive modelling this model will correspond to an all-poles model, whose power spectrum will be a series of peaks, corresponding to the AR coefficients $a^{(M)}$, among a general flat noise level, whose power is defined by the residual r_p. The received signal of the emitted pulse can be modelled very closely by an all-poles model, as its power spectrum is, by design, a Gaussian bell-curve. Any noise in the received signal will be due to non-linear interaction of this pulse with tissue, hence corresponding to the signal of the tissue harmonics, as described in the previous section. Using this model, the power of the received tissue harmonic signal is defined by the residue r_p, which can be directly displayed to yield an estimation of the tissue harmonic signal without the noise introduced by harmonic leakage.

In its current implementation $a^{(M)}$ and r_p are estimated from RF-data using Burg's iterative algorithm [14], using an AR-model order of seven, following the recommendation from [12]. Experiments have shown that the results are very robust to the chosen AR-model order. As the resulting residue signal has still a very high dynamic range, the dynamic range has been compressed by applying the Gamma-Brightness correction function, as it is used in analogue screens, to the data before displaying it:

$$P_{yy_{new}} = P_{yy} \exp(1/\gamma) \qquad \text{with } \gamma = 4.5 \tag{3}$$

The value for the parameter $\gamma = 4.5$ was experimentally chosen to give, in the experiment, a sufficient contrast to detect subtle image details in the filtered image. The value for this variable is image dependent and in our experiments can be adjusted between 3.5 and 5.0 depending on the desired contrast and viewing preferences of the user.

3 Methods for Evaluation

Unfortunately, due to the cross-patient variability of ultrasound data one cannot compare directly the image statistics of ultrasound data across individual image sets. In order to evaluate any improvements the filtering technique has on the ability to distinguish lesions from healthy tissue, the following procedure was used: various image classifiers computed from a region within a cancer mass and a region of healthy tissue with fully developed speckle were compared in each image. For each image, the contrast to noise ratio (CNR) of each classifier was computed as:

$$CNR = \frac{\|\mu_c - \mu_h\|}{\sqrt{\sigma_c^2 + \sigma_h^2}} \tag{4}$$

where μ_c and μ_h are the mean of the classifier in the cancer and healthy tissue regions, and σ_c and σ_h are the standard deviation of the classifier in the cancer and healthy tissue regions, respectively. Having calculated both the CNR of the filtered image CNR_f and the original B-mode data CNR_b, one can calculate the percentage fractional improvement of the classifier $frac_{\text{imp}}$ as:

$$frac_{\text{imp}} = (\text{CNR}_f/\text{CNR}_b) * 100 \tag{5}$$

3.1 Intensity Based Statistics

Using the above method, one can directly compute and compare the general gray-scale contrast between cancer masses and healthy tissues. Furthermore, one can also compute the SNR of both the cancer region SNR_c and healthy region SNR_h by [1]:

$$\text{SNR} = \overline{I}/\sigma_I \tag{6}$$

with $\overline{I}$ being the mean intensity and σ_I being the standard deviation of the intensity of the region of interest. However, as the physical principles causing speckle cannot be removed, one should not expect any significant improvement of the SNR by filtering the image, and it will remain by definition close to unity. The main purpose of the SNR is to quantify the amount of speckle in the region of interest.

3.2 Texture Measures

Due to the texture created by speckle, any trained clinician will not evaluate a region of interest solely on its image intensity, but rather on the general appearance of image texture. Hence, the 1st order statistics give only a very incomplete picture of the performance of the filtered image. One of the texture based metrics chosen for the validation of this method is the Neighbourhood Gray-Level Difference Matrix (NGTDM) [15] and the metrics derived from this statistic. The details on how to populate the NGTDM matrix $s(i)$ can be read in [15]. Using this matrix one can calculate the contrast f_{cont}, busyness f_{busy}, complexity f_{comp} and coarseness f_{coar} of a texture.

4 Results and Discussion

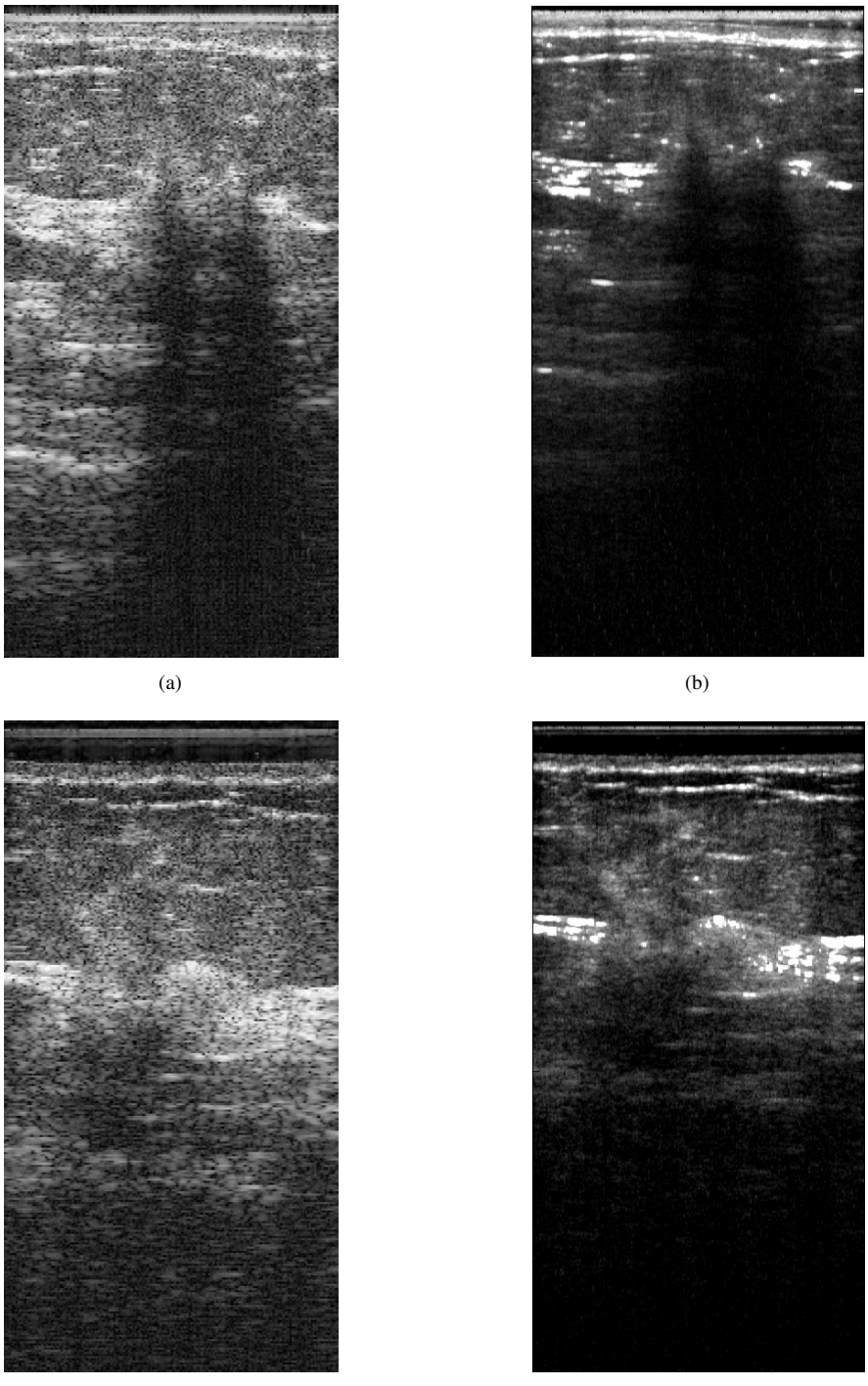
4.1 Experimental Set-up

Since tissue phantoms cannot reproduce the subtleties of real diagnostic ultrasound data, and the main advantage of harmonic imaging is the ability to identify subtle features, all experiments were done on an existing database of breast ultrasound RF-data with histological proven cancer cases. All data has been recorded during a study in breast cancer screening, on an Analogic AN-2300 with a BK-Medical 8805 probe using a centre frequency of 4.0 MHz, and recording the RF-data at a sampling frequency of 40 MHz. A subset of 25 cancer cases were randomly chosen from the database and analysed using the methods described in Section 3. As the filtered image will be displayed in a similar resolution to that of B-mode data and has to be compared to B-mode data, the filtered image has been downsampled in axial direction to the same resolution as that of the B-mode data used for the experimental analysis. In all datasets analysed, the regions of interests have been manually selected for both in-lesion and healthy tissue regions, having an extent of 30 pixels in lateral direction and 50 pixels in axial direction.

4.2 Results

As can be seen in Fig. 1, when comparing the filtered image to classical B-mode, the quality of the filtered image is significantly better. One cause of the improved image quality is the increased axial resolution of the image due to the much higher sampling rate of the RF-data sets. Besides of the increased image resolution, the contrast has been significantly improved and speckle texture is much finer grained. This visual impression is also strongly supported by the image intensity based statistics shown in Table 1, with the fractional improvement of the CNR being on average 145%. This large improvement in the CNR was found for almost all images. However there were still some images, such as the first set displayed in Figure 1, where there was no improvement. Surprisingly, the SNR improved with an average fractional improvement of the SNR_c being 105% and the SNR_h being 118%. This increase of the SNR appears to be still in a similar order of magnitude as the variance within the data, showing that the filter could not significantly reduce the general presence of speckle. From Table 2, depicting the performance of the texture classifiers, only the contrast and coarseness classifiers have shown consistent results, as the individual values for the busyness and complexity measures have shown a too large variance across the set of images to make any conclusions possible. It appears that these two measures do not well describe the image textures found in ultrasound data. The improvement of the contrast measure of the NGTDM in the filtered was caused by the reduction of speckle inside most cancer regions,

producing a very low contrast texture compared to the surrounding tissue. The reduction of speckle inside most cancer regions is also responsible for the improvement of coarseness CNR, as the coarseness of the cancer texture in the filtered image has been reduced by a factor of 10, while the coarseness of the surrounding tissue still remains similar to the values for the B-mode data.



(a) (b)

(c) (d)

Figure 1. the image quality improvement is easily noticable between classical B-Mode images a) and c) and the filtered images b) and d)

Table 1. Experimental Results: Intensity based metrics

	mean_{bm}	$\text{mean}_{\text{filt}}$	σ_{bm}	σ_{filt}	frac_{imp} %
CNR	0.82	1.19	0.052	0.13	145
SNR_c	0.94	0.99	0.017	0.014	105
SNR_b	0.98	1.16	0.008	0.025	118

Table 2. Experimental Results: Texture based metrics

	CNR_{bm}	CNR_{filt}	frac_{imp} %
f_{busy}	0.23	0.49	216
f_{cont}	2.7	7.9	297
f_{coar}	0.094	2.4	2539
f_{comp}	1.33	1.27	95

5 Conclusion

Both visual inspection and quantitative analysis have shown that the filtered image has in the majority of the clinical datasets a much higher contrast between the cancer masses and its surrounding tissue. Especially the improvement of the image intensity CNR is a notable result, and encourages the possibility of developing a segmentation algorithm for segmenting lesions and malignant masses using this filtered image data. The algorithm is not computationally demanding, and could be readily implemented in real time on current hardware. Hence, the next step of our research is to implement this filter on to an ultrasound machine, to assess the degree to which the measured increase in contrast effects the diagnosis of breast cancer in clinical practice.

This work is funded by the Engineering and Physical Sciences Research Council (EPSRC);

EPSRC Reference: GR/S94575/01

References

1. C. B. Burckhardt. "Speckle in ultrasound b-mode scans." *IEEE Transactions on Sonics and Ultrasonics* **25(1)**, 1978.
2. Y. Yu & S. Acton. "Speckle reducing anisotropic diffusion." *IEEE Transactions on Image Processing* **(11)**, pp. 1260–1270, 2002.
3. V. S. Frost, J. A. Stiles, K. S. Shanmuggam et al. "A model for radar images and its application for adaptive digital filtering of multiplactive noise." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **4(2)**, pp. 157–165, 1982.
4. J. S. Lee. "Digtial image enhancement and noise filtering by using local statistics." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2(2)**, pp. 165–168, 1980.
5. W. Yeoh & C. Zhang. "Constrained least squares filtering algorithm for ultrasound image deconvolution." *IEEE Transactions on Biomedical Engineering* **53(10)**, pp. 2001–2007, 2006.
6. T. Taxt. "Restauration of medical ultrasound images using two-dimensional homomorphic deconvolution." *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control* **42(4)**, pp. 543–554, 1995.
7. T. Taxt & J. Strand. "Two-dimensional noise robust blind deconvolution of ultrasound images." *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control* **48(4)**, pp. 861–866, 2001.
8. F. Tranquart, N. Grenier, V. Eder et al. "Clinical use of ultrasound tissue harmonic imaging." *Ultrasound in Medicine and Biology* **25(6)**, pp. 889–894, 1999.
9. H. C. Starritt, F. A. Duck, A. J. Hawkins et al. "The development of harmonic distortion in pulsed finite-amplitude ultrasound passing through liver." *Physics in Medicine and Biology* **31**, pp. 1401–1409, 1986.
10. D. H. Simpson & P. N. Burns. "Pulse inversion doppler: A new method for detecting nonlinear echoes from mircobubble contrast agents." *Proceedings IEEE Ultrasonics Symposium* pp. 1597–1600, 1997.
11. C. Shen & P. Li. "Harmonic leakage and image quality degradation in tissue harmonic imaging." *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control* **48(3)**, pp. 728–736, 2001.
12. J. Gorce, D. Friboulet, I. Dydenko et al. "Processing radio frequency ultrasound images: A robust method for local spectral features estimation by a spatially constrained parametric approach." *IEEE Transactions on Ultrasonics, Ferroelectrics, Frequency Control* **(12)**, pp. 1704–1719, 2002.
13. K. A. Wear, R. F. Wagner, M. F. Insana et al. "Application of autoregressive spectral analysis to cepstral estimation of mean scatterer spacing." *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control* **(1)**, pp. 50–58, 1993.
14. W. Press, S. Teukolsky, W. Vetterling et al. *Numerical Recipes in C++*. Cambridge University Press, 2002.
15. M. Amadasun & R. King. "Textural features corresponding to textural properties." *IEEE Transactions on Systems, Man, and Cybernectics* **19(5)**, pp. 1264–1274, 1989.

Automated Segmentation of Telomeres using Poisson MAP-MRF Labelling

Hugh Gribben[a1] Paul Miller[1]
Hongbin Wang[1] Mark Browne[2]

[1]The Institute of Electronics, Communications and Information Technology (ECIT)
Queen's University Belfast, Belfast, UK. BT3 9DT

[2]Andor Bioimaging Division, Morrisville, USA. NC 27560

Abstract. An automated unsupervised technique, based upon a Bayesian framework, for the segmentation of low light level imagery is proposed. Primarily, Mixture Modelling is used to provide a baseline estimate. This estimate is then refined to consider spatial correlations using Markov Random Field (MRF) Modelling. The technique has been implemented assuming low-light level Poisson statistics, and the results compared to the more widely used assumption of Gaussian statistics. Investigations revealed the Poisson technique quantitatively outperforms the Gaussian technique for synthetic low light imagery, both before and after avalanche multiplication via the multiplication register of a Low-light Level Charge Coupled Device (L3CCD). The technique was then applied to the task of segmenting a biomedical dataset obtained from a L3CCD, containing intra-cellular bodies, Green Fluorescent Protein (GFP) labelled telomeres, in an inter-phase cell nucleus. Qualitative results were promising, again showing improvement over the Gaussian technique.

1 Introduction

The recent development of L3CCD sensors has enabled biomedical experiments, which would have previously been impossible. One such experiment is the tracking of intra-cellular bodies, called telomeres, using a confocal microscope that employs an L3CCD. Tracking telomeres should help biomedical scientists gain insight into how and why they move in relation to each other, and the cell nucleus as a whole [1]. This knowledge could have implications in anti-aging and anti-cancer therapy. A novel aspect of the multidimensional datasets produced by an L3CCD is that they are Poisson distributed. Using this knowledge, we investigate the applicability of a Poisson Maximum a Posterior-MRF (MAP-MRF) approach to the segmentation of telomeres. Related works [2-5] typically adopt a Gaussian MAP-MRF approach to segmentation; however, very little previous work has been done on integrating Poisson models into MAP-MRF schemes [6]. To the authors' knowledge, no segmentation techniques have been previously applied to this class of imagery.The paper is organised as follows. In section 2, our Poisson MAP-MRF approach is defined. In section 3, results obtained are reviewed and analysed. Finally, conclusions and future work are discussed in section 4.

2 Poisson MAP-MRF Approach

2.1 Low Light Level Statistics and L3CCD Theory

At low light levels (up to approx. 50 photons per pixel per integration time), photon counts are distributed according to the Poisson distribution

$$p(y \mid \mu) = \exp\left(y \log(\mu) - \mu - \log(y!)\right) \tag{1}$$

where y is a non-negative number of occurrences, and μ the mean number of occurrences. L3CCDs contain a multiplication register which implements electron avalanche multiplication so that a large mean gain can be realised prior to the conventional readout amplifier. When the photon input level is small and the gain large, the L3CCD output can be estimated by providing the probability distribution of the L3CCD with (1), giving

$$p(x,u) = \sum_{n=1}^{\infty} \frac{\exp\left(-\mu - x/g\right)\mu^n \left(x/g\right)^{n-1}}{g\,(n-1)!\,n!} \tag{2}$$

This is the probability the output will be x when the mean light level is μ, the input number of photons is n, and the mean gain g [7].

[a] hgribben01@qub.ac.uk

2.2 Derivation of Poisson MAP-MRF Technique

Using Bayes estimation, a risk is minimized to obtain an optimal estimate. In terms of the segmentation problem, according to the Bayes rule, the posterior probability of a labelling estimate X given an observation Y can be computed by $P(X|Y) = P(Y|X)P(X) / P(Y)$ where $P(X)$ is the prior probability of X, $P(Y|X)$ is the likelihood function of X for fixed Y, and $P(Y)$ is the density of Y. The minimal risk estimate is therefore $X^* = \arg\max_X P(X \mid Y)$, the MAP estimate. As $P(Y)$ is a constant for fixed Y, the MAP estimate is equivalently found by

$$X^* = \arg\max_X \{P(Y \mid X) P(X)\} \tag{3}$$

In many applications, an initial estimate for X is obtained using mixture modelling, assuming a mixture density of Gaussian distributions. However, taking into account the Poisson nature of the low-light level imagery we wish to segment, we model the intensity distribution as a mixture of Poisson distributions, given by

$$f(Y_i \mid \theta, \lambda) = \sum_{k=1}^{K} \lambda_k f_k(Y_i \mid \theta_k) \tag{4}$$

where K is the number of assumed Poisson distributions, Y_i = observed value at pixel i, θ_k = the set of mean vectors $\{\mu_k\}$; $f_k(\cdot \mid \theta_k)$ is a Poisson density with mean μ_k; $\theta = (\theta_1, \ldots, \theta_k)$ and $\lambda = (\lambda_1, \ldots, \lambda_K)$ is a vector of mixture probabilities such that $\lambda_k \geq 0 \ (k = 1, \ldots, K)$ and $\sum_{k=1}^{K} \lambda_k = 1$ [2].

The algorithm used in practice to find the mixture of distributions that best model the dataset is the Expectation Maximisation (EM) algorithm, first introduced by Dempster et al. [8]. This is an iterative algorithm, which estimates the parameters via a Maximum Likelihood (ML) criterion. The mixture density returned can then be used to associate pixel observations with a Poisson density k using a simple ML estimation.

In MAP-MRF labelling, $P(X|Y)$ is the posterior distribution of a MRF. The prior model $P(X)$ takes into account spatial correlations present in an image, and is dependent upon the type of scene. Assuming our scene to be a piecewise constant surface, we consider an indicator function, $I(X_i,X_j) = 1$ if $X_i = X_j$ and otherwise = 0. The Potts model can be described by $P(X) \propto \exp\left(\phi \sum_{i,j} I(Xi, Xj)\right)$, where the sum is computed over all neighbour pairs. Spatial homogeneity in the model is expressed using a parameter ϕ, small values implying randomness, and large values implying uniformity [9]. The prior energy can then be defined as the negative of the sum of all clique potentials over X [10]

$$E(X) = -\sum_{i \in S} \phi \sum_{i,j} I(X_i, X_j) \tag{5}$$

where S is the set of all pixel sites, i is the site currently under consideration, and i,j the set of all pair-wise cliques with respect to i.

The likelihood model $P(Y|X)$ depends upon physical considerations. With the traditional approach, we assume Y to be degraded versions of a MRF realization X due to independent additive Gaussian noise. Taking a Gaussian distribution as a special form of a Gibbs distribution, we take advantage of a MRF's equivalence to a Gibb's distribution [10] to then define the likelihood as $P(Y \mid X) = \exp(-E(Y \mid X)) \Big/ \prod_{i \in S} \sqrt{2\pi\sigma_i^2}$ where σ_i^2 is the variance of the estimated Gaussian distribution at pixel i, and

$$E(Y \mid X) = \sum_{i \in S} (X_i - Y_i) / 2\sigma_i^2 \tag{6}$$

is the likelihood energy. Instead, we assume low-light statistics, taking each observed pixel value to be the sum of the true grey level value and independent Poisson noise. We then take (1) as a special form of the Gibbs distribution, as opposed to the Gaussian distribution; and again take advantage of a MRF's equivalence to a Gibb's distribution to define the likelihood as

$$P(Y \mid X) = \exp(E(Y \mid X)) \tag{7}$$

where

$$E(Y \mid X) = -\sum_{i \in S} (Y_i \log(X_i) - X_i - \log(Y_i)) \tag{8}$$

is the likelihood energy. Finally, we can add the prior and likelihood energies to yield the posterior energy

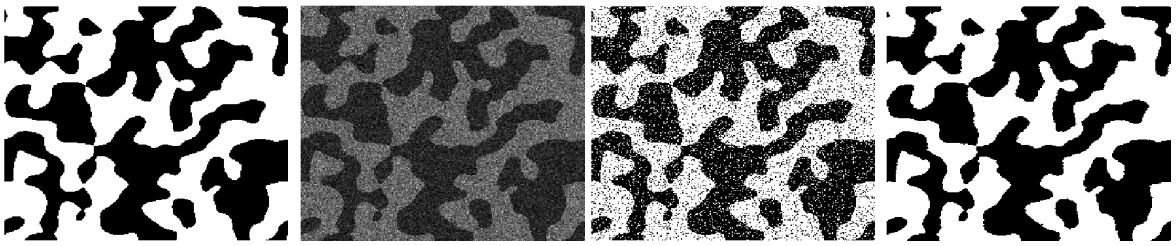
$$E(X \mid Y) = -\sum_{i \in S} (Y_i \log(X_i) - X_i - \log(Y_i)) - \sum_{i \in S} \phi \sum_{i,j} I(X_i, X_j) \tag{9}$$

The MAP estimate can then be found by then minimising the posterior energy. Practically, this can be performed by the use of the Iterated Conditional Modes (ICM) algorithm, originally introduced by Besag [11]. This is an iterative algorithm that begins with the observed scene *Y*, and the initial estimate of the true scene *X* from mixture modelling. It then proceeds to provide a new estimate of the true scene at each iteration, updating ϕ for each iteration until convergence is reached, or a maximum number of iterations complete.

3 Results and Analysis

3.1 Segmentation of Low-Light Data

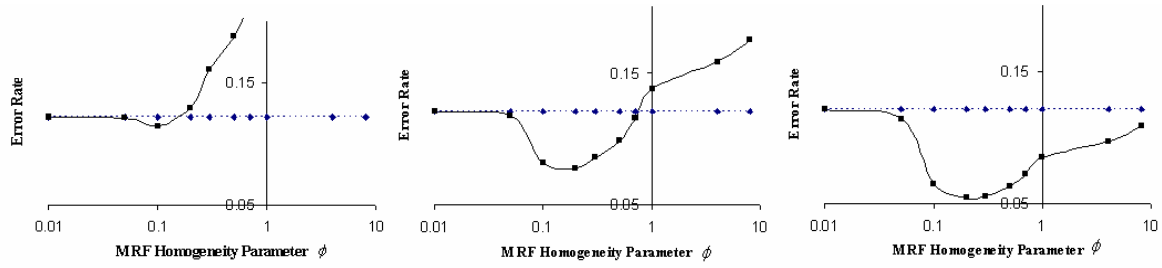
The proposed technique was initially applied to synthetic data with mean gain set to 1 for analytical purposes. In the first case, an image is produced by the creation of a binary random pattern, by convolving a normally distributed 256x256 random array with a Gaussian filter. This allows us to quantify the spatial dependencies within the data by comparison of the Full Width at Half Maximum (FWHM) parameter of the filter used (The greater the FWHM, the greater the spatial dependencies). Pixels are then labelled depending upon value. This provides us with our "true scene" $\hat{X}$ that we wish to recreate via our segmentation technique. Poisson noise is then added to each pixel, the mean Poisson value being dependent upon the initial labelling, to give low-light synthetic data. Fig. 1 shows examples of both, as well as qualitative results showing the initial estimate returned from the mixture modelling step, and the application of the MAP-MRF technique. Here, Poisson noise is added to the true scene with means equivalent to light levels of 4 and 10 photons/pixel/integration time, dependent upon the initial labelling. The application of the MAP-MRF technique shows considerable improvement over the initial estimate returned from the mixture modelling step, as well as a close approximation to the true scene we are attempting to recreate.



(a) (b) (c) (d)

Figure 1. Automated Segmentation of Synthetic Data (a) 'True Scene' with FWHM = 11.77 (b) 'Observed Scene' ('True Scene' with Poisson Noise Added) (c) Output - Mixture Modelling Segmentation of b (d) Output - Poisson MAP-MRF Segmentation of b

Fig. 2 shows quantitative results obtained from the MAP-MRF technique compared to the initial estimate returned from the mixture modelling step, as ϕ is increased over 10 iterations of the ICM algorithm for different initial true scenes with increasing spatial dependencies from 2(a) through 2(c). Quantitative results were found by comparing returned estimates *X* with the true scene using the error rate $err = \sum_{i \in S} (1 - M) \Big/ S_n$,where S_n is the total number of



(a) (b) (c)

Figure 2. Output Error Rates obtained from technique over 10 iterations varying ϕ (Solid Curve) compared to output from mixture modelling alone (Dotted Curve). Data created with Poisson distributed values of 4 and 10 photons/pixel/integration time (a). FWHM = 0.47. (b) FWHM = 2.4 (c) FWHM = 5.9

pixels and $M = 1$ if $X_i = \hat{X}_i$, $M = 0$ otherwise. The minimization of the error rate corresponds to the best input prediction, i.e. an ideal segmentation should give an error rate of $err = 0$. It can be seen in each case that an optimal value for the error rate improves upon the error rate returned from mixture modelling alone. Convergence of the ICM algorithm is shown in each case where the error rate is minimised, i.e. where the Potts model with a value of ϕ best represents the spatial dependencies present in the input image. The figures also show that as the spatial dependencies increase from 2(a) through 2(c), improvements in the corresponding results can be clearly seen. We can see that the optimal error rates returned are improved as the spatial dependencies within the original true scene are increased, and that the error rate is better than that returned with mixture modelling alone over a larger percentage of chosen ϕ. These results are intuitive with what we would expect, and were replicated for all synthetic imagery with average light levels less than 100 photons/pixel/integration time.

Fig. 3 shows the average percentage improvement of our Poisson MAP-MRF technique over the more widely used Gaussian MAP-MRF technique in terms of optimal error rate returned. Results were obtained from averaging optimal error rates returned from both approaches with varying synthetic images. The figure clearly shows that the Poisson approach outperforms the Gaussian approach as the average light level decreases. This follows intuitively, from our earlier investigation of low-light level statistics.

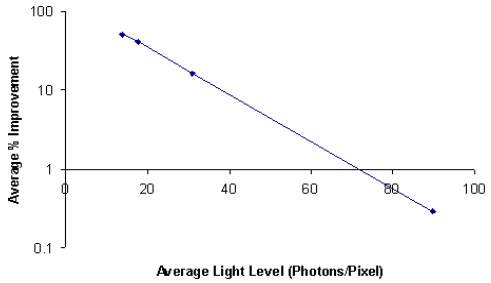


Figure 3. Average Percentage Improvement with Poisson MAP-MRF over Gaussian MAP-MRF as a function of average light level for low-light level Synthetic Imagery

3.3 Segmentation of L3CCD Imagery

Synthetic imagery was then created to approximate the output from a L3CCD sensor with varying mean gains. This was achieved by replicating the avalanche multiplication process applied by the multiplication register of a L3CCD on the low light synthetic imagery. Fig. 4 shows the average percentage improvement of the Poisson MAP-MRF technique over the Gaussian MAP-MRF technique in terms of optimal error rate returned; with differing mean gains applied. The figure again clearly shows that the Poisson approach outperforms the Gaussian approach for all mean gains shown, improving as the average light level decreases. We can also see that, in general, as the mean gain increases, the percentage improvement decreases somewhat due to the fact the distributions present in the data tend to Gaussian as the gain increases.

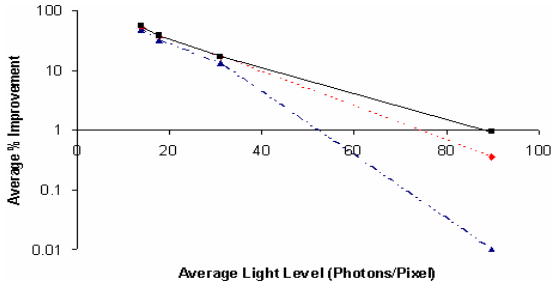


Figure 4. Average Percentage Improvement with Poisson MAP-MRF over Gaussian MAP-MRF as a function of average light level for Synthetic L3CCD Imagery created with varying mean gains of 100 (Solid Curve), 500 (Dotted Curve), and 700 (Dash-Dotted Curve).

Finally, the technique has been applied to the segmentation of data obtained from a L3CCD sensor; consisting of telomeres in a cell nucleus. The ability to track these components via this initial segmentation step should help gain insight into their implications in anti-aging and anti-cancer therapy [1]. In each automated case, the number of total components in an image was chosen via Bayesian Information Criterion analysis, and the value of ϕ chosen via

Pseudo-likelihood Information analysis. Both are automated Bayesian techniques that consider the ratio of likelihoods of output models to determine the optimal model. From left to right, Fig. 5 shows a cross-section of the original dataset, the output from manual segmentation, and the outputs from the MAP-MRF techniques. It clearly shows that the Poisson MAP-MRF technique provides a result closer to that obtained via the manual segmentation technique than the Gaussian MAP-MRF technique, with quantitative results comparing the automated outputs to the manual output reflecting this. The output of the Gaussian segmentation technique on a dataset consisting of the square root values of the original dataset is also included for comparison. We conclude that the assumption of Poisson statistics allows us to better model the true distributions present in the L3CCD datasets.

(a) (b) (c) (d) (e)

Figure 5. Segmentation of cross-section of L3CCD Dataset (a) Cross-section of original dataset (b) Output form Manual technique (c) Output from Gaussian MAP-MRF technique - Error Rate of 0.0036 compared to b (d) Output from Square Root Gaussian MAP-MRF technique - Error Rate of 0.0035 compared to b (e) Output from Poisson MAP-MRF technique - Error Rate of 0.0012 compared to b

4 Conclusions and Future Work

We have presented a novel technique, using a Poisson MAP-MRF approach for the segmentation of low-light imagery which, to our knowledge, has not previously been applied to the described classes of imagery. We have shown that the technique provides improvement over the simpler approach of mixture modelling, and that improved results are seen as the spatial dependencies within the data are increased. We have also shown that the technique shows improvement over a Gaussian MAP-MRF approach at low light levels. The technique has been applied to data output from a L3CCD, with successful results. Future work shall include further investigations to take fuller advantage of the distributions present within L3CCD datasets. Application of the technique to 3D and 4D imagery shall also be implemented.

Acknowledgements

Funding was provided by Andor Technology and DEL NI.

References

1. Y. Cong, W. E. Wright, & J. W. Shay, "Human telomerase and its regulation", *Microbiology Molecular Biology Review*, 66(3), pp. 407-425, September 2002.
2. D. Montgomery, A. Amira, & F. Murtagh, "An automated volumetric segmentation system combining multi-scale and statistical reasoning", *Proceedings of the IEEE International Symposium on Circuits and Systems, Kobe, Japan*, 4, pp. 3789-3792, May 2005.
3. Z. Peng, W. Wee & J. Lee, "MR brain imaging segmentation based on spatial Gaussian Mixture Model and Markov Random Field*", IEEE International Conference on Image Processing*, 2005.
4. H. Deng & D.A. Clausi, "Unsupervised Image Segmentation Using A Simple MRF Model with A New Implementation Scheme", In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2, pp. 691-694, IEEE, August 2004.
5. J. Goldberger & H. Greenspan, "Context-based segmentation of image sequences", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 28(3), March 2006.
6. M.K. Nguyen, H. Guillemin, & C. Faye, "Regularized restoration of scintigraphic images in Bayesian frameworks", *International Conference on Image Processing ICIP 99*, pp. 194-197, 1999
7. G. Basden, C. A. Haniff & C. D. Mackay, "Photon counting strategies with low-light level CCDs", *Mon. Not. R. Astron. Soc,* 345, pp. 985-991, 2003.
8. A. Dempster, N. Laird, & D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, 39(1), pp. 1-38, 1977.
9. D. Stanford & A. E. Raferty, "Approximate Bayes factors for image segmentation: The pseudo-likelihood information criterion (PLIC*)", IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(11), November 2002
10. S. Z. Li, *Markov Random Field Modelling in Computer Vision.* New York: Springer-Verlag, 2001.
11. J. Besag, "On the statistical analysis of dirty pictures,", *Journal of the Royal Statistical Society, Series B (Methodological)*, 48(3), pp. 259-302, 1996.

2-D to 3-D image registration of EPID and CT images for patient setup prior to radiation therapy treatment and evaluation with 4-D CT signal.

Mark Fisher[a] and Yu Su[a*]

[a]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ UK.

Abstract. During the delivery of radiation therapy the human body may move and this can compromise the treatment outcome. Margins are used in treatment planning to compensate for motion and other radiobiological uncertainties. Advances in imaging and the use of intensity-modulated radiation therapy has brought about renewed interest in understanding and compensating for the effects of motion. One of the goals of EU FP6 project MAESTRO is to investigate strategies for the management of intrafraction respiratory motion. This paper focuses on work in progress and investigates 2-D to 3-D image registration for patient localisation and setup prior to treatment. The approach is evaluated using a unique oscillating phantom captured using 4-D CT.

1 Introduction

The effectiveness of intensity modulated radiation therapy (IMRT) in delivering a high-dose distribution to the target while minimising the dose to surrounding organs at risk (see review by Bortfeld [1]) can be compromised by organ motion [2, 3]. Due to the relatively higher dose compared to conventional radiation therapy the need to understand and compensate for the motion of organs during treatment is acute. This has driven researchers to review existing protocols and develop new methods for motion management. Much current understanding of organ motion is based on research using fiducial markers, implanted in the organ and observed with fluoroscopy or electronic portal imaging devices (EPID). The most comprehensive publication on organ motion and its management by Langden and Jones [4] collated tables for liver, diaphragm, kidneys, pancreas, lung, bladder, rectum and prostate from 66 studies. Respiratory motion is a major cause of intrafraction variation and a particular problem for radiation therapy treatment of lung and breast tumours. Studies by Shirato *et al* and Shimizu *et al* [5,6] showed that with free breathing, implanted lung tumour markers moved up to 16 mm.

The most common type of respiratory motion compensation in clinical use are respiratory beam gating systems. The treatment beam is gated by a signal either derived directly from a spirometer or indirectly by tracking external markers placed on the skin. The success of the latter approach relies on an understanding of the correlation between internal and external motion. In a series of studies Vadam *et al* [7,8] have shown that for the diaphragm internal marker motion can normally be predicted from the motion of external markers fixed to the skin with an accuracy of about 3 mm., however, the authors report observing unacceptable variations of up to 9 mm.

Image-guided radiotherapy treatment (IGRT) attempts to track organ motion and compensate for it by moving the treatment beam. Neicu *et al* [9] demonstrated that accurate tumour localisation can be achieved by tracking implanted gold markers in real-time using gantry mounted X-ray systems. This information is used to gate and translate the beam in an approach called synchronized moving aperture radiation therapy (SMART). Recently, Rohlfing *el al* [10] published the first example of markerless real-time tracking using a Cyberknife radiosurgery system (Accuray, Inc., Sunnyvale, CA) employing orthogonal X-ray projection images. The system is used to track the motion of spinal vertebrate in X-ray image frames acquired at 1 min. intervals.

Accurate and fast localisation of a predefined target region inside the patient based on one or more X-ray images is an important component in IGRT that is referred to as the 2-D to 3-D registration problem. The 2-D to 3-D registration problem [11] involves taking one or more X-ray projection (2-D) images of the patient's anatomy and using those projections to determine the rigid transformation $\mathbf{T}$ (rotation and translation) that aligns the coordinate system of the CT (3-D) image with that of the X-ray projection images and thereby the coordinate system of the linac. A schematic overview of this process is shown in Figure 1. Unfortunately, the registration scheme shown in Figure 1 will not operate in real time (25 frames/sec.) as DRR images take about 1-2 seconds to compute. Novel approaches for fast computation of DRR images is the focus of current research [12].

EU FP6 project MAESTRO (Methods and Advanced Equipment for Simulation and Treatment in Radio-Oncology) is investigating a real-time markerless tracking approach for intra-fraction motion compensation. To avoid the need

*email: {mhf, sy}@cmp.uea.ac.uk

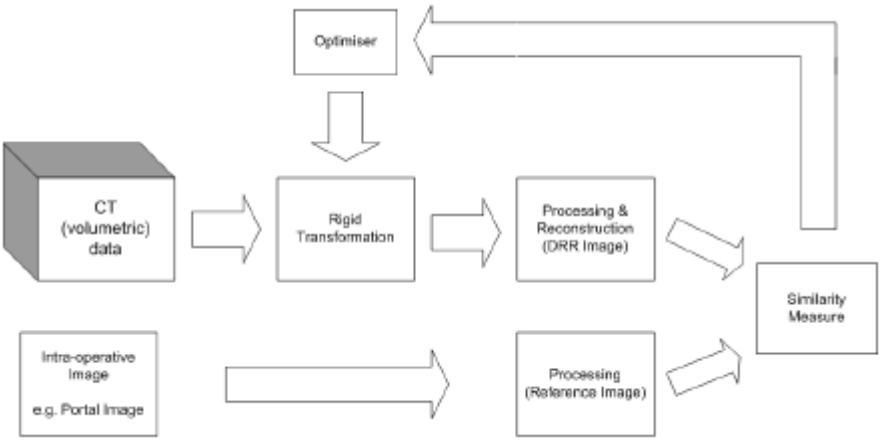


Figure 1. Schematic overview of the 2-D to 3-D registration process. For intensity-based 2-D to 3-D registration, the reference image is an intra-operative projection (2-D) image. It is used as is, with little or no processing. The floating image is a CT (3-D) image. It is processed by generating digitally reconstructed radiographs (DRR - an X-ray projection image derived from CT data) for various orientations of the CT image relative to the X-ray imaging system. The optimizer searches for the rigid transformation **T** that produces the DRR most similar to the real X-ray projection image. The optimal transformation is used to align the CT coordinate system with that of the operating room (from [12]).

to compute 2-D to 3-D registrations at the frame rate, a well documented processing bottleneck [10], we adopt a hybrid active shape model ASM-H [13]. However, there is still a need to verify patient set up and initialise the ASM search. This paper investigates the feasibility of 2-D to 3-D registration for this purpose. We evaluate several image similarity metrics within a 2-D to 3-D registration framework using X-ray DDR projections derived from 4D-CT of a reciprocating phantom simulating a respiratory cycle. The paper proceeds as follows: Firstly, we define the similarity metrics used and describe the registration algorithm, then we define the experimental methodology used to evaluate them and present results. Finally we draw conclusions.

2 Registration Algorithm

We use a 2-D to 3-D registration scheme to initiallise our ASM-H tracker. We have initially focused on intensity-based measures of image similarity since these are well documented and have been shown to perform well in another recently published study into the feasibility of markerless tracking of structures in portal images by Meyer *et al* [14]. The reader is referred to [15] Chapter 4 for a more extensive review of image similarity metrics. Following Meyer *et al* we have implemented four similarity measures: Mean-Sum-of-Squared-Difference (MSSD), Mutual Information (MI), Correlation Coefficient (CC) and Correlation Ratio (CR). These are defined as follows:-

$$MSSD(I_{ref}, I_j) = \frac{1}{N-1}\sum_{i=1}^{N}(B_i - A_i)^2 \quad (1)$$

$$CC(I_{ref}, I_j) = \frac{\sum_{i=1}^{N}(A_i - \overline{A})\cdot(B_i - \overline{B})}{\sqrt{(\sum_{i=1}^{N}(A_i - \overline{A})^2)\cdot(\sum_{i=1}^{N}(B_i - \overline{B})^2)}} \quad (2)$$

$$MI(I_{ref}, I_j) = \sum_{i=1}^{N} P(A_i, B_i) log \frac{P(A_i, B_i)}{P(A_i)P(B_i)} \quad (3)$$

where A is a region in the reference (template) image I_{ref} and B is a corresponding region in the floating image I_j. N is the total number of image pixels.

The correlation ratio (CR) proposed by Roche *el al* [16] is another similarity measure that attempts to improve on MI by taking account of the spatial relationships between pixels in the two images. CR is defined as:

$$CR(I_{ref}, I_j) = 1 - \frac{1}{N\sigma^2}\sum_{i=1}^{N} N_i \sigma_i^2 \quad (4)$$

where σ^2 and σ_i^2 represents the total and conditional variance of B.

Optimisation strategy: We currently use a simple best neighbour search strategy to find the optimum transformation $\mathbf{T}$. The 12 neighbours in parameter space $(x, y, z, \alpha, \beta, \gamma)$ are computed by varying each parameter by some given step size. Then each transformation $\mathbf{T_i}$ is evaluated by generating DRRs and computing the similarity between the DRR and the reference X-ray projection. The search is performed hierarchically, (coarse-to-fine) at four scales using a Gaussian filter with $\sigma = 0.5$ pixels (about 0.25 mm).

In the next section we evaluate the similarity metrics for object localisation and pose estimation using the 2-D to 3-D image registration scheme.

2.1 Experimental Method

Firstly DRR images from static 3-D pelvic CT data set (3 mm. inter-slice spacing) obtained from the Norfolk and Norwich University Hospital (NNUH) were registered to ROI cropped from the simulator X-ray of the same patient. Figure 2 shows the X-ray template and matching DRR. To evaluate the sensitivity of the system we perturb the CT volume by applying $\mathbf{T_{x,y}}$ and then recover the transform parameters (x,y) by image registration.

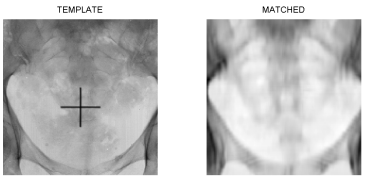


Figure 2. Pelvic CT data set: X-ray template (left); (ROI) matched to DRR (right)

Next, data from a multi-slice 4D CT scanner (1 mm. inter-slice spacing) Somatom Sensation Open (Siemens Medical Solutions, Erlangen, Germany) was provided by Institute Gustave Roussy, Paris. The data comprised 22 3-D CT frames of a phantom acquired at respiratory phases (0-100% exhale, 0-100% inhale). The phantom (Figure 3b), a homogeneous housing into which a combination of geometric shapes each with its own attenuation coefficient μ is inserted, was placed on a reciprocating platform (Figure 3a)), driven by a stepping motor controlled by a digitised respiratory signal (Figure 3a),c)). Hence, the displacement of the phantom in the y direction ($\mathbf{T_y}$) is directly proportional to the 'ground-truth' signal. A set of A-P DRR frames are derived from the 4-D CT sequence (an example frame is shown in

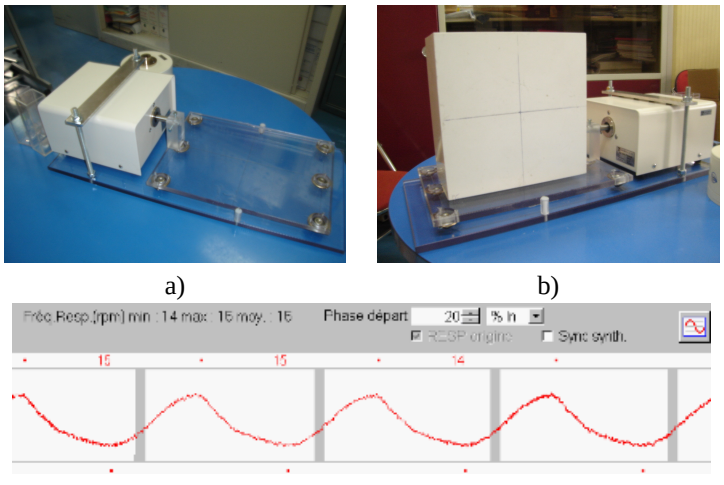


Figure 3. a) Reciprocating Platform, b) Phantom, c) Respiratory gating signal

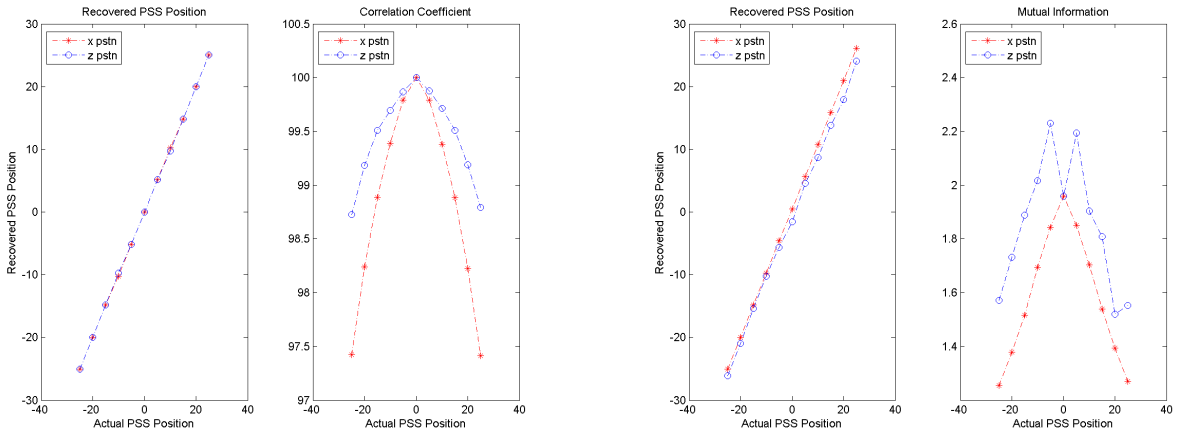
Figure 4a)) using ray-casting algorithm implemented in C++ [17]. The DRR simulates an image formed using a Varian as500 EPID (512x384 pixel) system (Varian Medical Systems, Palo Alto, CA). These are registered to a ROI cropped from the A-P X-ray image (Figure 4b)). The error between the recovered transform parameter $\mathbf{T_y}$ and the stepper motor control signal (ground-truth) is then used to evaluate the similarity measure used by the registration algorithm. Note: motion artifacts visible in some DRR frames are due to synchronisation errors in the 4-D CT acquisition ([18] provides a detailed discussion).

a) b) c)

Figure 4. a) A-P DRR, b) A-P X-ray, c) Template (ROI)

3 Experimental Results

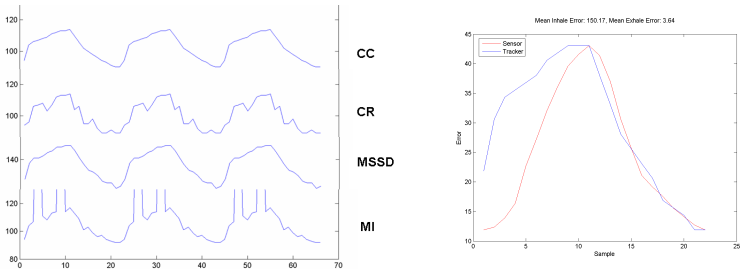
Figure 5 plots x,y values recovered using CC and MI similarity measures (Note: the y and z directions are transposed in the linac co-ordinate frame) and values of the similarity measure for neighbouring transform parameters (x,y) (i.e. to explore the correlation/similarity surface). These results confirm that the x-translation of the PSS can be recovered with a higher degree of accuracy than the y-translation, as y pixels are interpolated from more sparsely sampled z-slices. Interpolation artifacts are probably also responsible for the sidelobes evident in Figure 5b).



(a) (b)

Figure 5. Recovered PSS translation using (a) CC and (b) MI similarity measures.

Figure 6 shows the position of the ROI in each DRR frame recovered by registering the template ROI using CC, CR, MSSD and MI similarity metrics compared to the ground-truth drive signal. The results of an error analysis show that MSSD, and CC similarity metrics perform much better than CR, or MI (RMS errors: 2.99, 4.12, 32.9, 155.3 mm respectively).



(a) (b)

Figure 6. (a) displacement recovered by image registration using CC, CR, MSSD, MI similarity metrics; (b) error analysis (CC similarity metric)

3.1 Conclusions

Our results are consistent with the conclusions drawn by Meyer et al. [14] (i.e. MSSD is the best candidate for real-time object tracking in portal images), but because several seconds are needed to render DRR images the approach cannot

currently be used for real-time image registration, and other techniques (such as ASM-H) must be used for tracking. Nevertheless we can conclude the approach would be suitable in supporting/reinitialising the ASM-H tracker we have developed. Note: During the work we noticed that the motion vectors recovered depended on where we cropped the template. We concluded that motion artifacts present in the DRR were probably compromising the evaluation (i.e. the registration tool was evaluating the integrity of the 4D-CT).

3.2 Further Work

Approaches for fast DRR rendering recently proposed by Russakoff et al. [12] should enable 2D-3D image registration to be computed in less than a second. These will be incorporated in the next version of the software.

Acknowledgements

This work was partly funded by EU FP6 project No. LSHC-CT-2004-503564, Methods and Advanced Equipment for Simulation and Treatment in Radio-Oncology (MAESTRO). The authors wish to acknowledge their collaboration with the Colney Oncology Centre, Norfolk and Norwich University Hospital and Institute Gustave Roussy, Paris and thank them for supplying the CT data used in this study.

References

1. T. Bortfeld. "IMRT: a review and preview." *Physics in Medicine and Biology* **51**, pp. R363–R379, 2006.
2. BIR working party. *Geometrical Uncertainties in Radiotherapy: Defining the planning target volume*. British Institute of Radiology, London., 2003.
3. S. Webb. "Motion effects in (intensity modulated) radiation therapy: a review." *Physics in Medicine and Biology* **51**, pp. R403–R425, 2006.
4. K. Langden & D. Jones. "Organ motion and its management." *Int. J. Radiol. Oncol. Biol. Phys.* **50**, pp. 265–78, 2001.
5. H. Shirato *et al*. "Fluoroscope real-time tumour tracking radiotherapy." *Int. J. Radiat. Oncol. Biol. Phys.* **45(Suppl.)**, pp. 205, 1999.
6. S. Shimizu *et al*. "Fluoroscopic real-time tumor-tracking radiation treatment (RTRT) can reduce internal margin (IM) and set-up margin (SM) of planning target volume (PTV) for lung tumors." *Int. J. Radiat. Oncol. Biol. Phys.* **30(Suppl.)**, pp. 1346, 2000.
7. S.S. Vedam *et al*. "Quantifying the predictability of diaphragm motion during respiration with a non-invasive external marker." *Med. Phys.* **30**, pp. 505–513, 2003.
8. S.S. Vedam *et al*. "Predicting breathing motion for 4D radiotherapy (WIP)." *Med. Phys.* **30**, pp. 1472, 2003.
9. T. Neicu, H. Shirato, Y. Seppenwoolde et al. "Synchronized moving aparture radiation therapy (SMART): average tumour trajectory for lung patients." *Phy. Med. Biol.* **48**, pp. 587–598, 2003.
10. T. Rohlfing et al. "Markerless real-time 3-D target region tracking by motion backprojection from projection images." *IEEE Trans. Medical Imaging* **24(11)**, pp. 1455–68, 2005.
11. D. Tomaževič, B. Likar, T. Slivnik et al. "3D/2D Registration of CT and MR to X-Ray images." *IEEE Trans. on Medical Imaging* **22(11)**, 2003.
12. D. R. *el al*. "Fast generation of digitally reconstructed radiographs using attenuation fields with applications to 2d-3d image registration." *IEEE Trans. on Medical Imaging* **24(11)**, pp. 1441–54, 2005.
13. Y. Su & M. Fisher. "Marker-less intra-fraction organ motion tracking - A hybrid ASM approach." In *Proc. IEEE Int. Workshop on Imaging Systems and Techniques (IST'2007)*. Krakow, Poland, May 2007.
14. Juergen Meyer *et al*. "Tracking moving objects with megavoltage portal imaging: A feasibility study." *Medical Physics* **13(5)**, pp. 1275–1280, 2006.
15. R. Martí. *Image registration applied to multi-modality mammography*. Ph.D. thesis, School of Computing Sciences, UEA, Norwich, UK, 2002.
16. A. Roche, G. Malandain, X. Pennec et al. "The correlation ratio as a similarity measure for mutimodel image registration." In *Proc. MICCAI'98*, volume 1496, pp. 1115–24. Springer, Berlin, 1998.
17. J. P. T. Hernandorena, Y. Su & M. Fisher. "Efficiently computing and registering digitally reconstructed radiographs." In *Proc. Eighteenth International Conference on Systems Engineering*, p. Supplement. Coventry University, Coventry, UK, September 5–7 2006.
18. S.S. Vedam *et al*. "Acquiring a four-dimensional computed tomography dataset using an external respiratory signal." *Phys. Med. Biol.* **48**, pp. 45–62, 2003.

When Less is More: Improvements in Medical Image Segmentation through Spatial Sub-Sampling

P. A. Bromiley* and N.A. Thacker

Imaging Science and Biomedical Engineering, Stopford Building,
University of Manchester, Oxford Road, Manchester, M13 9PT.

Abstract. Segmentation is a common task in medical image analysis. It is frequently solved by fitting an intensity model, consisting of distributions for each pure tissue and each partial volume tissue combination, to the intensity histogram of the image data. However, this approach discards any spatial information present in the data. We present a method that recovers some of this information via regional sub-sampling during the fitting process. Experiments are performed on simple simulated data, simulated MR images from Brainweb, and real MR data from eight young normal subjects. The spatial sub-sampling procedure is shown to significantly improve the segmentation stability.

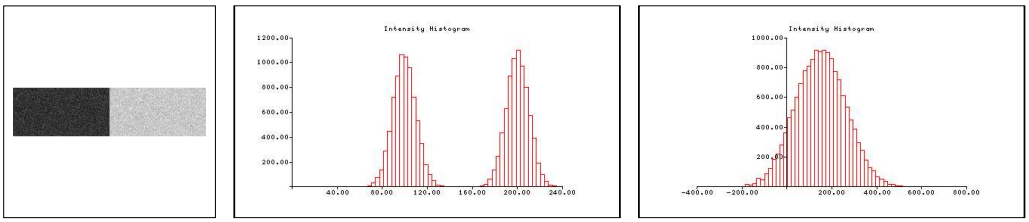
1 Introduction

One of the most common tasks in medical image analysis is segmentation: separating an image into individual tissue regions. This task is frequently performed (e.g. [1, 2]) by generating a model of the intensity distribution of the data, containing terms for each pure tissue and often terms for partial volume contributions (where a single voxel contains more than one pure tissue), and fitting this to the intensity histogram, using for example an Expectation-Maximisation (EM) algorithm [3]. However, this approach discards spatial information. Some researchers have investigated methods for recovering this information e.g. histograms built from local image phase [4] or Gaussian scale-space derivatives [5] have both been investigated in the context of image registration. In this work we propose using spatial maps of tissue locations to aid in the estimation of unstable parameters via a process of sub-sampling. The approach is similar to the use of spatial Bayesian priors, but bootstraps the maps from the data to avoid bias.

The segmentation algorithm used as a basis for the work described here has been presented previously at MIUA [6–10], and so we provide only a brief overview. The algorithm is based on a model of the intensity distribution for each tissue, consisting of pure tissue components represented by Gaussian distributions, and partial volume components (assumed uniform between pure tissue) represented for computational reasons by Gaussian distributions convolved with triangular distributions. Each pure tissue component contains parameters for the mean, the standard deviation, and a prior term representing frequency of occurrence, giving a total of four parameters. Unlike more conventional mixture models, the partial volume distributions are constrained by the parameters of the pure tissues they contain, which determine the mean and standard deviation, and therefore only contain only one additional free parameter, the frequency of occurrence. Partial volume can be set to zero for combinations of tissues that are known a-prior not to occur. This model is fitted to the intensity histogram of the data using the Expectation-Maximisation (EM) algorithm, providing a map of the volume of each tissue within each voxel.

Tissues tend to occur in contiguous blocks and for some sub-regions of the image pure tissue peaks can be easily identified. In principle the model we require is that which would provide a good fit to any sub-region of the image following estimation of the tissue normalisation parameters. Fitting a single combined histogram will not enforce this constraint, and the amount of information lost may be significant. We could attempt multiple simultaneous fits of disjoint regions, with common distribution parameters. Sampling would need to be done without replacement in order to avoid statistical over-fitting, and we can imagine that there will be some subsets which will maximally constrain some parameters. A simple experiment is provided here to illustrate the problem of regional contamination on the density fitting process. Logically, we would predict that if it were possible select a-priori all of the voxels containing one tissue this would provide the most accurate estimate of its parameters. For practical exploitation of this observation, we take a bootstrap approach, the aim being to generate sub-sampled data which consists primarily of one tissue type, providing a clearer peak to fit than is present in the histogram of the data as a whole. We believe that this approach is justifiable within the context of an EM fitting algorithm, as changing the sample selected for the estimation of parameters should not introduce bias, provided the model is still valid for the sub-sample. The standard proof of convergence for the EM algorithm therefore extends to scenarios in which parameters are iteratively updated individually, but for a set of separate likelihood functions defined according to their own data sub-set rather than a likelihood defined for the entire data set. It is the theoretical freedom this provides, together with the regional coherence of data, which we exploit for improved parameter stability.

*E-mail: paul.bromiley@man.ac.uk



(a) (b) (c)

Figure 1. The synthetic data (a) with 10% noise, histogram at 10% noise (b) and histogram at 100% noise.

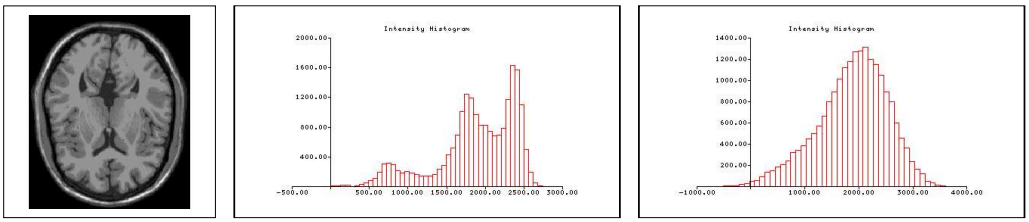
In order for sub-sampled regions to be suitable for parameter estimation they must satisfy the sampling requirements of the original density model. We can not therefore simply use a threshold on tissue probability to select data as this would impose hard cut-offs on the intensity histograms, which would not be described by the model. Instead we degrade an initial estimate of the tissue boundaries by blurring and quantisation in order to de-correlate the process of sampling with the grey-level values. The sample of a given tissue is therefore enhanced and close to a pure tissue sample, but the overall histogram distribution still conforms to the assumed model. The parameters of each pure tissue obtained from their individual sub-samples of data are then recombined to produce a complete tissue model, and the process iterated until convergence. Results from both Brainweb simulated MR images and real MR data, presented below, indicate that significant improvements in parameter stability can be gained through incorporating spatial information in this way.

2 Method

The mathematical formulation of the algorithm, and the parameter update equations, can be found in our previous publications [6–10]. The algorithm is capable of segmenting multi-dimensional data, and also of incorporating gradient information into the intensity histogram and model, although neither of these features were used here. In this work, we investigate the inclusion of spatial information into the above framework. At each step of the EM optimisation, the current tissue model can be used to generate tissue maps showing the volume of each pure tissue within each voxel. These can be binarised at the 50% level to produce maps of the spatial locations of each tissue type. These masks are blurred slightly by expanding each pixel to a 3x3 region, on the assumption that the main axes of the image do not align with any tissue boundary over the whole image plane and that over distances of a few pixels the tissue can switch entirely from one class to another. These masks are then used to sub-sample the original data set; tissue frequencies are maximised on each sub-sample to take account of the changing amount of data, and an expectation step is performed. The mean and standard deviation of each pure tissue is optimised from its respective sub-sample of the data. The parameters of each tissue are then recombined into a complete model, and estimation of the tissue frequencies is performed again prior to the next expectation step. The process is iterated until convergence.

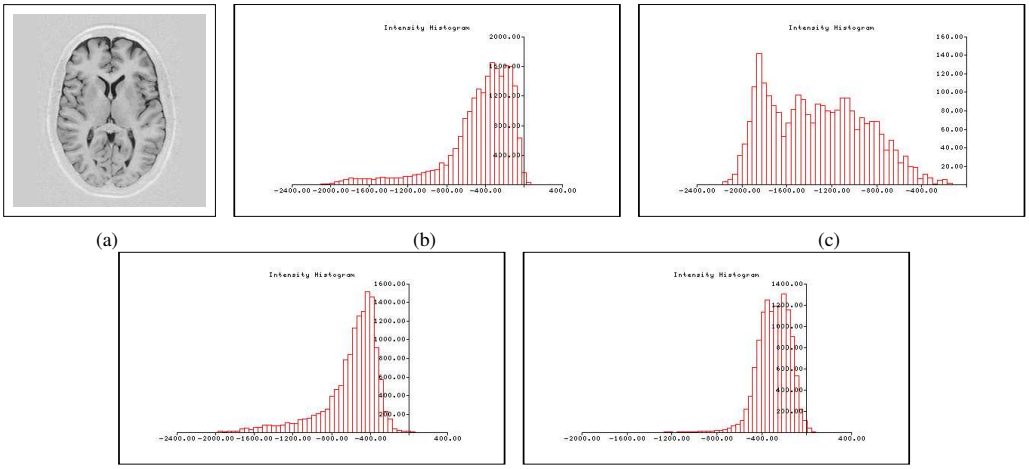
Three sets of experiments were performed in order to determine whether the incorporation of spatial information in the above framework improves the accuracy of the segmentation. In each set of experiments, a consistent starting point for the EM optimisation was enforced by using a hand-written initial tissue model, prepared through visual examination of the image histograms. 40 iterations of EM optimisation were applied in all segmentations; this was determined to be sufficient to ensure convergence to within 1% in all experiments.

In the first set of experiments, the synthetic data set shown in Fig. 1 was used, consisting of two blocks of uniform intensity, each 128 pixels wide by 64 pixels high, and having intensities of 100 and 200 grey-levels respectively (the darker region will be referred to as tissue 1, the brighter as tissue 2). 1000-iteration Monte-Carlo experiments were performed, in which independent Gaussian noise fields were added to the data prior to each iteration, and EM segmentation was performed, optimising only the parameters of tissue 1, holding the parameters of tissue 2 constant. Experiments were performed at each of ten levels of added noise, varying from $\sigma = 10$ to $\sigma = 100$ (i.e. from 10% to 100% of the separation in intensity between the two tissues. Sample histograms at the extremal noise levels are shown in Fig. 1. This procedure was then repeated a further 16 times, removing 8 columns from the right-hand side of the image each time, in order to simulate the effect of spatial sub-sampling of the data. Note that in the final experiment, the width of the tissue 2 region was reduced to 1 column rather than 0, to allow the use of the same, two-tissue, initial model. These experiments were intended as a form of calibration: they show the change in the accuracy of the optimised parameters for tissue 1 gained through reducing the contamination of the intensity histogram with tissue 2, for a range of different noise levels or, equivalently, separations between the mean intensities of the tissues. Note, however, that application of the spatial sub-sampling to real data would involve simultaneously masking-out and fitting each tissue.



(a) (b) (c)

Figure 2. The Brainweb data (a), histogram at 2% noise (b) and histogram at 10% noise.



(a) (b) (c) (d) (e)

Figure 3. An example of one of the real MR data sets (a), its intensity histogram (b), and the spatially sub-sampled histograms for CSF (c), GM (d) and WM (e).

This experiment does not simulate the simultaneous optimisation of the parameters of tissue 2, which would introduce errors into the combined process, and thus under-estimates the errors on the tissue 1 parameters and also therfore, the gains achieved through application of the spatial sub-sampling to real MR data.

Simulated MR images obtained from Brainweb [11] were used in the second set of experiments. A simulated T1 volume with 0% inhomogeneity and 0% image noise, consisting of 1x1x1 mm voxels, was obtained. In order to reduce the processor time required, a single slice was selected from the data set, as shown in Fig. 2, and skull-stripping was applied using masks prepared from the Brainweb tissue phantoms, such that the segmentation was limited only to the cerebro-spinal fluid (CSF), white matter (WM) and grey-matter (GM). 100 iteration Monte-Carlo experiments were then performed, in which independent Gaussian noise fields were added to the data prior to each iteration, and EM segmentation was performed both with and without spatial sub-sampling of the data. Five such experiments were performed, with the standard deviation varying from 2-10% of the intensity of the brightest tissue, in steps of 2%. Coincidentally, these values are also approximately 20-100% of the separation between the GM and WM peaks. Sample histograms are shown in Fig. 2.

The final set of experiments used real IRTSE MR scans from eight young (34-46 years, mean age 40) normal subjects, each consisting of 1x1x3 mm voxels. For reasons of speed, single structure-rich slices from approximately the same anatomical region were selected from each volume, and segmentation was limited to the CSF, GM and WM (located using hand-drawn binary masks). A sample slice and its intensity histogram are shown in Fig. 3. 100 iteration Monte-Carlo experiments were then performed, in which independent Gaussian noise fields with standard deviations equal to the that of the intrinsic image noise (estimated from the width of zero-crossings in horizontal and vertical gradient histograms) were added to the data prior to each iteration, and EM segmentation was performed both with and without spatial sub-sampling of the data. Fig. 3 shows example histograms produced by the spatial sub-sampling of the CSF, GM and WM at the start of the EM optimisation (i.e. prior to any optimisation of the hand-written model parameters).

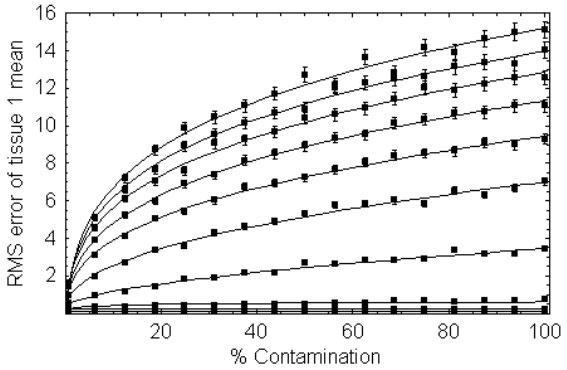
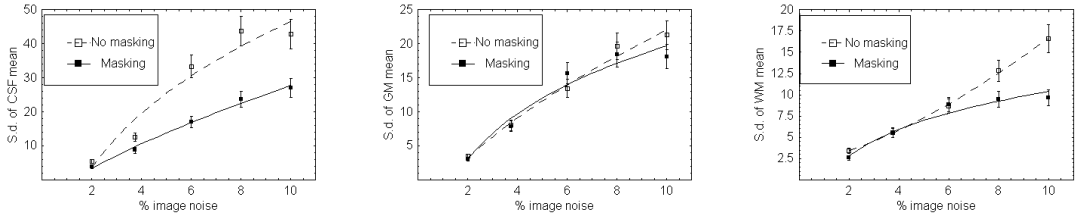
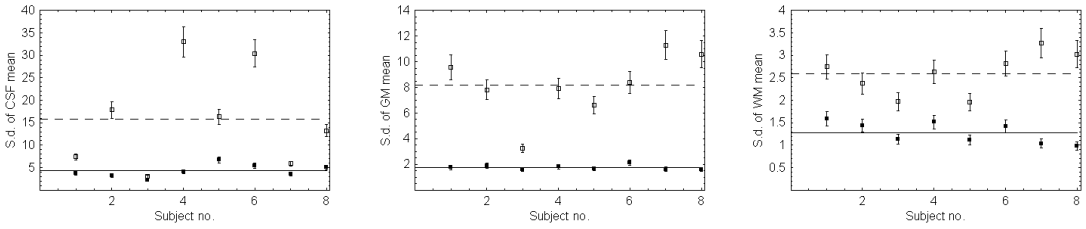


Figure 4. The RMS error on the tissue 1 mean in the experiments on synthetic data, plotted against the level of contamination of the intensity histogram with tissue 2. Power-law fits are shown. The lines are for 10-100% image noise, in steps of 10%, reading upwards from the bottom of the graph.



(a) (b) (c)

Figure 5. The standard deviations of the CSF (a), GM (b) and WM (c) mean intensities in a 100 iteration Monte-Carlo experiment on the Brainweb data, plotted against the total image noise (as a percentage of the dynamic range), for segmentations with and without the spatial sub-sampling of the data.



(a) (b) (c)

Figure 6. The standard deviations of the CSF (a), GM (b) and WM (c) mean intensities in a 100 iteration Monte-Carlo experiment on the real data, plotted against the subject no., for segmentations with and without the spatial sub-sampling of the data. The lines show the means across subjects, and the symbology is the same as in Fig. 5.

3 Results

Fig. 4 shows the RMS error on the estimated mean intensity of tissue 1 in the experiments on simulated data. When the noise level is low, and so the tissues are well separated in the histogram, sub-sampling the data has no effect; this indicates that, at least on this example, the sub-sampling procedure does not introduce bias. However, at higher noise levels the spatial sub-sampling begins to have a marked effect: significant improvements in segmentation accuracy are gained. The results from the Brainweb data (Fig. 5) are broadly consistent; as the noise level is increased and the tissue intensities begin to overlap, sub-sampling significantly improves the stability of the segmentation (note the Brainweb does not provide mean intensity values for the simulated tissues, preventing any absolute tests of accuracy beyond e.g. simply measuring the volumes of segmented tissues or the degree of overlap of the segmentations with the phantoms). The results from the real data (Fig. 6) are significantly better than those obtained with the Brainweb data: whilst absolute segmentation accuracy cannot be measured without a gold-standard, parameter stability is improved by a factor of 2-4 through the application of spatial sub-sampling. The effect is greatest in the GM, since this is the most obscured tissue in the intensity histogram, forming a shoulder on the low-intensity side of the WM. Similarly, the

effect is least marked in the WM, since this tissue forms a distinct peak in the histogram. The marked improvement of the sub-sampling performance on the real compared to the Brainweb data, particularly for the GM, may be due to sampling problems in the Brainweb simulation: spurious peaks were noted in the partial volume distributions in the histogram of the noise-free data, which are not consistent with typical partial volume distributions in real MR data.

4 Conclusion

Three main conclusions can be drawn from the results presented here. First, in situations where the intensity distributions of pure tissues overlap significantly, significant improvements in segmentation accuracy can be gained through recovering spatial information using the procedure described here. It should be noted that the initial tissue model, and thus the masks used to sub-sample the data, need not be highly accurate: hand-written models were used in the experiments, prepared simply from visual inspection of the intensity histograms.

Second, the spatial sub-sampling procedure discards large proportions of the data in the estimation of the tissue model parameters for each pure tissue, and furthermore the data samples change as the optimisation proceeds. Therefore, the results presented here demonstrate that the data sample can vary between the iterations of an EM optimisation without introducing biases or preventing convergence, as long as the sub-sampling procedure does not affect the goodness-of-fit of the model. This concept was adopted in our original segmentation algorithm, on which the current work is based, through the procedure of discarding the partial volume data in the parameter update step in order to reduce the sensitivity to the partial volume model used. The current work, using a much more aggressive sub-sampling of the data, demonstrates the general validity of this concept more fully. Furthermore, whilst not explored here, the potential computational benefits of sub-sampling the data and isolating specific model components for optimisation are clear.

Finally, it has been noted by many medical image analysis researchers that three of the main problems in the field, namely segmentation, registration, and shape modelling, can be viewed as three facets of an overall, unified problem, in that the results from any one of them can be used to perform the other two. The work presented here can be viewed as a first step towards the unification of segmentation and shape modelling: effectively it uses a crude, non-parametric shape model to sub-sample the data for segmentation. The improvements to segmentation accuracy gained indicate that further research in this direction could prove fruitful.

Acknowledgements

This work was supported by the the MIAS (Medical Images and Signals) IRC under EPSRC grant no. GR/N14248/01 and the MRC grant no. D2025/31. The software used in this study is freely available from www.tina-vision.net.

References

1. P. Santago & H. D. Gage. "Statistical models of partial volume effect." *IEEE Trans. Med. Imag.* **4**, pp. 1531–1540, 1995.
2. D. H. Laidlaw, K. W. Fleischer & A. H. Barr. "Partial-volume Bayesian classification of material mixtures in MR volume data using voxel histograms." *IEEE Trans. Med. Imag.* **17(1)**, pp. 74–86, 1998.
3. A. P. Dempster, N. M. Laird & D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Society* **39**, pp. 1–38, 1977.
4. M. Mellor & J. M. Brady. "Non-rigid multimodal image registration using local phase." In *Proceedings MICCAI'04*, pp. 789–796. 2004.
5. M. Holden, L. D. Griffin, N. Saeed et al. "Multi-channel mutual information using scale space." In *Proceedings MICCAI'04*, pp. 797–804. 2004.
6. D. C. Williamson, N. A. Thacker, S. R. Williams et al. "Partial volume tissue segmentation using grey-level gradient." In *Proc. MIUA*, pp. 17–20. 2002.
7. P. A. Bromiley, N. A. Thacker, M. L. J. Scott et al. "Bayesian and non-Bayesian probabilistic models for medical image analysis." *Image and Vision Comput.* **21(10)**, pp. 851–864, 2003.
8. M. Pokrić, N. A. Thacker, M. L. J. Scott et al. "Multi-dimensional medical image segmentation with partial voluming." In *Proc. MIUA*, pp. 77–80. 2001.
9. M. Pokrić, N. A. Thacker & A. Jackson. "The importance of partial voluming in multi-dimensional medical image segmentation." In *Proc. MICCAI*, pp. 1293–1294. 2001.
10. N. A. Thacker, M. Pokrić & D. C. Williamson. "Noise filtering and testing illustrated using a multi-dimensional partial volume model of MR data." In *Proc. BMVC*, pp. 909–919. Kingston, London, 2004.
11. C. A. Cocosco, V. Kollokian, R. K.-S. Kwan et al. "Brainweb: Online interface to a 3D MRI simulated brain database." *NeuroImage* **5(4)**, pp. S425, 1997.

Regional Left Ventricular Wall Motion Analysis using Hidden Markov Models

Sarina Mansor, J Alison Noble, Nicholas P Hughes

Wolfson Medical Vision Laboratory, Department of Engineering Science, University of Oxford, UK.

Abstract. In this paper, we examine the use of a hidden Markov model for regional heart function assessment based on stress echocardiography. We developed models for normal and abnormal hearts and performed the experiments on rest, stress and rest-to-stress sequences. We achieved a good accuracy of classification, particularly for the normal data but with also promising results for the abnormal case.

1 Introduction

Analysis of left ventricular (LV) regional function is important for diagnosing heart disease, especially to detect myocardial ischemia (a disease of heart muscle). In current clinical practice, the analysis mostly relies on visual assessment by experienced cardiologists. This makes the diagnosis of regional heart function disease a highly subjective and operator-dependent problem. While much has been done to attempt to automate the task of wall motion analysis on rest image sequences, there is a very small literature on analysis of stress echocardiography. Stress sequences (where the heart has been stressed by exercise or use of a pharmacological drug e.g. dobutamine) are more difficult to automatically analyse as the non-rigid motion of the heart is more challenging to track. Further, to our knowledge there has been no prior attempt to automatically classify heart motion based on the *combined information* derived from a rest and a stress sequence. This is the subject of this paper.

The first step of automated regional heart function analysis is to detect the heart wall borders. Many techniques have been developed to detect and automatically track both the interior (endocardial) and exterior (epicardial) borders of the LV. Current automated 2D echocardiography image tracking technology is now sufficiently well-developed for application on good-to-medium quality rest data [1]. However, further studies need to be done to show that this is also true for stress echocardiography. As a result, in the work reported here, we used manually derived contours as our initial focus has been on classification not automated tracking.

The second step of automated regional heart function analysis is classification of heart segmental function as either normal or abnormal based on the extracted contours (heart wall borders). This is the focus of this paper. Prior work on automated classification has looked, for example, at average motion of wall segments or using ICA to identify abnormal motion. In the literature we are only aware of work looking individually at either rest or alternatively stress data but not the combination of both [2].

This contribution investigates the use of a Hidden Markov Model (HMM) as a tool for stress echocardiography classification. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech [3], handwriting [4] and bioinformatics [5] because of their ability to successfully learn the time-varying characteristics of signals. Therefore, we would like to employ an HMM approach since the cardiac data inherits the time-varying and sequential properties. To the best of our knowledge, there is no other published work incorporating HMMs with regional heart function analysis. The most closely related work is the use of a HMM for shape analysis [6]-[8], particularly [7] which implemented a HMM for spatio-temporal pattern recognition.

2 Hidden Markov Models (HMMs)

A detailed tutorial on HMMs can be found in [3] and a brief description based on that paper is given here. A hidden Markov model (HMM) is a probabilistic model which describes the statistical relationship between an observable sequence *O* and a "hidden" state sequence *S*. The hidden state is discrete, but the observation values may be either continuous or discrete in nature [5].

An HMM is characterized by the following parameters:

1) The number of states of the model, N.
2) The state transition matrix, A
3) The observation probability distribution, $B = \{b_N\}$ for each state N
4) The initial state distribution, π

However, for convenience, the HMM λ is parameterized by A, B and π, with the notation: $\lambda = (A, B, \pi)$.

There are three basic problems associated with HMMs:

1. The classification/evaluation problem

   Given a model λ and a sequence of observations O, we would like to compute the probability that the observations are generated by the model, $p(O|\lambda)$. This problem can be solved by a 'forward-backward' procedure [3].

2. The decoding problem

   Given a model λ and a sequence of observations O, we would like to find the most likely sequence of hidden states that could have generated the observations. This problem is typically solved by the Viterbi algorithm [3].

3. The learning/training problem

   Given a set of observation sequences, we would like to adjust the model parameters, $\lambda = (A, B, \pi)$ to maximize the probability of the given dataset. This problem is typically solved by the EM algorithm [3].

In our work we deal with the first and third problems. A regional heart function model needs to be learned from a training set by the EM algorithm before using the forward-backward procedure for classification of a new dataset.

3 Methodology

3.1 Patient Data

The available database consisted of 20 studies of Contrast DSE (Dobutamine Stress Echocardiography) data acquired by two cardiologists as standard B-mode ultrasound image sequences. Each study contained:

- 4 planes: 2-chamber, 3-chamber, 4-chamber and short-axis views.
- 3 stress stages: rest (no dobutamine), intermediate (low-dose dobutamine) and peak (the maximum dose that a patient can take).

In this paper, we utilized the 2-chamber view (2CV) data at the resting phase and at the peak level of stress. Each data was recorded over a cardiac cycle, starting from the end of diastolic (ED) phase to the end-of-systolic (ES) and back to the ED phase.

Clinical regional wall functional assessment evaluates systolic thickening and endocardial wall motion. A normal (healthy) myocardial segment shows systolic thickening[1] and also endocardial movement towards the centre of the cavity. The clinical assessment describes each segment as being either normokinetic (healthy), hypokinetic (reduced contractility), akinetic (dead) or dyskinetic (moving outward).

In our work, the 17 segment model was used for wall scoring classification [9] and therefore the 2CV data had scoring reported on six segments. We selected 10 normal (healthy) cases and 10 cases identified as having regional function abnormalities as determined by wall scoring analysis. Epicardial and endocardial contours were manually traced by one of the 2 cardiologists in each case on each frame of the cardiac cycle.

3.2 Feature Vector / Observation Representation

Figure 1 shows the change of the myocardium thickness over a cardiac cycle for normal (top row) and abnormal (bottom row) patients. As we can see from Figure 1, the myocardial area increases in each segment from ED to ES for a normal heart. In contrary, there is not much heart wall contraction for an abnormal one. This is known as a wall thickening abnormality, which may be caused by an injured myocardium.

For this reason, for individual rest or stress sequence analysis (cases 1 and 2 respectively in our study) we used the six-segment myocardial area for each frame within a cardiac cycle as the feature/observation vector. This is computed as the area between the epicardial and the endocardial border.

We used a different feature vector for analysing combined rest-and-stress sequences (case 3). In this case we want to model the functional change of the heart from rest to stress. Before defining our choice, we first consider temporal alignment of rest and stress data.

There can be a large difference in heart rate between the rest and stress sequences, as the heart rate increases during stress testing. Consequently, the number of frames in a stress sequence is always smaller than that in the rest sequence since the rate of an ultrasound acquisition stays fixed. Temporal alignment needs to be done to generate the same number of frames for both rest and stress sequences. We used a non-linear interpolation method to do this as follows.

[1] The myocardial wall thickening during systole (the contraction of the heart) phase

The stress phases were manually matched to equivalently rest phases (refer to Figure 2), and the missing stress frames were then linearly[2] interpolated in between. After temporal alignment, the area between the two endocardial contours was extracted as the feature/observation vector. We chose this as our feature vector following discussions with our clinical collaborators who argued that endocardial motion is the strongest single cue that they use in visual stress echocardiography assessment.

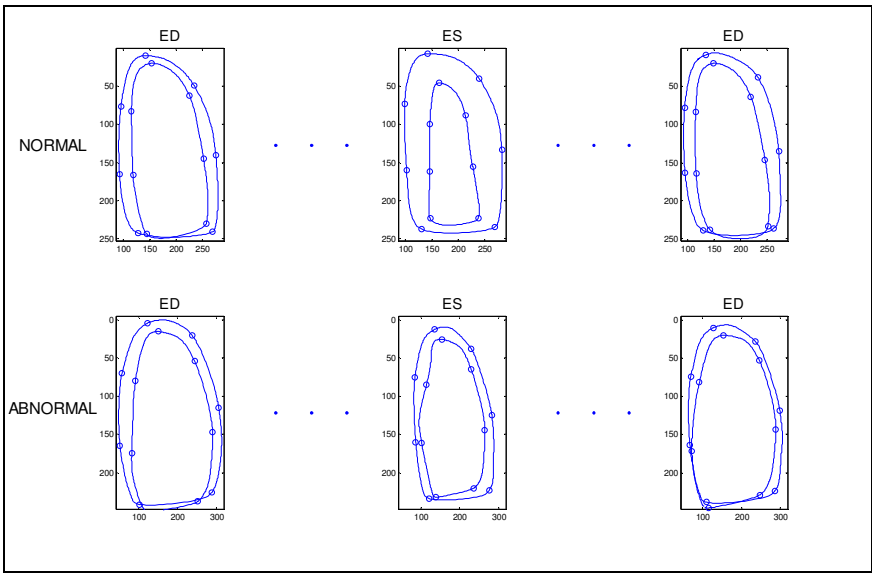


Figure 1: Epicardial and endocardial contours of LVat 2CV, showing the changes of the myocardium area over a cardiac cycle of normal (top row) and abnormal (bottom row) patients.

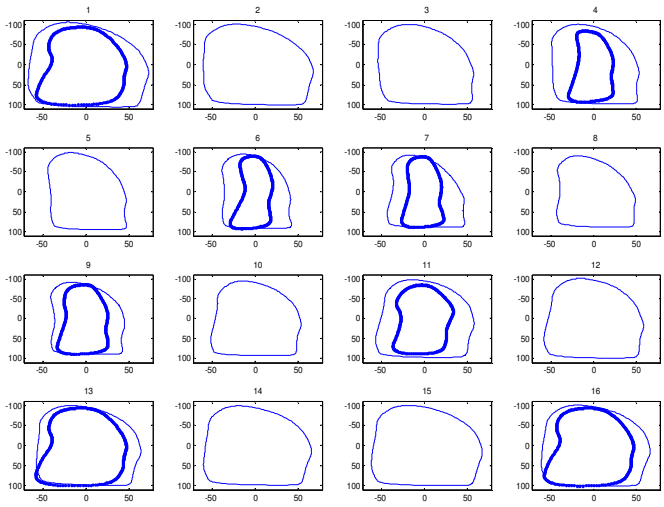


Figure 2: Phase matching between rest and stress sequences for the temporal interpolation (the thin and thick lines represent the endocardial contours of rest and stress data respectively)

[2] Non-linear interpolation might lead to better results in some cases and this will be looked at in future work.

3.3 Training

The sequential nature of the heart phases can be well modeled by a left-right HMM. We developed two models: one for normals and another for abnormals. Each model is comprised of three hidden states. In our case, the states of the HMM do not have a true physical meaning – they simply reflect common statistical properties of the observation vectors in the feature space. We could have employed more states but that would require more training samples for an accurate estimation of the model parameters [5].

We trained the two HMMs (one for normals, one for abnormals) in an *unsupervised* manner. Prior to training, the datasets (for normals and abnormals, respectively) were resampled to have 30 frames per patient. We then extracted the six-segment myocardial area feature vector for each frame, as previously described. Each dataset therefore consisted of 10 patients, with 30 six-dimensional feature vectors per patient.

In the first stage of training, we initialised the model parameters as follows. For each patient in the dataset, the first ten frames were assigned to state 1, the next ten to state 2 and the final ten to state 3. The parameters of the transition matrix a_{ij} were then initialised using the maximum likelihood estimates, based on the following equation (1)

$$a_{ij} = n_{ij} \Big/ \sum_k n_{ik} \qquad (1)$$

where n_{ij} is the total number of transitions from state i to state j over all of the label sequences.

The observation probability densities b_i (which were Gaussian) were initialised by fitting a multivariate Gaussian to the set of six dimensional features extracted from the frames assigned to each state i.

Following initialisation of the HMM parameters, we then ran the EM algorithm to update the model parameters until the change in the log likelihood was less than 0.01% (indicating convergence).

4 Results and Discussion

The new approach was evaluated using leave-one-out cross-validation. Recall that we have a total of 20 sequences, 10 for each model. In a leave-one-out cross-validation one of the 10 sequences for the model being tested is held out and the remaining 9 sequences are used to derive the HMM. The held-out sequence is then tested on this HMM and the other fully trained HMM. This procedure is repeated leaving out each sequence consecutively on its own. Classification results are presented using a confusion matrix. Tables 1 to 3 show the results for the rest, stress and 'rest-to-stress' cases. The diagonal entries (in bold) indicate the accuracy of each model.

Table 1. Percentage Confusion Matrix for 'rest' sequences (feature vector is myocardium area)

	Normal	Abnormal
Normal	**80%**	20%
Abnormal	40%	**60%**

Table 2. Percentage Confusion Matrix for 'stress' sequences (feature vector is myocardium area)

	Normal	Abnormal
Normal	**60%**	40%
Abnormal	50%	**50%**

Table 3. Percentage Confusion Matrix for 'rest-to-stress' sequences (feature vector is the area between the two endocardial contours)

	Normal	Abnormal
Normal	**70%**	30%
Abnormal	40%	**60%**

In each experiment, the classification accuracy for the normal model is higher than for the abnormal case. This is because the wall motion of healthy hearts is better-defined than for the abnormal case which really covers subcategories of abnormal functional behaviour i.e. hypokinetic, akinetic, dyskinetic. Since we had a limited number of patient datasets for the current study, we grouped all the abnormal cases into just one model. For a more accurate classification, they should be modelled according to each abnormal category.

Comparing experiments, the stress sequence has the lowest accuracy. This may be explained in part by the relatively large movement of the heart during stress. Furthermore, the image quality is often poor for this type of acquisition which affects the accuracy in tracing the myocardial borders. The analysis of rest-to-stress sequence had higher accuracy than that of the stress sequence which is encouraging evidence to further pursue the combined analysis of rest-and-stress data.

5 Conclusions

In this paper, a new approach to regional functional analysis of stress echocardiography data has been proposed based on HMMs. Two HMMs have been trained for the normal (healthy) and abnormal cases. A good accuracy of classification was achieved despite the limited number of data used. These encouraging results could be improved further by: 1) increasing the number of training data, for example at least 50 patients for each model, 2) developing different models for different types of abnormality and 3) investigating other possible observation vectors. The general approach would be equally applicable to 3D stress echocardiography which is receiving a lot of interest from the cardiology community at the current time. Finally, the methodology is not modality specific and could be equally applied to stress MRI.

References

1. J.A. Noble **"**Cardiology meets Image Analysis: just another application or can image analysis usefully influence clinical practice?", *IEEE International Workshop on Computer Vision in Biomedical Image Analysis, China*, 2005.
2. A. Suinesiaputra, A.F. Frangi, et al, "Automatic Prediction of Myocardial Contractility Improvement in Stress MRI Using Shape Morphometrics with Independent Component Analysis", *Information Processing and Medical Imaging*, 2005.
3. L. Rabiner "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of IEEE* **77**, 2, pp. 257-286, 1989.
4. T. Artieres, S. Marukatat & P. Gallinari "Online Handwritten Shape Recognition Using Segmental Hidden Markov Models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 29, Issue 2, pp. 205 – 217, 2007.
5. N.P. Hughes, L. Tarassenko & S. J. Roberts "Markov Models for Automated ECG Interval Analysis", *Advances in Neural Information Processing Systems* **16**, 2003.
6. Y. He & A. Kundu, "2-D Shape Classification Using Hidden Markov Model". *IEEE Trans. Pattern Analysis and Machine Intelligence* **13**, 11, pp. 1172 - 1184 , 1991.
7. K.H Fielding & D.W Ruck "Spatio-Temporal Pattern Recognition Using Hidden Markov Models", *IEEE Trans. Aerospace and Electronic Systems* **31**, 4, pp. 1292-1300, 1995.
8. M. Bicegon & V. Murino "Investigating Hidden Markov Models' Capabilities in 2D Shape Classification" *IEEE Pattern Analysis Machine Intelligence* **26**, 2, pp. 281-286, 2004.
9. M.D Cerquiera, N.J. Weismann, V. Dilsizian, V., et al. "Standardized myocardial segmentation and nomenclature for tomographic imgaging of the heart", *Circulation* **105**, 4, pp. 539-542, 2002.

Initial Results of an Application of Multiscale Line Operators for Blood-Vessel Segmentation in Digital Fundus Photographs

D.J.J. Farnell[1], F.N. Hatfield[2], P.C. Knox[3], M. Reakes[4], and S.P. Harding[5]

[1]Academic Radiation Oncology, Division of Cancer Studies, Faculty of Medical and Human Science, University of Manchester, c/o Christie Hospital NHS Trust M20 4BX, Manchester, United Kingdom
[2]School of Health Sciences, University of Liverpool, Liverpool L69 3GB, United Kingdom
[3]Division of Orthoptics, School of Health Sciences, University of Liverpool, Liverpool L69 3GB, United Kingdom
[4]Department of Physics, University of York, York YO10 5DD, United Kingdom
[5]St. Paul's Eye Unit, Royal Liverpool University Hospital Trust, Liverpool L7 8XP, United Kingdom

Abstract We employed a multiscale line operator (MSLO) approach in order to segment blood vessels in digital fundus images. Separately, a median filter technique was used in order to provide results that were compared to those of the MSLO. Both sets of results were further enhanced by subsequently employing a simple "randomly seeded" region-growing algorithm. We applied this approach to the images in the STARE retinal image archive (www.ces.clemson.edu/~ahoover/stare/). The STARE dataset contained both healthy (10 images) and diseased eyes (10 images). We determined receiver-operator characteristic (ROC) curves, and the areas under the curve (AZ) were obtained. A large increase in efficiency for our MSLO algorithm was observed for the entire dataset (AZ=0.953) compared to basic thresholding (AZ=0.753). Interestingly, the median filtering algorithm followed by region growing also performed well (AZ=0.947). The best results were found for the healthy subjects. Our results were found to compare very favourably to those results of previous segmentation procedures for the STARE dataset.

1 Introduction

Methods of blood-vessel segmentation in medical images [1] range from pattern recognition (multiscale approaches, skeletons, region growing, ridge-based approaches, blood-vessel tracking, matched filters, differential geometry, and mathematical morphology), model-based approaches (e.g. active snakes), tracking systems, to artificial intelligence and neural networks. Many of these approaches utilise image information at a variety of length scales. The multiscale line operator [2-4] (MSLO) has been used with great success in detecting linear structures in mammograms. In the context of blood-vessel segmentation in retinal images, one might expect it to work well because of the approximately linear nature of blood vessels locally. This method has the additional advantage of being conceptually straightforward. The MSLO algorithm is fast and is also relatively straightforward to implement. A number of algorithms have been used previously to carry out retinal blood-vessel segmentations [5-19]. For example, these methods include adaptive thresholding [6] and matched filter approaches that employ Gaussian profiles [8-12]. Subsequent adaptive threshold probing of a matched filter response has also been used [13] in order to track blood vessels with much success. Mathematical morphology and curvature estimation [14,15] and ridge-based vessel segmentation procedures [7,15] have also been used to segment blood vessels in fundus images. Model-based [16] methods and a scheme based on fuzzy logic [17] have been used for vessel segmentation. Indeed, many such vessel segmentation techniques often use image information at a variety of length scales (see, e.g., Refs. [5,18,19]).

2 Method

2.1 The STARE Retinal Image Archive

The STARE dataset is an archive of fundus images that is publicly available (see www.ces.clemson.edu/~ahoover/stare/). We use the same dataset as in previous treatments [5-7] that contains 20 images with healthy, diabetic, and AMD image features. According to information provided by the above website, 10 of these images were classified as normal and 10 as abnormal. The images were in colour and were 700×605 pixels in size. Two independent observers created manual tracings of the blood vessels using a graphical user interface, and these tracings were also downloaded from the STARE website. As in previous studies [7], we use the manual tracings of the first such observer as our "gold standard" for the STARE dataset. The images in the STARE dataset were used to test our segmentation algorithms only. The green channel of the fundus images is used in order to find the blood vessels.

2.2 Receiver-Operator Characteristic (ROC) Analysis

In order to quantify the accuracy of our results we performed an ROC analysis. We were able to determine the number of pixels classified as "positive" by our automated routines that were, in fact, either "true" or "false" compared to a "gold standard" provided by manual segmentations of the blood vessels. The entire vessel width was used here. We were able to find the true positive (sensitivity) and false positive (1-specificity) fractions. Hence, we

were able to perform an ROC analysis by plotting the true positive fraction on the ordinate axis against the false positive fraction on the abscissa. The area under the curve (AZ) was also determined.

2.3 Median Filter

Digital fundus photographs obtained using flash photography often exhibit unequal levels of illumination across the image. A simple and common way of countering this effect adaptively at each pixel location is to find the mean or median greyscale value over some locale and then use this value as our threshold. This is called adaptive thresholding. Here, we divided the greyscale value of a central pixel by the median value of the greyscale values of pixels in a square locale of size $L \times L$. A global threshold was then applied to the entire image in order to obtain an initial segmentation. Thus, by varying this threshold, we were able to form a succession of such binary images. The true positive and false positive fractions were obtained for each value of the threshold and thus an ROC curve could be formed. The results of this procedure shall be referred to here as those of "median filtering."

2.4 Multiscale Line Operator (MSLO)

We also implemented the MSLO [2-4] in order to enhance the retinal blood vessels in the image. For each pixel in the original image at point (i,j), the average value, P, of the greyscale values of a line of pixels centred on (i,j) is obtained (Figure 1). However, we also determine the average greyscale value, Q, of those pixels in a rectangular area surrounding this line (Figure 1). The response, S, of the line operator at point (i,j) at a given orientation of the line (and scale) is given by, $S=P-Q$. In order to keep our treatment fully consistent with those calculations outlined in Refs. [2-4], we used a line operator of length 5 pixels obtained at 12 orientations. The largest value of S for all 12 orientations is taken to be the value of the line operator at that pixel. The subtraction of the mean greyscale value of local environment, Q, from the mean value for the line, P, ensures that changes in background illumination should also be removed in the filtered image.

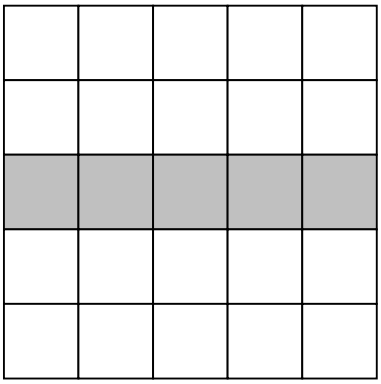


Figure 1: The line operator is applied by determining the difference, S, in mean greyscale values between the shaded squares and the unshaded squares. The result for the central pixel is the largest value of S over 12 orientations of the line.

The line operator is applied at varying levels of scale by constructing a "pyramid of images" at consecutively coarser scales via Gaussian sampling. The number of levels of the Gaussian pyramid was taken to be an explicit variable here. The final result for the MSLO was the sum of all line operator filtered images in the Gaussian pyramid. Equal weighting was given to each length scale in the final MSLO image. A threshold was then applied to the final image in order to provide an initial segmentation of the blood vessels. The line operator therefore enhances linear features in the image. Moreover, blood vessels of varying width are, in principle, treated on an equal footing. This algorithm was speeded by saving those configurations or "masks" for the line of central pixels and its neighbours in the "rectangular environment" (used in finding P and Q, respectively) at the 12 different orientations at the start of the entire run in a look-up table. The number of operations involved in the application of the line operators was thus kept to a minimum. The interested reader is referred to Refs. [2-4] for more detailed reviews of the MSLO algorithm.

2.5 Region Growing

The initial segmentations after thresholding of the filtered images demonstrated noise. The blood vessels tended to form long extended structures whereas the noise features were generally much smaller in size. A simple way of the decreasing the number of these noise features was to use a simple region-growing technique. If the initial binary segmentation from thresholding alone is a two-dimensional array of binary numbers denoted I and the final segmentation is a binary image called J (initially containing zeros only) then the algorithm is given by:

1. Choose a "positive" pixel in *I* randomly.
2. Trace recursively all "positive" pixels that are connected contiguously to this initial starting point and save the result to a new binary image array called *K*.
3. Let *J* be the union of *J* and *K*.
4. Let *I* be *I* minus *K*. (This removes the segmented feature in *K* from *I*).
5. Repeat from step 1 *M* times.

The contiguous "positive" pixels were traced recursively in Step 2 by using a predefined routine in the interactive data language (IDL) software (www.ittvis.com). The region-growing approach worked in practice because the probability of choosing a large extended structure (i.e. assumed to be a vessel) in *I* at any point is much greater than that of choosing a smaller feature (i.e. assumed to be due to noise). Clearly, this procedure reduced the false positive fraction at the cost of some small reduction in the true positive fraction.

3 Results

A colour fundus photograph of a normal subject is shown in Figure 2a. A blood-vessel segmentation (shown in white) found after applying the MSLO with three levels to the green channel is superimposed on the original colour image. The edge of the optic disk was sometimes found to lead to false classifications. The centre of the macula in some healthy subjects was found to be particularly dark and this lead to some false classifications; again, predominantly in "normal" subjects. A colour fundus image of an "abnormal" subject with a similar blood-vessel segmentation superimposed on it is shown in Figure 2b. Lesions were found to be the main cause of false classifications in the abnormal images; the edges of the bright disease feature caused false classifications to be made, as may be observed in Fig. 2b. Haemorrhages were another general and strong source of artefacts in the vessel segmentations. Microaneurisms were also seen in some images. These appear as small reddish features and were also found to lead to falsely classified pixels.

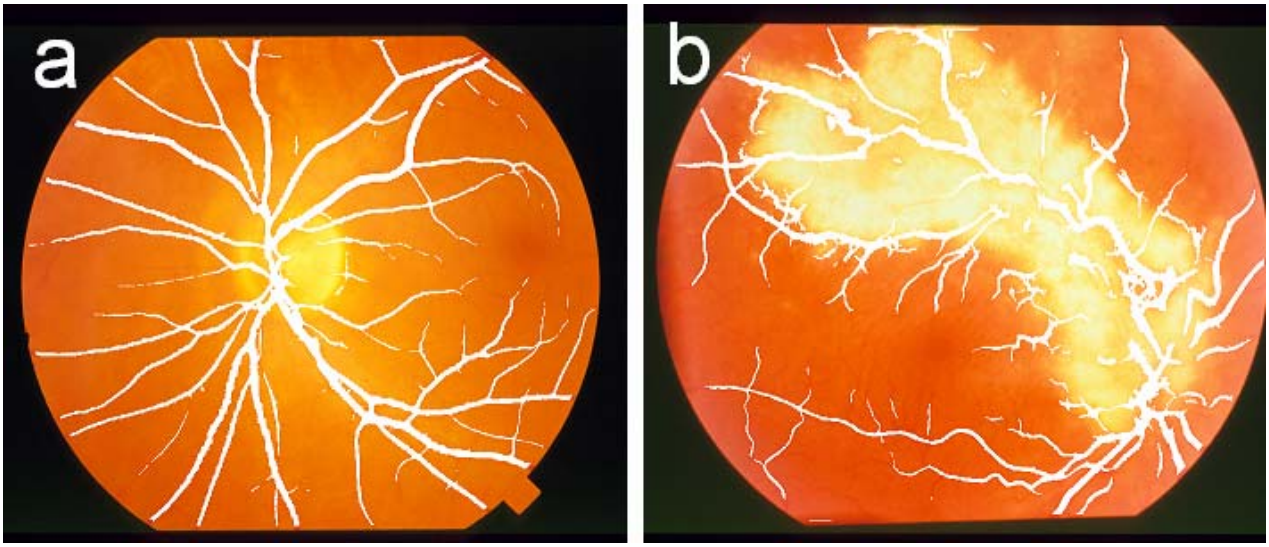


Figure 2: Extracted blood vessels (shown in white) using the MSLO superimposed on the original colour fundus image. a) Normal subject. b) Abnormal subject.

An ROC analysis was carried out for our blood-vessel segmentations compared to a "gold standard" of blood vessels traced out by a retinal image interpretation expert. The ROC figures for the entire STARE dataset are shown in Figure 3. The results of the MSLO were found to lie above those of basic thresholding. Median filtering also appeared to perform well.

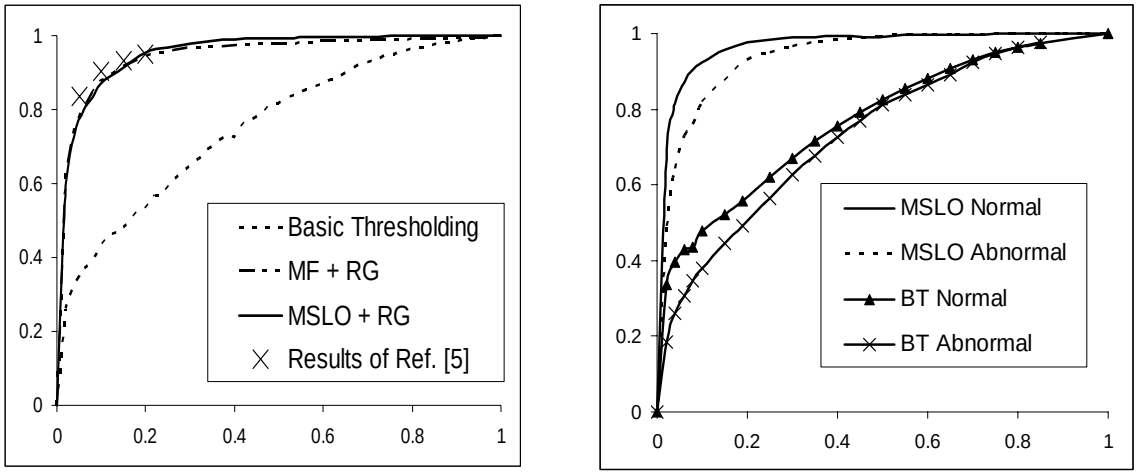


Figure 3: (a; left) ROC results for the multiscale line operator (MSLO) with 3 levels in the image pyramid and median filtering (MF) with L=20 are shown for the STARE dataset. Fifty region-growing (RG) steps were also used in both cases (M=50). ROC results are compared to results of Ref. [5]. (b; right) ROC results for the STARE dataset (normal images; abnormal images) for the MSLO with 3 levels in the image pyramid and basic thresholding (BT).

By visual inspection of Figure 3a, we see that our results compared well to those the results of another segmentation algorithm of Ref. [5] for the STARE dataset. The results for the area under the curve (AZ) for the entire STARE dataset are shown in Table 1. MSLO results provided larger AZ values than those of basic thresholding. The median filter also provided good results. The best results for region growing were found for M=50, although only small gains in the area under the curve were obtained. Table 1 indicates that our results compare well to those results of other methods [6,7] for the area under the ROC curve (AZ) for the STARE dataset.

RG Steps: M	0	50	100
MF L=10	0.930	0.938	0.939
MF L=15	0.942	0.945	0.946
MF L=20	0.945	0.947	0.947
MF L=25	0.944	0.945	0.946
MSLO 1 Level	0.865	0.876	0.876
MSLO 2 Levels	0.933	0.937	0.936
MSLO 3 Levels	0.953	0.953	0.953
MSLO 4 Levels	0.940	0.940	0.940
Basic Thresholding		0.753	
Ref. [6]		0.930	
Ref. [7]		0.961	

Table 1: Areas under the curve (AZ) for ROC analysis of the entire STARE dataset compared to results of Refs. [6] and [7] and basic thresholding. (MSLO = Multiscale Line Operator; MF= Median Filter; RG = Region Growing.)

	Normal			Abnormal		
RG Steps: M	0	50	100	0	50	100
MF L=15	0.956	0.958	0.960	0.928	0.933	0.934
MF L=20	0.960	0.961	0.961	0.929	0.932	0.933
MF L=25	0.959	0.960	0.960	0.929	0.930	0.931
MSLO 2 Levels	0.949	0.951	0.952	0.917	0.921	0.920
MSLO 3 Levels	0.966	0.966	0.966	0.939	0.940	0.939
MSLO 4 Levels	0.957	0.956	0.957	0.923	0.923	0.924
Basic Thresholding		0.769			0.737	
Ref. [7]		0.969			0.950	

Table 2: Areas under the curve (AZ) for ROC analysis of the STARE dataset for the Normal (10 images) and Abnormal (10 images) subsets compared to results of Ref. [7]. (MSLO = Multiscale Line Operator; MF= Median Filter; RG = Region Growing.)

ROC curves for the different image types (normal; abnormal) in the STARE dataset are presented in Figure 3b, and corresponding values for the areas under the curve are given in Table 2. Better results were obtained for the normal

images with respect to the abnormal images, as expected. The application of the median filter took approximately 30 seconds on a 2 GHz CPU with 256 MB of RAM. On the same machine and for the same images, the MSLO typically took 2 minutes to run for an image pyramid with 3 levels. The majority of the processing time was spent applying the line operator at the finest level of detail in the image pyramid. The region-growing step took another 5-10 seconds for approximately M=150 such steps. These calculations were carried out using the interactive data language (IDL) for ease of algorithm prototyping.

4 Conclusions

We have proven that the application of multiscale line operator (MSLO) techniques can lead to enhanced blood vessel segmentation in digital fundus images compared to basic thresholding. Interestingly, application of a median filter followed by a simple region-growing technique appeared to work well also. Our results were found to compare very favourably to those results of previous segmentation procedures [5-7] for the STARE dataset. We conclude that the new methods perform equally well to these previously described techniques. We note that speed of use is crucial to any practical real-world application of automated blood-vessel tracing at the point of image capture, e.g. as a screening tool for retinal eye disease. It is a non-trivial problem to decide at which point any algorithm is "good enough" in terms of accuracy of segmentation and is "fast enough" for clinical use.

References

1. C. Kirbas and F.H. Quek, "A Review of Vessel Extraction Techniques and Algorithms," *ACM Computing Surveys* Vol. 36, pp. 81—121, 2004.
2. R. Zwiggelaar and C.R.M. Boggis, "The benefit of knowing your linear structures in mammographic images," in *Proceedings of Medical Image Understanding and Analysis 2002*, pp. 73—76, 2002.
3. R. Zwiggelaar, S.M. Astley, C.R. Boggis, C.J. Taylor, "Linear structures in mammographic images: detection and classification," *IEEE Transactions on Medical Imaging* Vol. 23, pp. 1077—1086, 2004.
4. E.M. Hadley, E.R.E. Denton, R. Zwiggelaar, "Mammographic Risk Assessment Based on Anatomical Linear Structures," in *Digital Mammography*, Lecture Notes in Computer Science Vol. 4046 (Springer Verlag, Heidelberg), pp. 626—633, 2006.
5. N.M. Salem and A.K. Nandi, "Segmentation of retinal blood vessels using scale-space features and k-nearest-neighbour classifier," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2006* Vol. 2, pp. 1001—1004, 2006.
6. X. Jiang and D. Mojon, "Adaptive local thresholding by verification-based multithreshold probing with application to vessel detection in retinal images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 25, pp. 131—137, 2003.
7. J. Staal, M.D. Abràmoff, M. Neimeijer, M.A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging* Vol. 23, pp. 501—509, 2004.
8. S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Goldbaum, "Detection of Blood Vessels in Retinal Images Using Two Dimensional Matched Filters," *IEEE Transactions on Medical Imaging* Vol. 8, pp. 263—269, 1989.
9. O. Chutatape, L. Zeng, and S. Krishman, "Retinal Blood Vessel Detection and Tracking by Matched Gaussian and Kalman Filters," in *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* Vol. 20, pp. 3144—3149, 1998.
10. L. Gang, O. Chutatape, and S.M. Krishnan, "Detection and Measurement of Retinal Vessels in Fundus Images using Amplitude Modified Second-Order Gaussian Filter," *IEEE Transactions on Biomedical Engineering* Vol. 49, pp. 168—172, 2002.
11. H.L. Hsu, W. M.L. Lee T.Y. Wong, "Automatic grading of retinal vessel calibre," *IEEE Transactions on Biomedical Engineering* Vol. 52, pp. 1352—1355, 2005.
12. D. Satyarthi. M.R. Kumar, and S. Dandapat, "Gaussian Intensity Distribution Modelling of Blood Vessels in Fundus Images," in *Proceedings of IEEE INDICON 2005 Conference*, pp. 228—232, 2005.
13. A.D. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging* Vol. 19, pp. 203—210, 2000.
14. F. Zana and J.-C. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *IEEE Transactions on Image Processing* Vol. 10, pp. 1010—1019, 2001.
15. A.M. Mendonça and A. Campilho, "Segmentation of Retinal Blood Vessels by Combining the Detection of Centrelines and Morphological Reconstruction," *IEEE Transactions on Medical Imaging* Vol. 25, pp. 1200—1213, 2006.
16. K.A. Vermeer, F.M. Vos, H.G. Lemij, A.M. Vossepoel, "A model based method for retinal blood vessel detection," *Computers in Biology and Medicine* Vol. 34, pp. 209—219, 2004.
17. Y.A. Tolias and S.M. Panas, "A fuzzy vessel tracking algorithm for retinal images based on fuzzy clustering," *IEEE Transactions on Medical Imaging* Vol. 17, pp. 263—273, 1998.
18. M. Martinez-Perez, A. Hughes, A. Stanton, S. Thom, A. Bharath, and K. Parker, "Scale-Space Analysis for the characterisation of retinal blood vessels," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 1999*, pp. 90—97, 1999.
19. A.F. Frangi, W.J. Niessen, K.L. Vincken, and M.A. Viergever, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 1998*, pp. 130—137, 1998.

Three dimensional texture analysis of computed tomography images of the lung in the diagnosis of pulmonary disorders

Balaji Ganeshan [a], Kenneth A. Miles [b], Rupert C.D. Young [a] and Chris R. Chatwin [a]

[a] Industrial Informatics and Manufacturing Systems, Department of Engineering and Design, University of Sussex, Falmer, Brighton BN1 9QT, [b] Brighton & Sussex Medical School, University of Sussex, Falmer, Brighton BN1 9PX

Abstract. Three dimensional (3-D) selective-scale texture analysis has been developed for visual and quantitative assessment of ventilated and vascular lung tissue from computed tomography (CT). This may aid in the diagnosis of pulmonary disorders with altered vasculature. Prior to texture analysis, an initial automated segmentation procedure to include only lung parenchyma and generation of isometric volume were carried out. Furthermore, ventilated and vascular lung regions were separately selected. Texture analysis comprised two stages a) volume filtration using 3-D Laplacian of Gaussian filtering to highlight fine to coarse texture using different pass-band regions within ventilated and vascular lungs followed by b) quantification of texture. Quantification of texture was also performed on an unfiltered CT lung dataset. Volume rendering and image fusion of ventilated and vascular lung texture were employed for visualisation. 3-D filtered ventilated texture quantified as mean grey-level intensity was higher in emphysematous lung than non-emphysematous lung, possibly indicating ventilation defects. 3-D filtered vascular texture, in particular coarse texture, demonstrated a descending trend in entropy (or an ascending trend in uniformity) for normal lung, followed by pulmonary embolism and finally pulmonary disorder in the presence of both embolism and emphysema, possibly indicating the degree of vascularity associated with these different causes of altered pulmonary vasculature. Image fusion of ventilated and vascular lung texture highlighted areas of mismatched and matched defects in patients with pulmonary disorders. This feasibility study demonstrated that 3-D selective-scale texture analysis can potentially aid in the diagnosis of pulmonary disorders with altered vasculature.

1 Introduction

Recent advances in pulmonary imaging are having a pronounced impact on the ability to detect, characterize and quantify pulmonary disorders, providing both morphological and functional information. Furthermore with the development of digital chest radiography, computer-aided diagnosis (CAD) is now becoming feasible and increasingly important [1]. Several disorders such as pulmonary emphysema and pulmonary embolism (PE) are associated with alterations in the distribution of airways and blood vessels within the lungs. The diagnosis of PE depends on imaging findings as there are no reliable clinical features or laboratory tests [2, 3]. Computed tomography pulmonary angiography (CTPA) has become the initial imaging study of choice for evaluating patients with suspected PE. Diagnosis of these disorders from spiral CTPA however involves more than 100 sectional images, with additional images obtained using MPR and MIP. Also diagnosis frequently becomes difficult in sectional images due to anisotropy. Furthermore CT can demonstrate pathological changes in the lung associated with the presence of other disorders such as pulmonary emphysema, complicating the diagnosis of PE. CAD based techniques can provide visual and quantitative assessment of pulmonary disorders [1]. Automated volumetric visual analysis is a recognized CAD approach in the detection of PE from CTPA images [4]. Quantitative texture analysis (TA) of single planar perfusion lung scan indicated the presence of PE using ANNs [5, 6], while texture from computer simulated lung scan has shown to reflect homogeneity [7]. Adaptive TA technique outperformed simple mean lung density method and histogram analysis in classifying normal and emphysematous tissues from individual CT images [8]. 3-D selective-scale TA could selectively extract and quantify texture features based on scale and intensity variations to highlight areas of ventilated lung and enhance small and less conspicuous pulmonary vessels. In this feasibility study we employ this approach to identify potential texture correlates for ventilated and vascular lung on CTPA images. This may aid in the diagnosis of PE and distinction from other causes of altered pulmonary vascularity.

2 Texture Analysis Methodology

We studied CTPA examinations of 4 patients, of which one had normal lungs with no evidence of lung disease or PE and the remaining three were diagnosed with PE. Out of the three patients with PE, two were diagnosed with only embolism (PE1 and PE2) while the third patient had embolism and emphysema (PEE) as diagnosed by the radiologist.

Corresponding author: Balaji Ganeshan; Email address: b.ganeshan@sussex.ac.uk

2.1 Automated region of interest (ROI) selection containing only the lung

The chest CT image contains the heart, liver, spleen, kidney, fat, muscles and bones. Therefore an automated boundary detection algorithm was devised for selection of ROI containing only lung from each slice. Step-by-step implementation of this algorithm is shown in the Figure 1.
The generally employed thresholds for lung segmentation are ideally equal to the average lung density [9] which in our study was close to -200 HU. From the initial CT image (Figure 1A), this grey-level threshold was employed to highlight all pixels above this threshold and mask out the rest (binary image). The main boundaries were traced from this binary image but only the largest region i.e. the chest wall was selected and superimposed onto the original image (Figure 1B). This preliminary segmentation excluded the air outside the patient's body. The penultimate step was to trace only the lungs within this initial segmented region for which only pixels below -200 HU were highlighted. From this binary image boundaries were traced which included only the right and left lungs (Figure 1C). The final step involved masking out all the pixels except the right and left lungs (Figure 1D).

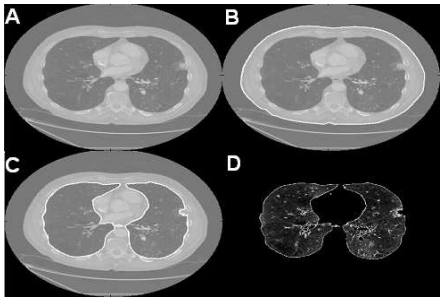


Figure 1. The process of automated segmentation of the lungs from the rest of the image is illustrated for the cross-section of the volume along the z-axis.

2.2 Re-sampling of whole lung for generating isometric volume

The 3-D LoG filter is a spherically symmetrical isometric filter, whereas the lung volume datasets for all the patients were not isometric with regard to both the physical size and sampling rate. Isometric voxels are critical for correct implementation of this filter to highlight only those features in the lung that match the filter tuning parameter (σ). The normal lung volume contained 222 slices while lungs with PE1, PE2 and PEE contained 200, 224 and 238 slices respectively, each having slice thickness of 1.25 mm. These lung volumes were first down-sampled in the axial plane (x-y direction) by a factor of 0.4923 (normal), 0.5625 (PE1), 0.6016 (PE2) and 0.664 (PEE) using cubic interpolation, which modified the sampling rate to give a sampling interval of 1.25mm in the x-y direction for all the lung volumes. The cubic interpolation works best, reducing sampling error, when interpolating for a factor less than 3, as was the case in this study. The modified volume sizes in voxels after down-sampling were 252x252x222 (normal), 288x288x200 (PE1), 308x308x224 (PE2) and 340x340x238 (PEE), with a uniform sampling rate of 1.25x1.25x1.25 (mm) in all the three axes, thus obtaining isometric voxels. Thus texture features highlighted by the 3-D LoG filter for a given filter parameter (σ) would be of the same scale and therefore comparable in each of the lung volume analysis. Even though the resolutions of the original volumes were decreased, this approach was employed instead of up-sampling in the coronal plane (z direction) due to the increased computing required to evaluate 3-D volume filtration. The next requirement is to obtain equality in the physical sizes of the lung volumes in the three orthogonal directions. This was achieved by adding zero-padding where necessary: 256x256x256 (normal) and 340x340x340 (PE1, PE2 and PEE).

2.3 Texture analysis of pulmonary airways and / or vasculature

PE is associated with thrombus obstructing a blood vessel [10] while emphysema is associated with abnormal enlargement of the air spaces accompanied by destructive changes to alveolar walls [11]. Separate texture analysis of ventilated and vascular lungs may provide additional and more precise assessment of patients with PE and emphysema. To achieve this segmentation, for each patient from within the whole lung, voxels having CT attenuation values between -500 and -950 HU were considered as ventilated lung while voxels above -500 HU were taken to contain pulmonary vessels. The lower threshold of -500 HU in the segmentation of pulmonary vessels was employed to include the effects of partial volume of small vessels that may have attenuation as low as -500 HU. 3-D texture analyses with and without filtration was carried out to independently assess texture within these segmented lung volumes highlighting ventilated and vascular lung. For filtered texture analysis of the ventilated lung, the grey-level contrast of unprocessed volume was reversed prior to filtration for enhancing airway defects or holes. Matched

defects occur predominantly in emphysema cases where airway defects are also accompanied by perfusion deficit whereas mismatched defects occur predominantly in embolism cases where perfusion deficit are generally not accompanied by ventilated defects.

2.4 Computation of 3-D LoG filtration of the lung volume

The 3-D Gaussian distribution (g) obtained from its 2-D representation [12] is formulated as follows:

$$g(x,y,z) = e^{-\frac{x^2+y^2+z^2}{2\pi\sigma^2}} \qquad \text{(Equation 1)}$$

where (x,y,z) is the spatial coordinates of the volume matrix,

σ is the standard deviation.

$\nabla^2 g$ is the volumetric Laplacian of Gaussian (LoG) filter, whose distribution in 3-D spatial domain is given below:

$$\nabla^2 g(x,y,z) = \frac{-1}{\pi\sigma^4}\left(1-\frac{x^2+y^2+z^2}{2\sigma^2}\right)e^{-\left(\frac{x^2+y^2+z^2}{2\sigma^2}\right)} \qquad \text{(Equation 2)}$$

From the expression for this circularly symmetric filter at different sigma (σ) values, the number of voxels representing the width between the diametrically opposite zero-crossing points in this filter can be calculated. The lower the sigma value, the smaller is the width of the filter in the spatial domain and the larger is the pass-band region of the filter in the frequency domain, highlighting fine details or features in the filtered volume in the spatial domain. Thus the filter can be tuned so as to highlight fine (σ = 0.5, filter width: 2 voxels or 2.5 mm) and coarse (σ = 1.5, filter width: 6 voxels or 7.5 mm) textures or features within the ventilated and vascular lung. This width can be considered as the scale at which the structures in the whole lung will be highlighted and enhanced whilst structures below this scale will become blurred. Frequency domain filtration reduces computational time. The 3-D Fourier transform of the LoG filter and the lung volume were computed by taking the Fourier transforms of the 3-D spatial filter ($\nabla^2 g$) and the lung volume (h) along the three orthogonal directions. Figure 2 shows the cross-section and sub-volume representation of the 3-D LoG filter. The inverse Fourier transform of the filtered spectrum gives the resultant filtered volume in the spatial domain. Expression for the 3-D filtered volume is given by equation 3.

$$h_f(x,y,z) = F^{-1}\left\{F\left(\nabla^2 g(x,y,z)\right) * F\left(h(x,y,z)\right)\right\} \qquad \text{(Equation 3)}$$

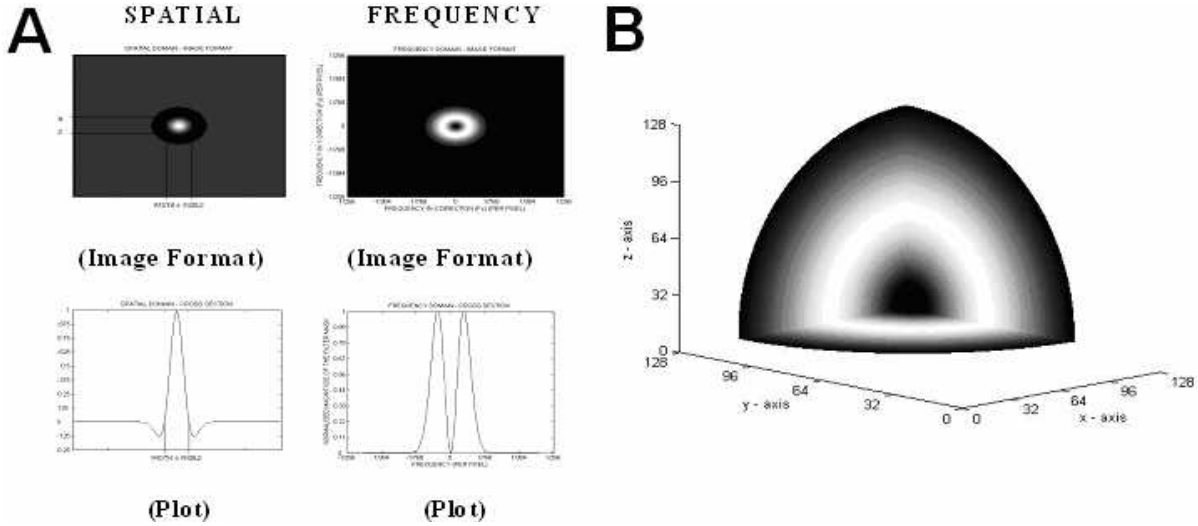


Figure 2 A) Cross-sections of the 3-D LoG filter in the spatial and frequency domain and **B)** Volume rendering of the sub-volume of the of the 3-D LoG filter in the frequency domain at a sigma (σ) value of 1.5.

2.5 Quantification of Texture

The following statistical and spectral parameters were used to evaluate the unfiltered and filtered texture within the volume of interest V in each filtered volume $a(x,y,z)$: (a)Mean grey-level intensity, (b)Entropy, and (c)Uniformity, indicating how close the volume is to a uniform distribution of the grey-levels. These parameters are defined below where N the total number of voxels in V, l is the grey-level (for example $l = 1$ to k indicates grey-

level from 1 to k) in V and $p(l)$ the probability of the occurrence of the grey-level l based on the histogram technique:

$$Mean\ Grey-Level\ Intensity\ (m) = \frac{1}{N}\sum_{(x,y,z)\in V}[a(x,y,z)] \quad \text{(Equation 4)}$$

$$Entropy\ (e) = -\sum_{l=1}^{k}[p(l)]\log_2[p(l)] \quad \text{(Equation 5)}$$

$$Uniformity\ (u) = \sum_{l=1}^{k}[p(l)]^2 \quad \text{(Equation 6)}$$

3 Results

Texture quantification of ventilated and vascular lung is shown in Table 1 in all the 4 patients. 3-D filtered texture quantification showed greater sensitivity, as compared to no filtration, in enhancing subtle variations in ventilated and vascular lung texture in patients with pulmonary disorders. Filtered ventilated texture quantified as mean grey-level intensity of emphysematous lung (PEE) was greatly different from that of normal while these textures were comparable in non-emphysematous lungs (normal, PE1 and PE2). Mean grey-level intensity value was highest for PEE possibly indicating ventilation defect. Filtered vascular texture, in particular for coarse texture quantified as entropy and uniformity, of patients with pulmonary disorders (PE1, PE2 and PEE) was different from normal. A descending trend in entropy (or ascending trend in uniformity) values was observed for normal tissue, followed by PE and finally for PEE, possibly indicating degree of vascularity (or perfusion) associated with different causes of altered pulmonary vasculature. Image fusion of ventilated and vascular lung texture highlighted areas of mismatched and matched defects in patients with pulmonary disorders (Figure 3).

Texture	Parameter	Normal	PEE	PE1	PE2
Without Filtration (Ventilated)	Mean	-836	-862 (3.11 ↑)	-788 (5.74 ↓)	-841 (0.6 ↑)
	Entropy	7.8125	7.8556 (0.55 ↑)	8.1036 (3.73 ↑)	7.8308 (0.23 ↑)
	Uniformity	0.0061	0.0057 (6.56 ↓)	0.0045 (26.23 ↓)	0.0058 (4.92 ↓)
Fine (Ventilated)	Mean	119.6	329.8 (175.75 ↑)	99.58 (16.74 ↓)	124.56 (4.15 ↑)
	Entropy	7.2884	9.3777 (28.67 ↑)	6.6164 (9.22 ↓)	7.3616 (1.00 ↑)
	Uniformity	0.2012	0.1109 (44.88 ↓)	0.2539 (26.19 ↑)	0.1979 (1.64 ↓)
Coarse (Ventilated)	Mean	175.88	536.18 (204.86↑)	132.16 (24.86↓)	172.68 (1.82↓)
	Entropy	5.9293	8.1417 (37.31↑)	4.9260 (16.92↓)	5.8950 (0.58↓)
	Uniformity	0.3609	0.2299 (36.30↓)	0.4552 (26.13↑)	0.3642 (0.91↑)
Without Filtration (Vascular)	Mean	-214	-267 (24.77 ↓)	-283 (32.24 ↓)	-259 (21.03 ↓)
	Entropy	0.6600	0.5380 (18.48 ↓)	0.8324 (26.12 ↑)	0.6929 (4.98 ↑)
	Uniformity	0.9160	0.9317 (1.71 ↑)	0.8898 (2.86 ↓)	0.9104 (0.61 ↓)
Fine (Vascular)	Mean	129.12	76.94 (40.41 ↓)	123.58 (4.29 ↓)	115.32 (10.69 ↓)
	Entropy	1.2401	0.8549 (31.06 ↓)	1.3382 (7.91 ↑)	1.1886 (4.15 ↓)
	Uniformity	0.8782	0.9174 (4.46 ↑)	0.8672 (1.25 ↓)	0.8833 (0.58 ↑)
Coarse (Vascular)	Mean	468.6	293.94 (37.27↓)	405.16 (13.54↓)	417.62 (10.88↓)
	Entropy	3.1621	2.2932 (27.48↓)	2.8722 (9.17↓)	2.8578 (9.62↓)
	Uniformity	0.6901	0.7737 (12.11↑)	0.7179 (4.03↑)	0.7202 (4.36↑)

Table 1. Texture quantification of ventilated and vascular lung volume on CT (% change in magnitude from normal).

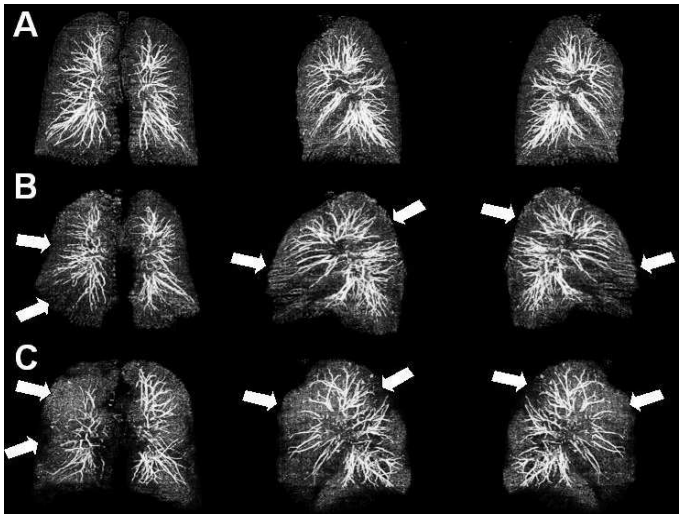


Figure 3 The coronal, left and right sagittal views of the fusion of ventilated and vascular lung texture extracted by 3D fine filtered texture analysis in (A) normal lung, (B) pulmonary embolism (PE1), and (C) pulmonary embolism and emphysema (PEE). The arrows point to areas of mismatched (PE1 and PEE) and matched (only PEE) defects.

4 Discussion

The results of this feasibility study demonstrate the potential of 3D selective-scale texture analysis of CTPA to provide texture correlates for ventilated and vascular lung, thereby demonstrating it to be a useful non-invasive tool in the assessment and distinction of PE from other causes of altered pulmonary vascularity. This technique higlighted and quantified areas of ventilated defects and enhanced small and conspicuous vessels associated with perfusion deficit, based on scale and variation in image intensity. This is otherwise lost when analysing global lung features in pulmonary disorders. Furthermore CTPA examinations, being readily available and more commonly employed for patients suspected with PE, are suitable for this 3D texture analysis on account of the generation of thin CT slices and use of contrast agents. Nevertheless, the findings in this feasibility study need to be confirmed in a larger series of patients for further clinical evaluation. In summary, 3D selective-scale filtered ventilated and vascular texture analysis has the potential to be a novel visual and quantitative tool in the diagnosis of pulmonary disorders.

References

1. D.D. Maki, W.B. Gefter and A. Alavi “Recent advances in pulmonary imaging”, *Chest* **116**, pp 1388-1402, 1999.
2. R.D. Hull, G.E. Raskob, C.J. Carter et al “Pulmonary embolism in outpatients with pleuritic chest pain”, *Arch Intern Med* **148**, pp 838-844, 1988.
3. S. Patil, J.W. Henry, M. Rubenfire and P.D. Stein “Neural network in the clinical diagnosis of acute pulmonary embolism”, *Chest* **104**, pp 1685-1689, 1993.
4. Y. Masutani, H. MacMahon and K. Doi “Computerized detection of pulmonary embolism in spiral CT angiography based on volumetric image analysis”, *IEEE Transactions on Medical Imaging* **21**(12), pp 1517-1523, 2002.
5. G.D.Tourassi, E.D. Frederick, C.E. Floyd Jr and R.E. Coleman “Multifractal texture analysis of perfusion lung scans as a potential diagnostic tool for acute pulmonary embolism”, *Computers in Biol. and Med.* **31**, pp 15-25, 2001.
6. E.B. Cargill, K. Donohue, G. Kolodny, J.A. Parker and R.E. Zimmerman “Analysis of lung scans using fractals”, *SPIE* **1092**, pp 2-9, 1989.
7. R. Cinotti, S. Edery, E. Kahn, H. Susskind, A.B. Brill and R. Di Paola “Lung scintigraphy clustering by texture analysis”, *Eur. J. Nucl. Med*. **16**, pp 353-359, 1991.
8. R. Uppaluri, T. Mitsa, M. Sonka, E.A. Hoffman and G. McLennan “Quantification of pulmonary emphysema from lung computed tomography images”, *Am J Respir Crit Care Med* **156**, pp 248-254, 1997.
9. G.J. Kemerink, R.J.S. Lamers, B.J. Pellis, H.H. Kruize and J.M.A. Van Engelshoven “On segmentation of lung parenchyma in quantitative computed tomography of the lung”, *Med. Phys.* **25**(12), pp 2432-2439, 1998.
10. Collaborative study by the Prospective Investigation of Pulmonary Embolism Diagnosis Investigators. Value of the ventilation/perfusion scan in acute pulmonary embolism: results of the PIOPED. *JAMA.* **263**, pp 2753-59, 1990.
11. American Thoracic Society. Chronic bronchitis, asthma, and pulmonary emphysema: a statement by the committee on diagnostic standards for nontuberculous respiratory diseases. *Am. Rev. Respir. Dis.* **85**, pp 762-768, 1962.
12. D. Marr “Representing the image: Zero-Crossings and the Raw Primal Sketch”, In: J. Wilson, P. Monsour, editor. *Vision*. San Francisco: W.H. Freeman and Company, pp 54-68, 1982.

3D skin texture analysis for early diagnosis of malignant melanoma

Yi Ding[a], Lyndon Smith[a], Melvyn Smith[a], Robert Warr[b],

[a]Machine Vision Laboratory, CIMMS, Faculty of CEMS, University of the West of England, Bristol, BS16 1QY

[b]Pigmented lesion clinic, Plastic surgery, North Bristol NHS Trust, Bristol, BS16 1LE

1.Abstract

Malignant melanoma (MM) is a life-threatening dermatological disease, the successful treatment of which relies heavily on early recovery and diagnosis. It is proposed that early MMs tend to have larger 3D topological magnitude and direction disruptions than benign lesions. With a view to capturing these 3D indications, a 6 light photometric stereo device has been developed to recover reliable 3D skin texture in the form of surface gradients. In order to characterise the 3D skin texture, two separate numerical methods have been proposed which quantify surface direction and magnitude disruption. A true 2D centroid which is the projection of the real 3D centroid on the image plane has been used to determine the reference angles and the surface direction disruption using minimum error estimation. This is a more accurate method than the conventionally defined centroid which is susceptible to intensity distortions. Preliminary studies on 30 lesions (of which 5 are MMs) have resulted in 80% specificity with all 5 MMs separated from benign lesions.

2.Introduction

Employment of non-invasive analysis comprises a potentially useful method for obtaining several important indications. Conventional non-invasive analysis has focused on finding the ABCD (Asymmetry, Border, Color and Diameter) features [1] of melanoma. Recently, some attention has been given to the analysis of skin line patterns [2]; where local transformations are used to find 2D skin line patterns from the image gradients. Variations in the skin line patterns are then used as a measure of the skin characteristics. The separation criterion is that both skin line 1) magnitude difference and 2) direction difference between skin and lesion tend to be larger in melanoma than in benign lesions. Similar criteria have also been sought using electrical impedance [3]. Other techniques include those employing spectral analysis from a dermatological imaging device such as SIAscope [4] where colour variations are used to analyse the distribution of melanin and blood in relation to skin thickness.Those techniques can provide reasonably good results for clinical diagnosis, however, it should be understood that due to the large range of dermatological diseases and various forms of MM, the complexity of the problem makes a clear distinction between melanoma and benign lesions a very challenging task, and so far no single technique has been able to achieve fully accurate diagnosis alone. Hence further improvement is always being sought as additional dimensional information or feature descriptors become available.

Photometric stereo was first described by Woodham in 1978 [5]. The theory of this technique is to use surface reflectance properties to recover the surface's physical properties (orientations and reflection rate) under the rule of a particular reflectance model (e.g. Lambertian or Phong). In order to recover the three dimensional surface normals (x- y- and z-), at least three images, each captured from a static viewpoint (camera) but under differently positioned illuminants, are required. In terms of computational complexity, the difficult problems of image registration and finding correspondences between pixels in different images are avoided. The surface normal vector can later be transferred to surface gradient vectors triangularly. By integrating the surface gradients, surface height can be recovered. However, this process can introduce errors and is subject to some limitations. So this work investigates direct analysis of the 3D information in the form of surface gradients which is another way of describing the 3D information.

3.3D skin texture analysis

3.1. Skin analyser

Figure 1 depicts a 6-light photometric stereo device known as skin analyser which is used as the data acquisition system. Its schematic configuration is illustrated in Figure 1(Left) and the developed device is shown in Figure 1 (Right). When used in clinical trials, it is placed with its axis perpendicular to the skin surface and a camera takes 6 color images with each under a different LED illumination, the operation takes less than 1 second, hence it meets the demand of the static setup required by photometric stereo.

For an ideal Lambertian surface, the image irradiance equation can be expressed as:

$$i(m,n) = \rho(m,n)\frac{(-p(m,n),-q(m,n),1)}{\sqrt{p^2(m,n)+q^2(m,n)+1}}(\cos\alpha\sin\beta,\sin\alpha\sin\beta,\cos\beta)^T \quad 1)$$

Where α and β are slant and tilt directions of the illuminants, the partial derivatives $p(m,n)$ and $q(m,n)$ are the x-axis (indexed by *m*) component and y-axis (indexed by *n*) component of the surface gradients at image position (*m*,*n*) respectively, and ρ is the surface reflection rate (albedo).

The triangular relation between unit surface normals and surface gradients is [6]

$$(n_x(m,n),n_y(m,n),n_z(m,n)) = \frac{(-p(m,n),-q(m,n),1)}{\sqrt{p^2(m,n)+q^2(m,n)+1}} \quad 2)$$

Where $n_x(m,n),n_y(m,n),n_z(m,n)$ are unit surface normals in x-, y- and z- axis. Substitute 2) into 1), we have

$$i(m,n) = \rho(m,n)(n_x(m,n),n_y(m,n),n_z(m,n)(\cos\alpha\sin\beta,\sin\alpha\sin\beta,\cos\beta)^T \quad 3)$$

Since the modular of unit surface normal is equal to 1, i.e. $n_x(m,n)^2+n_y(m,n)^2+n_z(m,n)^2=1$, so only three variables are unknown including the albedo and any two from the $n_x(m,n),n_y(m,n),n_z(m,n)$. So we need at least 3 lights to solve the equation. All other lights have abundant information which could be used for specular and shadow detection and recovery. The criteria used to determine whether a pixel is within shadows or specularities is based on the same on used by Barsky [8]. The accuracy of the system has been demonstrated [7] and its application for better lesion segmentation using albedo images has been described [9].

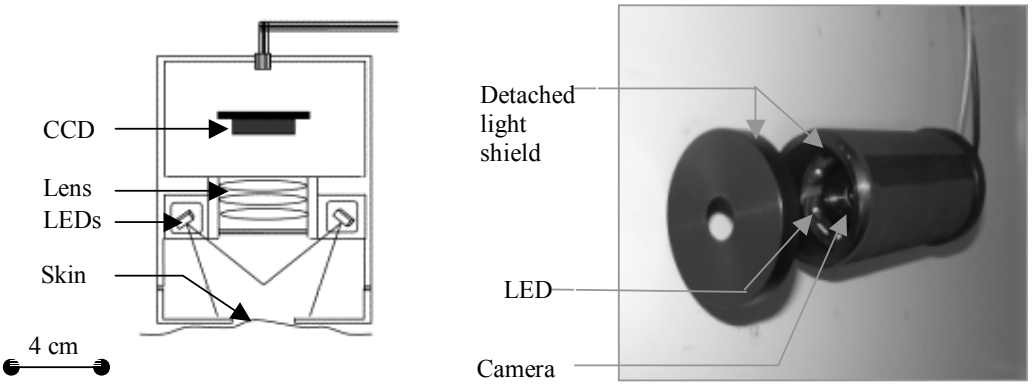


Figure 1 Left: Schematic to scale; Right: developed handheld colour photometric stereo device known as the 'Skin Analyzer'. Only two of the six LED light sources are shown in either picture.

The surface gradients acquired by skin analyser are 3D skin texture which belongs to surface texture [10] which should not be confused with image texture. Most conventional skin texture based techniques e.g. the skin line pattern technique, employ image texture, which inherently are the features formed by pixel intensities under a particular illumination direction, and are illumination-dependent. This is in contrast with surface textures, which are intrinsic physical properties of an object, (comprising the reflection pattern – which is a 2D surface texture described by albedo map and surface topology which is a 3D surface texture described by normal/gradient map), and it is illumination-invariant. Our 3D texture analysis is also rotation-invariant as no assumption regarding the rotational direction of the image is made as long as the imaging surface (skin) is perpendicular to the imaging device.

3.2. Statistical analysis

In this preliminary research, differences between some benign lesions and MM in both direction and magnitude of the 3D surface texture have been observed. As evident in Figure 2, the benign Seborrhoeic Keratosis (SK) has a regular directional distribution of surface gradients, while some regions within MM have abrupt changes in directional distribution; similarly, the magnitude distribution within the Benign Compound Nevus (BCN) is more regular (smaller divergence) than that of MM. Here we describe the separation criteria: Only the lesions with both irregular direction and magnitude gradient distributions are regarded as suspected melanoma, while lesions which only meet one or less irregular criterion are regarded as benign lesions.

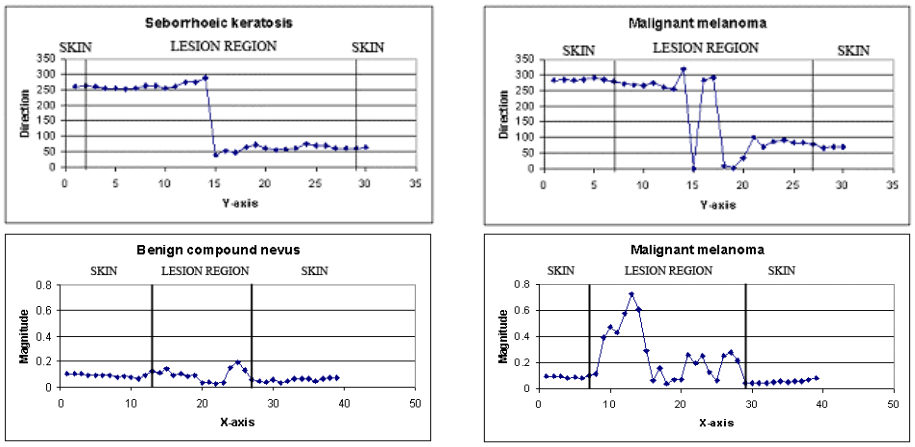


Figure 2 Top left: a sample column of surface gradient directions of a benign Seborrhoeic Keratosis; Top right: a sample column of surface gradient directions of a malignant melanoma; Bottom left: a sample row of surface gradient magnitude of a benign compound nevus; Bottom right: a sample row of surface gradient magnitude of a malignant melanoma

3.3. Surface direction disruption

To quantify the surface direction disruption, we have to establish a standard reference model to compare with current lesion in terms of surface direction errors. The reference model can be found by targeting a specific reference point or through neural training of numerous benign lesions. In this paper, we choose the former approach, where the centroid is chosen as this specific reference point. To formulate this problem, we define the direction deviation as the sum of directional errors between the estimated reference directions and gradient directions across all image positions.

$$(m_c^*, n_c^*) = \arg\min \sum_{(m,n)\in A_l} \| (\theta_r^*(m,n) - \theta_s(m,n)) \| \qquad 4)$$

where (m_c^*, n_c^*) is the estimated centroid, A_l denotes the lesion region, $\|\square\|$ denotes the Euclidean distance, θ_s is the surface gradient direction and θ_r^* is the estimated reference direction at position. The idea is to firstly find an initial centroid estimate (m_{c1}^*, n_{c1}^*), then compute the sum of directional errors between the reference directions and surface gradient directions within the lesion region A_l. This process will generate the directional errors for the current centroid estimate (m_c^*, n_c^*). This process will be repeated for all the other center estimates within the lesion, $\{(m_{c2}^*, n_{c2}^*), (m_{c3}^*, n_{c3}^*), ..., (m_{cN}^*, n_{cN}^*)\} \in A_l$. Finally, the centroid estimate that gives the minimum errors will be chosen as the determined center, (m_c^*, n_c^*).

To explain the reference direction more clearly, let's refer to Figure 3(Middle), the reference direction depends on the relative positions between the current position (m,n), and the estimated center of directions (m_c^*, n_c^*). Given the x-axis as the reference direction, then the reference direction θ_r^* at (m,n) can be computed as

$$\theta_r^* = \begin{cases} \delta & \text{if } (m - m_c^*) \geq 0 \text{ AND } (n - n_c^*) \geq 0 \\ 2\pi + \delta & \text{if } (m - m_c^*) \geq 0 \text{ AND } (n - n_c^*) \leq 0 \\ \pi + \delta & \text{all others} \end{cases} \quad 5)$$

where $\delta = \tan^{-1}(\frac{\| m - m_c^* \|}{\| n - n_c^* \|})$

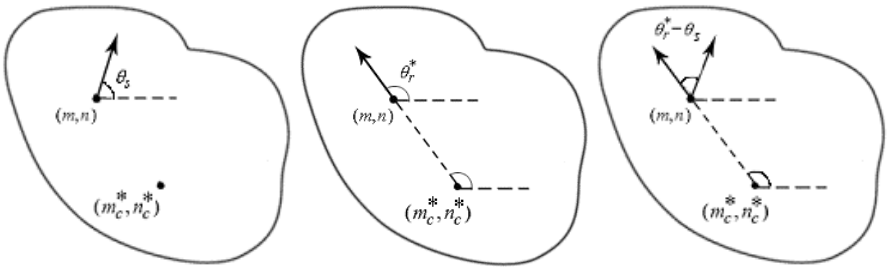


Figure 3 Left: surface gradient direction; Middle: Estimated reference direction; Right: The difference in direction between surface gradient direction and estimated reference direction

It should be noted that the estimated centroid using 4) gives a more accurate 2D centroid which is important to asymmetry analysis. This is because the image of a 3D object depends on its shape, its reflectance properties, and the distribution of light sources, large specular reflection distortion on the 2D lesion image surface is a problem for conventional 2D intensity-based techniques. As shown in Figure 4, the true 2D center can be accurately located by our proposed method which does not suffer from reflection problems and therefore is inherently more robust.

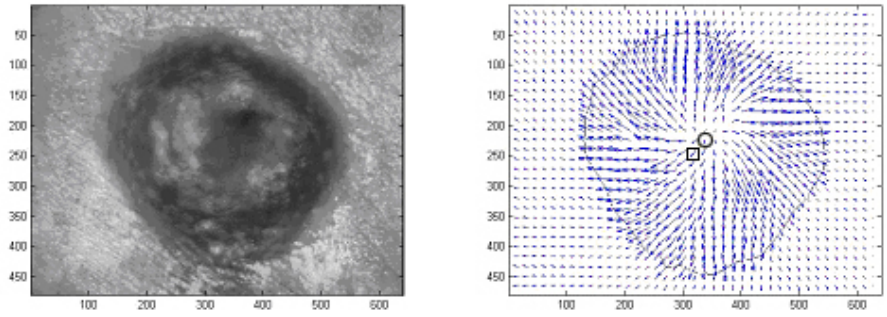


Figure 4 Left: a Seborrhoeic Keratosis image; Right: its gradient distribution and the estimated center of directions (true 2D centroid) marked as "○", while the conventionally calculated centroid using intensities is marked as "□"

Here the surface direction disruption E_d is defined as the average of the minimum errors associated with the determined center of directions.

$$E_d = \frac{\min \sum_{(m,n) \in A_l} \| (\theta_r^*(m,n) - \theta_s(m,n)) \|}{A} \quad 6)$$

where A is the number of pixels within the lesion.

3.4. Surface magnitude disruption

We take a different approach of defining the magnitude deviation, where firstly we calculate the mean magnitude within a lesion, Then the surface magnitude disruption C_R is calculated as the sum of absolute errors between current magnitude $R(m,n)$ and the mean magnitude $\overline{R}$

4.Experiments

Our experiments include 30 lesions which includes 25 benign lesions and 5 MMs. All of these lesions have been captured at our collaborating institutes using the skin analyser. The lesions are classified as either benign or malignant according to both surface direction and surface magnitude disruption. As shown in Figure 5, the best results that can be achieved with 3D skin texture analysis is 100% sensitivity and 80% specificity. Each lesion type's classification rate has also been listed in Table 1. We compare it with the SIAscope [4] which is widely employed in the UK. The difficulty of identifying seborrhoeic keratosis as benign using the SIAscope has been acknowledged during our clinical trials at collaborating clinics. The reason is that those lesions also have intradermal pigmentation which may be seen by SIAscope as melanin penetration in the dermis and be recorded as suspected melanoma. In contrast, 5 out of 7 seborrhoeic keratosis have been successfully recognised as benign lesions in our studies which will certainly reduce the number of suspected "wrong" melanomas identified by the SIA scope.

Type of lesion	No. of Successful classification as benign or malignant/No. of lesions
Malignant melanoma (MM)	5/5
Benign compound nevi (BCN)	6/8
Seborrhoeic keratosis (SK)	5/7
Intradermal nevi (IN)	4/4
Dermatofibroma (DF)	2/2
Hyperkeratotic squamous papilloma (1 HSP)	1/2
Blue nevus (BN)	1/1
Unidentified benign lesion	1/1

Table 1 Lesions with neural network classification rate

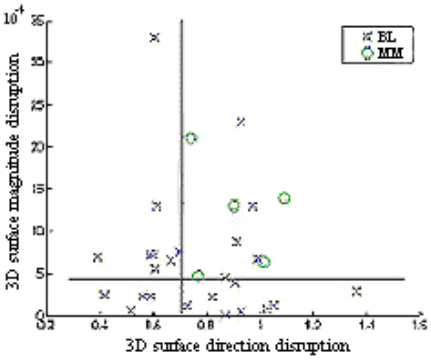


Figure 5 Scatter plot of lesions in 2D feature space (3D surface magnitude disruption vs 3D surface direction disruption)

5.Discussion

This paper has successfully demonstrated an interesting and important application of 3D skin features for automated MM diagnosis. This visualisation nature of this technique enables good interpretation of lesion anisotropy and regularity. The separation criteria based on surface gradient differences are independent of existing techniques and hence provide a distinct method for MM diagnosis. Building upon two existing classes of lesions with regular directional and magnitude distribution, further work is being carried out to study the difference in surface gradient anisotropy within each class to further improve classification.

6.References

1. Frieman, R. et al, "Early detection of malignant melanoma: The role of physician examination and self-examination of the skin". *CA: A Cancer Journal for Clinicians*, 35(3):130–151, 1985.
2. She, Z. et al, "Analysis of skin line pattern for lesion classification", *Skin research and technology*, 9(1): pp. 73-80, 2003.
3. Aberg, P. et al, "Skin cancer identification using multifrequency electrical impedance-a potential screening tool", *IEEE Trans. on biomedical engineering*, Vol. 51, Issue 12, pp. 2097 – 2102, Dec. 2004.
4. Moncrieff, M. et al, Spectrophotometric intracutaneous analysis: a new technique for imaging pigmented skin lesions, *British Journal of Dermatology*, Vol 146, No. 3, March 2002, pp. 448-457(10)
5. Woodham, R. 1978, "Photometric Stereo: A Reflectance Map Technique for Determining Surface Orientation from Image Intensity", *Proc. SPIE. Image Understanding System and Industrial Applications*, Vol. 155, pp. 136-143, 1978.
6. Horn, BKP, *Robot Vision*, MIT press, 1986.
7. Sun. J. et al, "Object surface recovery using a multi-light photometric stereo technique", accepted for *Image and vision computing*, *CVA special issue*, 2006.
8. Barsky, S. et al, "The 4-Source Photometric Stereo Technique for Three-Dimensional Surfaces in the Presence of Highlights and Shadows", *IEEE Trans. on PAMI*, 25(10), pp. 1239-1252, 2003.
9. Sun, J. et al, "Reflectance of human skin using colour photometric stereo - with particular application to pigmented lesion analysis", accepted for *skin research and technology*, 2007.
10. Smith, M. L, The analysis of surface texture using photometric stereo acquisition and gradient space domain mapping, *Image and Vision Computing*, No. 14, pp. 1009-1019, Nov 1999.

Lossless Predictive Algorithm for Medical Image Compression

Josip Knezović*, Mario Kovač, and Martin Žagar

Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia

Abstract. A new lossless image compression algorithm is proposed. Algorithm is based on a highly adaptive predictor that uses blending context obtained using simple classification technique. This allows modeling of more complex image structures that occur often in medical images such as non–trivially oriented edges and the periodicity and coarseness of textures. The prediction step is followed by contextual error modeling which removes remaining statistical redundancies in the residual image data. Finally, error image is entropy coded. Results obtained with proposed compression method on a large set of medical images are very encouraging, beating several well known lossless compression methods. Algorithm can be efficiently applied for natural images as well.

1 Introduction

New approaches in medical imaging such as 3D and 4D bio–modeling produce a great amount of image data. For efficient storage and transmission of those data, compression is mandatory. Lossy compression techniques generally achieve better compression, but reconstructed image data is not as same as the original. Furthermore, in medical imaging, at least certain parts of images are required to be stored and transmitted without any loss of information. Thus, we propose that lossless algorithm is employed for at least those vital parts of interest. [1].

Predictive image coding in which the prediction error of the current pixel is coded has shown to be the most effective technique in lossless image compression. If employed predictor can effectively model the spatial correlation among neighboring pixels, remaining prediction error will be mostly decorrelated and easily coded with an entropy coder. On the other hand, it is well known that image data are nonstationary, i.e. properties of image regions vary all over the image [2]. Accordingly, it is necessary to adapt the predictor to the changing image characteristics. Another assumption of local stationariness is very well applied to the image data. This means that for arbitrarily small image regions, the predictor adapted to the dominant local property will predict well inside the region. In this paper a new predictive model for image compression is proposed. The model is based on classification and blending of predictors and contextual error modeling. The classification is performed in order to capture and model higher order redundancies inside the local image region. On the causal set of classified neighboring pixels, the set of selected static predictors is dynamically blended to produce the prediction.

2 Proposed method

Proposed lossless compression algorithm is composed of three main steps as shown in Figure 1:(1)Prediction of the current pixel; (2) Contextual modeling of prediction error; and (3) Entropy coding. Image is treated as two-dimensional array $I(x, y)$ of pixel grey intensity values with the width W and the height H, where $0 \leq x < W$ and $0 \leq y < H$. Pixels are observed sample by sample in raster scan order, from top to bottom, left to right. In the assumed backward adaptive approach, the encoder is allowed to use only past information that is also available to decoder. This means that for forming the prediction only previously observed pixels are used, as shown in Figure 1. In fact, only small subset of previously encoded pixels is used to form the causal template. Predictor from Figure 1 uses causal context of surrounding pixels for the prediction of the current pixel :

$$\hat{I}(x, y) = f(\Omega(x, y)) \tag{1}$$

For simple predictors the compass point notation is used. For example N denotes North pixel, W denotes the West pixel from the current pixel, NE North-West, etc. Instead of coding the real pixel value, the pixel prediction $\hat{I}(x, y) = f(\Omega(x, y))$ is performed and the prediction error $e(x, y) = I(x, y) - \hat{I}(x, y)$ is produced. In our approach we treat image as a composition of regions with varying dominant properties such as edges, textures, smooth regions, noisy regions etc. Those properties pose different and conflicting constraints on the prediction function. The main properties of image regions and their requirements on the predictor are given in [3, 4]. If we consider linear prediction function:

$$\hat{I}(x, y) = f\big(\Omega(x, y)\big) = \sum_{I(i,j)\in\Omega} a_{i,j} \cdot I(i, j), \tag{2}$$

*E-mail:josip.knezovic@fer.hr, Tel: +385 1 6129 619

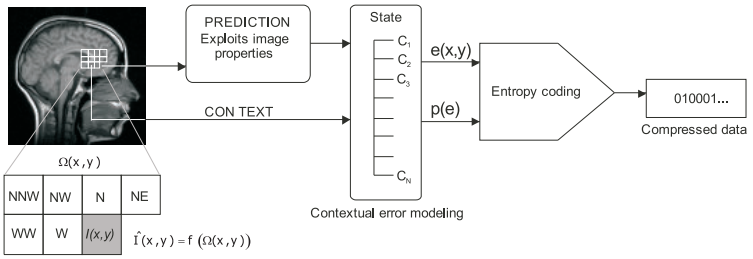
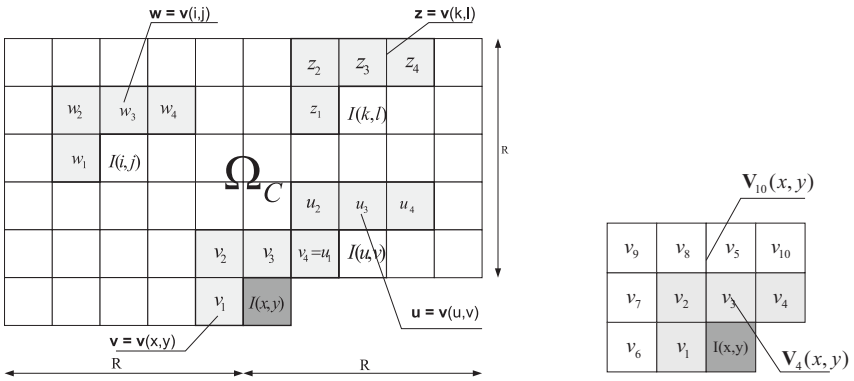


Figure 1. Block scheme of predictive image coding



(a) Causal context (b) Vector template

Figure 2. Contextual elements of proposed predictor

then we can put following constraints on the predictor coefficients $\{a_{i,j}\}$ depending on the dominant property of the current region: 1. **Smooth regions:** The intensity of pixel doesn't change and it is required that $\sum a_{i,j} = 1$. For planar region at least one of $a_{i,j}$ needs to be negative so that the gradient can be estimated; 2. **Noisy regions:** Require that the least magnitude of noise is introduced into prediction, i.e. $\sum |a_{i,j}|$ should be as small as possible. The best approach is to use of averaging prediction function; and 3. **Edges and textured regions:** The most important visual part of images [3]. Edges require adaptation in the predictor for edge detection and orientation. Textures are more difficult to model and can be considered as combination of noise and edges [4]. Image regions can be viewed as composed of structures mentioned before with one dominant property. The effectiveness of any prediction scheme depends on its ability to adapt to dominant properties in image regions. Heuristically tuned switching predictors such as GAP (CALIC algorithm) [5], MED (LOCO-I, JPEG-LS standard) [6] have the drawback which is the lack of robustness in the presence of nontrivial structures such as those found in medical images.

2.1 Proposed Predictor

In order to efficiently model different image structures we propose adaptive predictor based on the idea of predictor blends. The blending predictor is extended with dynamic determination of blending context on a pixel–by–pixel basis. The set of predictors $\mathcal{F} = \{f_1, f_2, \cdots, f_N\}$ is composed of *N* static prediction functions adjusted to predict well in the presence of specific property. For example simple predictor $f_W = I(x-1, y)$ is known to predict well in the presence of sharp horizontal edge. The classification process determines the set of neighboring pixels on which the blending of predictors from $\mathcal{F}$ is performed [4]. Figure 2 depicts basic elements of proposed prediction scheme. Ω_C denotes the causal context used by the predictor. It is a rectangular window of radius *R* composed of previously encoded pixels on which the search procedure for classification is performed. Each pixel from the Ω_C and currently unknown pixel $I(x, y)$ is associated with its vector template $\mathbf{v}(x, y)$ that is composed of *d* closest causal neighboring pixels as shown in Figure 2(b). As an example, the vectors of size $d = 4$ ($\mathbf{v}_4(x, y)$) and $d = 10$ ($\mathbf{v}_{10}(x, y)$) are shown. In Figure 2 the vectors of size four are used. The Euclidean distance between associated vectors will be used for classifications of pixels into cells of *similar* elements, just like in VQ design. In order to reduce the complexity of proposed scheme some basic simplifications are introduced. First, only the *current* cell in which the current pixel lies needs to be calculated. Next, the cell population, i.e. number of pixels that go into the cell together with the current pixel's vector, is set as constant *M* at the beginning of the coding process. Proposed prediction scheme operates as follows:

1. *Classification:* For each pixel $I(i,j) \in \Omega_C$ compute the Euclidean distance $D(i,j)$ between its corresponding vector $\mathbf{v}(i,j)$=$\mathbf{w}$ and the current pixel's vector $\mathbf{v}(x,y)$=$\mathbf{v}$:

$$D(i,j) = ||\mathbf{v}(i,j) - \mathbf{v}(x,y)|| = ||\mathbf{w} - \mathbf{v}|| = \sum_{k=1}^{d} |w_k - v_k|^2 . \tag{3}$$

   Based on the computed distances, determine *M* pixels from Ω_C that belong to the current cell, i.e. with the smallest vector distances from the current pixel's vector. The current cell will be used as blending context Ω_B for $\mathcal{F}$. This step is similar to nearest neighbor selection in VQ design.

2. *Blending:* On the blending context Ω_B perform the blending of the set $\mathcal{F} = \{f_1, f_2, \cdots, f_N\}$ of predictors, as described in [3]. For every predictor f_k the penalty G_k is calculated by the following equation:

$$G_k = \sum_{I(i,j)\in\Omega_B} (\hat{I}_k(i,j) - I(i,j))^2, \tag{4}$$

   where $\hat{I}_k = f_k(i,j)$ is the prediction of $f_k \in \mathcal{F}$ for the pixel $I(i,j) \in \Omega_B$. Based on the penalties we form the prediction for the current pixel $\hat{I}(x,y)$ as:

$$\hat{I}(x,y) = F(x,y) = \left[\frac{\left(\sum_{k=1}^{N} \frac{1}{G_k} \cdot \hat{I}_k(x,y) \right)}{\sum_{k=1}^{N} 1/G_k} \right]. \tag{5}$$

   The prediction for the current pixel is the weighted sum of predictions of all the predictors from $\mathcal{F}$ with weights inversely proportional to corresponding penalties. The penalty of predictor reflects its efficacy on the blending context. If the predictor predicts well, its contribution in the final prediction will be higher. The predictors that do not predict well on the current blending context will eventually be blended out by associated large penalties. The denominator in (5) normalizes the final prediction so that the sum of weights equals to 1.

3. *Error correction:* On the blending context Ω_B calculate *typical* error of the final predictor as:

$$\bar{e}(\Omega_B) = \frac{1}{M} \sum_{I(i,j)\in\Omega_B} (F(i,j) - I(i,j)). \tag{6}$$

   Based on the typical error of blending predictor F the final prediction for the current pixel is further refined as:

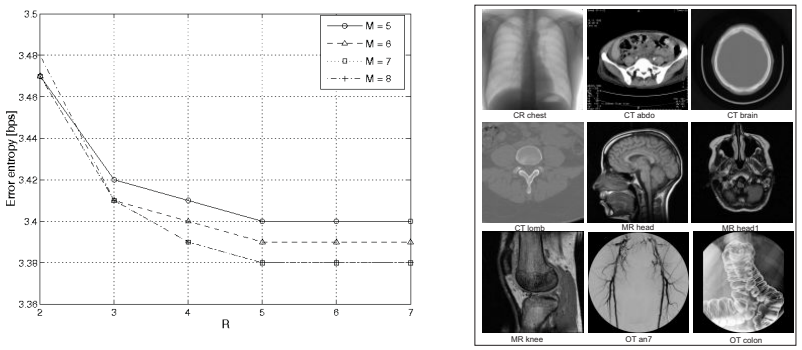
$$\dot{I}(x,y) = F(x,y) + \bar{e} = \hat{I}(x,y) + \bar{e}. \tag{7}$$

   This final step of proposed predictor captures typical bias of the blending predictor F on the classified set of pixels Ω_B that are the part of similar structure as current pixel.

Through the classification and blending process, proposed predictor adjusts itself to the dominant local property. The blending allows other non–dominant properties to be modeled in the prediction, although with less contribution. This is crucial difference compared with switching predictors that don't have the capability to model nontrivial image structures with mixture of properties. Pixels from the search window that do not belong to the region with the same dominant property as the region in which current pixel is will not be included in the current cell and thus they will not be part of the blending context.

2.2 Contextual Error Modeling

Although the prediction step removes statistical redundancies within image data, there are remaining structures in the error image [2]. Those structures are removed using contextual modeling of prediction error, where the context or the state is the function of previously observed pixels, errors or any other relevant variables. As reported by Wu, the heuristic method that uses both, previous pixels template and causal error energy estimate is best suited for this purpose [5]. Wu's contextual model is composed of two different submodels: (1) Model with large number of states that is used for prediction error feedback; and (2) Model with low number of states used for error probability estimation. On the other hand, Wu's predictor is a heuristic predictor with low degree of adaptation and our proposal is highly adaptive predictor with already built in error feedback mechanism (error correction step in the prediction). This implies that we need smaller and less complex contextual model for estimation of error probabilities. Therefore, our error model is built using error discriminant [5]: $\Delta = d_h + d_v + 2|e_w|$, where $d_h = |W - WW| + |N - NN| + |N - NE|$, and $d_v = |W - NW| + |N - NN| + |NE - NNE|$, are horizontal and vertical gradients around the current pixel, and e_w is the prediction error on the west pixel W from the current pixel. Δ is uniformly quantized into eight levels to produce the state of the model [5]. Every state contains the histogram table which is used for probability estimation of the prediction error in the current state.



(a) Zero order entropy vs. radius (b) Representative set of medical images

Figure 3. Average zero order entropy of prediction error

Radius R	Cell size M	Vector size d
5	7	4

Table 1. Working parameters for CBP

3 Experimental Results

In order to evaluate the performances of the proposed predictor we exhaustively run tests on the large set of various 2D medical images [7], [8]. For the purpose of demonstration, we chose a small set of 9 medical images with 8 bits per pixel precision shown in Figure 3(b). Proposed scheme has several parameters that can be varied in order to balance between the compression ratio and the computational complexity. Those are:(1) **Radius** R of the classification context that defines the size of the classification template on which the current cell is estimated (Figure 2). Increasing R, the algorithm has more chance to capture the structure of the current region, but at the cost of higher complexity; (2) **Population size** M of the current cell. Captures the similar pixels to the current pixel for blending of predictors. If this parameter is too small the resulting predictor is overspecialized. On the other hand, higher value of would result in pixels in the current cell that are not similar to the current pixel;(3) **The size of the vector template** d. Vector template is composed of closest already observed pixels. In proposed predictor we experimented with the vector size 1 to 10 as shown in Figure 2(b). The larger vector size, the better capability to capture more complex structures and the larger number of computations. (4) Set of predictors $\mathcal{F}$. Exhaustive tests on a large amount of various medical images had shown that the set $\mathcal{F}$ should contain simple static predictors that are suited to various regions, such as oriented edges, planar regions, smooth regions etc. Our experiments resulted in best choice for set $\mathcal{F}$ to be: $\mathcal{F} = \{N, W, NW, NE, N + W - NW, G_W, G_N\}$, where $G_W = 2N - NN$ and $G_N = 2W - WW$. The set includes simple predictors suited for edges, planar predictor and predictors with gradient modeling for planar regions. Through the blending process the final predictor has the form of averaging predictor which is suited for noisy regions and for modeling higher structures such as textures. Inclusion of additional, more complex predictors in $\mathcal{F}$ resulted in negligible increase of compression efficacy and substantial increase in complexity. Although the set $\mathcal{F}$ was a result from our empirical investigations on an extremely large set of medical images, it reveals the universality of the approach: every predictor from the resulted set is a representative for specific property such as horizontal and vertical edges, planar structures and gradient regions.

Figure 3(a) shows the average zero-order entropy of the proposed predictor on a selected set. We show zero-order entropy in bits per symbol (bps) versus radius of the classification template and with the cell population M as a parameter. Vector template is set to four, since it showed to be the best tradeoff between compression and complexity. As can be seen, the entropy saturates for $R > 6$ and $M > 6$ and increasing those values above results in negligible increase in compression.

Following the results shown in 3(a) and from our experiments, Table 1 shows experimentally chosen working parameters of the CBP predictor that showed to be the best tradeoff between compression and performances for exhaustive set of medical images (over one thousand). Using those values, we designed a complete lossless image compression algorithm that incorporates CBP predictor, contextual error modeling described section 2.2 and arithmetic coding. We call it Classification and Blending Predictive Coder (CBPC). Table 2 shows the results of our CBPC coder compared with the results of popular lossless coders for our test set. We show the bitrate as an average number of bits per pixel

Image	CALIC	JPEG–LS	JPEG200$_R$	CBPC
CR Chest	2.35	2.39	2.52	**2.27**
CT Abdo	2.27	1.89	2.59	**1.92**
CT Brain	1.24	1.29	1.42	**1.10**
CT Lomb	2.21	2.34	2.38	**2.17**
MR Head	4.27	4.44	4.47	**4.18**
MR Head1	4.44	4.62	4.71	**4.33**
MR Knee	4.98	5.08	5.11	**4.92**
OT An7	3.70	3.67	3.97	**3.62**
OT Colon	3.21	3.20	3.45	**3.10**
Average	3.18	3.21	3.40	**3.07**

Table 2. Comparison of coders in achieved bitrate (bpp)

or symbol (bps) required to encode the image. The results correlate well with a complete set of medical images used in our experiments. The first column shows the bit-rates of the CALIC algorithm [2], the second column the JPEG-LS results [6] and the third column shows the results of JPEG 2000 lossless compression that uses reversible wavelet transform [9]. We didn't include the results of dictionary based compression methods because those techniques can't give comparable results for medical image data [10]. The CBPC coder outperformed all other coders by at least 0.11 bps on average. Although this decrease in bitrate seems negligible, it shows its benefit on a large amount of image data.

4 Conclusions

The results obtained with our proposed compression algorithm are encouraging. Although we showed the results on whole images, proposed method can be applied on image regions of interest as well for lossless compression of selected regions [1]. Proposed approach of modeling the image as composed of regions with the mixture of dominant and non–dominant properties has shown to be efficient for lossless image compression. Our predictor is moderately more complex than predictors used in contemporary lossless compression algorithms such as CALIC and JPEG-LS. Also, computational complexity of proposed method can be tuned for both goals: better compression and faster time by changing its parameters on a image by image basis.

References

1. M. Zagar, M. Kovac & I. Bosnic. "Lossless and lossy compression in 4d bio-modeling." In *Proceeding of the First International Conference in Information and Communication Technology & Accessibility*, pp. 271–276. 2007.
2. N. Memon & X. Wu. "Recent developments in context-based predictive techniques for lossless image compression." *IEEE Computer* **10**, pp. 127–136, 1999.
3. T. Seeman & P. Tischer. "Generalized locally adaptive DPCM." In *Proceeding of IEEE IEEE International Conference on Image Procesing*, pp. 473–488. IEEE, 2000.
4. J. Knezovic, M. Kovac & H. Mlinaric. "Classification and blending prediction for lossless image compression." In *Proceeding of IEEE Melecon 2006 Conference*, pp. 486–489. IEEE, 2006.
5. X. Wu. "Lossless compression of continuous-tone images via context selection, quantization, and modeling." *IEEE Transactions on Image Processing* **6**, pp. 656–664, 1997.
6. M. Weinberger, G. Seroussi & G. Sapiro. "he loco-i lossless image compression algorithm: Principles and standardization into JPEG-LS." *HP Laboratories Technical Report HPL-98-198* 1999.
7. B. Meyer. "Tmw page." http://www.csse.monash.edu.au/ bmeyer/tmw/index.html, 2005.
8. S. Barre. "Medical image samples." http://www.barre.nom.fr/medical/samples/, 2006.
9. D. Santa-Cruz & T. Ebrahimi. "An analytical study of jpeg 2000 functionalities." In *Proceeding of IEEE Data Compression Conference*, pp. 49–52. IEEE, 1997.
10. D. Clunie. "Lossless compression of grayscale medical images - effectiveness of traditional and state of the art approaches." In *Proceeding SPIE*. 2000.

Medical Image Segmentation by Water Flow

Xin U Liu, Mark S Nixon

ISIS Group, School of ECS, University of Southampton, UK

Abstract. We present a new image segmentation technique based on the paradigm of water flow and apply it to medical images. The force field analogy is used to implement the major water flow attributes like water pressure, surface tension and adhesion so that the model achieves topological adaptability and geometrical flexibility. A new snake-like force functional combining edge- and region-based forces is introduced to produce capability for both range and accuracy. The method has been assessed qualitatively and quantitatively, and shows decent detection performance as well as ability to handle noise.

1. Introduction

Segmenting anatomical structures from medical images is a fundamental step in analysing medical data. For example, the retina vessel structures can provide useful information like vessel width, tortuosity and abnormal branching which are helpful in diagnoses. Manual delineation of the structures is ineffective and inefficient especially when the number of images is large. Therefore, automatic analysis is needed. However, the complexity and variability of the anatomical features, together with the imperfections of typical medical images such as intensity inhomogeneities and imaging noise which cause the boundaries of structures of interest discontinuous or indistinct, make the automatic segmentation very challenging.

Many methods have been proposed in medical image segmentation. Active contours or snakes [1] are one of the powerful segmentation techniques. An active contour is essentially a parameterized curve which evolves from an initial position to the object's boundary following some rules so that a specified energy functional can be minimized. The methods achieve many desirable features including inherent connectivity and smoothness that counteract object boundary irregularities and image noise. Therefore, they provide an attractive solution to the medical image segmentation problem. However, there are still many limitations. Classical parametric snakes use edge information and need good initialization for a correct convergence. Moreover, they cannot handle topological and geometrical changes like object splitting or merging and boundary concavities. Many methods have been proposed to overcome these problems. The balloon models [2], distance potentials [3], and gradient vector flow (GVF) [4] have been developed to solve the problems of initialization and concave boundary detection. Snake energy functionals using region statistics or likelihood information have also been proposed [5, 6]. A common premise is to increase the capture range of the external forces to guide the curve towards the boundaries. For complex topology detection, several authors have proposed adaptive methods like the T-snake [7] based on repeated sampling of the evolving contour on an affine grid. Geometric active contours [8, 9] have also been developed where the planar curve is represented as a level set of an appropriate 2-D surface. They work on a fixed grid and can automatically handle topological and geometrical changes. However, many methods solve only one problem whilst introducing new difficulties. Balloon models introduce an inflation force so that it can "pull" or "push" the curve to the target boundary, but the force cannot be too strong otherwise "weak" edges would be overwhelmed. Region-based energy can give a large basin of attraction and can converge even when explicit edges do not exist but it cannot yield as good localization of the contour near the boundaries as can edge-based methods. Level set methods can detect complex shapes well but also increase the complexity since a surface is evolved rather than a curve.

Many region merging and growing techniques rely on the assumption that adjacent pixels in the same region have similar characteristics such as intensity, texture or colour. They test the statistics inside different adjacent regions and then decide whether or not they can be merged according to the specified homogeneity criterion. The methods are free of topological changes since they are pixel-wise techniques without smoothness constraints [10]. However, this feature tends to yield irregular boundaries and small holes, especially for noisy images [11]. Besides, the region statistics comparison standards on which they are based can lead to inaccurate contour detection.

Another shortcoming of active contours and region merging methods, both of which are mathematical models, is the difficulty of explaining the abstract operation principles, model variables and parameters. In comparison, physical models have the fundamental advantage that the working principles, input and output are all explicit and material and hence are easy to interpret and debug. The physical model we propose is based on water flow analogy because the features of water like fluidity and surface tension can lead to topological adaptability and geometrical flexibility, as well as contour connectivity and smoothness. In this paper, the force field analogy [12], which is highly robust to noise, is adopted to define the resultant force of water pressure, surface tension and adhesion, instead of the binary convolution method and the gradient of 'clipped' edges [13, 14]. This provides a more solid and complete theoretical

basis for the framework. The method shows encouraging performance on synthetic and real images. A quantitative performance evaluation is also presented to show noise robustness of the new operator, compared with the level set method based on region statistics [6].

2. Water flow model

The new technique aims to detect objects with closed shapes (simple or complicated) accurately whilst retaining a smooth shape. We thus use the analogy to water flow, which is a compromise between several factors: the position of the leading front of a water flow depends on pressure, surface tension, adhesion/capillarity. There are some other natural properties like turbulence and viscosity, which are ignored here. Image edges and some other characteristics that can be used to distinguish objects are treated as the "walls" terminating the flow. The final static shape of the water will describe the related object's contour.

The flow is determined by pressure and the resistance. The relationship between the flow rate f_r, the flow resistance R and the pressure difference, is given by:

$$f_r = (P_i - P_o)/R \tag{1}$$

where P_i and P_o are pressure of the inflow and outflow, respectively. The pressure difference drives the flow and

$$f_r = AV_{effective} \tag{2}$$

where A is the cross-sectional area and $V_{effective}$ is the effective flow velocity. Hence the velocity can be related to force and resistance through equations (1) and (2).

There are small discontinuities or weak regions existent on the contours which may lead to "leakage" of water. The surface tension, which can form a water "film" to bridge gaps, is then applied to overcome the problem. There is an attractive force, named adhesion, existing between water and walls. It is also adopted in the new technique so as to assist surface tension to bridge edge gaps and allow flow into narrow braches.

2.1 Framework of the water flow analogy

One pixel in the image is considered to be one basic unit of the water. The water pressure is defined as the resultant force of the repulsive forces between the water elements. The elements on the water contour, however, are considered to attract other contour elements, and hence generate surface tension. The image edges that are not flooded produce attractive forces and the total attraction is defined as the adhesion.

The flow process is assumed to be made up of two separable steps. The first stage is *acceleration*: the contour element achieves a velocity due to the presence of the water pressure, surface tension and any adhesive force, and the ultimate value is given by equations (1) and (2). The next step is *exterior movement* where the moving element is now free from the influence of other water elements and suffers only external image forces. This is not consistent with real action but is sufficient for the digital image analogy and greatly simplifies the algorithm.

The water element can move outwards in any direction for which the component of velocity is positive. However, only if the velocity is sufficiently large, can the element break through the image resistant forces and 'flood' an adjacent 'dry' position. To reconcile the flow velocity with forces, dynamical formulae are used. Here, the definitions of kinetic energy and resistant work are used. Given an element with mass m, a positive velocity v on a particular direction and is acted by the force F during the process, if the element can arrive at the direction-related position after a displacement S, then this inequality must be fulfilled:

$$FS + mv^2/2 \geq 0 \tag{3}$$

where force F is scalar that is positive when it is consistent with velocity v, and negative otherwise. The inequality means that only if the initial kinetic energy given by water forces could exceed the resistant work given by the image force, can the element finally flow to the target position.

2.2. Force field, water forces, and flow velocity

In this new water flow model, each water element is treated as a particle exhibiting attraction or repulsion to other ones, depending on whether or not it is on the surface. The adhesive force, essentially, is determined by the attractive forces from the adjacent edge points. These facts demonstrate a link between the new model and the force field analogy [12]. Both water elements and the image pixels that are not flooded are treated as arrays of mutually attracted or repelled particles acting as the source of a Gaussian force field. *Gauss's law* is used as a generalization of the inverse square law which defines the gravitational and electrostatic force fields. Denoting the mass value of pixel with position vector $\mathbf{r}_i$ as $L(\mathbf{r}_i)$, we can define the total attractive force at $\mathbf{r}_i$ from other points within the area $\mathbf{W}$ as

$$\mathbf{F}(\mathbf{r}_i) = \sum_{j \in \mathbf{W},\, j \neq i} L(\mathbf{r}_j) \, (\mathbf{r}_j - \mathbf{r}_i) \Big/ |\mathbf{r}_j - \mathbf{r}_i|^3 \tag{4}$$

The equation can be directly adopted in the water flow model, provided the mass values are properly defined. The magnitude of a water element is set to 1, and that of an image point equals to the corresponding edge strength. The water contour points and the image pixels should be set positive, and the interior water elements should be negative because equation (4) is for attractive forces.

From equations (1) and (2), the flow velocity is inversely proportional to the resistance of water. In a physical model, the resistance is decided by the viscosity, the flow channel and temperature etc. Since this is a physical analogy which offers great freedom in selection of parameter definitions, we can relate the resistance definition to certain image attributes. For instance, in retina vessel detection, if the vessels have relatively low intensity, we can define the resistance to be proportional to the intensity of the pixel. If we derive the resistance from the edge information, the process will become adaptive. That is, when the edge response is strong, resistance should be large and the flow velocity should be weakened. Thereby, even if the driving force set by users is too "strong", the resistance will lower its influence at edge positions and the problem in balloon models [1], where strong driving forces may overwhelm "weak" edges, can be suppressed. We first write an equation of velocity by (1) and (2):

$$\mathbf{V}(\mathbf{r}_i) = \mathbf{F}(\mathbf{r}_i) / A \cdot R \tag{5}$$

where $\mathbf{V}(\mathbf{r}_i)$is the resulting flow velocity. In this paper, A is set as a constant, and R at position (u, v) is determined by

$$R = \exp\{k \cdot \mathbf{E}(u, v)\} \tag{6}$$

where $\mathbf{E}$ is the edge strength matrix and k controls the fall of the exponential curve. If we assign a higher value to k, the resistance would be more sensitive to the edge strength, and a lower k will lead to less sensitivity.

2.3. Image force

During the *exterior movement* step, the moving element is acted solely by the image force, which is a combination of edge-based and region-based force functionals. The edge potential force is defined as the gradient of an edge response map since it gives rise to vectors pointing to the edge lines [4]. The force is large only in the immediate vicinity of edges and always pointing towards them. The potential force on a contour element (x_c, y_c) is given by:

$$\mathbf{F}_P = \nabla \mathbf{E}(x_t, y_t) \tag{7}$$

where $\nabla \mathbf{E}$ is the gradient of the edge map, and (x_t, y_t) are the coordinates of the flow target because the potential force is presumed to act during the second stage of flow where the element has left the contour and is moving to the target.

The forces defined above work well as long as the gradient of edges pointing to the boundary is correct and meaningful. However, as with corners, the gradient can sometimes provide useless or even incorrect information. Unlike the method used in the inflation force [2] and T-snake [7], where the evolution is turned off when the intensity is bigger than some threshold, we propose a *pixel-wise* regional statistics based image force. The statistics of the region inside and outside the contour are considered respectively and thus yield a new image force:

$$F_S = -\frac{n_{int}}{n_{int}+1}(\mathbf{I}(x_t, y_t) - \mu_{int})^2 + \frac{n_{ext}}{n_{ext}-1}(\mathbf{I}(x_t, y_t) - \mu_{ext})^2 \tag{8}$$

where subscripts "*int*" and "*ext*" denote inner and outer parts of the water, respectively; μ and n are the mean intensity and number of pixels of each area, separately; $\mathbf{I}$ is the original image. The equation is deduced from the Mumford-Shah functional used by Chan's level set operator [6] and the derivation has been shown in [13, 14].

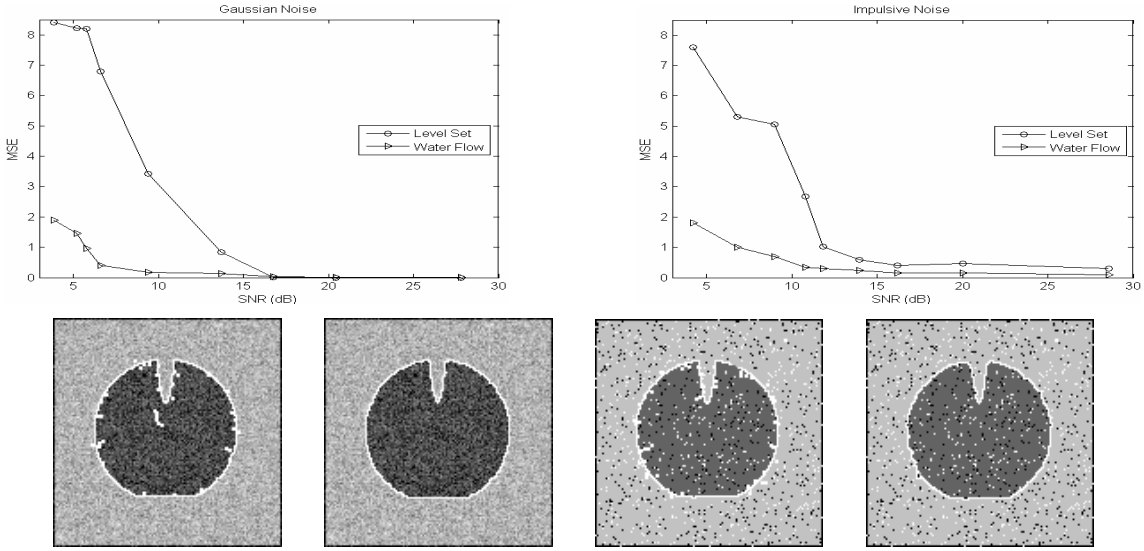
Edge-based forces provide a good localization of the contour near the real boundaries but have limited capture range whilst region-based forces have a large basin of attraction and relatively low detection accuracy. A convex combination method is chosen to unify the two functionals:

$$F = \alpha F_P + (1 - \alpha) F_S \tag{9}$$

where all terms are scalar quantities where a positive value means the direction is from the origin to the target flow position, and α ($0 \le \alpha \le 1$) is chosen by the user to control the balance between them.

2.4. Movement decision process

If the *acceleration stage* gives an initial velocity pointing outwards, inequality (3) will be used to determine whether or not the element could reach the target position in the second flow stage. Each possible target position will be tested separately. As all the terms in inequality (3) are scalars, the velocity given by equation (5) will be decomposed. The scalar terms will be positive if they point from the origin to the target. In inequality (3) and equation (5), there are parameters like m, S and A which are defined as constant. For simplicity, we divide the two sides of inequality (3) by S, combine all the constant parameters and hence get the movement decision processor:



(a) LS (MSE: 0.83) (b) WF (MSE: 0.13) (c) LS (MSE: 1.02) (d) WF (MSE: 0.31)

Figure 1. Quantitative evaluation and detection examples for level set method (LS) and water flow operator (WF).

$$\lambda v^2 + F \geq 0 \tag{10}$$

where λ, the combination of mass m, displacement S and area A, is a regularization parameter set by users controlling the trade-off between the two energy terms. In total, if the initial flow velocity v is positive, and inequality (10) can be satisfied, the water element will move to the corresponding direction and flood the target position.

3. Experimental Results

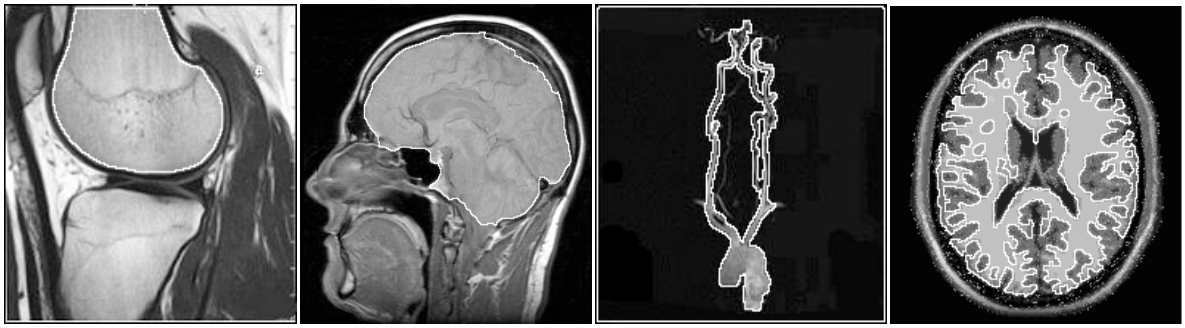
3.1. Quantitative evaluation

A quantitative result assessment is performed as shown in figure 1. The level set method based on region statistics [6] is chosen for comparison. The test image is synthetic in order that the ideal result can be achieved. The object of interest is designed with a boundary concave to increase the detection difficulty. Different levels of Gaussian and Impulsive noise are added. The mean square error (MSE) is used to measure the performance under noise.

$$MSE = \sum_{k=1}^{I_D} d_k^{\,2} \Big/ \max(I_D, I_I) \tag{11}$$

where I_I and I_D are the number of ideal and detected contour points respectively and d_k is the distance between the kth detected contour point and the nearest ideal point. For both noise, the performance of the water flow model is markedly better than the level set operator especially when the noise contamination is severe (SNR less than 10dB). The performance superiority of the new operator under noisy conditions is further illustrated qualitatively by the segmentation results for Gaussian noise (SNR: 13.69) and impulsive noise (SNR: 11.81), see figure 1 (a) to (d). This robustness to noise is desirable for medical image segmentation as discussed in Section 1.

3.2. Natural medical images



(a) (b) (c) (d)

Figure 2. Segmentation results in real medical images: a) femur in MR knee image, b) brain in a sagittal MR image, c) carotid artery in a MR carotid MRA image and, d) grey/white matter interface in MR brain image slice.

(a) (b) (c)

Figure 3. Segmenting vessels in retinal images with low resolution and quality (k=50, α=0.5, λ=1)

We applied the new model to natural medical images. Figure 2 presents the segmentation results for MR images. The water sources are all set inside the object of interest and parameter are chosen as k=50, α=0.5, λ=1. The resultant contours are relatively smooth by virtue of surface tension. The operator can find weak-contrasted boundaries as shown by figure 2 (b) where the indistinct interface between the brain and the spine is detected. This is achieved by combining a high value of k that gives the operator a high sensitivity to edge response and the region-based forces. The fluidity of water leading to topological adaptability and geometrical flexibility is fully realised. Figure 2 (c) and (d) illustrate this – the complex structures and irregular branches are segmented successfully.

Retinal vessel segmentation plays a vital role in medical imaging since it is needed in many diagnoses like diabetic retinopathy and hypertension. The irregular and complex shape of vessels requires the vessel detector to be free of topology and geometry. Furthermore, digital eye fundus image often have problems like low resolution, bad quality and imaging noise. The water flow model is a natural choice. Figure 3 shows the segmentation results. In figure 3 (a), only one water source inside the vessels within the optic disk is initialised. We can see that the major vessel structure has been detected, and the result is a single, connected structure. By setting multiple initialisations, more complete results can be achieved as shown in figure 3 (b) and (c). In figure 3 (b), multiple flows of water merged, leading to a single vessel structure. In 3 (c), some flows merged and some remained separated. This can be improved by post-processes like gap-linking techniques.

4. Conclusion

This paper combines the water flow model with force field analogy. This provides the physical framework a more theoretically sound and complete basis. The desirable features of water such as fluidity and surface tension are gained, leading to topological adaptability, geometrical flexibility, and segmentation smoothness. The immunity to noise which is enhanced by embodying force field paradigm also helps to make it suitable for medical image processing. The application of the model to MR and eye fundus images illustrates great potential of the method. Note that this is a general framework and can be combined with more refined edge/region detectors to achieve better performance, and the physical feature of the model makes it straightforward to be extended to the segmentation of 3D volumes.

References

[1] M. Kass, A. Witkin and D. Terzopoulos, "Snakes: Active contour models," *Int'l. J. of Comp. Vision,* 1(4):321-331, 1988

[2] L. D. Cohen, "On active contour models and balloons," *CVGIP, Image Understanding,* 53(2): 211-218, 1991.

[3] L. D. Cohen and I. Cohen, "Finite element methods for active models and balloons for 2-D and 3-D images," *IEEE Trans. PAMI,* 15: 1131-1147, 1993.

[4] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Trans. Image Processing,* 7(3): 359-369, 1998.

[5] M. Figueiredo and J. Leitao, "Bayesian estimation of ventricular contours in angiographic images," *IEEE Trans. Medical Imaging,* 11: 416-429, 1992.

[6] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Processing,* 10: 266-276, 2001.

[7] T. McInerney and D. Terzopoulos, "Topologically adaptive snakes," *Int'l Conf. Computer Vision,* pp. 840-845, 1995.

[8] V. Casselles, R. Kimmel, and G. Spiro, "Geodesic active contours," *Int'l Journal of Computer Vision,* 22(1):61-79, 1997.

[9] R. Malladi et al., "Shape modeling with front propagation: A level set approach," *IEEE Trans. PAMI,* 17: 158-174, 1995.

[10] R. Adams, and L. Bischof, "Seeded region growing," *IEEE Trans. PAMI,* 16(6): 641-647, 1994.

[11] S.C.Zhu and A.Yuille, "Region competition: unifying snakes, region growing, and Bayes/MDL for multi-band image segmentation," *IEEE Trans. PAMI,* 18(9): 884-900, 1996.

[12] D. J. Hurley, M. S. Nixon, and J. N. Carter, "Force field feature extraction for ear biometrics," *CVIU,* (98): 491-512, 2005.

[13] X. U. Liu and M. S. Nixon, "Water flow based complex feature extraction," *Acivs 2006,* pp. 833-845, 2006.

[14] X. U. Liu and M. S. Nixon, "Water flow based vessel detection in retinal images," VIE 2006, pp. 345-350, 2006.

Spatio-Temporal Segmentation of Left Ventricle in Real Time 3D Echocardiographic Images Using Phase Information

Weiwei Zhang[a], Chunming Li[b], J. Alison Noble[a] and J. Michael Brady[a] *

[a]Wolfson Medical Vision Lab, Department of Engineering Science, University of Oxford
Parks Road, Oxford OX1 3PJ, United Kingdom,
[b]Institute of Imaging Science, Vanderbilt University, Nashville, TN 37232, USA

Abstract. We present a novel variational level set method to automatically segment the left ventricle (LV) in real time 3D echocardiographic images. The motivation for our research is to assist a clinician to track the motion of the LV in a cardiac sequence for the diagnosis of heart disease. The image phase gradient (rather than the intensity gradient) is employed as the edge indicator function. The motion continuity across the cardiac sequence is constrained by a region-weighted parameter in the external force. Incorporating image local phase into the variational level set method without re-initialization proposed in [1] enables an implementation with a simple finite difference scheme and allows for flexible and efficient initialization. The preliminary results on one volunteer dataset are shown in a qualitative and quantitative manner.

1 Introduction

Real-time three-dimensional echocardiography (RT3DUS) is receiving a lot of attention at the current time for cardiac diagnosis and therapy. Unlike previous 3D echocardiography modes: rotational probes and free hand probes, third generation 3D echocardiography is a truly real-time acquisition modality which generates a pyramidal volume data set with a 2D array of transducer elements. Although RT3D has lower spatial resolution than cardiovascular MR (CMR), ultrasound machines are portable and relatively cheap; acquisition times are fast, hence there is a drive to find roles for this imaging modality in clinical practice.

Tracking the structure and motion of the left ventricle (LV) is of great clinical interest in the quantitative assessment of heart function. Early segmentation approaches focused on single frame (2D) endocardial border tracking to derive measures such as the ejection fraction for the diagnosis of heart diseases. However, since the heart moves non-rigidly and the speckle pattern associated with the deforming tissue decorrelates across the cardiac cycle, it is natural and logical to view segmentation of the LV as a spatio-temporal problem [2]. For example, Mulet-Prada and Noble [3] proposed to use a feature asymmetry measure derived from the local phase to automatically detect the endocardium in a spatio-temporal space. Friedland et al. [4] represented the contour using polar coordinates and modeled the temporal continuity using a Markov random field (MRF). The tracking problem was formulated as the minimization of an energy function and optimized by a simulated annealing algorithm. Dias et al. [5] proposed a similar algorithm but took into account the speckle statistics in joint endocardial and epicardial border detection. A spatio-temporal version of Deriche edge detection was developed by Herlin and Ayache [6] to track the cardiac boundary. Active contour has been proposed as an alternative approach for tracking the LV boundaries. Mikic et al. [7] proposed a knowledge-guided segmentation approach in which the optical flow estimates were used to guide the propagation of the fitted curve from one frame to another. Chalana et al. [8] employed an active surface model to represent the LV endocardial and epicardial borders as two planar curves. The image gradient was used as the attracting force, while the motion between consecutive frames was constrained by an external energy term. Lin et al. [9] developed a novel variant of level set method that combined the region and edge information across spatial scales but required a well located curve extraction at a high pyramid level.

In this paper, we propose a novel variational level set approach to segment the LV in a RT3D echocardiograhic sequence. The image phase gradient is used as the external force for the contour to converge to the exact LV boundaries and motion continuity across the cardiac cycle is constrained by a region-weighted parameter, provided that the maximum contraction frame is detected by cross-correlation. A penalizing energy that can completely eliminate the re-initialization step is employed as the internal energy. The proposed method has been implemented in a simple numerical solution with flexible initialization of the level set function. Experimental results show its ability to spatio-temporally segment the LV boundaries in both short axis (SA) and long axis (LA) views of RT3D echocardiographic images.

* {weiwei,noble,jmb}@robots.ox.ac.uk and chunming.li@vanderbilt.edu

2 Segmentation Method

The level set method [10] is a useful tool for image segmentation. The key idea behind the method is to represent the moving front $\mathcal{C}$ as the zero level set $\mathcal{C}(t) = \{(x, y)|\phi(x, y, t) = 0\}$ for a higher dimensional level set function $\phi(x, y, t)$. In the variational level set method [11], the curve evolution equation is the gradient flow of the level set function that minimizes certain energy functionals:

$$\frac{\partial \phi}{\partial t} = -\frac{\partial \mathcal{E}}{\partial \phi} \tag{1}$$

where $\frac{\partial \mathcal{E}}{\partial \phi}$ denotes the Gateaux derivative (or first variation) of the functional $\mathcal{E}(\phi)$. The variational method is superior to the pure PDE-driven method not least with respect to incorporating prior knowledge (region-based or shape-prior information) into the functional. In this paper, we propose the following total energy functional to be minimized:

$$\mathcal{E}(\phi) = \mathcal{E}_\psi(\phi) + \mathcal{P}(\phi) \tag{2}$$

where $\mathcal{E}_\psi(\phi)$ is the external term based on local phase information to attract the level set towards the object boundaries and $\mathcal{P}(\phi)$ is the internal term that penalizes the deviation of ϕ to a signed distance function. Due to the penalizing effect of the internal term, the gradient flow that minimizes the functional $\mathcal{E}(\phi)$ is able to enforce the signed distance property of ϕ during the evolution and the re-initialization step can be completely eliminated (see section 2.2).

2.1 Phase Gradient as the Edge Indicator Function

Most edge-based segmentation approaches depend on an external force to stop the level set evolving at the nearest maxima of the input image gradient. This requires that the image be presmoothed (by a Gaussian kernel) at a certain spatial scale in order to create a large basin of attraction for the level set to converge, which is usually defined by an edge indicator function [1]. However, the Gaussian convolution blurs the location of edges so that the contour is not well attracted to the exact object boundaries. In particular, in RT3D echocardiographic images, the LV boundary is usually not well identified due to speckle orientation dependence of acquisition and intensity dropouts in the shadow zones. A purely intensity-based approach is rarely successful in segmenting the LV. The local phase is an alternative image descriptor that better describes a detected image feature, such as an edge or ridge in a signal (rather than signal magnitude) [12]. It has previously been advocated as a good representation for ultrasound image analysis [2, 3]. In this paper, we propose to use the image phase gradient instead of the intensity gradient as the edge indicator function g_ψ:

$$g_\psi = \exp(-|\nabla \psi(x, y)|), \qquad \psi(x, y) = \tan^{-1}\left(\frac{b}{\sqrt{(h_1 \otimes b)^2 + (h_2 \otimes b)^2}}\right) \tag{3}$$

where ψ is the local phase estimated using the monogenic signal [13] and b is the bandpass filtered image $b = f * I$. In our work, the log-Gabor filter is chosen as the bandpass filter f. h_1 and h_2 are a pair of vector valued filters that are odd and distributed with isotropic energy in the frequency domain [13].

The external energy based on the local phase information can be defined as:

$$\mathcal{E}_\psi(\phi) = \lambda \mathcal{L}_{g_\psi}(\phi) + \nu \mathcal{A}_{g_\psi}(\phi) = \lambda \int_\Omega g_\psi \delta(\phi) |\nabla \phi| dxdy + \nu \int_\Omega g_\psi H(-\phi) dxdy \tag{4}$$

where $\lambda > 0$ and ν are the constants; δ and H are the Dirac function and Heaviside functions respectively. The geometric meaning of the energy $\mathcal{L}_{g_\psi}(\phi)$ is the length of the zero level curve of ϕ in the conformal metric. The energy functional $\mathcal{A}_{g_\psi}$ can be considered as the weighted region of $\Omega_\phi^- = \{(x, y)|\phi(x, y) < 0\}$. Depending on the relative position of the initial contour, the coefficient ν can be defined as a positive or negative value to control the movement of the contour. In our case, the motion of the heart in the cardiac cycle can be divided into two main stages: diastole (expansion) and systole (contraction). In order to track the LV boundaries automatically in a sequence, we first use cross-correlation to detect the maximum contraction frame [14]. If the initial contour is placed outside the LV, the coefficient ν can be set to a positive value, so that the contour can shrink toward the LV boundaries from the beginning of the sequence to the maximum contraction frame. After the maximum contraction frame, the coefficient ν should be set to a negative value in order that the contour can expand till the end of the sequence.

[1] $g = \frac{1}{1+|\nabla G_\sigma * I|^2}$ where I is the input image and ∇G_σ is the Gaussian kernel with standard deviation σ (or scale).

2.2 Level Set Evolution Without Re-Initialization

In the traditional level set method, a re-initialization step is necessary and unavoidable in order to maintain a stable evolution. The solution is to keep the evolution of the level set function close to a signed distance function,[2] especially in a neighborhood around the zero level set. However, in practice, the re-initialization process is computationally expensive and breaks the agreement between the theory of the level set method and its implementation [15]. Li et al. [1] proposed an energy functional with an internal energy term that penalizes the deviation of the level set function to a signed distance function. Therefore, the signed distance property can be maintained in the evolution, which is the gradient flow that minimizes the energy functional, and hence there is no need for re-initialization. It is well known that the key property of a distance function is $|\nabla\phi| = 1$. Any function that satisfies $|\nabla\phi| = 1$ is equal to a signed distance function plus a constant. Therefore the penalizing energy can be defined by:

$$\mathcal{P}(\phi) = \mu \int_{\Omega} \frac{1}{2}(|\nabla\phi| - 1)^2 dxdy \tag{5}$$

where μ is a positive constant. Equation (5) is actually a metric that characterizes how a function ϕ approximates a signed distance function in $\Omega \in \Re^2$. The effect of equation (5) is to automatically maintain the level set function as an approximate signed distance function during the evolution of ϕ, so that the re-initialization is no longer necessary.

In summary, the total variational functional $\mathcal{E}$ is now updated to combine the external and internal energy as outlined in the above sections. To minimize the energy functional $\mathcal{E}(\phi)$, we use the standard gradient descent (also called steepest descent) method by solving the gradient flow equation:

$$\frac{\partial\phi}{\partial t} = \lambda\delta(\phi)\text{div}(g_\psi \frac{\nabla\phi}{|\nabla\phi|}) + \nu g_\psi \delta(\phi) + \mu[\triangle\phi - \text{div}(\frac{\nabla\phi}{|\nabla\phi|})] \tag{6}$$

where $\triangle$ is the Laplacian operator, div is the divergence operator and $\text{div}(\frac{\nabla\phi}{|\nabla\phi|})$ computes the mean curvature of the level contours.

3 Implementation

3.1 Initialization of Level Set Function

Since the re-initialization step is eliminated in our approach, the level set function is no longer required to be initialized as a signed distance function. Thus it allows for a more flexible initialization scheme. Following the efficient initialization scheme proposed in [1], we initialize the level set function as:

$$\phi_0(x,y) = \begin{cases} -\rho, & (x,y) \in \Omega_0 - \partial\Omega_0 \\ 0 & (x,y) \in \partial\Omega_0 \\ \rho & \Omega - \Omega_0 \end{cases} \tag{7}$$

where Ω_0 is a subset of the image domain Ω and $\partial\Omega_0$ is the boundary of Ω_0. The ρ is a positive constant larger than 2ε, where $\varepsilon = 1.5$ is the width of the regularized Dirac function: $\delta_\varepsilon = 0$, when $|x| > \varepsilon$ and $\delta_\varepsilon = \frac{1}{2\varepsilon}[1 + \cos(\frac{\pi x}{\varepsilon})]$, when $|x| \leq \varepsilon$.

3.2 Numerical Implementation

Equation (6) can be numerically implemented by defining the iterative update equation:

$$\phi_{i,j}^{k+1} = \phi_{i,j}^k + \tau L(\phi_{i,j}^k), \qquad L(\phi_{i,j}^k) = \frac{\phi_{i,j}^{k+1} - \phi_{i,j}^k}{\tau} \tag{8}$$

where τ is a time step, and $L(\phi_{i,j})$ is the approximation of the right hand side in equation (6) by using the central difference scheme to estimate spatial partial derivatives. There is a tradeoff between the time step and accuracy in boundary tracking. The larger the time step, the more quickly the contour evolves and the more errors in the boundary location. Moreover, in the numerical implementation of equation (6) the time step τ and the coefficient μ must satisfy $\tau\mu < 1/4$ in order to ensure a stable iteration. We have empirically determined that the combination of the parameters $(\lambda = 6, \mu = 0.2, \nu = \pm 2.5, \tau = 1)$ yields a good segmentation result.

[2] $\frac{\partial\phi}{\partial t} = sign(\phi_0)(1 - |\nabla\phi|)$ where ϕ_0 is the function to be re-initialized, and $sign(\phi)$ is the signed distance function.

4 Results

One volunteer dataset has been used to date in the experiments. The RT3D echocardiography was acquired using a Philips Sonos 7500 and contained 18 frames in one cardiac cycle for the volunteer. The mid short axis (SA) and mid long axis (LA) slices were taken from a RT3D apical acquisition.

Figure 1 (top row) shows the results for the SA view over one cardiac cycle (Due to the paper space constraint, we selected number 1, 4, 8, 11, 15 and 17 frames in the figure). The initial contour was roughly placed outside the LV. The final contour from one frame is automatically applied to the next frame as the initial contour. The ν in equation (4) is defined as $\nu = 2.5$ for the first frame to the eighth frame and changed to a negative value after the maximum contraction frame. Figure 1 (bottom row) shows the results on the LA view over one cardiac cycle. Figure 2 shows the 3D segmentation results on the ED and ES volumes, which were obtained from multi-slice segmentation by the proposed method.

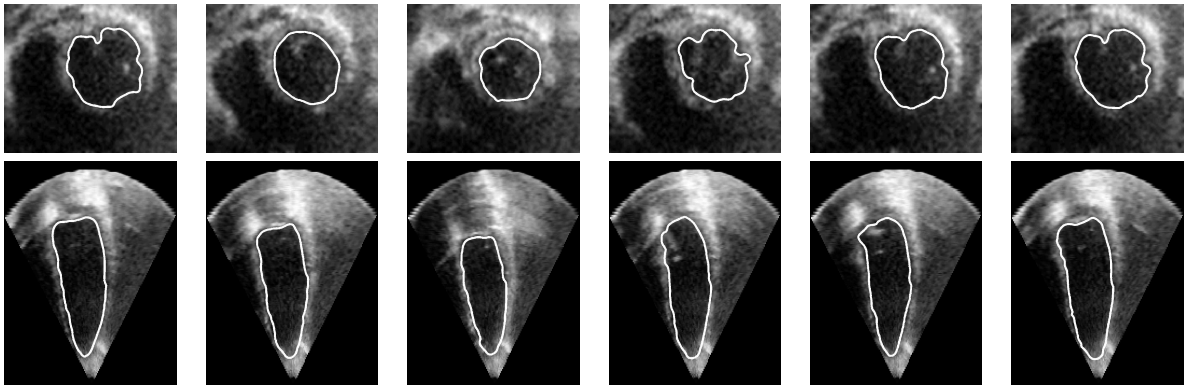


Figure 1. Result on the LV segmentation of SA and LA views over one cardiac cycle.

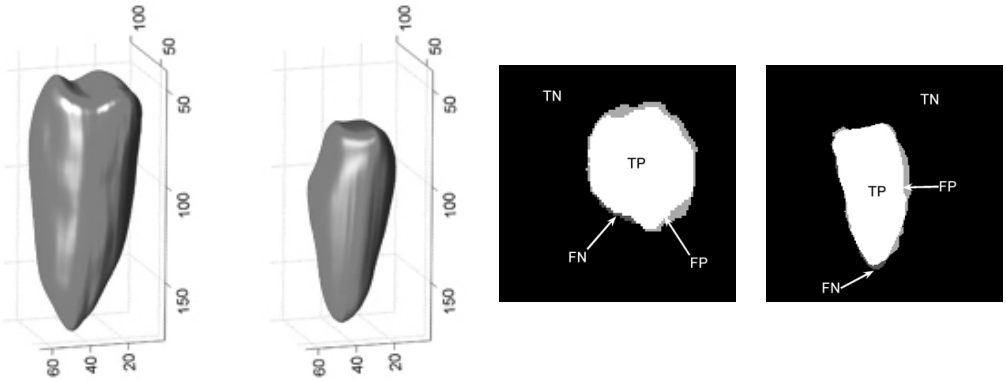


Figure 2. Extension to the 3D segmentation of the ED and ES volumes and an example of the validation method.

5 Validation

We validated our proposed segmentation method by comparing the results with a set of the LV contours in the cardiac cycle that were manually traced by an expert cardiologist. After the segmentation, the image domain can be divided into two regions: $k = \{\Omega_{seg} \in \Omega\}$ where the pixels are inside the LV and $q = \{\Omega - \Omega_{seg}\}$ where the pixels are outside the LV (or the image background). Therefore, the Dice Similarity Coefficient (DSC) can be used to evaluate the performance of the local phase based segmentation method [16]:

$$DSC = \frac{2 \cdot |\mathcal{C}_{Auto} \cap \mathcal{C}_{Manual}|}{|\mathcal{C}_{Auto}| + |\mathcal{C}_{Manual}|} = \frac{2 \cdot TP}{(TP + FN) + (TP + FP)} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{9}$$

where $\mathcal{C}_{Auto}$ denotes the LV area segmented by the proposed automatic approach and $\mathcal{C}_{Manual}$ denotes the LV area segmented by the expert manual segmentation. We also define the true positive (TP) as the pixels correctly assigned

to k by both methods; the true negative (TN) as the pixels correctly assigned to q by both methods; the false positive (FP) as the pixels correctly assigned to k by the manual segmentation but assigned to q by the proposed segmentation and the false negative (FN) as the pixels correctly assigned to q by the proposed segmentation but assigned to k by the manual segmentation. The mean values of DSC are 92.83% and 93.85% for the SA and LA views respectively.

6 Discussion

In this paper, a novel variational level set method is presented to automatically track the LV boundary in a RT3D echocardiographic sequence. The preliminary results have shown that the edge indicator function using the local phase gradient can well preserve the edge information despite the strong presence of speckle noise so that the contour can arrive at the exact boundary. Apart from the elimination of re-initialization step, one advantage of incorporating the penalizing energy is to allow for a flexible initialization scheme, such as registration-guided initialization of the level set function. Since the segmentation of other modality cardiovascular image such as cardiovascular MR (CMR) is relatively easy, co-registration of CMR image with echocardiographic image can provide a rough estimate for the initial level set function to efficiently converge to object boundaries. In the future, we will investigate the co-registration of CMR and RT3D echocardiographic images to assist the LV segmentation and the clinical validation and application.

References

1. C. Li, C. Xu, C. Gui et al. "Level set evolution without re-initialization: a new variational formulation." *In. Proc. CVPR'05* **Vol 1**, pp. 430–436, 2005.
2. J. A. Noble & D. Boukerroui. "Ultrasound image segmentation: A survey." *IEEE. Trans. Med. Imag.* **25**, pp. 987–1010, 2006.
3. M. Mulet-Parada & J. A. Noble. "2d+t acoustic boundary detection in echocardiography." *Med.Imag.Anal* **40**, pp. 21–30, 2000.
4. N. Friedland & D. Adam. "Automatic ventricular cavity boundary detection from sequential ultrasound image using simulated anneal." *IEEE. Trans. Med. Imag.* **8**, pp. 344–353, 1989.
5. J. Dias & J. Leitao. "Wall position and thickness estimation from seqences of echocardiograms images." *IEEE. Trans. Med. Imag.* **15**, pp. 25–38, 1996.
6. I. L. Herlin & N. Ayache. "Features extraction and analysis-methods for sequences of ultrasound images." *Med. Imag. Anal.* **10**, pp. 673–682, 1992.
7. I. Mikic, S. Krucinski & J. D. Thomas. "Segmentation and tracking in echocardiographic sequences: active contour guided by optical flow estimates." *IEEE. Trans. Med. Imag.* **17**, pp. 274–284, 1998.
8. V. Chalana, D. T. Linker, D. R. Haynor et al. "A multiple active contour model for cardiac boundary detection on echocardiagraphic sequences." *IEEE. Trans. Med. Imag.* **15**, pp. 290–298, 1996.
9. N. Lin, W. C. Yu & J. S. Duncan. "Combinative multi-scale level set framework for echocardiograhic image segmentation." *Med. Imag. Anal.* **7**, pp. 529–537, 2003.
10. O. Osher & J. A. Sethian. "Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulation." *J. Com. Phys.* **79**, pp. 12–49, 1988.
11. T. Chan & L. Vese. "Active contour without edges." *IEEE. Trans. Imag. Proc.* **10**, pp. 251–269, 2001.
12. M. Mellor & J. M. Brady. "Phase mutual information as a similarity measure for registration." *Med. Imag. Anal.* **9**, pp. 330–43, 2005.
13. M. Felsberg & G. Sommer. "A new extension of linear signal processing for estimating local properties and detecting features." *In Proc. DAGM.* pp. 195–202, 2000.
14. D. Perperidis, R. H. Mohiaddin & D. Rueckert. "Spatio-temporal free-form registration of cardiac mr image sequences." *Med. Imag. Anal* **9**, pp. 441–456, 2005.
15. J. Gomes & O. Faugeras. "Reconciling distance functions and level sets." *J. Visiual Commuinc. and Imag. Representation.* **11**, pp. 209–223, 2000.
16. L. R. Dice. "Measures of the amount of ecologic association between species." *Ecology* **26(3)**, pp. 297–302, 1945.

Automated detection of hyper-fluorescent leakage in retinal angiographic sequences by temporal Analysis

C. R. Buchanan[a] and E. Trucco[b*]

[a]Joint Research Institute in Signal and Image Processing,
School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK
[b]School of Computing, University of Dundee, UK

Abstract. We present an image analysis algorithm to detect vascular leakage in retinal fluorescein angiographic image sequences. The key contribution is a two-step spatio-temporal detection technique exploiting the evolution of intensity levels in hyper-fluorescent leakage regions. The rationale is that where haemorrhaging is present there will be a characteristic evolution of the intensities over time, in the simplest case an increase of fluorescein in later frames in comparison to areas of normal perfusion. The specific nature of the intensity profiles is determined using an *AdaBoost* learning algorithm. Such a classifier is trained using a varied set of ground-truth image sequences and tested on unseen images.

1 Introduction

This paper presents an algorithm for detecting vascular leakages in retinal fluorescein angiographic (FA) sequences by exploiting the evolution of intensity levels over time. Vascular leakage in the retina is a serious condition that may be caused by vascular disorders such as diabetic retinopathy and retinal vein occlusions. An automated method to detect the resulting haemorrhages could have the advantage of being more objective than manual inspection and being repeatable and reproducible.

Our algorithm adopts the *plausible detection* paradigm introduced by Trucco and Kamat [1], whereby a detection task is approached by first collecting a set of acceptable (plausible) candidates, including possible false positives but avoiding false negatives, then applying specific knowledge to filter out false positives. Our study makes use of image sets obtained from the Optos Panoramic 200MA scanning laser ophthalmoscope in fluorescein angiography mode. The crucial advantage of this imaging modality is that it captures up to a 200°field of view of the retina, much greater than conventional fundus cameras. The images have dimensions of 3900×3072 pixels with 256 grey levels. The wide field of view is valuable to clinicians, particularly when identifying regions of interest in the retinal periphery [2].

This paper is organised as follows. Previous work is summarised in Section 2, our algorithm is outlined in Section 3, results are presented and discussed in Section 4 and finally we conclude with a summary of our work.

2 Related Work

To our knowledge, all previous approaches to leakage detection in retinal fluorescein angiograms have followed a simple concept; that of analysing the temporal change in intensity between an *early* and a *late* frame. The Ophthalmic Imaging Research Group at Aberdeen University developed an automated method to detect macular leakage in fundus images by this principle [3–5]. An image pair is aligned by semi-automatic registration and the macular region analysed on a pixel-by-pixel basis using grey level thresholding and region growing. They detect and quantify leakages into five levels of severity. A majority voting scheme is also proposed using three pairs of early and late frames [4, 5]. The most recent paper [5] incorporates a principled method for correcting uneven background illumination, which respects macular shading and significant retinal structures to allow for more robust results. Likewise, the work of Martinez-Costa et al. [6] aligns an early-late image pair and then detects pixels with a high difference in grey level. This work exclusively analyses the central foveal area for macular leakage. The fovea centre is manually located in order to define a region-of-interest, with concentric circular areas representing severity of leakage centred on the fovea.

However, only using a pair of frames (or three pairs in the case of majority voting) overlooks considerable information. As a result such methods are limited by the initial image pair selection, requiring that both images are of good quality with minimal artefacts or occlusions; should this not be the case, the result could be misdetections. An approach deriving information from multiple frames should be more robust to intermittent occlusions and artefacts. Whereas previous studies have used models of fluorescein decay in their analysis, our approach is to estimate the characteristic

*e-mail: c.r.buchanan@hw.ac.uk

intensity change over time from real data. Cree et al. [5] constructed a realistic exponential model of fluorescein decay, whereas Phillips et al. [4] employed a simple linear model. In contrast, our technique captures the changes in intensity from real FA sequences with and without leakage. A learning algorithm automatically selects the criteria to distinguish between leakage and non-leakage regions. Furthermore, previous work has assumed a clear increase of fluorescence in the late image compared to the early frame. However, this simple criterion will not necessarily allow to distinguishing between hyper-fluorescent leakage and potential false objects such as late filling vasculature, scars from laser surgery and late 'staining' of the optic disc. Our proposed scheme to exploit the evolution of intensity over time should be able to distinguish better between these false objects.

3 Methodology

We aim to locate regions of hyper-fluorescent leakage on the surface of the retina. We do not intend to distinguish between different sources of leakage, e.g. incompetent vessels, neovascularisation. The algorithm is outlined in Figure 1 and detailed in the following sections.

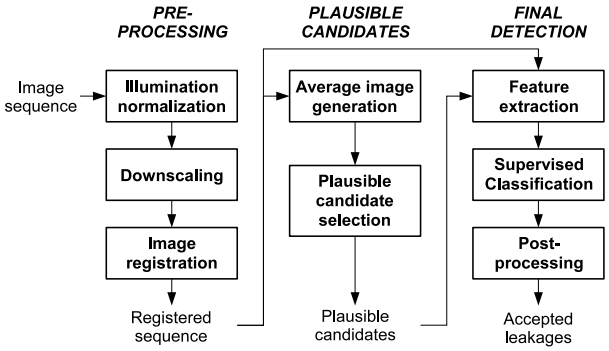


Figure 1. System architecture

In summary, the pre-processing phase consists of downscaling frames followed by alignment. Then plausible detection acts in two steps. The first step consists of computing the average intensity frame and locating plausible leakage candidates. The purpose of the second step is to apply contextual knowledge (intensity change over time) to filter out as many false positives as possible. The evolution of the image intensity of leakage regions follows a characteristic pattern which is, however, difficult to capture with a mathematical model. This pattern can be learned by a supervised classifier and used to distinguish true leakage regions from false positives included by the first step.

An FA image sequence $(I_1, \ldots, I_n)$ has associated capture times $(T_1, \ldots, T_n)$ from injection of the fluorescein dye.

3.1 Pre-processing

Illumination normalisation

Simple illumination normalisation is performed to compensate partially for effects of varying illumination and instrument adjustment. Each pixel value is normalised, $I(x, y) = (I(x, y) - \mu)/\sigma$, where μ is the mean value of the image pixels and σ the standard deviation over the whole frame. The normalised frames are then used in all further analysis.

Downscaling

To reduce computational cost, frames are downscaled by a factor of four using bicubic interpolation. Fourfold downscaling does not destroy the intended targets. In addition, the impact of sensor noise is reduced.

Image registration

The frame alignment algorithm considers each pair (I_i, I_{i+1}) of consecutive frames and matches control points (vessel bifurcations) between image pairs. In brief, a vessel mask is constructed for each frame by applying matched filtering to enhance linear structures followed by grey-level thresholding. Bifurcations are located by skeletonisation and application of rules in an 8-connect pixel neighbourhood over the skeleton image [7]. A bifurcation matching algorithm then searches for corresponding pairs of control points. Correspondences are computed by contextual feature matching; attributes include branching angles, vessel widths, vessel cross-profiles and the displacement vectors to neighbouring bifurcations. Finally, a global affine transform identified by the correspondences is estimated by least squares and applied. Rigid registration provides an approximate solution sufficient for this initial study but which ought to be replaced

by a more sophisticated scheme offering improved registration accuracy.

Any frame which cannot be reliably registered with its predecessor (i.e., the number of matched correspondences does not exceed a minimum threshold) is discarded from the sequence. This type of registration failure occurs with large patient motion or in the presence of significant ciliar occlusions. This strategy helps to remove such unusable frames from further analysis.

3.2 Step 1: Plausible candidates

Average image generation

In this step we consider only frames occurring after 30 seconds, after which hyper-fluorescence begins to appear. A pixel-wise average over such frames $(I_1, \dots I_n)$ is performed to obtain a composite image, C. We record the mean intensity value rather than the maximum to avoid bright, intermittent artefacts (eyelids). Each pixel of C is calculated

$$C(x, y) = \sum_{i=1}^{n} I_i(x, y)/n \tag{1}$$

Plausible candidate selection

Any regions of hyper-fluorescent leakage and possible false targets will appear with high intensity in the composite image. An initial set of plausible candidates $\hat{C}$ is obtained by simple grey-level global thresholding applied to C. Following the plausible detection paradigm, this step captures all real leakages and possibly false targets. Given the dataset used, it was determined that the mean value of C is a suitable threshold level ensuring that all true hyper-fluorescent regions are kept while discarding much of the retinal background.

3.3 Step 2: Final detection

In this final step we identify the most likely leakage candidates in the plausible set. As a result of dye dilution in the bloodstream pixel intensities will change over time. The assumption is that areas of leakage should have a characteristic intensity evolution distinguishing it from other areas. In normal areas the intensity over time would be expected to peak, level off and then gradually decrease. In leakage areas the intensities show a slower rate of decrease and may even continue to increase due to the excess fluorescein pooling in that area. Notice that the intensity profiles of ciliar occlusions and other possible false targets are inconsistent with those of leakages; for instance intensity changes in areas of ciliar occlusion tend to be erratic.

In the following steps, we first extract features of the time-intensity profile at certain absolute times and a learning algorithm finds the criteria to distinguish between leakage and non-leakage intensity profiles.

Feature definition

Each pixel of a candidate region has corresponding pixels throughout the sequence $(T_1, \dots T_n)$ which define the *time-intensity profile* capturing the change in intensity over time. Additionally, as each pixel of $\hat{C}$ is equivalent to an area of 4×4 pixels in the original images (before downscaling) we reduce the effects of noise on these profiles.

The features we consider are simply the averages of the profile values in five adjacent intervals (bins) spanning the profile. We observed that capture times vary for each session, i.e., the k-th frame of different sessions may be captured at different times from injection. Consequently, before further analysis we must obtain a measure of the signal at certain absolute times. For this reason we perform binning of the frames into defined time windows. The chosen time windows are (in seconds): before 30; 30-59; 60-149; 150-299; after 300. Each bin may contain from 3 to 10 frames approximately, depending on the sequence. Then for each bin b an average value $x(b)$ is calculated. The result for each pixel of the candidate image $\hat{C}$ is a 5-dimensional feature vector, $x(1), \dots, x(5)$. Additionally, we use the differences between $x(b)$ and $x(b-1)$ for $b = 2, \dots, b = 5$, capturing the relative intensity change between windows. As a result, the final feature vector is a 9-dimensional vector.

Training the classifier

We address leakage detection as a binary classification task assigning a candidate profile (feature vector) to one of two classes, leakage or non-leakage. Ground truth data with leakage regions accurately marked out was used for training. The training data consists of a set of 9-dimensional feature vectors $\mathbf{x}_1, \dots, \mathbf{x}_N$ with corresponding target

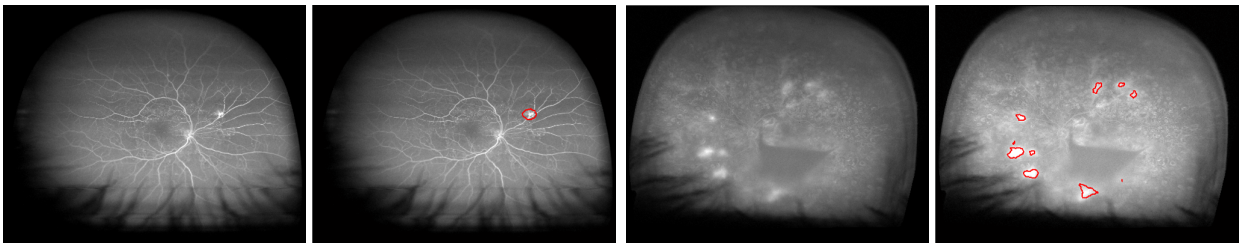


Figure 2. Two detection examples showing: late FA frame from input sequence; overlay of final predictions

labels $t_1, \ldots, t_N$ where $t_j \in \{-1, +1\}$.

AdaBoost [8], an adaptive boosting scheme is used to combine a series of weak classifiers into a powerful committee classifier. For this study each weak classifier is a decision stump, equivalent to a simple threshold on one of the input variables. In brief, during training AdaBoost creates a series of weak classifiers over a number of rounds $m = 1, \ldots, M$. For each round a distribution of weights is updated indicating the importance of examples in the data set where the weights of incorrectly classified examples are increased. In this way the next classifier focuses more on those examples. The result is a series of weak classifiers $h_1, \ldots, h_M$ and corresponding weighting coefficients $\alpha_1, \ldots, \alpha_M$, with greater weights assigned to more accurate classifiers. Effectively we employ AdaBoost to select the most informative criteria with which to distinguish between the two classes given the supplied features.

Classification

An unseen test sequence must undergo the steps above including pre-processing, candidate selection and feature extraction to derive a set of compatible feature vectors $\mathbf{y}_1, \ldots, \mathbf{y}_N$. This data is supplied to the trained classifier which queries each of the weak learners, effectively making a weighted majority vote. The result is a prediction $H(\mathbf{y}_i)$ for each test example $\mathbf{y}_i$, where the sign denotes the class as positive or negative (Equation (2)). Changing the parameter θ (normally set to zero) makes the prediction more conservative or more speculative. Setting θ negative will decrease false negatives, but increase false positives.

$$H(\mathbf{y}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{y}) - \theta\right) \tag{2}$$

Post-processing segmentation

Predictions are recorded in a final image, where regions of positive predictions are considered as regions of leakage. One-pixel regions are considered as outliers and discarded. Morphological closing is then employed to minimize small areas of background noise. Accepted regions can be characterised by area and level of severity defined by the magnitude of their score from Equation (2).

4 Experimental validation

For the purpose of this preliminary study we acquired ten anonymous FA sequences (eight with haemorrhages and other pathologies, and two without pathology). The sequences contain 15 to 60 frames captured over approximately 6 minutes from injection of the fluorescein dye. Ground truth data for all sessions was obtained by manually tracing leakage regions on registered frames. The annotation procedure was performed with guidance from a retinal specialist after observation of sequences in whole. This dataset was used to train and test the classifier using a *leave-n-out* procedure so that testing was always performed on unseen image sequences. This involved training on seven randomly selected sequences from the set and testing on the three held out sequences. The procedure was repeated five times and the classification results on the test sequences averaged. Training of the classifier proceeds as described in Section 3.3. We observed that there is no real benefit in further training after approximately 25 training rounds. The best generalization performance on unseen images is obtained within the first few rounds. This indicates that the decision boundary in the feature space is relatively simple.

Two examples of final detections are presented (Figure 2). All test predictions were evaluated against the ground-truth data. A test pixel with the same class as its corresponding ground truth is regarded as correct and counts were made of error rate, false positive and false negative predictions. The average classification error (averaged over the five rounds)

is 0.42%. The average sensitivity is 85.11% (ranging between 64.60% and 99.14% over the test cases). The average specificity is 99.90%. Higher sensitivity could be achieved by decreasing the parameter θ (Equation (2)). In this context it may be preferable to decrease the rate of false negatives, with the rationale that it is better to raise a limited number of false alarms (to be checked by clinicians) than missing some pathology altogether.

In comparison, Phillips et al. [3, 4] evaluate their results on simulated data and on example angiograms and report their method to be robust with large areas but less dependable with multiple small areas (accurately estimating the area to within 3.6%) tested over ten FA sequences. They do not report sensitivity or specificity measures. It must be noted that the method is also limited by the rudimentary image capture, digitization and computational restrictions of the time. Martinez-Costa et al. [6] report sensitivity between 82% and 98% when comparing manual and automatic predictions tested over 21 FA sequences.

After inspection of test results we conclude that some errors arise from image acquisition (e.g. incorrect patient position, eyelids), others from the subjective nature of the ground truth, e.g., noisy region boundaries. It is crucial to achieve accurate alignment of frames - any significant misalignment can result in missed detections. Severe cases have required manual intervention. In the worst cases some small hyper-fluorescent regions are missed entirely. Inspection of the results for the two healthy retinal sequences reveals no false alarms for one sequence, but some false positives for the other in a particularly difficult area to classify (late 'staining' surrounding the optic disc).

5 Conclusion

We presented a preliminary image analysis algorithm to detect vascular leakage in retinal angiographic sequences by exploiting the evolution of intensities over time. Our system pinpoints the location of leakages, making it possible to overlay angiograms with areas of hyper-fluorescent leakage. Additionally, whereas previous studies have only analysed the macular region, by the use of wide-field imaging our study also allows detection in peripheral regions. It is apparent that wide-field imaging is crucial for observing such peripheral pathology.

Results compare well with the current literature. The principle of analysing intensity change over time allows for clear discrimination between areas of increasing fluorescence and non-leakage areas. We expect to improve performance by exploring alternative feature representations and with larger sets of representative FA image sequences.

Acknowledgements

This work was supported by Optos plc and a Knowledge Transfer Partnership grant (KTP000684) of the UK Department of Trade and Industry. Thanks to D. Cairns, G. England and C. Mazo of Optos plc for valuable input. Thanks to Prof B. Dhillon for assistance with ground truthing of FA sequences.

References

1. E. Trucco & P. Kamat. "Locating the optic disk in retinal images via plausible detection and constraint satisfaction." *Image Processing, 2004. ICIP'04. 2004 International Conference on* **1**, 2004.
2. A. Manivannan, J. Plskova, A. Farrow et al. "Ultra-Wide-Field Fluorescein Angiography of the Ocular Fundus." *American Journal of Ophthalmology* **140(3)**, pp. 525–527, 2005.
3. R. Phillips, P. Ross, P. Sharp et al. "Use of temporal information to quantify vascular leakage in fluorescein angiography of the retina." *Clinical Physics and Physiological Measurement* **11(4 a)**, pp. 81–85, 1990.
4. R. Phillips, P. Ross, M. Tyska et al. "Detection and quantification of hyperfluorescent leakage by computer analysis of fundus fluorescein angiograms." *Graefe's Archive for Clinical and Experimental Ophthalmology* **229(4)**, pp. 329–335, 1991.
5. M. Cree, J. Olson, K. McHardy et al. "The preprocessing of retinal images for the detection of fluorescein leakage." *Phys Med Biol* **44(1)**, pp. 293–308, 1999.
6. L. Martinez-Costa, P. Marco, G. Ayala et al. "Macular edema computer-aided evaluation in ocular vein occlusions." *Comput Biomed Res* **31(5)**, pp. 374–84, 1998.
7. H. Azegrouz & E. Trucco. "Max-min central vein detection in retinal fundus images." *Proc. IEEE Int Conf on Image Processing (ICIP'06)* pp. 1925–1928, 2006.
8. Y. Freund & R. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences* **55(1)**, pp. 119–139, August 1997.

The Analysis of Effects on Breast Cancer Diagnosis Using the Mammographic Semantic Information

Da Qi[a]*, Erika R.E. Denton[b], Joanna M.E. Leason[b], Diaa Othman[c], and Reyer Zwiggelaar[a]†

[a]Department of Computer Science, University of Wales Aberystwyth, Aberystwyth, UK
[b]Department of Radiology, Norfolk and Norwich University Hospital NHS Trust, Norwich, UK
[c]Bronglais Hospital, Breast Surgery, Ceredigion and Mid-Wales NHS Trust, Aberystwyth, UK

Abstract. In this paper, we describe the analysis of the effects of mammographic semantic information in breast cancer diagnoses. A brief description of relations between semantic information and image features are given. We demonstrate the experiments based on mammographic semantic information and MIAS database. Each mammogram was annotated by expert radiologists with semantic information and a an NHSBSP five-point score. Two kinds of classifiers were applied to classify the mammogram according to NHSBSP five-point score by the semantic information and radiologists also classified them by their annotated semantic information. The analysis of the experiments results provided further understanding the effects of using the mammographic semantic information on breast cancer diagnosis. It also provided a common knowledge base can be both understand by computers and human experts.

1 Introduction

Breast cancer is still a significant cause of cancer death in women in the world. The likelihood of survival is increased when detected at an early stage and mammography is currently the most effective method for breast cancer detection [1]. This process is conducted by expert radiologists through reading and interpreting the image information of mammograms. Currently, Computer Aided Diagnostic (CAD) techniques are being developed and are aimed at improvement of the sensitivity and specificity of the diagnostic process [2]. However, these CAD techniques are using different kinds of extracted image features/textures in classification [3], little sematic information is used in these CAD techniques, and the effect of semantic information on the breast cancer diagnosis has not been evaluated. The *semantic information* covers the vocabulary and/or concepts in the mammographic domain and the relations between them and can be viewed as an additional higher concept level on top of the level representing the image features. Including the semantic information in the diagnostic process is, potentially, a way to increase the sensitivity and specificity of mammographic CAD systems. In this paper, we demonstrate experiments based on mammographic semantic information and the MIAS [4] database. Each mammogram was annotated by expert radiologists with semantic information and an NHSBSP five-point score [5]. Two kinds of classifiers (linear and non-linear) were applied to classify the mammogram according to the NHSBSP five-point score by the semantic information. These results were compared with those of radiologists using by their annotated semantic information (so without image information) to classify the mammograms. The analysis of the results provided further understanding of the effects of using the mammographic semantic information on breast cancer diagnosis.

2 Material and Methods

It should be noted that the aim of our proposed experiments is to evaluate the effects of semantic information on breast cancer diagnosis, not to improve the sensitivity and specificity of the diagnosis performance. We are trying to understand how human, radiologists in particular, execute the diagnosis procedure and the information they use in this process.

2.1 MIAS Data

The Mammographic Image Analysis Society (MIAS) database is used in the described experiments [4]. The images in this database were scanned with a Joyce-Loebl microdensitometer SCANDIG-3, which has a linear response in the optical density range 0-3.2. Each pixel is 8-bits deep and at a resolution of $50\mu m \times 50\mu m$. The total database contains 322 images and 321 are viewable (file mdb295ll is damaged). Each image has a confirmed ground truth of either benign or malignant.

*ddq04@aber.ac.uk
†rrz@aber.ac.uk

2.2 Semantic Information and Image Features

The mammographic semantic information is based on a developed mammographic ontology [6, 7]. The mammographic ontology contains most of the vocabulary in the domain of mammography and is a mixture of *is-a* and semantic relations. *Is-a* relation represents the relationship between class and subclass and semantic relation represents the relationship between class/subclass and its features. The final part of a mammographic ontology consists of semantic aspects, which include a) a description of how the values of the attributes, in combination with specific abnormality classes, lead to classification of the abnormality [8], b) spatial relationships and associations between abnormality classes [9], c) synonyms for abnormality classes and attributes [10], d) spatial relationships between abnormalities and image location [11]. The specific values of the attributes and the association between the various abnormalities in combination with some suitable logic will determine the classification of the mammographic images. According to NHSBSP guidelines for breast cancer screening assessment [5], a five point score scale would be: B1. normal tissue, B2. benign lesion, B3. lesion of uncertain malignant potential, B4. suspicious, and B5. malignant. Such classification will determine the subsequent process.

2.3 Experimental Design

The overall aim of this set of experiments is to evaluate if the mammographic semantic information can be beneficial in the mammographic decision process, i.e. for use of both CAD systems and radiologists. Three experiments were designed, which are described as:

Experiment 1: Classify MIAS images according to NHS Breast Screening Programme [5]

This experiment is aimed at obtaining the performance benchmark of the observers for the MIAS database. Each observer views the images through the image display interface (Fig. 1) and assigns an NHSBSP five-point score to each image. The results provide the performance of each radiologist when compared to the ground truth. Moreover, the results from subsequent experiments will be compared to these benchmark results.

Figure 1. MIAS image display interface.

Experiment 2: Provide detailed abnormality features using the ontology

This experiment is aimed at obtaining detailed semantic information about the type and morphology of the abnormalities based on the provided ontology. The resulting morphological specification will be used to assess if the provided information is sufficient for automatic classification approaches to distinguish between the various types of abnormalities and their NHSBSP categories. To simplify the case at this stage, all features were assigned either 0 or 1, which resulted in a binary vector for each image. These vectors will be used in Equation 1.

Experiment 3: Classify MIAS images based on semantic information

This experiment is aimed at obtaining NHSBSP five-point score for the MIAS images simply based on all semantic

information from **Experiment 2**. No images are provided and the assessment is based on text records only. This experiment will be based on the information as provided by the same observer.

2.4 Automatic Classification

As mentioned above, we are not trying to improve performance of diagnosis, so the two kinds of basic classifiers used in this paper are not expected to be optimal (see also Discussion). Firstly, the linear classifier calculated the distance between test data and the mean of each training set and the test data is assigned to the closest training set. Secondly, the nearest neighbor rule [12] is used, which is a non-linear classifier. In both cases a Euclidean distance was used, which is given by

$$D_{ij} = \sqrt{(\mathbf{x}_j - \mathbf{m}_i)^{\mathbf{T}}(\mathbf{x}_j - \mathbf{m}_i)} \tag{1}$$

where $\mathbf{x}_j$ represents the jth data vector, $\mathbf{m}_i$ represents mean vector of the ith category from the training data, for the linear classifier, and $\mathbf{m}_i$ represents the ith training sample for the non-linear classifier. A leave one out methodology was used to obtain the experimental results. The training set came from the results of **Experiment 2** with a class label to each image from those of **Experiment 1**.

3 Results

Three expert radiologists completed the experiments. They are all mammographic experts and were working in mammography when conducting these experiments. They have been working in mammography from 10 years to 6 months. It should be noted that the image display interface (Fig. 1) that was used for **Experiments 1** and **2** deviated from the image viewing conditions in mammographic screening and as such the experimental results are no indication of the performance of the radiologist in such a programme.

The first row in Table 1 shows the percentage of correctly classified cases when comparing the results from **Experiment 1** for each radiologist with the MIAS ground truth. The percentage of correct cases is determined by the number of correctly classified cases divided by the total numbers of cases. As mentioned in Sec. 2.3, the MIAS ground truth has only two categories - benign and malignant, we therefore grouped the NHSBSP categories into two, where category 1-3 map to benign and category 4-5 map to malignant. The experiments were carried out sequentially. From Table 1, expert 1 has the highest correct classification rate while expert 3 has the lowest. This is strongly correlated with the experience of the experts, as Expert 1 has 10 years of experience while Expert 3 has only 6 months and Expert 2 has experience situated between the other two.

	Expert 1	Expert 2	Expert 3
% of correct cases from **Experiment 1**	89.41%	84.74%	76.64%
% of correct cases from **Experiment 3**	81.62%	78.50%	77.57%
Kendall tau rank CC	0.78 ($p < 0.001$)	0.83 ($p < 0.001$)	0.87 ($p < 0.001$)

Table 1. Percentage of correctly classified cases from **Experiment 1 and 3** when comparing the results with the MIAS ground truth and Kendall tau rank CC between two experiments

The performance correlation between each radiologist can be valued by Kendall tau rank correlation coefficient (CC) [13, 14]. Its value ranges from -1 to 1, which indicates the relation between two variables is from perfect inverse concordance to perfect concordance. Zero means the two variables are independent. Table 2 shows these coefficients for the radiologists' performance. The performance between expert 1 and expert 2 has the highest CC value (0.56) of all, it indicates these two experts gave the most similar results in **Experiment 1**. Meanwhile, the performance between expert 1 and expert 3 produces less than 0.5 correlation coefficient, which indicates these two experts hold most differences between each other in the results of **Experiment 1**.

The results of **Experiments 3** from each radiologist indicate the effects of mammographic semantic information on the diagnostic procedure. The second row of Table 1 shows the correct classification rate of the results from **Experiments 3** compared with the MIAS ground truth. The performance simply based on ontology without viewing images was decreased in the cases of expert 1 and expert 2. The rate of expert 3 is slightly, but not significantly, increased. Table 3 also shows the Kendall tau CC of the radiologists' performance in **Experiment 3**. The highest CC value is the correlation of performance between expert 2 and expert 3, which is 0.47. In the other hand, the correlation of performance between expert 1 and expert 2 is less than 0.40. All the CC are less than 0.50. The performance of the experts in **Experiment 3** deviated from each other more than those in **Experiment 1**.

	Expert 1	Expert 2
Expert 2	0.56 ($p < 0.001$)	
Expert 3	0.48 ($p < 0.001$)	0.51 ($p < 0.001$)

Table 2. Kendall tau rank CC among radiologists' performance on **Experiment 1**

	Expert 1	Expert 2
Expert 2	0.39 ($p < 0.001$)	
Expert 3	0.40 ($p < 0.001$)	0.47 ($p < 0.001$)

Table 3. Kendall tau rank CC among radiologists' performance on **Experiment 3**

The third row of Table 1 shows the CC of each expert's own performance in the **Experiment 1** and **Experiment 3**. All the CC indicate high correlation between each expert's performance in both experiments. The highest CC value is from expert 3, who has the lowest correct classification rate among the three experts. Meanwhile, the lowest CC value is from expert 1, who has the highest correct classification rate of all. This situation could be explained as: a) the semantic information provided is more in line with expert 3 than with expert 1, b) an increase in experience can be linked to the use of a mental examples of abnormalities library that determine the classification process instead of a simple description, c) some subtle aspects of the classification process are not captured by the semantic information which tend to be used by expert 1 (hence the decrease in classification results) and less so by expert 3 (and hence less influenced by their lack).

As mentioned in Sec. 2.3, the results of **Experiment 1** can also be considered as a benchmark of the results from other experiments and automatic classification. Tables 4 and 5 show the confusion matrices between the benchmark of expert 1 and the result based on the two classifier using the semantic information provided by the same expert in **Experiment 2** (the full results from expert 2 and expert 3 are not shown due to page limitations but are summarized in Table 6). The diagonal of the matrices indicates the correctly classified cases, the part under diagonal indicates the number of under-scored cases classified by automatic classifiers and the part above diagonal indicates the number of over-scored cases.

		Linear Classifier results - Expert 1				
		1	2	3	4	5
Expert 1 - Benchmark	1	146	1	2	0	0
	2	7	26	6	0	0
	3	28	3	61	5	4
	4	9	0	4	0	14
	5	1	0	0	0	4

Table 4. The confusion matrix providing a comparison between the benchmark and linear classifier result from Expert 1

		Non-linear Classifier results - Expert 1				
		1	2	3	4	5
Expert 1 - Benchmark	1	146	1	2	0	0
	2	4	32	2	1	0
	3	14	2	80	4	1
	4	4	0	21	2	0
	5	0	0	5	0	0

Table 5. The confusion matrix providing a comparison between the benchmark and non-linear classifier result from Expert 1

Table 6 shows the summary of percentage of correctly classified cases and Kendall tau CC when the results from **Experiment 3**, linear classifier and non-linear classifier respectively compared with the benchmark of the same expert. We can see the effects of mammographic semantic information on the manual and automatic classification process. Given the mammographic semantic information without image information, the classification of radiologists and non-linear classifier have the similar percentage of correct cases and CC for both expert 1 and expert 2. On the other hand, the manually classification in **Experiment 3** is more similar to the linear classification than non-linear classification for expert 3. The reasons could be: a) with increased experience, radiologist tends to behave much more like a non-linear classifier than a linear classifier, b) the non-linear classifier used in this paper did not reflect the diagnosis procedure for expert 3.

4 Conclusion and Discussions

We have demonstrated the experiments completed by mammographic radiologists and automatic classifiers and analyzed the results, which provides a thorough understanding of the effects of using mammographic semantic information on breast cancer diagnosis procedure. Based on the mammographic semantic information, manual classification results show strong correlation to those with image information (Table 1). Automatic classification results show a strong correlation with the benchmark (Table 6) and indicate that they provided an appropriate level of descriptiveness in most areas. Areas that were not completely sufficient will be re-iterated and re-evaluated in subsequent steps of the development process. The difference between automatic classification results indicated that the performance does depend on the experts' assessment, which will need further investigation.

Classification based on semantic information by Expert 1			
	Experiment 3	Linear Classifier	Non-linear Classifier
% of correct cases	80.06%	73.83%	81.00%
Kendall tau CC	0.78 ($p < 0.001$)	0.68 ($p < 0.001$)	0.78 ($p < 0.001$)
Classification based on semantic information by Expert 2			
	Experiment 3	Linear Classifier	Non-linear Classifier
% of correct cases	72.59%	69.78%	77.26%
Kendall tau CC	0.83 ($p < 0.001$)	0.73 ($p < 0.001$)	0.84 ($p < 0.001$)
Classification based on semantic information by Expert 3			
	Experiment 3	Linear Classifier	Non-linear Classifier
% of correct cases	80.69%	75.70%	71.34%
Kendall tau CC	0.87 ($p < 0.001$)	0.83 ($p < 0.001$)	0.79 ($p < 0.001$)

Table 6. Percentage of correctly classification and Kendall tau correlation coefficient for all the results compared with the **Experiment 1** of the same expert

It should be noted that we are not expecting to gain high accuracy in the diagnostic performance, because classifing mammograms purely based on semantic information is not expected to be sufficient. However, as we mentioned in Sec. 2.2, there exist relations between semantic information and image features. Take micro-calcification as example, the image features include the positions of calcification in the image, the area and perimeter of each calcification, their roughness, compactness, length, width, elongation, the distribution of calcifications, and so on. Our future work will focus on finding the link between the mammographic sematic information and the extracted mammographic image features. Mammographic image feature extraction and feature selection methodology will be used to obtain the relevant image features. In next stage, we will apply this process to Digital Database for Screening Mammography (DDSM) [15] as well, because DDSM provides more categories than MIAS does. Moreover, different automatic classifiers will be applied to the linked information and their discriminate power will be investigated in the future work.

References

1. H. d. Koning. "Mammographic screening: evidence from randomised controlled trials." *Annals of Oncology* **14(8)**, pp. 1185–1189, 2003.
2. R. Birdwell, D. Ikeda, K. O'Shaughnessy et al. "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection." *Radiology* **219**, pp. 192–202, 2002.
3. H. Cheng, X. Cai, X. Chen et al. "Computer-aided detection and classification of microcalcification in mammograms: a survey." *Pattern Recognition* **36**, pp. 2967–2991, 2003.
4. D. H. Davies. "Digital mammography - the comparative evaluation of film digitizers." *British Journal of Radiology* **66**, pp. 930–933, 1993.
5. R. Wilson, D. Asbury, J. Cooke et al. (editors). *Clinical guidelines for breast cancer screening assessment*. Number 49. NHSBSP Publication, April 2001.
6. E. Manley, D. Qi, E. R. Denton et al. "Development of a Computer Aided Mammographic Ontology from Multiple Sources." In 7^{th} *International Workshop on Digital Mammography*, pp. 225–230. June 2004.
7. D. Qi, E. R. Denton & R. Zwiggelaar. "Linking Image Structures with Medical Ontology Information." In *Lecture Note in Computer Science*, number 4046, pp. 399–406. July 2006.
8. http://www.cancerscreening.nhs.uk/breastscreen/index.html. NHS *Breast Screening Programme*. Accessed 24/11/06.
9. J. Stell. "Part and complement: fundamental concepts in spatial relations." In *Annals of Mathematics and Artificial Intelligence*, pp. 1–17. 2004.
10. R. Birdwell, E. Morris, S. Wang et al. *Pocket Radiologist: Breast*. Amirsys Inc., 2003.
11. J. A. S. Caulkin, S. Astley & C. Boggis. "Sites of occurrence of malignancies in mammograms." In *4th International Workshop on Digital Mammography*, pp. 279–282. Nijmegen, The Netherlands, 1998.
12. D. Aha & D. Kibler. "Instance-baced learning algorithms." *Machine Learning* **6**, pp. 37–66, 1991.
13. H. C. Kraemer. "Correlation coefficients in medical research: from product moment correlation to the odds ratio." *Statistical Methods in Medical Research* **15**, pp. 525–545, 2006.
14. P. Valz & M. Thompson. "Exact inference for kendall s and spearman rho." *Journal of Computational and Graphical Statistics* **3**, pp. 459–472, 1994.
15. M. Heath, K. Bowyer, D. Kopans et al. "The Digital Database for Screening Mammography." In M. J. Yaffe (editor), 5^{th} *International Workshop on Digital Mammography*, pp. 212–218. Medical Physics Publishing (Madison, WI), Toronto, Canada, June 2000.

Registration of Preoperative MR to Intraoperative Ultrasound Images for Guiding Minimally Invasive Prostate Interventions

Dominic Morgan[a,*], Hashim Uddin Ahmed[b], Doug Pendsé[c], Mark Emberton[b], David Hawkes[a] and Dean Barratt[a]

[a]Centre for Medical Image Computing, Dept. of Medical Physics & Bioengineering,
[b]Institute of Urology and [c]National Medical Laser Centre,
University College London, London, UK

Abstract. MR-to-transrectal-ultrasound (TRUS) image registration achieved by matching features, such as the capsule surface, and anatomical landmark points provides a useful benchmark for the achievable accuracy of rigid registration in the prostate. However, this approach is impractical in the intraoperative situation given the unacceptable length of time required to accurately segment TRUS images. In this study, the accuracy of this registration approach was evaluated and compared with a gradient-based feature-to-image registration technique, which would require minimal user interaction during a procedure. The results indicate that the gradient-based technique is very accurate when the assumption of rigidity is satisfied, but is highly sensitive to gland deformation, mainly due to the presence of the TRUS probe. We conclude that although this technique shows promise, further development is required to achieve clinically acceptable registration accuracy in the presence of deformation.

1 Introduction

In the United Kingdom prostate cancer is the most common cancer among males with over 30,000 new cases each year (Cancer Research UK, 2007). Conventional treatment for prostate cancer is somewhat binary, in that either watchful waiting or full gland treatment, such as radiotherapy or radical prostatectomy, is the current standard of care. Recently, a number of minimally invasive therapies have become available that make focal treatment realisable. These include high-intensity focused ultrasound (HIFU), photodynamic therapy (PDT) and cryotherapy [1–3]. Potential advantages of such interventions include reduced morbidity and the possibility of repeat treatment, although more data is needed to fully determine the efficacy of targeted focal treatment [4].

Transrectal ultrasound (TRUS) is the standard method for guiding biopsy and treatment. The modality provides high temporal resolution during surgery but lacks diagnostic information in the majority of cases. Preoperative magnetic resonance (MR) imaging, on the other hand, can indicate the spatial distribution of cancer and identify areas most likely to contain disease [5]. However, MR is acquired hours, days or even weeks before the patient enters the operating theatre. Augmentation of real-time TRUS data with pathological information from the MR would enable targeting of suspicious anatomy during biopsy and focal therapy. Central to achieving this is accurate, fast registration of the two datasets.

Kaplan et al. [6] attempted rigid MR-TRUS registration in the prostate using a set of six corresponding points located at the superior, inferior, anterior, posterior, left and right extremes of the gland. Later Reynier et al [7] presented both rigid and elastic registration using point cloud surfaces to define the border of the prostate capsule. Both experiments were performed using very sparse TRUS datasets (5mm slice spacing) and MR images were acquired using an endorectal coil. We present two registration methods: The first uses a combination of surfaces and points [8], whilst the second is our implementation of a more automated technique, originally proposed by Wu et al. [9], that does not require ultrasound segmentation. This both saves time and reduces interobserver variability, which can be significant [10]. Semi-automatic segmentation methods have been investigated previously but still require an significant level of user-intervention [10–12]. Other relevant studies have focused on MR to interventional-MR registration [13, 14].

2 Method and Materials

2.1 Patients and MR Imaging

Patients were recruited to either PDT or HIFU focal ablation trials, both approved by the local ethics committee of University College Hospitals NHS Foundation Trust. TRUS data was acquired prior to a transperineal template-guided biopsy or PDT using a Hitachi EUB500 US scanner with the EUP-U533 5-9 MHz transrectal probe (Hitachi Medical Systems, Wellingborough, UK) with an in-plane resolution of 0.16 mm. Corresponding T2-weighted turbo-spin-echo transverse MR images were obtained on a Siemens Avanto (Siemens Medical Solutions, Bracknell, UK) with a slice thickness of 3.0 mm and resolution of 0.39 mm in-plane. No endorectal coil was used.

*d.morgan@ucl.ac.uk

2.2 TRUS Protocol

Three-dimensional TRUS datasets were acquired using a mechanical stepper mechanism (Tayman Medical Inc., MO, USA) to translate the TRUS probe and image capture used a frame grabber connected to a laptop computer. A series of transverse images was obtained, typically with a 2 mm spacing between slices. Care was taken to collect all images in one sweep due to the possibility of hysteresis in the stepper thread. The setup was such that movement is along the long axis of the TRUS probe in order to prevent inconsistent deformation.

2.3 Feature-based Registration

The surface-based approach required both datasets to be manually segmented by defining control points on the prostate capsule border using a custom-designed software tool written in MATLAB (The MathWorks Inc., MA, USA). Segmentation was carried out independently for each slice in which the observer deemed the border to be adequately visualised. Between 25 and 40 control points per slice were identified and between 6 and 12 slices were contoured; the exact number varied according to the gland volume, complexity of shape, and imaging modality. The points were resampled using a periodic, cubic interpolating spline to provide a smooth continuous contour. Five paired datasets were segmented by both an expert and a relatively inexperienced observer.

At the base and apex (the superior and inferior extremes) of the gland it is difficult to confidently delineate the capsule due to the complexity of the anatomy in these regions. This confined the segmentation into a 'band' around the mid-gland rather than over the whole surface introducing non-physical symmetries into the problem and affecting the robustness of the registration.

To overcome this, and to improve the accuracy in the superior-inferior direction, points defined at the base and apex of the prostate were additionally included in the registration. These are the locations at which the urethra enters the superior and inferior extremes of the gland, respectively.

Registration was performed in the TRUS co-ordinate space by minimising the distance between the set of resampled MR contour points and the TRUS surface in a least-squares sense. A Euclidean distance transform (DT) was applied to the TRUS volume to give the MR-point-to-US-surface distance and a cost function then calculated by taking the value of the DT at each of the transformed MR points using trilinear interpolation. The locations of the points are determined by the rigid transformation, T, relating MR to TRUS co-ordinates.

The cost function, C, defined as

$$C = \sum_{i=1}^{N} (D[T(x_i)])^2 + w(d_A^2 + d_B^2)$$

was minimised using the non-linear least-squares optimiser provided within the MATLAB environment's Optimisation Toolbox and based upon the interior-reflective Newton method. w is the point weighting and shifts the emphasis to either points or surfaces. d_A and d_B are the apex and base distances. $D[x]$ is the Euclidian distance transform, computed using the method of Maurer et al. [16], at a point x. x_i are the locations of the N contour points.

To reduce the registration time and improve robustness, the centroids of the two sets of contours were automatically aligned before registration, assuming zero rotation. This is justified by the fact that relatively little rotation occurs about the axis perpendicular to the sagittal plane between the acquisition of each set of data.

2.4 Gradient-based Feature-to-image Registration

In preoperative to intraoperative registration, all significant time constraints are effective only after intraoperative data is acquired; processing is acceptable on the MR image but must be kept to a minimum on the TRUS. Accurate contouring by hand can take up to 30 minutes, far longer than is acceptable within the operating theatre. This motivated us to look for a method in which manual segmentation is performed only on MR, with registration automatic thereafter.

We have implemented the method of Wu et al. [9] in which an organ surface is registered directly to the greyscale intensity information within an ultrasound image. Registration was achieved through alignment of the MR surface normal vectors with the gradient vectors of the TRUS image. Image gradients were calculated using Gaussian derivatives filters. These filters include a level of smoothing that must be specified through choice of the width of the Gaussian kernel.

Due to the sparseness of slices, only gradients in the x and y directions, i.e. in-slice, were considered. Surface normal vectors were approximated as the radial vectors from the MR contour centroid to a point on that contour and normalised to unit length. TRUS image gradient vectors were not normalised.

The cost function, S, defined as

$$S = \sum_{i=1}^{N} \vec{n_i} \cdot \nabla I(x_i)$$

was minmised, where $\nabla I(x_i)$ is the gradient of image I, $\vec{n_i}$ the surface normal, and x_i is the ith pixel location, determined by the current pose. Gradients were approximated using trilinear interpolation. S was computed over the larger resampled set of points from a periodic, cubic interpolating spline. Numerical optimisation was performed using the pattern search (PS) and genetic algorithm (GA) methods. When using PS a starting estimate must be provided. This is optional for GA.

2.5 Validation

A target registration error (TRE) was computed for each original and final rigid transformation by calculating the distance between one or more anatomical landmarks, defined manually in both the TRUS and MR images. The differing nature of the modalities, and the two most readily identifiable points being used in the registration algorithm, made definition of landmark pairs challenging. Consequently, we relied upon idiosyncrasies in a given individual's anatomy by identifying, for example, cysts and calcification as points for TRE calculation.

3 Results

3.1 Feature-based Registration

Table 1 shows a comparison between the expert and inexperienced observers in both mean TRE and success rate, based on a visual assessment of registration accuracy. Registration was deemed to be successful if both the position and orientation were such that TRUS contours were very close to the capsule border when viewed overlaid on an interpolated MR volume. All slices were looked at and a decision was made based upon the volume as a whole It should be noted that, due to deformation, the TRS contours will rarely perfectly delineate the MR gland boundary and hence 'success' is inevitably based upon observer interpretation. The residual refers to the final RMS value of the distance between the points and the surface. Results are the 'unregistered' TRE from centroid alignment and the corresponding post-registration TREs associated with varying the point weighting w.

	Method	Mean TRE (Range)	Residual	Success Rate	Success TRE
Inexperienced Observer	Unregistered	9.93 (3.59 - 23.9)	-	-	-
	$w = 0$	2.84 (1.70 - 3.78)	1.77	60%	2.21
	$w = 20$	3.24 (2.11 - 4.36)	1.87	80%	2.96
	$w = 50$	4.22 (2.35 - 5.57)	2.13	60%	3.95
Expert Observer	Unregistered	8.77 (2.79 - 22.5)	-	-	-
	$w = 0$	3.03 (1.68 - 3.90)	1.78	80%	2.84
	$w = 20$	3.66 (1.21 - 4.85)	2.02	100%	3.31
	$w = 50$	4.93 (1.32 - 7.19)	2.66	40%	4.94

Table 1. Summary of results averaged for 5 patients. All distances are in mm.

It can be seen that although the inclusion of points with an intermediate weighting of 20 has increased the mean TRE, the success rate has been improved and hence this should be considered the most successful method.

3.2 Gradient-based Registration

In our first experiment to test the gradient-based method, TRUS-derived surfaces were registered back to the images from which they were segmented. Eliminating deformation and providing a known ground truth in this way acts as a useful test of the principle. Results using an unseeded GA optimisation and a Gaussian derivative filter of width 3 pixels are presented in Table 2. The base and apex points were additionally used to calculate the TRE as they were no longer used for registration. The very low TRE values confirm that the gradient-based cost-function was a valid choice.

Using the same method for intermodal registration proved less successful. A variety of numerical optimisation schemes

Patient No.	TRE / mm	Apex Distance	Base Distance
1	0.14	0.40	0.16
2	0.11	0.34	0.40
3	0.10	0.15	0.21
4	0.24	0.16	0.27
5	0.26	0.25	0.29
Mean	**0.17**	**0.26**	**0.26**

Table 2. Results of TRUS surface to TRUS image registration. All distances in mm.

were tested including PS and GA, varying the Gaussian width, seeding the GA with a starting estimate and interpolating the contours to increase the number of points. Table 3 shows the results for the two most successful techniques - GA with Gaussian width 3 pixels, interpolated contours and an aligned centroid starting estimate. The results were obtained after running the registration 10 times. The corresponding values for a PS optimisation with the same parameters are given for comparison.

	Genetic Algorithm		Pattern Search	
Patient No.	Success TRE (Range)	Success Rate	TRE / mm	Visual Assessment
1	7.30 (3.80 - 12.6)	10%	23.6	Failure
2	4.33 (1.10 - 6.95)	50%	3.84	Success
3	4.81 (1.28 - 10.4)	100%	3.88	Success
4	5.36 (2.90 - 7.98)	60%	4.10	Success
5	4.34 (2.83 - 5.55)	90%	4.77	Failure

Table 3. Results of MR surface to TRUS image registration. All TREs are in mm.

The results in Table 3 indicate that the MR-TRUS registration using the gradient-based technique does not perform as well as the surface-to-surface method. We found that the registration tended to fail for datasets where significant deformation had occurred between MR and TRUS imaging. It is visually apparent that the majority of deformation is caused by the TRUS probe and is localised in the posterior region of the gland, near to the rectal wall.

4 Discussion and Conclusions

When interpreting the results obtained for intermodal registration care must be taken to put the TRE values in context. It should be noted that there is deformation in the gland and as such a localistation error is introduced into the points used to calculate the TRE. Further to this, the thickness of the MR slices and also the marked difference in intensity characteristics of the same anatomy between the two modalities reduce the TRE accuracy further. More work needs to be done to quantify these errors. The rigid-body assumption also does not account for errors due to localised deformation. Deformations affecting registration accuracy are caused by the US probe, bladder filling and difference in anatomical position (supine for MR versus lithotomy for TRUS) [15].

Work is currently underway to find a more accurate method of validation. This could be similar to that of Reynier et al. [7] who inserted a catheter in order to render the urethra visible in both MR and TRUS images. In our study, a catheter was already present during TRUS acquisition, but not during MR imaging. Consequently, the urethra was not visualised in MR in most cases. Using an MR-visible catheter would allow urethral displacement to be measured but sensitivity to rotation in the axial plane and displacement in the superior-inferior direction may be diminished. A more invasive solution would be to implant seeds within the prostate providing a number of accurate TREs spread throughout the gland. However, this is undesirable for ethical reasons and it can be difficult to localise seeds in the ultrasound volume.

Another source of inaccuracy is in identifying base and apex points and the relative weighting used in the optimisation step should reflect this. One possible solution is to compute a weighting automatically, possibly as a function of the difference in base-to-apex distances between the two modalities.

As highlighted earlier, accurately contouring the superior and inferior extremes of the prostate was found to be very challenging in the majority of patients. This forces contours into a band around the mid-gland. Incorporating coronal acquisitions might provide a more complete surface, and how best to combine two such sets of orthogonal contours is currently being investigated.

Feature-to-image registration removes the time-consuming contouring of the prostate TRUS images intraoperatively. Interobserver inconsistency is also eliminated. If deformation or movement of the gland occurs during the procedure (for example, due to repositionng of the TRUS probe, gland swelling as a response to therapy, or patient movement), invalidating the registration, a new registration can be readily performed without the need to resegment the TRUS images. The method is, however, more sensitive to deformation than the feature-based approach. The accuracy of interpolating gradient vectors between slices is also an area for improvement. In order to optimally use the TRUS gradient information, it is desirable to reslice the surface through any given plane. A parametric surface (for example, using spherical harmonics) would allow this and work is ongoing to incorporate such a representation to improve accuracy.

The robustness in the feature-based registration was excellent - 100% success rate with expert defined contours. The centroid alignment gives a reliable starting estimate. The lack of robustness is currently a significant obstacle in the gradient-based registration; a 60% overall success rate needs to be improved upon. For gradient-based registration, a starting estimate could be simply obtained in the clinical setting by including minimal user-intervention to define the approximate centroid of the gland. This would not add significant time to the registration procedure.

Acknowledgements

The authors would like to thank Enrico DeVita for his assistance in providing the MR scans for this study.

References

1. C. Chaussy, S. Thuroff, X. Rebillard et al. "Technology insight: High-intensity focused ultrasound for urologic cancers." *Nature clinical practice urology* **2(4)**, pp. 191–198, 2005.
2. C. Moore, T. Nathan, W. Lees et al. "Photodynamic therapy using meso tetra hydroxy phenyl chlorin (mthpc) in early prostate cancer." *Lasers in surgery and medicine* **38(5)**, pp. 356–363, 2006.
3. V. Mouraviev & T. J. Polascik. "Update on cryotherapy for prostate cancer in 2006." *Current opinion in urology* **16(3)**, pp. 152–156, 2006.
4. B. Akduman, A. B. Barqawi & E. D. Crawford. "Minimally invasive surgery in prostate cancer: current and future perspectives." *Cancer J* **11(5)**, pp. 355–361, 2005.
5. A. P. S. Kirkham, M. Emberton & C. Allen. "How good is mri at detecting and characterising cancer within the prostate?" *European urology* **50(6)**, pp. 1163–1175, 2006.
6. I. Kaplan, N. E. Oldenburg, P. Meskell et al. "Real time mri-ultrasound image guided stereotactic prostate biopsy." *Magn Reson Imaging* **20(3)**, pp. 295–299, Apr 2002.
7. C. Reynier, J. Troccaz, P. Fourneret et al. "Mri/trus data fusion for prostate brachytherapy. preliminary results." *Med Phys* **31(6)**, pp. 1568–1575, Jun 2004.
8. C. R. Maurer, R. J. Maciunas & J. M. Fitzpatrick. "Registration of head ct images to physical space using a weighted combination of points and surfaces." *IEEE Trans Med Imaging* **17(5)**, pp. 753–761, Oct 1998.
9. R. Wu, K. V. Ling, W. Shao et al. "Registration of organ surface with intra-operative 3d ultrasound image using genetic algorithm." In *MICCAI*, pp. 383–390. 2003.
10. H. Ladak, F. Mao, Y. Wang et al. "Prostate boundary segmentation from 2d ultrasound images." *Medical Physics* **27(8)**, pp. 1777–1788– EF, 2000.
11. S. D. Pathak, V. Chalana, D. R. Haynor et al. "Edge-guided boundary delineation in prostate ultrasound images." *IEEE Trans Med Imaging* **19(12)**, pp. 1211–1219, Dec 2000.
12. I. B. Tutar, S. D. Pathak, L. Gong et al. "Semiautomatic 3-d prostate segmentation from trus images using spherical harmonics." *IEEE Trans Med Imaging* **25(12)**, pp. 1645–1654, Dec 2006.
13. B. Fei, J. L. Duerk & D. L. Wilson. "Automatic 3d registration for interventional mri-guided treatment of prostate cancer." *Comput Aided Surg* **7(5)**, pp. 257–267, 2002.
14. B. Fei, J. Duerk, D. Sodee et al. "Semiautomatic nonrigid registration for the prostate and pelvic mr volumes1." *Academic Radiology* **12(7)**, pp. 815–824, 2005.
15. T. E. Byrne. "A review of prostate motion with considerations for the treatment of prostate cancer." *Medical dosimetry* **30(3)**, pp. 155–161, 2005.
16. J. Calvin R. Maurer, R. Qi & V. Raghavan. "A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25(2)**, pp. 265–270, 2003.

Sub-Voxel Reconstruction of Fibre Orientations

Shahrum Nedjati-Gilani[a], Geoff J. M. Parker[b] and Daniel C. Alexander[a*]

[a]Centre for Medical Image Computing, University College London, Gower Street, London, WC1E 6BT,
[b]Imaging Science and Biomedical Imaging, University of Manchester, Oxford Road, Manchester, M13 9PT

Abstract. In this paper, we present a new method for finding accurate fibre orientations and volume fractions of fibre populations on a sub-voxel scale from a 3D diffusion MRI acquisition in order to distinguish between various fibre configurations such as crossing, kissing, fanning and bending. We use our method with differently sub-sampled datasets to find the best way to accurately reconstruct fibre population information.

1 Introduction

Diffusion MRI provides an insight into the microstructural architecture of tissue by observing the restricted and hindered displacement of water molecules undergoing Brownian motion in vivo [1]. By looking at the probability density function p of displacements over a fixed period of time t, inferences can be made about the tissue microstructure. In some regions of the brain, the barriers to diffusion have no preferred orientation, resulting in isotropic diffusion. An example of an isotropic region in the brain are the ventricles which contain cerebrospinal fluid (CSF). As there are few barriers to water-molecule mobility, water displacement is equally likely in all directions. Grey matter consists of dense tissue containing many barriers to water mobility, such as cell walls and membranes. However, the barriers in grey matter often have no preferred orientation and so hinder the water displacement equally in all directions, again resulting in isotropic diffusion. In other regions of the brain, water molecules are constrained by organised fibrous structure, causing diffusion anisotropy. Water molecules in white matter (such as the pyramidal tracts and the corpus callosum) move on average further along fibres than across them, resulting in anisotropic diffusion.

Diffusion-Tensor MRI (DT-MRI) is the most common diffusion MRI technique [2]. DT-MRI computes the apparent diffusion tensor based on the assumption that p is a zero-mean tri-variate Gaussian distribution with covariance proportional to the diffusion tensor $\mathbf{D}$, a positive-definite, symmetric and second-order tensor. DT-MRI is often used for mapping fibre orientations. However, DT-MRI is only capable of recovering a single fibre orientation in each voxel. Other reconstruction algorithms, such as PAS-MRI [3], $\mathbf{q}$-ball [4] and spherical deconvolution [5] have appeared that can resolve crossing fibres; however, these more complex algorithms are generally less successful in correctly identifying the orientation of single fibres and are more computationally demanding than DT-MRI. Also, existing reconstruction methods identify multiple fibre populations, but not how they relate to each other; fibre-orientation reconstruction at each voxel is independent of other voxels, and does not use information available from the neighbourhood environment.

We present a new method for finding accurate fibre orientations and volume fractions of fibre populations on a sub-voxel scale from a 3D diffusion MRI acquisition which can be used to distinguish between various fibre configurations such as crossing, kissing, fanning and bending. We use our method with differently sub-sampled datasets to find the best way to reconstruct accurate fibre population information.

2 Method

We treat the task of finding accurate fibre orientations and volume fractions on a sub-voxel scale as a general inverse problem, which we solve by regularization and optimization.

2.1 Problem Outline

For a set of voxels $l_i, i = 1 \dots L$ and wavenumbers $\mathbf{q}_k, k = 1 \dots M$, we have measurements $A(l_i, \mathbf{q}_k)$. From these measurements, we want to find $\mathbf{p}(s_h)$, a set of model parameters in each of a set of high-resolution voxels s_h, where $h = 1 \dots H$. The forward problem is to estimate the measurements $A(l_i, q_k)$ from $\mathbf{p}(s_h)$. Measurement estimates on the high-resolution grid come directly from the model parameters $\mathbf{p}(s_h)$, and we can estimate the measurements at l_i by:

$$\tilde{A}(l_i, \mathbf{q}_k) = \sum_{h=1}^{H} \mu_{hi} A(s_h, \mathbf{q}_i) \tag{1}$$

*Email:S.Nedjati-Gilani@cs.ucl.ac.uk.

where μ_{hi} is a weighting coefficient that accounts for partial overlap between s_h and l_i and factors such as the point-spread function and the slice profile. The inverse problem finds the model parameters from $A(l_i, \mathbf{q}_k)$. We solve the inverse problem using optimization to minimize a distance metric between the true and estimated measurements, subject to spatial coherence. For image I, we minimize the objective function:

$$J(I) = \alpha T(I) + E(I) \tag{2}$$

where $T(I)$ is a smoothing term ensuring that transitions of fibre populations of neighbouring sub-voxels are smooth, α is a weighting coefficient, and $E(I)$ is the error component defined as

$$E(I) = \sum_{i=1}^{L} \sum_{k=1}^{M} \Big(\frac{A(l_i, \mathbf{q}_k) - \tilde{A}(l_i, \mathbf{q}_k)}{A(l_i, \mathbf{q}_k)} \Big)^2. \tag{3}$$

2.2 Model Fitting and Initialization

We use Behrens' model [6], which consists of two components: one modelling the isotropic diffusion of free water in the voxel, and another modelling the anisotropic diffusion along the fibres. We assume a maximum of two anisotropic fibre populations in each voxel, so that

$$A(l_i, \mathbf{q}_k) = f_0 e^{-td|\mathbf{q}_k|^2} + f_1 e^{-td(\mathbf{e}_1 \cdot \mathbf{q}_k)^2} + f_2 e^{-td(\mathbf{e}_2 \cdot \mathbf{q}_k)^2} \tag{4}$$

where $(f_1, \mathbf{e}_1)$ and $(f_2, \mathbf{e}_2)$ are the respective volume fractions and orientations of the fibre populations in the voxel, d is the diffusivity, t is the diffusion time, and $f_0 = 1 - (f_1 + f_2)$. The model parameter set for s_h is $\mathbf{p}(s_h) = \{d, f_1, f_2, \mathbf{e}_1, \mathbf{e}_2\}$. We fit the model to the data to find initial values for $\mathbf{p}(l_i)$, and interpolate linearly for $\mathbf{p}(s_h)$, such that

$$\mathbf{p}(s_h) = \sum_{i=1}^{L} \rho_{hi} \mathbf{p}(l_i) \tag{5}$$

where ρ_{hi} is the partial overlap fraction of sub-voxel s_h in voxel l_i.

2.3 Smoothing

For each sub-voxel s_h, we define the smoothing function T_1 as:

$$\begin{aligned} T_1(s_h) = \sum_{\forall n \in N(s)} \Big[& \sqrt{f_{s_{h1}} f_{n_1}} |\mathbf{e}_{s_{h1}} \cdot \mathbf{m}_{s_h,n}|^2 |\mathbf{e}_{s_{h1}} \cdot \mathbf{e}_{n_1}|^2 \\ & + \sqrt{f_{s_{h2}} f_{n_2}} |\mathbf{e}_{s_{h2}} \cdot \mathbf{m}_{s_h,n}|^2 |\mathbf{e}_{s_{h2}} \cdot \mathbf{e}_{n_2}|^2 \\ & + \frac{1}{3} \sqrt{f_{s_{h0}} f_{n_0}} (1 - |f_{n_0} - f_{s_{h0}}|) \Big] \end{aligned} \tag{6}$$

where $N(s_h)$ is the 6-neighbourhood of s_h, $\mathbf{m}_{s_h,n}$ is the vector between the centres of voxels s_h and n, and $f_{s_{h0}}$ and f_{n_0} are the isotropic volume fractions of s_h and n respectively (shown in Figure 1).

The value of $T_1(s_h)$ is invariant to the model parameters, and is constant when the fibre populations and orientations of all neighbouring voxels of s_h are identical to that of s_h, i.e. $\mathbf{p}(n) = \mathbf{p}(s_h)$ for $\forall n \in N(s_h)$. There are three factors contributing to $T_1(s)$:

- Similarity of fibre orientations of the fibre populations of s with those of its neighbours, quantified by $|\mathbf{e}_{s_1} \cdot \mathbf{e}_{n_1}|^2$ and $|\mathbf{e}_{s_2} \cdot \mathbf{e}_{n_2}|^2$.
- Similarity of volume fractions in s with those of its neighbours, quantified by $1 - |f_{n_0} - f_{s_{h0}}|$.
- The alignment of the voxels relative to the fibre orientation, quantified by $|\mathbf{e}_{s_{h1}} \cdot \mathbf{m}_{s_h,n}|^2$ and $|\mathbf{e}_{s_{h2}} \cdot \mathbf{m}_{s_h,n}|^2$. If a neighbouring voxel is more closely aligned to the fibre orientation, we penalize differences in the fibre orientation more. This reflects the idea that it is more important for voxels following the path of the fibre-orientation to have a similar fibre-orientation.

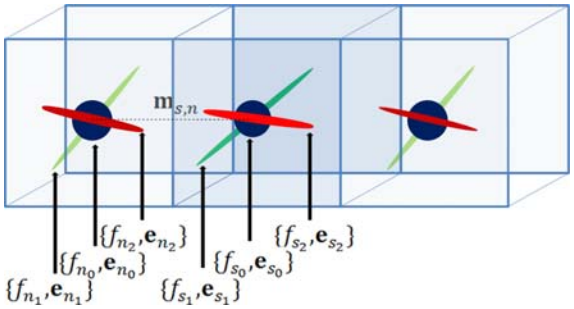


Figure 1. Illustration of a sub-voxel s with two voxels of its 6-neighbourhood, and associated fibre populations.

We also define a second smoothing function based on the normalized difference of d compared to neighbouring values:

$$T_2(s_h) = \frac{|d_{s_h} - d_n|}{\sqrt{d_{s_h} d_n}}. \tag{7}$$

We can now define the smoothing term:

$$T(I) = \sum_{h=1}^{H} \big(T_1(s_h)\big)^{-2} + \beta\big(T_2(s_h)\big)^2 \tag{8}$$

where β is a weighting coefficient, which allows us to control the relative contribution of $T_2(s_h)$ to the smoothing term.

2.3.1 Fibre Matching

To compute $T_1(s_h)$, we need to associate the fibre populations of s_h with those in the neighbouring voxels to decide which pairs of fibres match. If $\{(f_{x_1}, \mathbf{e}_{x_1}), (f_{x_2}, \mathbf{e}_{x_2})\}$ are the fibres in voxel 1 and $\{(f_{y_1}, \mathbf{e}_{y_1}), (f_{y_2}, \mathbf{e}_{y_2})\}$ are the fibres in voxel 2, we consider the two available possibilities:

- Possibility 1: $(f_{x_1}, \mathbf{e}_{x_1})$ corresponds to $(f_{y_1}, \mathbf{e}_{y_1})$ and $(f_{x_2}, \mathbf{e}_{x_2})$ corresponds to $(f_{y_2}, \mathbf{e}_{y_2})$.
- Possibility 2: $(f_{x_1}, \mathbf{e}_{x_1})$ corresponds to $(f_{y_2}, \mathbf{e}_{y_2})$ and $(f_{x_2}, \mathbf{e}_{x_2})$ corresponds to $(f_{y_1}, \mathbf{e}_{y_1})$.

We evaluate both possibilities with:

$$\begin{aligned} H_1 &= (1 - |f_{x_1} - f_{y_1}|)(1 - |f_{x_2} - f_{y_2}|)(f_{x_1} f_{y_1} |\mathbf{e}_{x_1} \cdot \mathbf{e}_{y_1}| + f_{x_2} f_{y_2} |\mathbf{e}_{x_2} \cdot \mathbf{e}_{y_2}|) \\ H_2 &= (1 - |f_{x_2} - f_{y_1}|)(1 - |f_{x_1} - f_{y_2}|)(f_{x_2} f_{y_1} |\mathbf{e}_{x_2} \cdot \mathbf{e}_{y_1}| + f_{x_1} f_{y_2} |\mathbf{e}_{x_1} \cdot \mathbf{e}_{y_2}|). \end{aligned} \tag{9}$$

In Equation 9, the first and second terms in parentheses increase the value of the result if the associated voxel fractions are close together. The third term will have a high value if the corresponding fibre orientations from the pair of voxels are similar; the similarity is evaluated by finding the vector product of the orientations. As the importance of the similarity of fibre orientations depends on the contributions of the associated volume fractions, we also multiply the vector product by the volume fractions. If $H_1 > H_2$, we assume that possibility 1 is true, otherwise we assume possibility 2 to be true.

2.4 Optimization

For computational tractability, we break down the objective function into L patches corresponding to voxels l_i. We now solve the inverse problem by minimizing the objective function

$$J(l_i) = \alpha T(l_i) + E(l_i) \qquad \text{for } i = 1, \cdots, L \tag{10}$$

separately in each voxel, with the smoothing function defined as

$$T(l_i) = \sum_{h=1}^{H} \mu_{hi} \Big[\big(T_1(s_h)\big)^{-2} + \beta\big(T_2(s_h)\big)^2 \Big] \tag{11}$$

and the error function defined as

$$E(l_i) = \sum_{k=1}^{M} \Big(\frac{A(l_i, \mathbf{q}_k) - \tilde{A}(l_i, \mathbf{q}_k)}{A(l_i, \mathbf{q}_k)} \Big)^2 \tag{12}$$

We minimize the objective function for each voxel l_i by least-squares minimization using a Levenberg-Marquardt algorithm with respect to $\mathbf{p}(s_h)$. Once we have performed the optimization for all l_i, we repeat this process to refine the parameters further until convergence, which is usually within a few iterations.

3 Experiments and Results

We test our method on three datasets that are sub-sampled in different ways from a $128 \times 128 \times 32$ image with 61 diffusion weighted images with a b-value of 1200 s mm^{-2} and one measurement at $\mathbf{q} = \mathbf{0}$, with eight repeats of each measurement, acquired on a Philips 3T Achieva with 8 element head coil. Each of the datasets contain the same number of measurements, and simulate acquisitions that we expect to require approximately the same time. We run our optimization method on each dataset to determine which set provides us with the most accurate information about fibre populations in voxels. From one repeat, we select a $18 \times 18 \times 8$ ROI (shown in Figure 2) and sub-sample to create three datasets:

- For dataset 1, we double the thickness of each slice to create a $18 \times 18 \times 4$ voxel array with 61 diffusion weighted images and one measurement at $\mathbf{q} = \mathbf{0}$ to create a low-spatial high-angular resolution (LSHA) dataset.
- For dataset 2, we only use 30 diffusion weighted images and one measurement at $\mathbf{q} = \mathbf{0}$ while retaining the spatial resolution, creating a high-spatial low-angular resolution (HSLA) dataset. We choose the subset of 30 directions from the original set of 61 that are optimally spread over the unit sphere, using the method proposed by Cook et al [7], which is implemented in the Camino package [8]. This is done by modelling each direction as the axis of a pair of identically charged particles on the sphere, and searching for a configuration that minimizes the electrostatic energy of pairs in the same subset using simulated-annealing optimization.
- Dataset 3 consists of two subsets; both have the slice thickness doubled to create a $18 \times 18 \times 4$ voxel array, and consist of 30 diffusion weighted images and one measurement at $\mathbf{q} = \mathbf{0}$. The two subsets are offset by half a slice width in the z-direction, and each of the 'jittered' subsets uses 30 of the 61 original directions with no overlap of directions used in the subsets. We expect that the combination of the two subsets contains more information about fibre populations than the previous two datasets.

We run our optimization algorithm on all three datasets to reconstruct fibre population information over the $18 \times 18 \times 8$ voxel grid. We evaluate the reconstruction from each dataset using the metric

$$C = \sum_{h=1}^{H} \sum_{k=1}^{M} \frac{|A(s_h, \mathbf{q}_k) - \tilde{A}(s_h, \mathbf{q}_k)|}{A(s_h, \mathbf{q}_k)} \tag{13}$$

where $A(s_h, \mathbf{q}_k)$ is now the average measurement of all 8 repeats; lower C is better. We show the reconstructed fibre populations and orientations on a slice from each dataset in Figure 2. After five iterations, the scores for each reconstruction are 46.17, 36.78 and 57.56 respectively, indicating that the HSLA dataset provides more accurate information.

4 Discussion and Conclusion

We have presented an algorithm that can find fibre populations and their respective orientations and volume fractions accurately on a subvoxel scale in a 3D diffusion MRI acquisition. By comparing results from different sub-sampled acquisition datasets, we have found that a high-spatial low-angular resolution acquisition provides better reconstruction compared to a low-spatial high-angular acquisition. The jittered acquisition appears to offer no advantage, although further work is required to test the dependence on starting point and algorithm parameters. Other future work could include looking at the effect of alternative sub-voxel division schemes, and to demonstrate differences of this method with that obtained by interpolating the results from fitting Behrens' model using the fibre matching method discussed previously.

Acknowledgements

This work was conducted with the aid of funding by EPSRC and Philips Medical Systems.

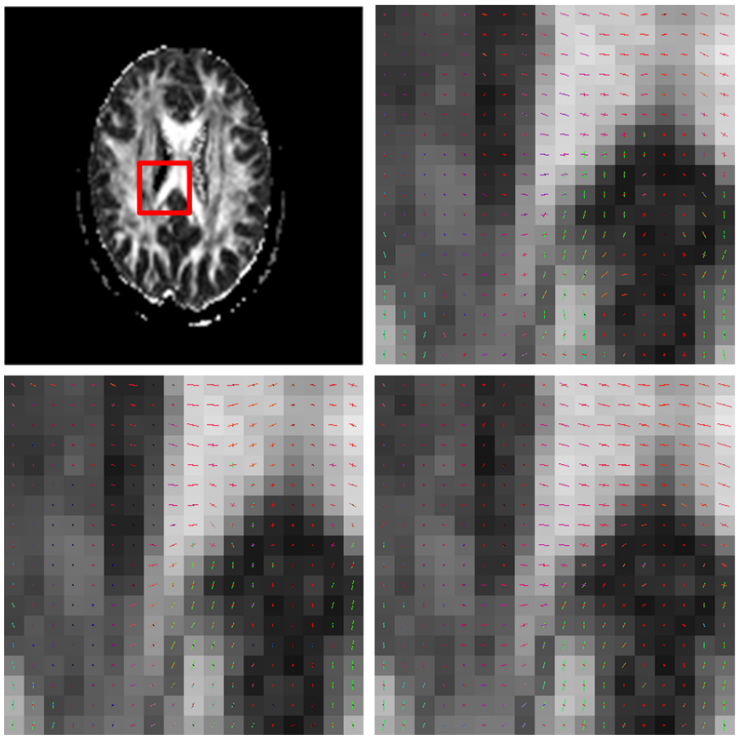


Figure 2. FA map with the ROI highlighted (top-left), and the resulting fibre population estimates in the LSHA (top-right), HSLA (bottom-left), and 'jittered' (bottom-right) datasets. The orientations are colour coded with red indicating mediolateral, green anteroposterior and blue superoinferior, and the length of the orientations are in proportion to the corresponding volume fraction.

References

1. P. J.Basser & C. Pierpaoli. "Microstructural and physiological features of tissues elucidated by quantitative diffusion tensor MRI." *J Magn Reson B* **111**, pp. 209–219, 1996.
2. P. J.Basser, J. Matiello & D. L. Bihan. "MR diffusion tensor spectroscopy and imaging." *Biophys J* **66**, pp. 259–267, 1994.
3. K. M. Jansons & D. C. Alexander. "Persistent Angular Structure: new insights from diffusion MRI data." *Inverse Problems* **19**, pp. 1031–1046, 2003.
4. D. S. Tuch. "Q-ball imaging." *Magnetic Resonance in Medicine* **52**, pp. 1358–1372, 2004.
5. J. Tournier, F. Calamante, D. Gadian et al. "Direct estimation of fibre orientations in partial volume contaminated regions using spherical deconvolution." In *Proc. 12th Annual Meeting of the International Society for Magnetic Resonance in Medicine*, p. 88.
6. T. E. J. Behrens, M. W. Woolrich, M. Jenkinson et al. "Characterization and propagation of uncertainty in diffusion weighted MR imaging." *Magnetic Resonance in Medicine* **50**, pp. 1077–1088, 2003.
7. P. A. Cook, M. Symms, P. A. Boulby et al. "Optimal acquisition orders of diffusion-weighted MRI measurements." *Journal of Magnetic Resonance Imaging (in press)* .
8. P. A. Cook, Y. Bai, S. Nedjati-Gilani et al. "Camino: Open-source diffusion-MRI reconstruction and processing." In *Proc. 14th Annual Meeting of the International Society for Magnetic Resonance in Medicine*, p. 2759.

Supervised MS Lesion Segmentation

T. Shepherd and D. C. Alexander.[a*]

[a]Centre for Medical Image Computing, University College London, Malet Place Engineering Building, WC1E 6BT

Abstract. We have developed an interactive method for supervised segmentation of MS lesions in MRI. The method combines a boundary tracking algorithm with SVM classifiers designed for texture discrimination. We perform experiments to evaluate classifier performance and to assess the accuracy and intra-/inter-rater variability of our segmentation method. Results show that the texture descriptors are an improvement over MRI intensity in classifying lesions and boundaries, and the semiautomatic tool is a promising method of exploiting the classifier in segmentation that extends naturally to a broad range of applications.

1 Introduction

The main challenge to the future of conventional MRI in Multiple Sclerosis (MS) studies is to provide better means of lesion segmentation [1].

The semiautomatic boundary extraction method in [2], which follows high intensity gradients to define lesion boundaries, is confounded by a partial volume effect at the edge of lesions and a general tendency noted in [3] for the intensity profiles of MS lesions to fall gradually from the centre outwards. At the Institute of Neurology in London, 4 experts who routinely segment lesions using the tool in [2] reported that $40-60\%$ of contours require subsequent editing by freehand drawing, of which roughly half require complete replacement. Automated lesion segmentation is confounded by global intensity variations caused by magnetic field inhomogeneities, which require separate correction [4]. Intensity classification is also limited as lesion histograms do not follow an easily parametrizable distribution [5]. Also, the morphology, size and internal structure of lesions vary both over time, due to the complex histology of the disease [6] and scan/re-scan variability [7] and over space, as a single brain contains lesions at different stages [8]. For a given stage in the disease, the pathology of a lesion also varies between patients [9]. Classification problems involving this level of within-class variability can be addressed using a Support Vector Machine (SVM) classifier [10]. We design SVMs for texture discrimination, which also overcomes the limitations of intensity alone in lesion classification. In addition, we propose that MS lesion segmentation should make efficient use of expert knowledge in a supervised framework. This is in-line with the two requirements stated in [11], that a segmentation tool must **1.** provide as complete control as possible to the user, and **2.** minimise the user involvement and the total user time without compromising precision and accuracy. Thus, the central aim of this work is to design a semiautomatic segmentation tool that minimises demand on the operator without compromising the quality of human defined contours. We extend the 'Jetstream' algorithm of Perez *et al.* [12] and tune it for our application. We test the efficacy of the approach for MS lesion segmentation.

In this paper we present an interactive segmentation tool based on Jetstreams. After a summary of classical Jetstreams we describe our adaptations to the algorithm and some new image priors used for the MS lesion application. We then present experiments performed by experts segmenting PD and T2 weighted MR images.

2 Background

Perez *et al.* introduced Jetstreams, which use particle filtering to track points $\mathbf{x} = \{x, y\}$ along boundaries in the image plane [12]. A particle is defined as a chain of i points $\mathbf{x}_{0:i}$ and at any time there exists a set of M particles $(\mathbf{x}^m_{0:i})_{m=1:M}$. Tracking relies on the iterative computation of posterior densities. Given the image data Y the posterior probability for the next point along the boundary is

$$p_{i+1}(\mathbf{x}_{0:i+1}|Y) \propto p_i(\mathbf{x}_{0:i}|Y) \times q(\mathbf{x}_{i+1}|\mathbf{x}_{i-1:i}) \times l(Y(\mathbf{x}_{i+1})), \tag{1}$$

where q and l are shape and image priors respectively. The Jetstream algorithm involves recursive computation of p_{i+1} in a three-stage procedure. A *prediction* stage randomly selects M locations for the next point $\mathbf{x}_{i+1}$, at a fixed distance from $\mathbf{x}_i$, using a normal distribution $q = N(\theta; 0, \sigma_\theta)$, where θ is the angle between the previous step $\mathbf{x}_{i-1:i}$ and the proposed step $\mathbf{x}_{i:i+1}$. The prior angular distribution gives the Jetstream a smoothness constraint governed by the parameter σ_θ. Next, a *weighting* stage weights these M proposed steps by the product $q \times l$ where $l(Y(\mathbf{x}_{i+1}))$

*{T.Shepherd,D.Alexander}@cs.ucl.ac.uk

is the likelihood that a proposed step keeps the particle on the true boundary. This likelihood is given by the ratio of probabilities p_{on} and p_{off}, that a point lies on and off the boundary respectively, given by

$$(a) \quad p_{\text{on}} \propto N\big(\psi; 0, (\sigma_\psi / |\mathbf{G}_I|)\big) \qquad \text{and} \qquad (b) \quad p_{\text{off}} \propto e^{\left(\frac{-|\mathbf{G}_I|}{<|\mathbf{G}_I|>}\right)}, \tag{2}$$

where $\mathbf{G}_I$ is the image intensity gradient vector, $< \cdot >$ denotes the mean calculated over the image and ψ is the acute angle between the direction of the last step taken and the local boundary direction estimated via $\mathbf{G}_I$. The spread of p_{on} in equation 2 (a) is modified by $\frac{1}{|\mathbf{G}_I|}$ so that a particle adheres to the boundary direction more for stronger edges. Equation 2 (b) attracts particles to areas of high $|\mathbf{G}_I|$. After normalising, the M weights give a discrete approximation of the posterior density, which is used in the third stage of *importance sampling*. By selecting M points with replacement from the posterior distribution, those with higher weighting have higher probability of remaining in the sampled set. The M sampled points may not be unique, as the most likely steps $\mathbf{x}_{i:i+1}$ may be duplicated. This concludes one iteration of the boundary tracker and the end points of the new particle set become the starting points for the next M predictions.

After a fixed number of iterations N, there are M particles $(\mathbf{x}_{0:N}^m)_{m=1:M}$ that are likely to share many pixel locations. This completion of N steps will be referred to as a single 'run' of the Jetstream contour. After each run the desired boundary section is defined as the mean path over the final particle set.

3 Algorithm and Framework

We present four technical adaptations to the algorithm above. First, we replace the normal prior angular distribution with the Von-Mises distribution $V(\theta; 0, \kappa_\theta) \propto e^{\kappa_\theta cos(\theta - 0)}$, which is periodic with period 2π. By choosing this distribution we maintain a general framework wherein the distribution is circularly symmetric for large absolute angles where the tails of the normal distribution fall monotonically. Second, the Jetstreams are constrained to avoid self-intersection. Third, we define a Jetstream as the path that, after importance sampling, has the maximum total weight. Unlike the mean path chosen in [12], this last modification retains the intended smoothness constraint and avoids false paths being created where the particle set is multi-modal such as at forks in the boundary. Fourth, when the boundary is nearly complete, the user can invoke an extra prior designed to create a closed contour as follows. Holding down a key activates an attraction force, comprising an extra likelihood term in the posterior weighting, proportional to the inverse exponential distance to the start of the contour. This final run waits until enough particles terminate at the starting point, then choses from these particles, the one with the highest weighting *independent* of the attraction force.

The user initialises a contour by drawing a small straight line section anywhere on a lesion boundary. The program then displays the Jetstream that results from tracking N points, referred to as one 'run'. At all times the contour is defined as the set of points from successive runs, interpolated using Bresenham's line algorithm. Between runs, the user selects a pixel as an 'anchor' point, from which the next run begins. An anchor placed at the 'tip' of the Jetstream causes the tracking to continue as normal. An anchor placed *beyond* the tip creates a straight bridge between runs. An anchor placed *anywhere along* the current contour will remove the Jetstream beyond that point. Interaction is also able to guide the Jetstream by placing an anchor to one side of the contour. This is used as a manual method of corner handling, replacing numerical methods presented in [12] and [13]. During boundary extraction the user can adjust the smoothness, 'detail' (distance between points) and length of a single run using slide-bars.

4 Image Priors for Regions and Boundaries

We introduce two new image priors to complement intensity information with local texture. Both rely on SVMs, using a radial basis kernel function to map feature vectors into a high dimensional feature space for classification. This enables local texture to be classified without explicit feature calculation [14].

We define three variants of the Jetstream algorithm. The first, referred to as **Jetstream A**, uses a *boundary-trained* SVM. We provide the SVM with 25-dimensional input feature vectors, each made up of 25 pixel intensities from a sampling window. The sampling window is a star-shape as used in the autoregressive texture model suggested in [15]. This gives larger scale texture information than a square window, for the same dimensionality. The sampling window is centred on lesion boundary pixels for the positive class in a binary classification of 'on-boundary' vs 'off-boundary'. For the negative class the sampling window is centred on pixels chosen at random locations weighted by the inverse exponential distance to the nearest boundary (with a margin of one pixel). In this way, the SVM learns to discriminate between boundaries and the nearby tissue, inside and outside a lesion, that Jetstreams are most likely to encounter. Upon classification, the decision value output by this boundary-trained SVM is denoted d_b and replaces $\mathbf{G}_I$ in equations 2.

We separately infer local boundary direction by convolving the image of d_b with two orthogonal 'ridge' templates to give the x and y components of a direction vector.

In the second case, referred to as **Jetstream B**, we use a 5×5 pixel square sampling window to capture smaller scale texture. The windows are centred on all pixels inside MS lesions, giving feature vectors for the positive class in a binary classification of 'lesion' vs 'non-lesion'. For the negative class, the sampling window is centred on non-lesion pixels, at random locations weighted as before to be close to, but outside MS lesions. Classification by this region-trained SVM yields the decision value d_r. We use directional Sobel filters to calculate the gradient vector $\mathbf{G}_{d_r}$ of the classified image, which replaces $\mathbf{G}_I$ in equations 2.

We also implement Jetstreams using intensity gradients $\mathbf{G}_I$ as in [12], using directional Sobel filters to calculate $\mathbf{G}_I$. This *classical* Jetstream will be referred to as **Jetstream C**. Figure 1 (a),(b) and (c) show the magnitudes of the image priors used in Jetstreams **A**, **B** and **C** respectively. Lesion boundaries are apparent in all cases. False positives are also

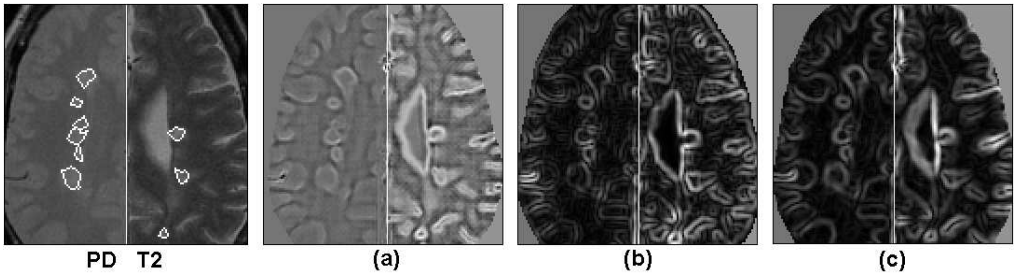


Figure 1. *(far left)* An axial slice showing PD (left hemisphere) and T2 (right hemisphere) intensity. White contours show the ground-truth segmentation of lesions. (a) Classified using the boundary-trained SVM, (b) edge map of the images classified by the region-trained SVM, (c) edge map of the raw images.

present, particularly at the boundaries between grey and white matter, as these regions are largely omitted from the training data used to create (a) and (b). However, user-guided Jetstreams are unlikely to encounter these regions.

5 Experiments and Results

We performed experiments using 40 brains containing MS lesions, labelled with ground truth by an expert using the tool in [2]. In all experiments, any one SVM is trained using data from all but one image and used to classify data in the remaining image. All experiments are repeated for PD and T2 data.

5.1 Texture Classification

To evaluate the success of the SVMs we performed Receiver Operator Curve (ROC) analysis on the results of classifying ground truth throughout a whole brain. By thresholding the range of decision values at 500 equal increments, we build up a ROC curve and indicate classifier performance by the Area under the Curve (AUC). We repeat for all images to give an average AUC. Table 1 shows the mean AUC ($\pm$ one standard deviation) for the two SVMs, along with results of ROC analysis performed on the intensity of all ground truth pixels throughout all 40 volumes, either inside lesions ('region intensity') or on lesion boundaries('boundary intensity') with the same negative-class locations as defined for

Image parameter	**Region intensity**	**Boundary intensity**	**Region-trained SVM (d_r)**	**Boundary-trained SVM (d_b)**
PD	0.836	0.610	0.928 ± 0.055	0.875 ± 0.043
T2	0.854	0.624	0.927 ± 0.021	0.858 ± 0.039

Table 1. Comparison of AUC for intensities and SVM classification.

the SVMs. These results show that the texture SVMs identify lesions and boundaries better than T2 or PD intensity.

5.2 Contouring

We asked 4 expert raters to segment lesions using the Jetstreams, plus the freehand drawing tool provided by ImageJ. All raters followed a randomised sequence of 24 tasks from 6 lesions and the 4 methods. To preserve intensity non-

uniformity, scanner inconsistencies and patient variation, we chose 3 slices from different patients, wherein we chose 2 lesions in different hemispheres of the brain. We evaluate the segmentation accuracy and intra-/inter-rater variability of the Jetstreams and the freehand tool using 2 measures of dissimilarity between any pair of contours: The Hausdorff distance [16], to measure the difference in shape between one contour and another (relevant to studies of lesion morphology) and the absolute difference in the areas enclosed by 2 contours (relevant to studies of lesion load).

To measure the **accuracy** of each Jetstream, we take the dissimilarity between contours created by a rater using that Jestsream and the same rater's freehand contour. We then take the mean dissimilarity over all 6 lesions. To give an idea of the accuracy of the freehand method, we measure the dissimilarity between the same freehand contour and a second one created days later. Figure 2 compares the mean dissimilarity of each method used on T2 images. We observed similar results for PD images. Figure 2 suggests that SVM Jetstreams are more accurate than classical Jetstream C and that Jetstream accuracy is comparable to the freehand method.

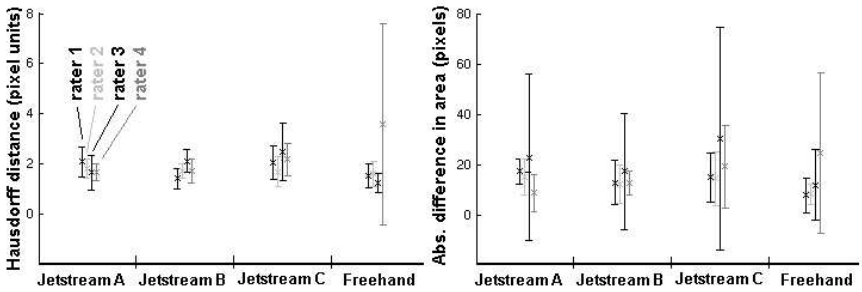


Figure 2. Accuracy of all methods. Mean dissimilarity is shown for each rater with error bars at $\pm$ one s.d.

To compare the **intra-rater variability**of each method, we measure the dissimilarity between contours created by the same rater and method at different times, days apart, and take the mean dissimilarity over all 6 lesions. Figure 3 shows the results for T2 images. Intra-rater variability does not differ significantly between methods, similarly for PD images.

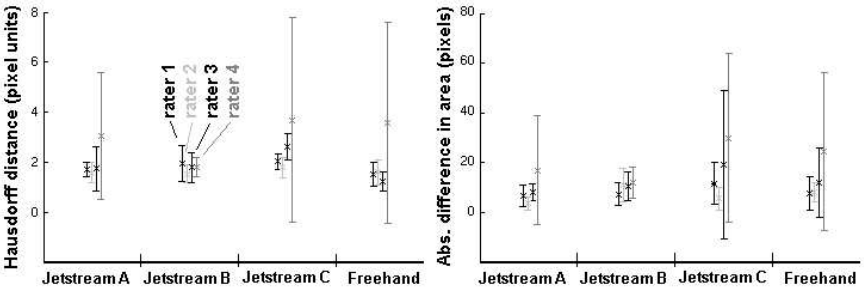


Figure 3. Intra-rater variability of all methods. Mean dissimilarity is shown for each rater with error bars at $\pm$ one s.d.

The freehand contours overlain in figure 4 *(left)* reveal that, while some lesions such as (i) and (ii) are unambiguous, others such as (iii) and (iv) are perceived differently by at least one rater, who groups nearby lesions together. This ambiguity is distinct from **inter-rater variability**, whereby lesions that are perceived the same by two raters are segmented differently due to limitations of the method. We choose two unambiguous lesions (i) and (ii) to compare the inter-rater variability of each method. We measure the dissimilarity between the two contours created by each *pair* of raters. We then take the mean dissimilarity over all 6 pairs. The plots on the right of figure 4 compare the dissimilarity measures for the 4 methods used on T2 images. These plots suggest that the SVM Jetstreams improve on the classical Jetstream C, and have comparable inter-rater variability to freehand contouring.

6 Discussion and Conclusions

We have presented a user-guided tool for MS lesion segmentation. The underlying method reduces demand on the operator meaning that interaction time should be reduced when the user is familiar with the tool. At the same time the results are comparable to the gold standard of expert manual drawing, in terms of accuracy and variability.

The texture-based SVMs offer more discriminating power than PD or T2 intensity, either in classifying lesion regions to give an improved edge map, or in a one-shot method of boundary localisation. Visual inspection (figure 1) shows

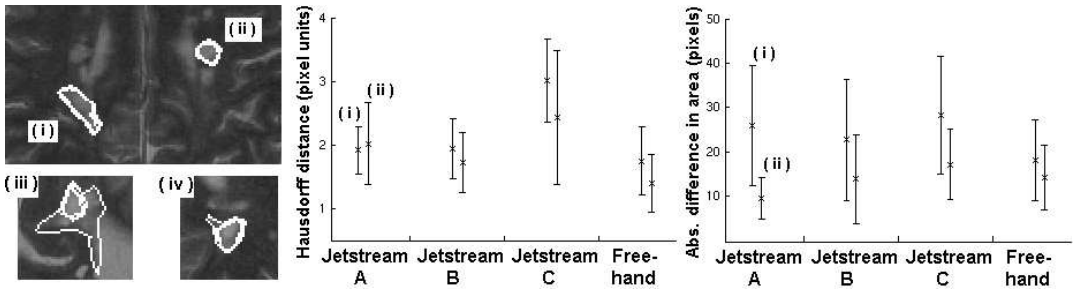


Figure 4. Inter-rater variability of all methods. *(left)* 4 lesions shown with freehand contours from the 4 expert raters overlain in white. *(right)*Mean dissimilarity shown for lesions (i) and (ii), with error bars at ± one s.d.

that the boundary-trained SVM produces better resolved boundaries but at lower contrast than the edge maps. These new classifiers should generalize well, and do not rely on multispectral data or intensity normalisation. The SVM itself can be improved over time using incremental learning, as more ground truth becomes available. The SVM Jetstreams out-perform the classical method based on intensity gradients.

The semiautomatic framework balances well the requirements of maximum user-control with minimal user-involvement. The combination of machine learning and run-time user-guidance extends beyond the MS lesion problem, and we expect greater benefits in terms of reduced user-engagement and improved segmentation results, in applications where regions of interest are typically larger than MS lesions.

Acknowledgements

The authors are grateful to D.Tozer, J.Swanton, J.Furby, T.Hayton and L.Fisniku at the Institute of Neurology, London, for providing data and advice, and performing contouring tasks for experiments. This work was funded by the EPSRC.

References

1. Y. Ge. "Multiple sclerosis: The role of MR imaging." *American Journal Of Neuroradiology* **27**, pp. 1165–1176, 2006.
2. D. Plummer. "Dispimage: Un mezzo di analisi e presentazione per iconografia medica." *Riv. Neuroradiol* **5**, pp. 489–495, 1992.
3. B. Johnston, M. Atkins, B. Mackiewich et al. "Segmentation of multiple sclerosis lesions in intensity corrected multispectral MRI." *Medical Imaging, IEEE Transactions on* **15**, pp. 154–169, 1996.
4. D. Wicks, P. Tofts, D. Miller et al. "Volume measurement of multiple sclerosis lesions with magnetic resonance images." *Neuroradiology* **34**, pp. 475–479, 1992.
5. B. Sajja, S. Datta., R. He et al. "Unified approach for multiple sclerosis lesion segmentation on brain MRI." *Annals of Biomedical Engineering* **34**, pp. 142–151, 2006.
6. H. Lassmann, W. Bruck & C. Lucchinetti. "Heterogeneity of multiple sclerosis pathogenesis: Implications for diagnosis and therapy." *Trends in Molecular Medicine* **7**, pp. 115–121, 2001.
7. J. Simon, A. Sherzinger, U. Raff et al. "Computerized method of lesion volume quantitation in multiple sclerosis: Error of serial studies." *American Journal Of Neuroradiology* **18**, pp. 580–582, 1996.
8. C. Pachai, Y. Zhu, J. Grimaud et al. "A pyramidal approach for automatic segmentation of multiple sclerosis lesions in brain MRI." *Computerized Medical Imaging and Graphics* **22**, pp. 399–408, 1998.
9. C. Lucchinetti, W. Bruck, J. Parisi et al. "Heterogeneity of multiple sclerosis lesions: Implications for the pathogenesis of demyelination." *Annals of Neurology* **47**, pp. 707–717, 2000.
10. V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
11. A. Falcão, J. Udupa & F. Miyazawa. "An ultra–fast user–steered image segmentation paradigm: live wire on the fly." *Medical Imaging, IEEE Transactions on* **19**, pp. 55–62, 2000.
12. P. Pérez, A. Blake & M. Gangnet. "Jetstream: Probabilistic contour extraction with particles." In *Proceedings, IEEE International Conference on Computer Vision*, pp. 524–531. 2001.
13. Y. Famao, S. Lin, L. Shukai et al. "Extraction of complex object contour by particle filtering." In *IEEE Int. Geoscience and Remote Sensing Symposium*, pp. 3711–3713. 2005.
14. T. Shepherd & D. Alexander. "A support vector machine for 3D texture-based classification of bone in CT images." In *Proceedings, Medical Image Understanding and Analysis*, pp. 221–225. 2006.
15. J. Mao & A. K. Jain. "Texture classification and segmentation using multiresolution simultaneous autoregressive models." *Pattern Recognition* pp. 173–188, 1992.
16. D. Huttenlocher, G. Klanderman & W. Rucklidge. "Comparing images using the Haussdorff distance." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **15**, pp. 850–863, 1993.

Automatic Segmentation of the Left Atrium

Rashed Karim[a], Raad Mohiaddin[b] and Daniel Rueckert[a]

[a]Department of Computing, Imperial College London, London, UK.
[b]National Heart and Lung Institute, Imperial College London, London, UK.

Abstract. Segmentation of the left atrium is vital for pre-operative assessment of its anatomy in radio-frequency catheter ablation surgery, which is commonly used for treating atrial fibrillation. In this paper we present an automatic approach for segmenting the left atrium and its pulmonary veins from MR angiography (MRA) data sets. Based on the notion that in MRA the atrium becomes connected to surrounding structures via partial volume affected voxels and narrow vessels, the atrium can be separated if these regions are characterized and identified. The blood pool, obtained by subtracting the pre- and post-contrast scans, is first segmented using a region-growing approach. The segmented blood pool is then subdivided into disjoint subdivisions based on its Euclidean distance transform. These subdivisions are then merged automatically starting from a seed point and stopping at points where the atrium connects to a neighbouring structure. The resulting merged subdivisions produce the segmented atrium.

1 Introduction

One of the most common causes of deaths in patients with cardiovascular related illnesses are heart strokes and attacks. The American Heart Association reports that 15% of all heart strokes are caused by a life-threatening condition called atrial fibrillation (AF) [1]. AF is a condition involving the left atrium of the heart. The left atrium is one of the four chambers of the heart. It receives oxygenated blood from the lungs and pumps it into the left ventricle. This blood is then circulated to the rest of the body. In a healthy adult the left atrium pumps blood into the ventricle in a regular rhythm. In AF, the left atrium quivers in an abnormal rhythm and is no longer able to pump blood into the left ventricle efficiently. This may cause possible pooling and clotting of blood in the left atrium which can lead to a stroke. Radio-frequency catheter ablation (RFCA) has become the treatment of choice for patients suffering from AF [2]. The objective of the ablation procedure is to eliminate sources of *ectopic foci* by charring tissues with high radio-frequency energy. These sources are identified pre-procedurally by examining electroanatomical maps that correlate the eletrophysiological characteristics with the endocardial anatomy. Medical literature indicates pulmonary venous drainages of the left atrium to be strong sources of ectopic focal activity [8]. Thus, a correct segmentation of the left atrium and its pulmonary venous drainages is important at this stage for planning the surgical procedure and also to allow the integration of the electroanatomical data with cardiac imaging data.

The left atrium is a highly anatomically variable structure where the number and sizes of the pulmonary venous drainages can vary significantly across patients. Variations to the right and left drainage patterns of the atrium have been documented in [3]. This high degree of variability makes left atrium segmentation difficult. Little work has been done on automated left atrium segmentation. Berg et. al. [4] uses a shape-constrained deformable surface model to segment the left atrium from multi-slice CT images. However, the model is restricted to a mean surface model which is built from training samples. Although such a technique may perform well on the commonly occurring variants of the atrium, the under-representation of the rarely occurring variants in the mean model makes it difficult for it to fit to these instances. John et. al. [5] uses a data-driven approach for segmenting the atrium from MR Angiography (MRA) datasets. The segmented blood pool is subdivided into regions which are later merged but separated at narrowings. This technique is based on the notion that the atrium is commonly connected to neighbouring structures through a narrowing. Although the method is fast and robust, it suffers from several limitations such as its inability to segment in cases where there is a non-narrow connection between a pulmonary vein and the pulmonary artery caused by a string of partial volume affected voxels. Our method is similar to the one described in [5]. An important difference is the manner in which our system can correct for over and under-segmentations using very little user-interaction. A second difference is the calculation of saddle points in [5] which is replaced with finding a point with the highest Euclidean distance value in the separating surface between any two adjacent subdivisions. This gives the true diameter of the separating surface. A third difference is the use of automatic threshold selection for segmenting the blood pool which has greatly improved the quality of our results.

2 Segmentation

The images acquired for this study are MRA images. In these images, the atrium is enhanced along with the pulmonary veins, pulmonary artery and aorta. Surrounding structures such as bones and other organs can also appear

in the MRA sequences and these are easily removed by subtracting the post- from the pre-MRA scans. The input to our segmentation system is thus the subtracted image showing only the enhanced vessels.

2.1 Segmenting the blood pool

Although, subtracting the post- from the pre-MRA scans removes bone and other irrelevant structures, it is still necessary to extract the blood pool and further remove neighbouring enhanced structures as much as possible in order to reduce the computational steps involved in determining the subdivisions. The left atrium is surrounded by neighbouring structures such as the pulmonary arteries, ascending aorta and the left ventricle. During acquisition, blood which is contrast-enhanced perfuses into these regions very quickly making it difficult to avoid imaging these neighbouring structures. These neighbours can be partly removed by using a bounding box which selects a region of interest around the left atrium, thereby excluding the unwanted structures as much as possible. The blood pool within the bounding box is extracted using a region-growing segmentation technique where a seed point inside the atrium is selected. The lower and upper thresholds for blood are automatically computed using the Otsu method [6]. In this method an optimal threshold value which maximizes the *between-class* variance between blood and non-blood tissue classes is determined. The output at this stage is a binary image with the segmented blood pool that contains the left atrium, its pulmonary venous drainages and segments of neighbouring vessels and structures. Although it is not vital to remove the ventricle blood pool in the context of RFCA, the bounding box can cut off the atrium from the ventricle blood pool along a suitable user-defined plane.

2.2 Computing subdivisions

The blood pool is subdivided into regions using a scheme described below. Neighbouring subdivisions are later merged automatically to give the segmented atrium. The process of subdividing the blood pool is done in three steps: 1) Computing the Euclidean distance transformation of the image. 2) Finding local maximums on the EDT map. 3) Determining subdivision membership for each voxel in image.

In the first step we compute the Euclidean distance transformation (EDT) of the binary image. Given that we have a distance metric defined, a distance transformation is an assignment of each voxel to a distance. This distance is the closest distance of a voxel to the background voxel. Defining the metric to be the Euclidean distance between any two points, the EDT can be computed for our binary image. It has recently been possible to compute the exact EDT of a binary image in linear time and we use the algorithm described in Maurer et al. [7]. Under the EDT transformation, a voxel deep inside the blood pool gets a high distance value whereas a voxel in the proximity of a boundary gets a relatively smaller value. Voxels outside the blood pool is assigned a distance value of zero.

The next step is to compute the local maximum points on the EDT map. A voxel is a local maximum if its EDT value is greater than the EDT value of all its neighbours. We use a 26-voxel neighborhood, for calculating local maximums. The spatial distribution and frequency of these local maximum points is of interest. There are a large number of local maximum points in structures that are narrow and small such as the blood vessels which includes the pulmonary venous drainages to the left atrium. Wider structures such as the body of the atrium have a relatively smaller number of local maximums. This is due to the geometric nature of the EDT transform. Once the local maximum points are determined, for each voxel we compute the subdivision it belongs to. Starting from the voxel we search along the path with increasing EDT values until we reach a local maximum. The voxel belongs to the subdivision that is centred by this local maximum. In this way each local maximum produces a subdivision which is generally composed of voxels surrounding it. Figure 1(ii) shows an MRA slice subdivided using the described subdivision scheme.

2.3 Subdivision Merging

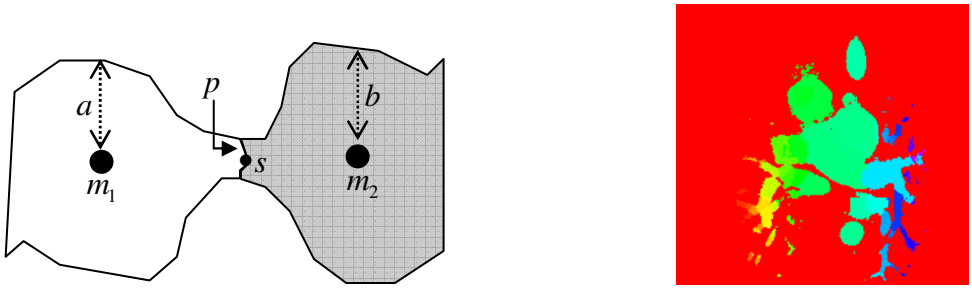
Once the subdivisions are computed, they can be merged interactively and manually by the user. Assuming that the user has some knowledge about the shape of an atrium, subdivisions inside the atrium can be manually chosen by selecting the local maximums which uniquely identify them. This is a semi-automatic approach for segmenting the left atrium. The user selects the local maximums based on a visual assessment of whether it lies inside or outside the atrium. This visual assessment is possible by overlaying the local maximum points on the original MRA of the segmented blood pool. In a next step, the selected subdivisions are merged, giving the segmented atrium. The complexity of this manual procedure depends on the number of local maximums or subdivisions. As described earlier, due to the way the local maximums are spatially distributed, more subdivisions occur within the pulmonary venous drainages than around the body of the atrium. This makes the task of selecting the subdivisions easier and quicker within the body of the atrium and relatively harder and slower at the pulmonary veins. Although this manual technique can be slow; however, when performed correctly it gives good segmentation results. Alternatively, the

subdivisions can be merged without any user intervention by using an automatic scheme described in the following section.

2.4 Automatic Subdivision Merging

Physically the left atrium is only connected to the left ventricle through the mitral valve. However, in MRA images, due to the partial volume effect the blood pool in the atrium becomes frequently connected to the blood in the ascending aorta. Additionally, the pulmonary venous drainages also appear connected to the pulmonary artery. Any left atrium segmentation scheme must be robust against such artefacts. It can be safely assumed that these connections are very narrow regions connecting either sides of a blood pool. The subdivision scheme described in the previous section will produce subdivisions on either side of such narrowings. It now merely remains to describe a merging scheme which merges subdivisions not separated by such narrowings and stops merging subdivisions that are separated by a narrowing.

To detect a narrowing, we first define a size metric for the subdivisions. The size can be attributed by its diameter which is the Euclidean distance value of its centre. These centres, as described in previous sections, are local maximum points. There is also a separating surface between any two adjacent subdivisions. The size of the separating surface is proportional to its diameter value. This is the deepest point within the surface which also has the largest Euclidean distance value.



(i) (ii)

Figure 1: (i) Two subdivisions with their diameters and separating surface (ii) An MRA slice subdivided into disjoint subdivisions.

Consider the *diameters* of two neighbouring subdivisions m_1 and m_2 which are separated by a separating surface *s* (figure 1(i)). The diameter of a subdivision is the Euclidean distance value of its local maximum. Assuming that *a* and *b* are the diameters of two neighbouring subdivisions with the separating surface diameter denoted by *p*. A merging criterion can be described in terms of a merging value which is defined as:

$$\text{merging value} = \min(a,b) - p \tag{1}$$

The merging value can be controlled using a user-defined threshold. The merging criterion allows any two adjacent subdivisions to be merged when their merging value satisfy a threshold. It prevents merging of subdivisions where the merging value falls below the threshold. In this way, the diameter of the separating surface at a narrowing will have a smaller diameter compared to the diameters of the adjacent subdivisions, thus yielding a large merging value. On the contrary, at non-narrow connections, the diameter of the separating surface will be comparable to that of its adjacent subdivisions, thus giving a smaller merging value. Selecting an appropriate threshold value will thus stop the merging process at possible narrowings.

The merging process starts from a user-selected seed point, preferably located close to the centre of the atrium. The subdivision containing the seed point is our seed subdivision. Similar to a region-growing approach, all subdivisions connected to the seed subdivision are merged on the condition that they fulfil the merging criterion. Symbolically, merging two subdivisions is equivalent to replacing them with a single subdivision with a diameter equal to the larger of the two diameters. At the end of the merging process the resulting merged subdivisions is the segmented atrium separated from the rest of the connected structures.

3 Results

We have tested our segmentation technique on 12 patient MRA datasets. The datasets acquired were diverse in terms of the anatomy of the left atrium and its pulmonary veins. One of the datasets had an abnormal enlarging of the left atrial body (figure 2(i)), and one of them had an un-documented and unusual left drainage where the two left pulmonary veins joined into one vessel (figure 2(iii)). The technique was robust against these cases and all cases were segmented successfully. The segmentation results were evaluated by an expert clinician by overlaying the segmentation on the original MRI and noting the differences. An important criterion for a good left atrial segmentation for RFCA is that it must include all the pulmonary veins to the left atrium. Missing a pulmonary vein can lead to an unsuccessful RFCA procedure where the patient can complain of recurring episodes of AF caused by the overlooked pulmonary vein [9]. The inclusion of all pulmonary veins in the segmentation was thus easily validated using this overlaying technique.

Once the EDT and subdivisions are computed, the segmentation process is almost instant taking less than a second for each MRA image. Computation of the EDT for the image is also very fast taking less than a second. However, subdivision computation can take a few seconds depending on the size of the dataset. The segmented images were visualized using a marching cubes isosurface reconstruction where the iso-value for blood was automatically selected using the Otsu method [6].

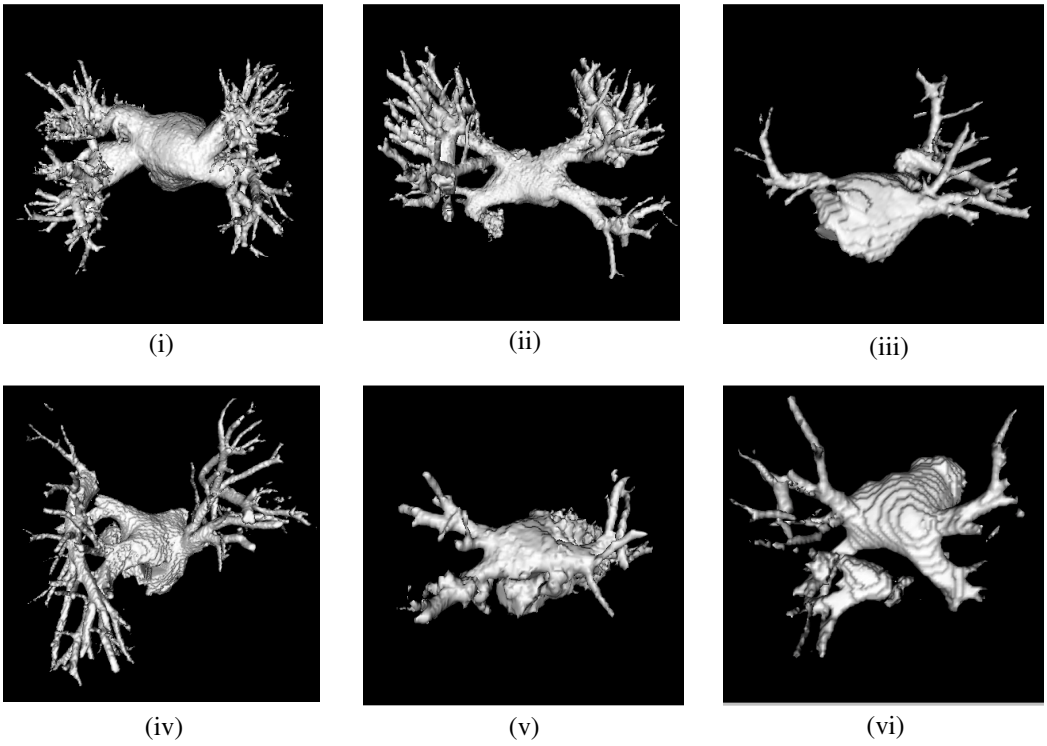


Figure 2: Marching cube iso-surface reconstruction of the segmented left atriums of 6 different patients. Note the variation in shape, size and anatomy of the atrium across patients.

.

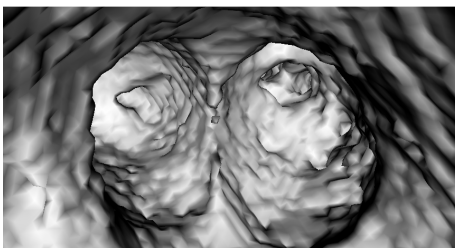


Figure 3: An endocardial of the segmented atrium of figure 2(i). The camera was positioned to face the ostia of the two pulmonary venous drainages to the right side of the atrium.

Electrophysiologists sometime also prefer an endocardial view where a clear view of the drainage openings are visible. This view is important for surgical planning allowing the circumferential path of ablation to be traced easily. Figure 3 shows one such view of the segmented atrium of figure 2(i) where the camera was positioned facing the two drainage openings to the right side of the atrium.

4 Discussions

Due to the high degree of anatomic variability in the left atrium un-predictable shape variations may occur and can make such a data-driven technique susceptible to mis-segmentation. In at least one of the patient data we have noticed narrow regions *within* the atrium (figure 2(i)). This causes the algorithm to cut part of the atrium off at the narrowing producing an under-segmented left atrium. If these narrowings are within an opening of a drainage, an entire pulmonary drainage will be excluded from the segmentation leading to possible surgical failure. However, mis-segmentations can be easily verified by overlaying the segmented image on top of the original MRA dataset.

Our system can easily correct under-segmentations. We do this by identifying a seed point in the missed region and running the algorithm to segment this missed region separately. A different merging threshold is often required. The original segmentation can then be combined with the missed region segmented separately to give the final segmented atrium. In a similar way, possible over-segmentations can also be corrected by identifying a seed point within the over-segmented region and running the algorithm to subtract this region. We have found the subdivision merging process leaking into the pulmonary artery in some instances thus causing the over-segmentation. In such cases, a seed point selected inside the pulmonary artery and running the algorithm with a new merging threshold value allows the over-segmented artery to be segmented and removed from the final segmentation. It is also important to select the right value for the merging threshold (eq. (1)). Selecting a slightly higher threshold than what is appropriate can easily lead to over-segmentation and cause the merging process to leak into neighbouring structures. There are yet no means of automatically selecting this threshold value, and we currently rely on a trial-and-error approach.

5 Conclusions

We have presented a technique for automatically segmenting the left atrium from MRA datasets. The technique exploits the fact that partial volume affected voxels in MRA can cause the atrial blood pool to become connected to the blood pool of neighbouring structures via narrow regions. The Euclidean distance transformation of the blood pool is produced and is used to schematically subdivide the image into disjoint subdivisions. These subdivisions are later merged automatically starting from a seed point and stopping at the partial volume affected narrowings, yielding the segmented left atrium. The system has been tested successfully on 12 diverse MRA patient datasets and the segmentation results have been validated qualitatively by an expert clinician. In future research, we will present our results on a larger collection of MRA datasets and also include a quantitative means for validating the segmentation results. We also intend to investigate the possibility of automatically locating the pulmonary veins in the segmented atrium and determining the diameters of their ostia using endocardial views (figure 3). This can be useful in RFCA where the eletrophysiologist is required to make such accurate measurements before circumferentially ablating around the tissues of the ostia.

6 References

1. "Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation: analysis of pooled data from five randomized controlled trials" *Archives of Internal Medicine* **154**. pp. 1449-1457, 1994.
2. M. D. Lesh, G. F. Van Hare, L. M. Epstein et. al. "Radiofrequency Catheter Ablation of Atrial Arrhythmias: Results and Mechanisms" *American Heart Association* **89**. pp. 1075-1089, 1994.
3. E. M. Marom, J. E. Herndon, Y. H. Kim et. al. "Variations in Pulmonary Venous Drainage to the Left Atrium: Implications for Radiofrequency Ablation" *Radiology* **230.** pp 824-829, 2004.
4. J. V. Berg & C. Lorenz. "Accurate left atrium segmentation in multislice CT images using a shape model" In *Proceedings of the Society of the Photo-Optical Instrumentation Engineers* **5747**. pp. 351-360, 2005.
5. M. John & N. Rahn. "Automatic-left Atrium Segmentation by Cutting the Blood Pool at Narrowings" In *Proceedings of MICCAI*. pp. 798-805, 2005.
6. N. Otsu. "A Threshold Selection Method from Gray-Level Histogram" *IEEE Transactions on Systems, Man. and Cybernet* **9**. pp 62-66, 1978.
7. C. R. Maurer, R. Qi, V. Raghavan "A Linear Time Algorithm for Computing Exact Euclidean Distance Transforms of Binary Images in Arbitrary Dimensions" *IEEE Trans. on Pattern Analysis and Machine Intelligence* **25**. pp. 265-270, 2003.
8. M. Haissauguerre, P. Jais, D. C. Shah et. al. "Spontaneous initiation of AF by ectopic beats originating in the pulmonary veins" *The New England Journal of Medicine* **339.** pp. 659-666, 1998.
9. H. Nakashima, K. Kumagai, H. Noguchi et. al. "Evaluation of the recurrence of Atrial Fibrillation after pulmonary venous ablation" *Japanese College of Cardiology* **40(3).** pp. 87-94, 2002.

Longitudinal Voxel-Based Morphometry with Unified Segmentation: Evaluation on Simulated Alzheimer's Disease

Gerard R. Ridgway[a], Oscar Camara[a], Rachael I. Scahill[b], William R. Crum[a], Brandon Whitcher[c], Nick C. Fox[b], and Derek L. G. Hill[a*]

[a]Centre for Medical Image Computing, University College London.
[b]Dementia Research Centre, UCL. [c]GlaxoSmithKline.

Abstract. The goal of this work is to evaluate Voxel-Based Morphometry and three longitudinally-tailored methods of VBM. We use a cohort of simulated images produced by deforming original scans using a Finite Element Method, guided to emulate Alzheimer-like changes. The simulated images provide quite realistic data with a known pattern of spatial atrophy, with which VBM's findings can be meaningfully compared. We believe this is the first evaluation of VBM for which anatomically-plausible 'gold-standard' results are available. The three longitudinal VBM methods have been implemented within the unified segmentation framework of SPM5; one of the techniques is a newly developed procedure, which shows promising potential.

1 Introduction

Voxel-Based Morphometry [1] is a method for automated whole-brain analysis of local structural differences, using Statistical Parametric Mapping (`http://www.fil.ion.ucl.ac.uk/spm/`); Longitudinal variants have been developed for application to cohorts with serial imaging [2, 3]. VBM necessitates preprocessing of the images, including spatial normalisation and tissue-segmentation. There is great difficulty in evaluating the performance of VBM methods due to the lack of ground truth. To the best of our knowledge, no previously published VBM studies of realistically complex data have had gold-standard maps of the regions that should be detected.

We have developed Finite Element Methods (FEM) which can structurally alter images, producing finely-controllable, clinically realistic changes [4]. Such simulated images have known underlying deformation fields and volume changes, which can form a gold standard for evaluating atrophy-measurement techniques.

Alzheimer's Disease is a progressive neurodegenerative disorder, of great clinical and socio-economic importance. AD causes a loss of brain tissue which can be visualised and quantified using serial Magnetic Resonance Imaging [5]. Using a cohort of AD patients with MR images at baseline and one year later, we simulated new approximate year-on scans from the original baselines, guided by semi-automated measures of whole-brain, hippocampal, and ventricular volume changes [6]. The original baseline and simulated follow-up images then constitute a data-set with known FEM ground truth; we use this to derive a gold standard suitable for evaluating longitudinal VBM, and compare four such techniques, one of which is novel.

2 Methods

2.1 Voxel-Based Morphometry and Longitudinal VBM

VBM in SPM5 involves unified tissue-segmentation and spatial normalisation [7], followed by spatial smoothing and voxel-wise statistical testing. With serial data, statistical analysis can take advantage of reduced within-subject variability. To capitalise on the longitudinal information, changes should also be made to the VBM preprocessing methods. In this work, we evaluate standard VBM against two longitudinal methods from the literature (which we have adapted to be compatible with the unified segmentation framework) and our own newly developed SPM5 method.

All SPM analyses were performed within an explicit mask derived from the smoothed ground-truth grey-matter segmentation. All smoothing was done with an 8mm FWHM Gaussian kernel. A one-sample t-test was performed on subtraction images; single-tailed contrasts for atrophy (increase$<$0) and 'reverse contrast' of tissue-gain (increase$>$0) were evaluated and thresholded with multiple comparison correction using Random Field Theory ($p_{FWE} < 0.01$).

*Derek.Hill@ucl.ac.uk

2.1.1 Standard VBM

Here, 'Standard' VBM refers to simple application of unified preprocessing independently to each scan of each subject; only the statistics differ from the non-serial case. 'Standard' should not be contrasted here to 'optimised' VBM [8], which the unified segmentation model aims to supersede [7].

2.1.2 Tied-normalisation

The preprocessing step of spatial normalisation should take advantage of the fact that multiple time-points for a single subject can be registered much more accurately than scans of different subjects, and that initial rigid alignment already reveals a great deal about within-subject change [5]. Using the non-unified model of SPM2, Gaser (in Draganski et al. [2]) developed a method with longitudinally tied spatial normalisation, in which repeat scans are transformed using the parameters determined for their corresponding baselines, then independently segmented.

Following the introduction of SPM5's unified framework, an extended generative model for unified longitudinal segmentation and normalisation should ideally be developed. As a simpler alternative, we have implemented an approach which applies the baseline normalisation parameters to the native-space baseline and follow-up grey matter images from separate unified segmentations.

2.1.3 Pre-averaged

More advanced techniques can combine inter-subject spatial normalisation with precise intra-subject registration using High-Dimensional Warping (HDW). One such method (designed by Ashburner, and implemented in [3]), creates low-noise averaged images of HDW-registered longitudinal sets, before inter-subject spatial normalisation and segmentation in SPM2. (i.e. averaging is 'pre' segmentation.)

We have adapted this approach to the SPM5 framework, with unified segmentation and inter-subject normalisation following the intra-subject warping and averaging. The intra-subject volume changes from HDW must be taken into account to generate the follow-up data, which can be elegantly done by modulating the native-space segmented average-images with the HDW Jacobian fields before applying the predetermined inter-subject transformations. This avoids the interpolation error due to the transformation of the Jacobians in [3].

2.1.4 Post-averaged

We propose a technique similar to pre-averaging, but novel, and well-suited to SPM5's unified segmentation. The new method should be better for subjects with large longitudinal change that might not be fully recovered by HDW, as in this case, the pre-averaged images may be too blurred to segment well. Each time-point is first segmented, and SPM5's bias-corrected version is saved; HDW transformations are then determined on the corrected images and applied to their native-space segmentations. The warped segmentations are then averaged; i.e. averaging is 'post' segmentation of sharp original images. Each average segmentation is modulated with the HDW volume changes to create follow-up equivalents, then each set of original and modulated segmentations is spatially normalised with the baseline parameters.

2.2 Finite Element Modelling of Atrophy

The atrophy simulation process is based on that described in [4]. It consists of four main steps: (1) Generation of a reference mesh; (2) Warping to a subject-specific mesh; (3) Deformation of the mesh using a FEM solver; (4) Application of the deformations to the baseline image of each subject, to produce a new simulated follow-up image. The reference mesh was built using the BrainWeb atlas labels of these structures [9] (`http://www.bic.mni.mcgill.ca/brainweb/`). The adaptation of the reference mesh to each subject was achieved with a mesh warping procedure guided by a fluid registration algorithm [10].

We used a cohort of 18 probable AD patients (7 female; ages from 55 to 86, mean 70) with baseline and 12-month follow-up MRI scans [6]. The FEM simulation was driven using values of the subjects' volume changes in the brain, hippocampi, and ventricles (from semi-automated segmentation-based measurements). Simulated mean (standard deviation) percentage volume increases were: brain, -2.43 (1.18); hippocampi, -4.74 (3.24); ventricles, 11.49 (5.35). Figure 1(a-c) shows a single-subject example of atrophy simulation; ventricular expansion, cortical thinning, and opening of CSF spaces can be observed.

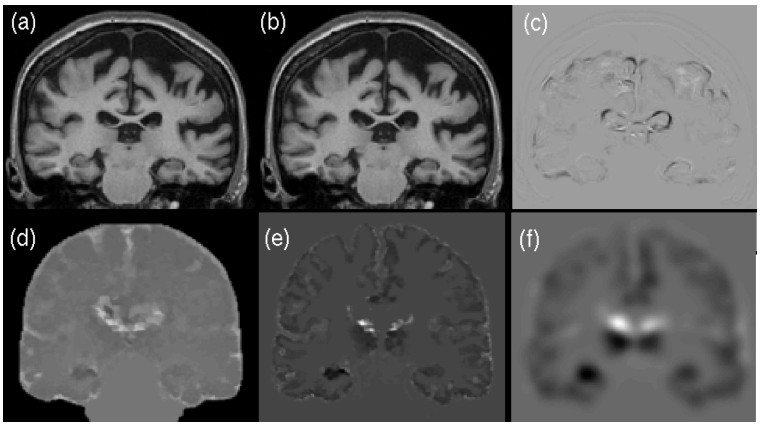


Figure 1. Example case of simulated atrophy: (a) Original baseline; (b) Simulated +1yr follow-up; (c) Subtraction image. The same subject's gold-standard volume changes in BrainWeb space: (d) Volume gain (VG=1yr/orig); (e) GM-increase = (GM*VG)-GM; and (f) Smoothed GM-increase, as entered into the analysis.

2.3 Generation of a Gold Standard

Because the same mesh is warped from the BrainWeb template to each individual patient, there is a known correspondence between elements of the warped meshes for the different subjects; therefore the volume change of each element can be mapped back to the common space. By converting the element-wise volume changes to a voxel-wise representation, an image of the ratio of follow-up to original volume is created. These volume gain ratio images can be used to modulate the BrainWeb Grey Matter Segmentation, resulting in perfectly aligned effective follow-up segmentations, similar to those in the two HDW-based longitudinal VBM methods. The original BrainWeb GM is then subtracted from each follow-up and the result smoothed. Figure 1(d-f) illustrates this process for one subject. The gold-standard smoothed subtraction images could be entered into an identical one-sample t-test as the actual sets of VBM subtraction images. For reasons discussed in section 4.1, we instead use contrast (negative mean) images, thresholded at different values for visualisation purposes.

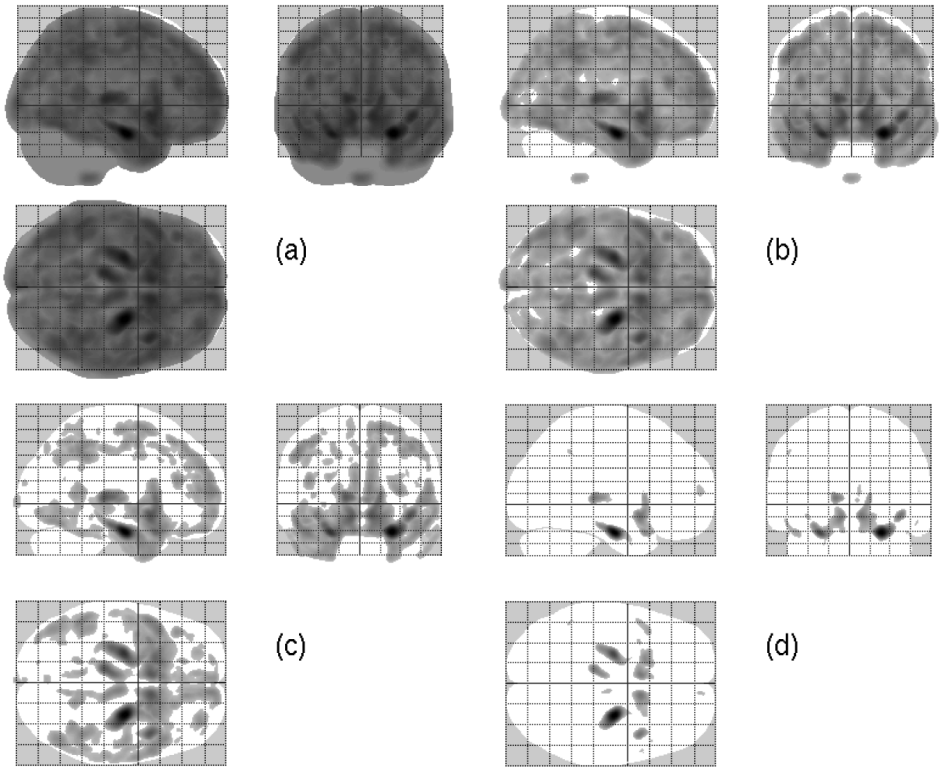


Figure 2. Gold-standard average atrophy, Maximum Intensity Projections thresholded at: (a) 0, (b) 0.01, (c) 0.02, (d) 0.03.

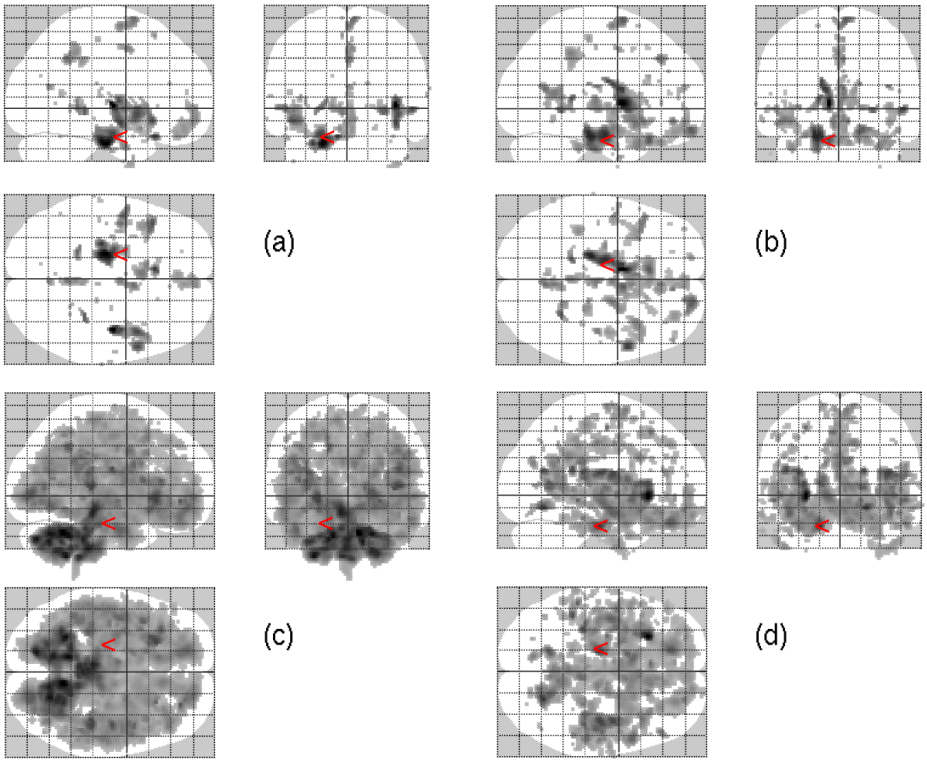


Figure 3. Maximum Intensity Projections of significant atrophy ($p_{FWE} < 0.01$) for VBM methods: (a) Standard; (b) Tied-normalisation; (c) Pre-averaged; (d) Post-averaged.

3 Results

Gold-standard maximum intensity projections, at varying thresholds, can be seen in figure 2. Statistical results from the four VBM methods are presented in figure 3 as maximum intensity projections. In both cases the atrophy t-contrast (increase$<$0) is shown. For the 'reverse contrast' (i.e. gain of GM over time), none of the four methods detected any voxels at the corrected level. Table 1 shows correlations between the ground truth contrast image and the contrast or t-value images for the four methods:

Method	std	tied	pre	post
contrast	0.24	0.17	0.63	0.65
t-values	0.16	0.10	0.47	0.36

Table 1. Image-wise Spearman rank-correlations (over in-mask voxels).

4 Discussion

4.1 From Simulation Ground Truth to VBM Gold Standard

The atrophy simulation method gives the ground-truth volume change over the nodes of the deformed mesh. However, several steps are required to convert this data into a gold standard for VBM. Following the obvious approach of performing the same statistical analysis of the ground-truth GM-increase images as of the VBM subtraction images, we obtained unrealistic t-maps (not shown). Unreasonably large t-values occur outside the regions in which FEM volume changes were introduced. This may be due to the low spatial variance of the volume change maps outside these areas. Inter-subject spatial variability is far lower for these images than it is for natural anatomical variation in real patient images, even after spatial normalisation.

Instead, we threshold the 'contrast image' — the numerator of the t-statistic. The maximum intensity projections (see figure 2) now look sensible, but this method leaves open the question of at what level the contrast image should be thresholded. On the other hand, there is also a degree of arbitrariness in the choice of threshold (α) for statistical tests. The approach taken here of presenting differently-thresholded versions of the gold standard allows a multi-scale evaluation of the pattern of atrophy. Note that the purpose of the gold standard is to indicate the spatial pattern of simulated atrophy; its significance is not of innate interest.

4.2 Longitudinal VBM

The gold-standard results shown in figure 2 indicate the presence of diffuse global atrophy, with greater focus on the temporal lobes and strongest change in the hippocampi; the cerebellum and brain-stem are spared. The key advance of this work is that the VBM methods may be compared directly to this desired pattern, as well as to each other.

We first note that all four methods appear to perform better at detecting the hippocampal and temporal lobe atrophy compared to the more diffuse cortical atrophy. This is probably due to the greater natural anatomical variation in the pattern of cortical folding. Standard VBM detects the least atrophy of the four methods, though there are no obvious false-positive regions. Longitudinally tied normalisation seems to give only minor improvements, though there is some evidence that the less variable tied registration preserves more of the cortical atrophy. Intriguingly, the correlation with ground truth (table 1) appears worse. Both HDW-based methods appear much more sensitive, though some of the atrophy they report is not apparently well-matched to the gold standard (e.g. the insula). In addition, some areas present in the gold standard appear to be missed despite the greater apparent sensitivity (e.g. temporal horns, and the focal nature of the hippocampal atrophy). The correlations in table 1 reaffirm the superiority of the HDW-based methods, but don't allow a clear preference for either.

The pre-averaging method [3] seems to produce false-positive results in the cerebellum and brainstem. Our new post-averaging method appears to avoid this, at the expense of detecting less true cortical atrophy. Additionally, the post-averaging method has better detected the hippocampal atrophy. Reasons for the differences are not entirely clear, as both methods use the same HDW transformations. The pre-averaging method segments (and normalises) an image with higher signal-to-noise ratio but potentially significant blurring; while post-averaging of original (lower SNR segmentations) also improves the SNR of the results. The relative merits of the two alternatives need further investigation.

We note that there are statistical objections to the comparison of t- or p-values, as a difference in significance is not equivalent to a significant difference. However, with fundamentally different methods such as the standard and HDW-based VBM evaluated here, there is a risk of registration problems if the ANOVA interactions between atrophy and method are tested. The findings shown here are intended to allow comparison between distributions of detected atrophy of the methods and gold standard, with the aims of informing choices between different VBM methods, and guiding further comparative studies. In the future, we hope to perform more quantitative analysis, including a direct ground-truth based investigation of segmentation performance. The evaluation may be complemented by testing the methods on real data.

Acknowledgements

GRR is supported by an EPSRC CASE Studentship sponsored by GSK. RIS and NCF are supported by the MRC.

References

1. J. Ashburner & K. J. Friston. "Voxel-based morphometry–the methods." *Neuroimage* **11(6 Pt 1)**, pp. 805–821, June 2000.
2. B. Draganski, C. Gaser, V. Busch et al. "Neuroplasticity: changes in grey matter induced by training." *Nature* **427(6972)**, pp. 311–312, Jan 2004.
3. G. Chételat, B. Landeau, F. Eustache et al. "Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study." *Neuroimage* **27(4)**, pp. 934–946, Oct 2005.
4. O. Camara, M. Schweiger, R. Scahill et al. "Phenomenological Model of Diffuse Global and Regional Atrophy Using Finite-Element Methods." *Medical Imaging, IEEE Transactions on* **25(11)**, pp. 1417–1430, Nov. 2006.
5. N. C. Fox & J. M. Schott. "Imaging cerebral atrophy: normal ageing to Alzheimer's disease." *Lancet* **363(9406)**, pp. 392–394, Jan 2004.
6. J. M. Schott, S. L. Price, C. Frost et al. "Measuring atrophy in Alzheimer disease: a serial MRI study over 6 and 12 months." *Neurology* **65(1)**, pp. 119–124, Jul 2005.
7. J. Ashburner & K. J. Friston. "Unified segmentation." *Neuroimage* **26(3)**, pp. 839–851, July 2005.
8. C. D. Good, I. S. Johnsrude, J. Ashburner et al. "A voxel-based morphometric study of ageing in 465 normal adult human brains." *Neuroimage* **14(1 Pt 1)**, pp. 21–36, July 2001.
9. D. Collins, A. Zijdenbos, V. Kollokian et al. "Design and construction of a realistic digital brain phantom." *Medical Imaging, IEEE Transactions on* **17(3)**, pp. 463–468, June 1998.
10. W. R. Crum, C. Tanner & D. J. Hawkes. "Anisotropic multi-scale fluid registration: evaluation in magnetic resonance breast imaging." *Phys Med Biol* **50(21)**, pp. 5153–5174, Nov 2005.

Classification of Colon Biopsy Samples by Spatial Analysis of a Single Spectral Band from its Hyperspectral Cube

Khalid Masood and Nasir Rajpoot

Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK
email: {khalid, nasir}@dcs.warwick.ac.uk

Abstract.
The histopathological analysis of colon biopsy samples is a very important part of screening for colorectal cancer. There is, however, significant inter-observer and even intra-observer variability in the results of such analysis due to its very subjective nature. Therefore, quantitative methods are required for the analysis of histopathological images to aid the histopatholgists in their diagnosis. In this paper, we exploit the shape and structure of the gland nuclei cells for the classification of colon biopsy samples using two-dimensional principal component analysis (2DPCA) and Support Vector Machine (SVM). We conclude that the use of textural features extracted from non-overlapping blocks of the histopathological images results in a non-linear decision boundary which can be efficiently exploited using a SVM with appropriate choice of parameters for its Gaussian kernel. The SVM classifier outperforms all the remaining methods by a clear margin.

1 Introduction

Colon cancer is one of the leading forms of cancer and second most fatal, after the lung cancer, in England and Wales [1]. The disease can be treated very effectively if detected in its early stages. Routine screening (ie, colonoscopy in this case) can save lives by nipping the problem in the bud. However, the screening process relies heavily on the accuracy of the judgement made by histopathologists analysing the biopsy samples taken from suspected polyps. Unfortunately, the histopathological analysis is very subjective in nature and is marred by both intra- and inter-observer variability, potentially leading to different treatment regimes. The purpose of this study in the grand scheme of things is to increase the reliability of the screening process by introducing objectivity to the screening process.

Several researchers have looked at the problem of classification of biopsy slides. However, due to the limited space here, it is not possible to give a comprehensive literature review. In [2], classification of tumors is done using expression levels of gene patterns in the tissue samples. In [3], it is proposed that metrics based on the responses of receptive field operators modelling the orientation selectivity of the neurons found in the early visual pathway are capable of discriminating between images of normal, dysplastic (transitional) and cancerous samples. A related computational model of light interaction with colon tissue and classification using the tissue reflectance spectra is analyzed in [4].

In this paper, we address two important questions: First, does the spatial analysis of a single spectral band (as opposed to the spatial-spectral analysis of the corresponding hyperspectral image cube) suffice for efficient classification of biopsy sample? An obvious advantage of using spatial analysis on a single band is its reduced computational and storage complexity. Spectral analysis has been used in [5] and our goal is to determine whether the spatial analysis can achieve similar or even better classification performance. The second question addressed in this paper is: How effective are the features extracted using linear subspace projection of raw image blocks as compared to textural features from the same blocks?

Our approach is based on the idea that development of colon cancer alters the macroarchitecture of the tissue glands. The cancerous stimuli cause cells to adapt by altering their pattern of growth. This phenomena results in the increase in the size of existing nuclei and also considerable increase in their number. The nice tubular structure for a normal tissue changes to deformed structure for malignant tumors. The malignant tumor shows considerable variation in nuclei size and shape. The following sections present materials and methods used in our experimentation, followed by quantitative results and their discussion. The paper ends with some concluding remarks and future directions.

2 Materials and Methods

A tissue micro-array of stained biopsy samples for several different patients is prepared using standard Haematoxylin and Eosin (H&E) staining procedures. Standard digital colour or greyscale images for histopathological analysis of each of the biopsy samples are normally obtained by placing the micro-array under a microscope equipped with a

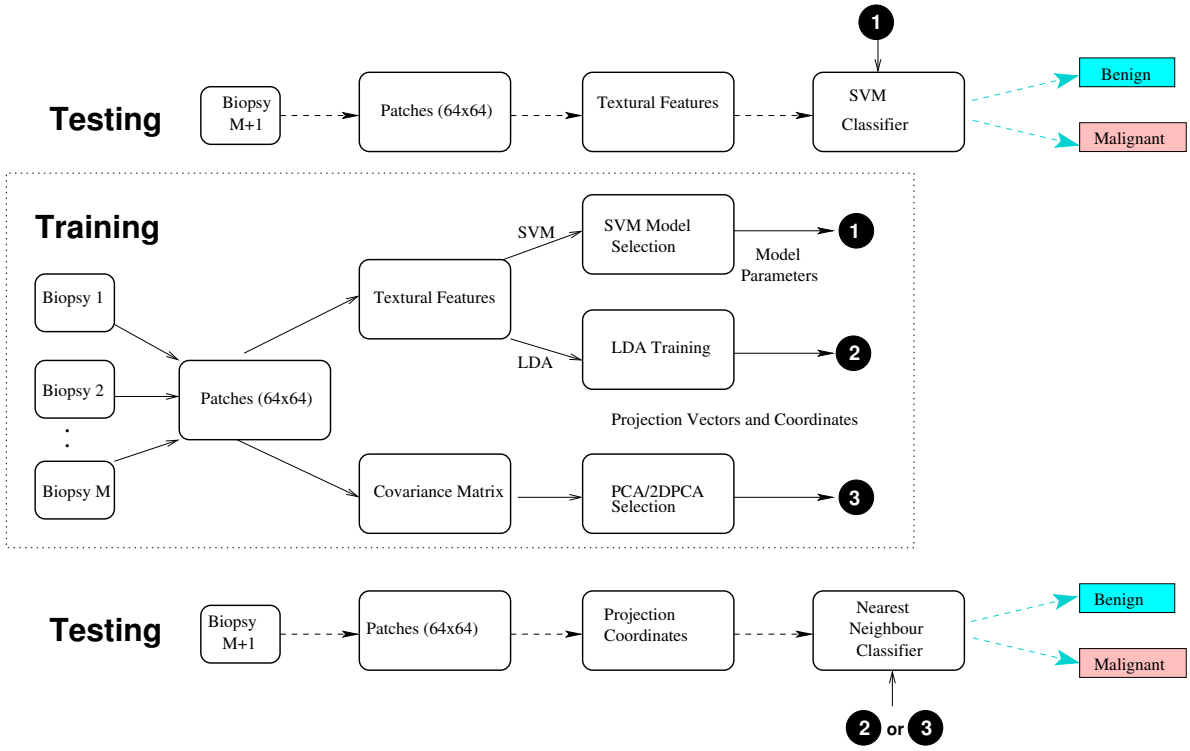


Figure 1. Block Diagram

visible light source and a CCD camera. In our case, however, the imaging setup (same as that used in [5]) consisted of a Nikon Biophot microscope and a unique tuned light source based on a digital mirror device (DMD). The DMD based light source is capable of transmitting any combination of light frequencies. Hyperspectral image cubes of the biopsy sample at a magnification of 400X are captured at 128 different wavelengths, with narrow bandwidths, in the range of 400–800*nm*. In this way, each image cube consists of 128 spectral bands (*spectral resolution*) while the spatial resolution of each spectral band is 448×691. The image cubes are spatially cropped to obtain spectral bands with a spatial resolution of 448×640, in order to facilitate our local spatial analysis which operates by analysing 64×64 blocks of a single hand-picked spectral band for each cube. After experimentation with number of bands, observing the effect of a particular band on the classification performance, we found that a number of spectral bands in the middle part of the visible light spectrum for most datasets seemed to contain sufficient textural information for biopsy classification purposes. We picked one of the middle bands, the 75th band, for all our experiments reported in this paper.

A unified block diagram of the methods investigated is shown in Figure 1. As depicted in the Figure, there are two types of classification paradigms employed for comparison purposes: first type using raw image blocks and subspace projection methods (classical principal components analysis, PCA, and two-dimensional PCA), and the second type using textural features for the image blocks and utilising LDA or SVMs. The methodology is described in more detail in the following sections.

2.1 Classification using Raw Image Patches

In order to address the second question in Section 1, we investigate whether features obtained by maximising the overall scatter of the raw image blocks contain sufficient information for discrimination purposes. Such features can be obtained by the classical linear subspace projection (or dimensionality reduction) method of PCA. We also investigate the use of two-dimensional PCA (2DPCA) [6] for feature extraction from raw image blocks. First, we give a brief description of the two methods.

2.1.1 Subspace Projection of Raw Patches with PCA

The subspace projection methods operate by projecting a given $n \times n$ image block, or patch, $P_i (i = 1, 2, \ldots, M)$, where M is the total number of training image blocks) onto a small number d of projection vectors $X_k^{(1)}$, for $k = 1, 2, \ldots, d$.. In case of PCA, the projection vectors are the eigenvectors corresponding to the largest d eigenvalues of a covariance

matrix $C^{(1)}$ obtained from all the M training image blocks as follows,

$$C^{(1)} = \frac{1}{M}\sum_{i=1}^{M}(v_i - \bar{v})(v_i - \bar{v})^T$$

where v_i denotes the patch P_i rearranged into a one-dimensional (1D) vector (normally, in a row-by-row fashion) and $\bar{v}$ denotes the average of all such vectors for the training patches. The above equation results in a potentially large covariance matrix $C^{(1)}$ having dimensions $n^2 \times n^2$, although the so-called transpose trick can be used to reduce its dimensions to $M \times M$, nevertheless requiring extra computations for calculating the projection coordinates.

2.1.2 Two-dimensional Principal Component Analysis (2DPCA)

2DPCA is a new linear subspace projection method particularly developed for image classification. It was recently shown to be successful for face recognition [6]. The basic idea behind 2DPCA is that the covariance matrix is computed using the training images without requiring to first convert them to 1D vectors. This is done as follows,

$$C^{(2)} = \frac{1}{M}\sum_{i=1}^{M}(P_i - \bar{P})^T(P_i - \bar{P})$$

As a result of the above computation, the covariance matrix $C^{(2)}$ is of the size $n \times n$, which could be much smaller than its conventional counterpart $C^{(1)}$. Furthermore, the linear subspace projection using 2DPCA is a more accurate reflection of the spatial relationship between image pixels in all the training samples as it maximises the overall scatter of all the training images. Just as matrices for training images are used for computation of $C^{(2)}$, the computaion of projection coordinates for P_i is also done by projecting the corresponding matrix onto the first d eigenvectors of $C^{(2)}$ as follows,

$$a_i k = P_i X_k^{(2)}$$

where $X_k^{(2)}$ is the $n\times$ 1-dimensional eigenvector of $C^{(2)}$ corresponding to its kth largest eigenvalue λ_k. One fundamental difference between 2DPCA and conventional PCA is that the projection coordinates (or the 2DPCA feature matrix) for P_i in case of 2DPCA can be represented in the form of a $n \times d$ matrix A_i given by

$$A_i = [a_{i1}, a_{i2}, \ldots, a_{id}]$$

as opposed to a d-dimensional feature vector in case of the classical PCA obtained by projecting v_i onto $X_k^{(1)}$, for $k = 1, 2, \ldots, d$. The classification of a test sample is done by computing the sum of Euclidean distances between the columns of feature matrices of the test image and those stored in the database for training samples. A nearest neighbour classifier can be used to assign label to the test sample.

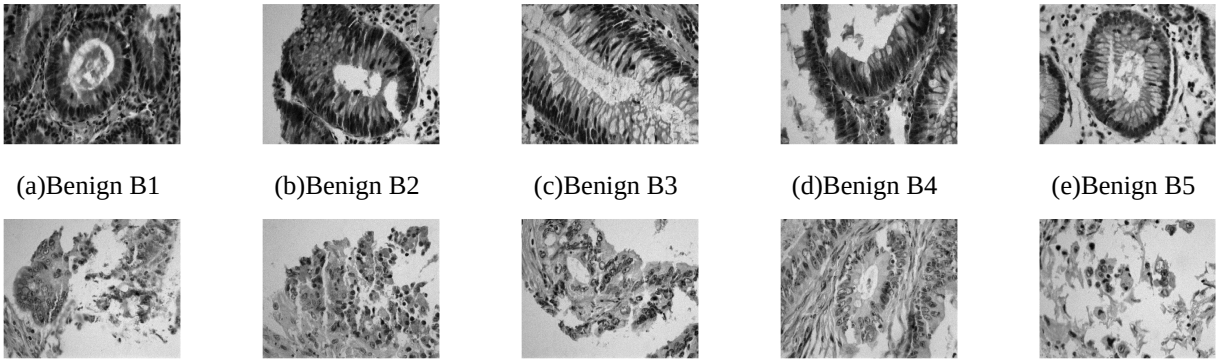
2.2 Classification using Textural Features

Although 2DPCA has been shown to be more efficient than classical PCA for image classification, it has some drawbacks too, such as the large dimensionality of the features (ie, the feature matrices) and the high computational complexity of classification of a test sample due to potentially large number of distance calculations. Therefore, we also investigate the effectiveness of local textural features for biopsy classification using a single spectral band.

Grey level co-occurrence matrix features are calculated for every patch P_i of the image corresponding to a single spectral band of the given hyperspectral cube for a biopsy sample. A co-occurrence matrix is computed using second order joint conditional probability density function $f(i, j|d, \theta)$ by counting all pairs of pixels separated by distance s in direction θ and having grey level i and j. By using two distance values and four values of θ (0, 45, 90, 135 degrees), we can compute a relatively small feature vector, which we used with a classical linear discriminant analysis (LDA) classifier and also appropriately tuned SVMs.

2.2.1 Linear Discriminat Analysis (LDA)

One of the major drawbacks of the two kinds of PCA discussed above is that they maximise the overall scatter but do not take into account the variability between samples belonging to the same class and those belonging to different classes. LDA overcomes this limitation by making use of both within-class and between-class scatter. For a more detailed coverage of LDA, the reader is referred to [7]. It suffices to mention here that the LDA projection vectors can be used in the same way the projection coordinates or feature vectors for training samples can be obtained using the classical PCA.



(a)Benign B1 (b)Benign B2 (c)Benign B3 (d)Benign B4 (e)Benign B5

(f)Malignant M1 (g)Malignant M2 (h)Malignant M3 (i)Malignant M4 (j)Malignant M5

Figure 2. Images of Colon Biopsy Samples

2.2.2 Support Vector Machine (SVM)

SVMs utilise a nonlinear mapping from the input feature space to an implicit high-dimensional feature space, where the nonlinear boundary between patterns in the input space is linearized [8]. The SVM kernel function allows one to avoid the explicit evaluation of mapping by using the so-called kernel trick. The choice of kernel depends on the data and its clustering. Gaussian kernel is defined below

$$K(x_i, x_j) = e^{-\gamma(x_i - x_j)^2} + C,$$

where x_i and x_j denote feature vectors corresponding to two patches P_i and P_j, is used in most cases because of its widely reported superior classification performance, also in this context [9]. In the above equation, C is a constant which normally does not significantly affect the classifier's performance, whereas the choice of γ (width of the Gaussian basis function) can have a significant influence on the performance. A variety of selection methods have been used including grid search and Newton's bisection methods [9]. Once a kernel is tuned properly, classification of test data can be performed effectively.

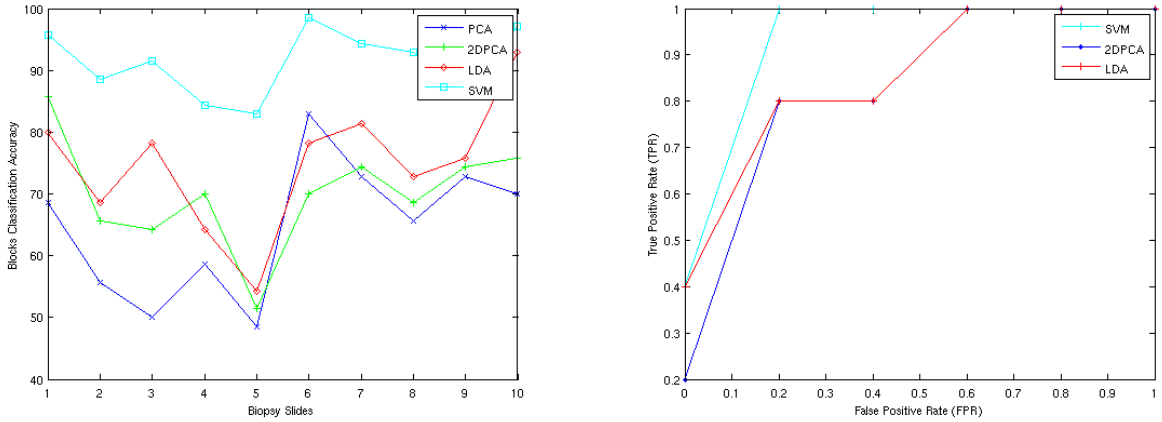
3 Results and Discussion

We conducted experiments with ten biopsy samples using a leave-one-out (LOO) testing strategy. As shown in Figure 2, five of these were benign and the remaining five malignant. The 75th spectral band of the hyperspectral image cubes for these biopsies was each divided into non-overlapping blocks of 64×64. For spectral band selection, we perform several experiments on different bands using SVM and found that middle bands have more textural information than initial and final bands, whereas the 75th band gives highest classification accuracy. In a 10-fold cross-validation setting, each time the training set consisted of patches from nine images, while the remaining tenth image was used for testing. The PCA and 2DPCA classifiers used 60 and 40 projection coordinates, respectively. For textural features, two values of s (namely, 1 and 2) and four values of θ as mentioned above are used and three features (namely, energy, variance and homogeneity) are used, yielding a 24-dimensional feature vector for each of the biopsies. LDA further used only 20 projection vectors for computing the linearly discriminating feature vectors. The Gaussian kernel for the SVM was tuned with a bisection search. A summary of block-wise classification accuracy results are shown in Table 1.

Correct classification accuracy of image blocks by all four methods for all the biopsy slides is shown in Figure 3(a). Each biopsy slide is given a label according to a cutoff threshold on the accuracy of labels assigned to its blocks. If we choose a 50% threshold, we get 10 slides out of 10 (except PCA) correctly labelled for each of the classifiers, although by doing that we risk having too many false positives or false negatives in practise. At 60% threshold, PCA performance degrades and only 6 slides are correctly classified while 2DPCA labels 9 slides in the right class. At this threshold, SVM has 100% classification accuracy. When we select the threshold to be 70%, LDA accuracy is down to 70% while SVM still performs at 100% accuracy. A plot of true positive rate (TPR) versus false positive rate (FPR), also known as the receiever operating characteristic (ROC) curve, is shown in Figure 3(b). SVM with textural features is the clear winner, as it has the largest area under the convex hull (AUCH) of all the ROC curves. It is our view that more than 70% is a reasonable value for the threshold if we consider that upto one third of some of the images contain only background information. It is worth noting that SVM yields 100% accuracy with a 80% threshold.

Non-Overlapping Blocks Classification Accuracy (%)							
Biopsy No.	**Raw data**		**Textural data**				
	PCA	**2DPCA**	**LDA**	**SVM(65)**	**SVM(70)**	**SVM(75)**	**SVM(80)**
B1	68.57	85.71	80.00	77.14	95.71	95.71	95.71
B2	55.71	65.71	68.57	61.43	81.43	88.57	84.29
B3	50.00	64.29	78.14	48.57	64.29	91.43	78.57
B4	58.57	70.00	64.29	54.29	51.43	84.29	75.71
B5	48.57	51.43	54.21	30.00	30.00	82.86	41.43
M1	82.86	70.00	78.14	98.57	98.57	98.57	98.57
M2	72.86	74.29	81.43	94.29	81.43	94.29	84.29
M3	65.71	68.57	72.86	94.29	97.14	92.86	94.29
M4	72.86	74.29	75.71	84.29	84.29	85.71	80.00
M5	70.00	75.71	92.86	98.57	92.86	97.14	92.86
Average	64.57	70.00	74.63	74.14	77.72	91.14	82.57

Table 1. SVM (with different Bands) and PCA, 2DPCA, LDA (Band 75) Results



(a) Classification Accuracy (b) ROC Curve

Figure 3. Performance Curves for 75th Spectral Band

4 Conclusions and Future Work

Our experimental results demonstrate that by taking only single band from a hyperspectral data cube, it is possible to perform a reasonable classification for histopathological analysis of colon biopsy samples. An appropriately tuned SVM coupled with textural features yields the best performance as compared to other subspace projection methods with and without textural features. This is because SVM introduces a nonlinear mapping where the decision boundary is linearized and hence separation is performed much more efficiently. The segmentation of glands and the measurement of features related to glandular shapes will be subjects of our future study.

References

1. N. S. Office. "Cancer statistics and registrations for england and wales." *HMSO* 1999.
2. T. Furey, N. Cristianini & N. D. et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." *Bioinformatics* 2000.
3. A. Todman, R. Naguib & M. Bennen. "Visual characteristics of colon images." *IEEE CCECE* 2001.
4. D. Rowe, E. Claridge & T. Ismail. "Analysis of multispectral images of the colon to reveal histological changes characteristic of cancer." *MIUA* 2006.
5. G. Davis, M. Maggioni & R. C. et al. "Spectral/spatial analysis of colon carcinoma." *Journal of Modern Pathology* 2003.
6. J. Yang, D. Zhang & A. Frangi. "Two-dimensional pca: A new approach to appearance-based face representation and recognition." *IEEE transactions on PAMI* 2004.
7. K. Fukunaga. "Introduction to statistical pattern recognition." *Academic Press* 1990.
8. V.Vapnik. "The nature of statistical learning theory." *Springer-Verlag, New York* 1995.
9. K. M. Rajpoot & N. M. Rajpoot. "Hyperspectral colon tissue cell classification." *SPIE Medical Imaging (MI)* 2004.

Comparative Analysis of Discriminant Wavelet Packet Features and Raw Image Features for Classification of Meningioma Subtypes

Hammad Qureshi[a], Nasir Rajpoot[a], Roland Wilson[a], Tim Nattkemper[b] and Volkmar Hans[c] *

[a]Department of Computer Science, University of Warwick, Coventry, CV4 7AL, UK
[b]Applied Neuroinformatics, University of Bielefeld, 33615 Bielefeld, Germany
[c]Institute of Neuropathogy, Evangelisches Krankenhaus, 33617 Bielefeld, Germany

Abstract. The idea of multiresolution analysis has been around for over two decades now. In this paper, we explore a multiresolution analysis based technique for histopathological image classification and compare it with raw image analysis. The principle idea for the former is to derive an optimal wavelet representation, called Adaptive Discriminant Wavelet Packet Transform (ADWPT), of the images in order to achieve the largest discrimination power. Our results show that the textural features combined with ADWPT yield a significant improvement in classification accuracy.

1 Introduction

One of the most important issues that has hampered the solution to the problem of image classification has been efficient pattern representation. Many researchers have argued that to achieve good image classification results, first an efficient representation of the image classes under consideration must be derived. It is widely accepted that the human visual system performs a transformation of the input image into a spatial frequency decomposition before the brain actually performs the recognition task. The idea of optimal representation has been researched widely and various techniques have been proposed. In this paper, we present an optimal, in terms of its discriminatory power, wavelet decomposition for the problem of histopathological image classification and demonstrate that it performs better than image statistics based analysis.

Histopathological diagnosis of tumours, especially of the brain and spinal cord, even today requires decision making by human experts. Manual diagnosis and decision making is hampered by two limitations: First, reviewing histological slides by humans is time consuming and the human experts are not always available. Secondly, although a lot of effort has been made to exactly define diagnostic criteria for all tumour entities within the World Health Organization (WHO) Classification of Tumours [1] but the inter-rater variability is still considerable (see e.g. [2]). This consequently influences further therapy regimens greatly and hence a bias is introduced. Due to the fast progress in digital image retrieval and analysis technologies, machine based decision making may be used to substantially support histopathologists by providing more objective diagnostic parameters and allow for high-throughput analysis. A first step, however, towards developing new algorithms for image classification (i.e., 'diagnosis') is to test whether the automated technique is able to reproduce human assignment of single tumour samples to diagnostic classes. In order to develop such a technique, we have focussed on meningiomas which are tumours of the brain and the spinal cord arising from cells of the surrounding (i.e. meningeal) coverings of the brain. Meningiomas account for 20% of all brain tumours and exist in three different grades of malignancy (WHO Grad I-III), most being benign (over 80%), but some showing an increased propensity to recurrence and rare cases being malignant. Most benign WHO Grade I meningiomas belong to one of four subtypes: Meningiothelial, fibroblastic, transitional, and psammomatous. Figure 1 shows the four meningioma subtypes and their salient features. Correct histopathological diagnosis can be made in most cases by a trained human expert, i.e. a neuropathologist. Therefore, this tumour is well suited for testing diagnostic properties of new machine based pattern recognition algorithms used in histopathological analysis.

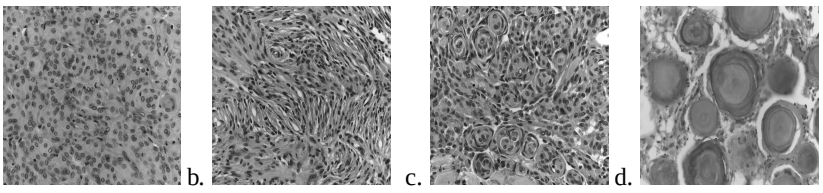


Figure 1. Meningioma Images at 40x magnification for each subtype a. Meningiothelial (cells form synctium), b. Fibroblastic (spindle shaped cells in collagen-rich matrix), c. Transitional (cells form whorls with psammoma bodies), d. Psammomatoes (high number of psammoma bodies)

Numerous studies [5–7] have successfully demonstrated the utility of wavelets and other transforms for image classification. Lessman et al. [3] studied the problem of content-based visualization of meningioma images to aid in

*contact author: hammad@dcs.warwick.ac.uk

characterization of the database contents using the wavelet transform. In our previous work [4], we presented the results of preliminary investigation into meningioma subtype classification using wavelet packets. Saito and Coifman [8] showed how wavelet packet transform may be used for local features extraction using relative entropy as the criterion for basis selection. Rajpoot [9] developed discriminant wavelet packets for texture classification. Support Vector Machines (SVMs) have been employed frequently in the literature to perform pattern classification. In [10], Nattkemper et al. demonstrated the ability of SVM by proving that they perform better than decision trees and nearest neighbor (k-NN) classifiers in breast tumor classification.

This paper presents a technique called the Adaptive Discriminant Wavelet Packet Transform (ADWPT) for feature extraction and image classification and compares its performance with classification using Haralick [11] features from raw images. The AWPT subbands can be thought of as descriptors [12] which can be used for image compression but they may also be employed as features for texture-based image classification. Our experimental results demonstrate the strength of an ADWPT based image representation for meningioma subtype classification.

2 Materials and Methods

The image acquisition process involved the analysis of the routinely stained histopathological slides on a Zeiss Axioskop 2 plus microscope with a Zeiss Achroplan 40x/0,65 lens to obtain 1300×1030 pixels, 24 bit, true color RGB pictures. Representative images of the four meningioma subtypes considered in this study are shown in Figure 1. These images were converted from their RGB values to 8-level grayscale images for analysis and cropped to 1024×1024 pixels. Each image is further divided into 4 parts of 512×512. In this paper, we compare two approaches first using raw image features and then wavelet packet features (refer Figure 3) to classify amongst patterns.

2.1 Raw Image Based Analysis

In image based analysis, we obtain statistical features described in section 2.3 after dividing a given image into non-overlapping slabs of 32 by 32, 64 by 64 and 256 by 256 pixels. Each slab is taken as a separate entity for features extraction. There were a total of 256 features for 32×32, 64 features for 64×64 and 4 features for 256×256. A divided image is shown in Figure 2a.

2.2 ADWPT-based Analysis

The first step to any statistical image classification technique is the computation of features to be fed into a classifier. It is of paramount importance that the features acquired are relevant and characterize the underlying image classes. As opposed to the ordinary wavelet transform, the ADWPT adaptively decomposes both the high frequency and the low frequency components of a signal at multiple resolutions. ADWPT is based upon the wavelet packet transform [13]. The computation of ADWPT decomposition is explained in the following sections.

2.2.1 Adaptive Discriminant Wavelet Packet Transform (ADWPT)

The first stage of our technique is the computation of the best wavelet packet basis in terms of the discrimination power of the four meningioma subtypes. This is important because our aim is to obtain a representation of the image which is effective in localizing the spatial frequencies and comparing the most different aspects of the images.

Let X be a set of all the WPT subbands at all possible resolutions. For an ADWPT based image classification, the problem can be posed as the selection of an optimal combination of subbands from a subset Y of X such that (*a*) it yields a complete basis for any given image and (*b*) it has the largest discriminatory power of all other combinations that can yield a complete basis. The aim of ADWPT basis selection process then is to obtain Y of cardinality n which achieves the best classification performance such that $n < N$ and $Y \subset X$. The process is divided into two steps:

Full Wavelet Packet Transform (FWPT)

The wavelet transform is computed by applying a highpass and lowpass filter upon the input signal to acquire the high frequency and low frequency subbands. The wavelet packet transform is performed iteratively on every subband to a certain level to obtain the high and low frequency coefficients at each stage. Without loss of generality, the transform for a 1D discrete signal $\mathbf{x} = \{x_i\}, i = 0, 1, ..., N-1$ may be computed as follows:

$$S_0^0(i) = x_i \quad (1)$$

$$S_j^{2p}(i) = \sum_k g_{k-2l} S_{j-1}^p(k) \quad (2)$$

$$S_j^{2p+1}(i) = \sum_k h_{k-2l} S_{j-1}^p(k) \quad (3)$$

where $i = 0, ..., N-1$, $l = 0, ..., 2^{-j}N-1$, g_n and h_n are lowpass and highpass filters respectively, $j = 1, 2, ..., J; J = log_2 N$, $S_j^p(l)$ is the transform coefficient corresponding to the wavelet packet function having relative support size 2^j, frequency $p2^j$ and is located at $l2^j$. Hence, j, p and l are regarded as the scale, frequency and position indices of the wavelet packet function. The transform is invertible if appropriate dual filters $\tilde{g}_n$ and $\tilde{h}_n$ are applied on the synthesis side.

A simple wavelet transform decomposes only the low level frequencies iteratively whereas wavelet packets involves decomposing all the subbands. The wavelet packet decompositions are maintained in a quadtree structure, with the parent being the original subband or image and the children being the wavelet decompositions of the parent. First the image is decomposed into its respective subbands and then each subband is decomposed further until a predefined maximum depth of the tree is reached. This results in a combinatorial explosion of possible wavelet packet bases that can be used to completely represent the image and its called the Full Wavelet Packet bases. The next stage is the selection of the best bases which is achieved with the help of dynamic programming.

Computation of Discrimination Power

First, a pseudo probability density function (pdf) is obtained for each subband using the normalized energy for the subband coefficients. A psuedo probability density function is computed by dividing the square of a coefficient by the sum of the square of the coefficients in a subband. The next step is to compute the average pseudo probability density functions (apdf) by summing the pseudo pdfs pairwise of the training images belonging to the same class and dividing them by 2.

$$\mathcal{A}(X_a, X_b) = (x_i^a + x_i^b)/2$$

X_a and X_b are the pseudo pdfs of two samples belonging to the same class. The process is repeated for all the subbands of the training images and it is important to note that the average of two sample images is computed per iteration. The objective is to acquire a basic model of localized frequency values for each class so that the difference between the classes may be estimated. Then the discriminating power of each subband is obtained, using the Hellinger distance between different classes, as follows.

$$\mathcal{D}(X, Y) = \sum_i (\sqrt{x_i} - \sqrt{y_i})^2$$

where x_i and y_i are the normalized energy of the ith subband coefficient for two training images of different classes. This distance is used as the discriminatory power $\mathcal{P}$ in the best basis selection process. The pairwise distance between the 4 different kinds of textures is obtained based upon the following formulae.

$$\mathcal{P}_k = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \mathcal{D}(X_i^k, X_j^k)$$

where i and j represent the different class indices.

Best Basis Selection

1. Compute the J-level full wavelet packet tree decomposition.
2. Initialize $j = J - 1$.
3. For all $0 \leq p < 2^j$, $0 \leq q < 2^j$, do the following:

   a If $\mathcal{P}_{p,q}^j < max[\mathcal{P}_{2p,2q}^{j+1}, \mathcal{P}_{2p,2q+1}^{j+1}, \mathcal{P}_{2p+1,2q}^{j+1}, \mathcal{P}_{2p+1,2q+1}^{j+1}]$ keep the four child subbands at depth $j+1$ where $\mathcal{P}_{p,q}^j$ represents the discrimination power of a node at position i, j

   b otherwise keep the parent at depth j and remove the child subbands.

4. Decrement j by 1.
5. If $j < 0$, then stop, otherwise goto step 3.

One important fact to mention here is that when we apply the ADWPT on different training or input data, in most cases we get the same decomposition. Figure 2b shows the decomposition obtained for four different data sets which implies that the decomposition remains consistent even when different subsets of texture data are used.

2.3 Statistical Features Extraction

Once the ADWPT of the images and the divided raw images are obtained, the next step is the computation of the statistical features. Gray co-occurrence based features proposed by Haralick [11] are used in our technique. We use texture correlation, contrast, homogeneity and energy.

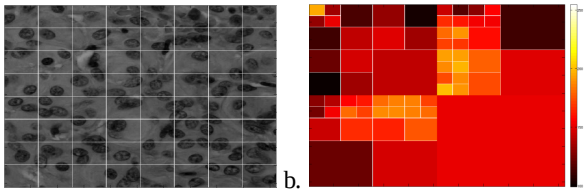


Figure 2. a. Raw image divided into 64x64 pixel slabs, b. Best basis decompositions for 4 different data sets with color denoting the discrimination power of the subbands

2.4 Classification using Support Vector Machines

SVM is a supervised classifier, which approximates the decision surfaces of the theoretical Bayes classifier. SVM has found broad area of applications since its invention in 1995 by Vapnik [14]. SVM uses various kernels to map the input space into a higher dimensional feature space to make the non-linear hyperplane linear. To achieve this without increasing computational complexity a kernel trick is employed.

In the training phase, the statistical features based on the gray level co-occurrence matrix of the subbands (in case of ADWPT, image at each resolution is called a subband) and slabs (in case of image blocks) are used as input vectors. Multiple SVMs are trained and feature vectors are used as training data. The Matlab version of LibSVM provided by Chang and Lin [15] was used in our analysis. The results were obtained using a Gaussian SVM. Figure 3 depicts the algorithms employed in this study.

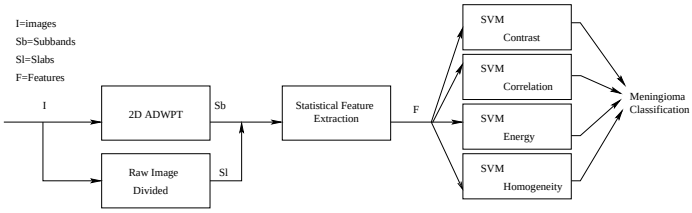


Figure 3. Block diagram of our technique

3 Results and Discussion

The experimental setup consisted of employing the Daubechie 8 tap filter for obtaining the ADWPT upto four levels. The averaged results of the various test runs are shown in the Table 1. The results presented have been cross-validated with four different test trial runs using different training and testing data sets. A total of 64 length features set is derived using ADWPT and the raw image based analysis yields feature sets of length 256 (32×32 slabs), 64 (64×64 slabs) and 4 (256×256 slabs).

In all 320 image samples were available as input data for training and testing. 80 image samples for each subtype of meningioma were available for experimentation. The data is divided into five batches per meningioma subtype with each batch containing 16 images. During the training and testing phase a leave one out mechanism is observed which means that 1 batch is kept for testing while the rest are used for training.

Feature	F_r	M_r	P_r	T_r	$Overall_r$	F_a	M_a	P_a	T_a	$Overall_a$
Contrast	32.8	29.7	93.8	28.1	46.1	90.7	62.5	98.5	76.6	82.1
Correlation	0	0	0	100	25	75	68.7	96.9	73.4	78.5
Energy	54.7	34.4	89.1	29.7	51.9	56.2	73.4	100	82.8	78
Homogeneity	46.8	39	89.1	23.5	49.6	54.7	67.2	96.9	81.3	75

Table 1. Raw image based (denoted by subscript r) vs. ADWPT features (denoted by subscript a) %age classification accuracy results (correct classification %age averaged over 4 trials) of meningioma images using SVM and gray-level co-occurrence features (F=Fibroblastic,M=Meningiothelial,P=Psammomatous,T=Transitional)

Table 1 shows the average results for various trial runs using statistical features over raw images and ADWPT based representation. The results for 32×32 and 256×256 slabs were also obtained but there was no significant improvement

in performance, hence, only the results for 64×64 are shown. Our results show that ADWPT performed consistently better than raw image features. It is also notable that for the psammomatous subtype both techniques perform consistently well. The primary reason for this, is that the psammomatous texture is significantly different from the other textures included in the study since it contains dispersed psammoma bodies and therefore, is most easily differentiable. It would be important to comment upon the performance of the statistical features computed and why they were chosen for analysis. The idea is to capture the degree of contrast, correlation, energy and the degree of homogeneity of pixels in case of image slabs and wavelet coefficients in case of wavelet packet subbands. From the results it can be seen that the contrast feature performs well for ADWPT representation and not too well for raw image features. The main reason for this is that with wavelet representation the whole image is analyzed at different scales and frequencies represented by each subband. Whereas incase of raw image features the image is divided and features are computed for each slab separately which means only local features embodied by the slab are analyzed. The ADWPT representation performs better as each correlation feature derived from it represents the whole image information at various scales and frequencies whereas in case of raw textural analysis only correlation features at a certain location in the image are computed (due to this very reason the contrast feature is not performing well in case of raw image analysis). Hence, ADWPT provides us with a powerful tool to find inherent textural properties which cannot be achieved with raw image analysis.

It is important to note here that all these features are taken separately and results computed exclusively. If they are combined in some novel way such as envisaged in our previous work [4], the results may improve. However, in this paper our focus has been to demonstrate that the classification accuracy improves significantly when ADWPT based image representation is used instead of raw images.

According to Table 1, there are certain meningioma subtypes for which the classification accuracy is not as good as it is in the case of other meningioma subtypes. Particularly fibroblastic and meningiothelial subtypes have been found difficult to differentiate as they resemble each other to a certain extent. Furthermore, the statistical features being employed at this instance have only been able to capture properties that are better able to differentiate between the subtypes showing better accuracy. Our future work would focus on improving these results by using various more sophisticated features, experimenting with different classifiers and deriving better more optimal representations using the ADWPT.

References

1. P. Kleihues & W. K. Cavenee. *World Health Organization Classification of Tumours. Pathology and Genetics. Tumours of the Nervous System*. IARC Press, 2000.
2. P. Burger. "What is an oligodendroglioma?" *Brain Pathol (2002)* **12**, pp. 257–259, 2002.
3. B. Lessmann, V. Hans, A. Degenhard et al. "Feature space exploration of pathology images using content-based database visualization." In *Proceedings SPIE Medical Imaging*. 2006.
4. H. Qureshi, N. Rajpoot, K. Masood et al. "Classification of meningiomas using discriminant wavelet packets and learning vector quantization." In *Proceedings of Medical Image Understanding and Analysis*. 2006.
5. M. Unser & M. Eden. "Multiresolution feature extraction and selection for texture segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11(7)**, pp. 717–728, 1989.
6. A. M. P. Michael M. Leung. "Scale and rotation invariant texture classification." In *Record of The Twenty-Sixth Asilomar Conference on Signals, Systems and Computers*. 1992.
7. R. Porter & N. Canagarajah. "Gabor filters for rotation invariant texture classification." In *Proceedings of 1997 IEEE International Circuits and Systems*. 1997.
8. N. Saito & R. R. Coifman. "On local feature extraction for signal classification." In *Applied Analysis (O. Mahrenholtz and R. Mennicken, eds.), special issue of Zeitschrift fur Angewandte Mathematik und Mechanik, Akademie-Verlag, Berlin*. 1996.
9. N. Rajpoot. "Local discriminant wavelet packet basis for texture classification." In *Proceedings SPIE Wavelets X, San Diego, California*. 2003.
10. T. W. Nattkemper, B. Arnrich, O. Lichte et al. "Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods." *Artificial Intelligence in Medicine* **34(2)**, pp. 129–139, 2005.
11. R. Haralick. "Statistical and structural approaches to texture." In *Proceedings of the IEEE*. May.
12. N. Rajpoot, R. Wilson, F. Meyer et al. "A new basis selection paradigm for wavelet packet image coding." In *Proceedings IEEE International Conference on Image Processing (ICIP)*. 2001.
13. R. R. Coifman & M. V. Wickerhauser. "Entropy-based algorithms for best basis selection." *IEEE Transactions on Information Theory* **38**, pp. 713–718, 1992.
14. V. Vapnik. *Statistical Learning Theory*. Springer N.Y., 1998.
15. C.-C. Chang & C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at url-http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Improving Insight in Medical Volume Rendering

Ken Brodlie[a] and Marcelo Cohen[b*]

[a]School of Computing, University of Leeds, UK, [b]PUCRS, Brazil

Extended Abstract

Volume rendering is now a well established technique for the display of three-dimensional volumetric data. The fundamental idea is to model the dataset as a heterogeneous gel-like material whose colour and opacity at any point is determined by a mapping, or transfer function, of the data value. The gel is then rendered using computer graphics techniques such as ray casting.

The technique has had some success in medical visualization although it remains quite common for clinicians to prefer sequences of 2D images. This challenge has spurred the visualization community to seek improvements to the basic volume rendering technique, and some of these improvements will be discussed in the talk.

A difficulty until recently has been the speed at which volume rendered images can be created. Ray casting using CPU technology is slow and does not allow real-time interaction by changing the viewing direction for example. This has dramatically changed in the last few years with the advent of GPU technology. The data can be stored as a 3D texture and fast GPU-based texture rendering now allows interactive volume rendering on commodity hardware. An excellent overview of this area is provided by the IEEE Visualization tutorial [1].

The improved hardware performance has been accompanied by an investigation of new algorithmic techniques. Much of this comes from study of traditional medical illustration, where special effects are used to enhance the communication of important detail. Rather than try to reproduce reality, medical illustrators focus on a more abstract approach this can be traced back to Leonardo da Vinci who used exploded views to illustrate a baby in a womb.

Thus illustrative volume rendering has become an important research topic. It has been explored from two directions: geometric effects, where the picture is distorted to highlight some area, or feature, of interest; and rendering effects, where effectively the colour and opacity of the gel is modified in order to highlight some particular features. In fact most of the ideas combine elements of each approach. A simple but vital geometric effect is an ability to clip away parts of the volume that are not of interest.

The talk will review these algorithmic developments in general. Pioneering work on rendering effects was carried out by Rheingans and Ebert [2], where they showed that non-photorealistic effects such as silhouette enhancement can increase our ability to comprehend medical image data. Again an excellent starting point for study in this area is an IEEE Visualization tutorial [3]. An important recent paper is by Bruckner et al [4] where the opacity is modulated so as to make important regions more opaque, and less important more transparent.

For a study of geometric effects we look in detail at one particular piece of work by ourselves at Leeds. This adopts the focus and context principle first put forward for two-dimensional presentation in information visualization. The idea is simple: often there is some area within the image that one wants to explore in detail, but by zooming in we lose the overall picture. In focus and context, we enlarge the focus region but retain a reduced size surrounding region to maintain context. This idea transfers quite readily to 3D, and has been found useful in the visualization of cerebral aneurysms from CTA data. A box can be defined around the aneurysm as the focus, and the data distorted so as to enlarge the box area and compress the remainder. Three styles of distortion can be applied: bifocal (as just described); fish-eye, where there is a continuous decrease in magnification from a specified focus point; and a volume lens, simulating a magnifying glass, where the focus region is magnified, the surrounding area is left unchanged, and there is a steep transition between the two.

This work is fundamentally a geometric approach. However it is enhanced by a number of rendering effects, such as modifying the opacity of a voxel according to its position relative to the focus region, or its relation to the viewing direction. Another useful effect is to highlight the focus region by colorisation. An evaluation by the neurosurgery team at Leeds General Infirmary was very positive perhaps surprisingly the geometric distortion did not cause any concern to the surgeons (although an interventional radiologist was less enthusiastic about this aspect).

*kwb@comp.leeds.ac.uk; marcelo.cohen@pucrs.br

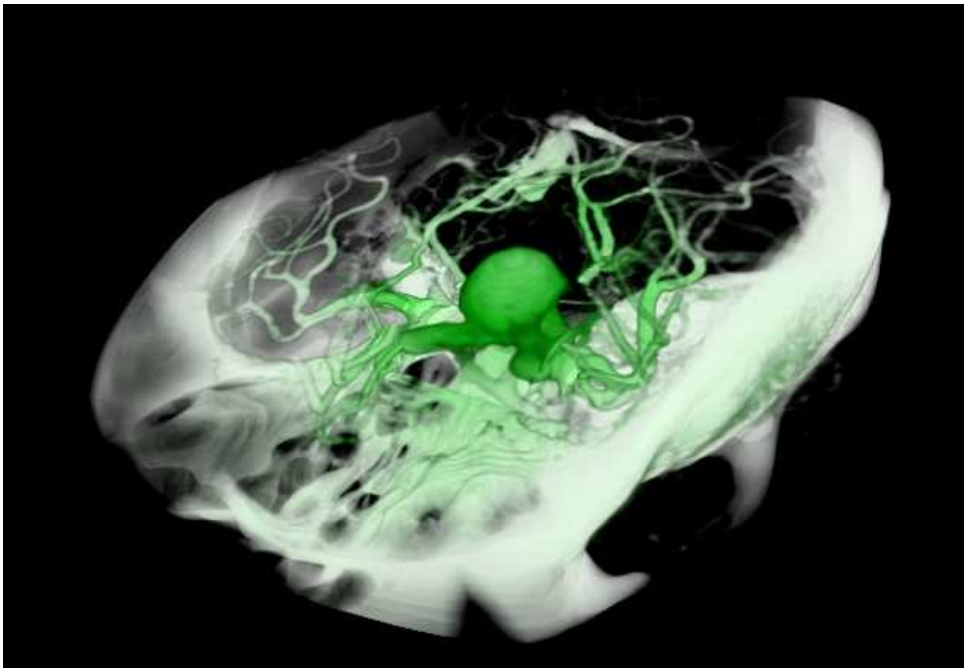


Figure 1. Focus and context visualization of a cerebral aneurysm

A key feature of the work is its careful exploitation of the GPU. The distortion is calculated by the CPU but stored as a texture and then used by the GPU in rendering. The figure shows an example of the work. Full details are in Marcelo Cohen's thesis [5]

References

1. M. Hadwiger. "Introduction to GPU volume rendering.", 2003. See http://www.vis.uni-stuttgart.de/vis03_tutorial/hadwiger_gpu_volrend2.pdf.
2. P. Rheingans & D. Ebert. "Volume illustration: Nonphotorealistic rendering of volume models." *IEEE Transactions on Visualization and Computer Graphics* **7(3)**, pp. 253–264, 2001.
3. I. Viola, M. E. Groller, K. Buhler et al. "Illustrative visualization.", 2005. See http://www.cg.tuwien.ac.at/research/publications/2005/Viola-vistutillustrativevis/.
4. S. Bruckner, S. Grimm, A. Kanitsar et al. "Illustrative context-preserving exploration of volume data." *IEEE Transactions on Visualization and Computer Graphics* **12(6)**, pp. 1–11, 2006.
5. M. Cohen. *Focus and context in volume visualization*. Ph.D. thesis, School of Computing, University of Leeds, 2006.

Contextual detection of ischemic regions in ultra-wide-field-of-view retinal fluorescein angiograms

C. R. Buchanan[a], E. Trucco[b], T. Aslam[c] and B. Dhillon[c] *

[a]School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK
[b]School of Computing, University of Dundee, UK
[c]Princess Alexandra Eye Pavilion, Edinburgh, UK

Abstract. We report a novel prototype algorithm using contextual knowledge to locate ischemic regions in ultra-wide-field-of-view retinal fluorescein angiograms. We use high-resolution images acquired by an Optos ultra-wide-field-of-view (more than 200 degrees) scanning laser ophthalmoscope. We leverage the simultaneous occurrence of ischemia with a number of other signs, detected automatically, typical for the state of progress of the condition in a diabetic patient. The specific nature of ischemic and non-ischemic regions is determined with an AdaBoost learning algorithm. Preliminary results demonstrate above 80% pixel classification accuracy against manual annotations.

1 Introduction and motivation

This paper reports a novel algorithm leveraging contextual knowledge to locate ischemic regions in ultra-wide-field-of-view retinal fluorescein angiography (FA). Our contributions are a context-based detection scheme for retinal ischemic regions in FA, a context framework based on a pathological (not just anatomical) knowledge, and the use of ultra-wide-field-of-view (UWFV) angiograms.

Our main clinical motivation is proliferative diabetic retinopathy (PDR), a major cause of blindness. Its aetiology involves the development of ischemic areas in the retina [1], which cause specific changes visible through angiography [2]. Other common conditions (e.g., retinal vein occlusions, arterial occlusions) involve development of ischemic areas. As PDR, they are identified by clinicians in angiograms and can be treated successfully by laser photocoagulation [3]. It is essential to identify appropriate signs of ischemic disease, as laser treatment must be applied at an appropriate time to prevent neovascular complications [4]. Computer analysis and identification of ischemic areas would be useful in a variety of ways. Firstly, objective analysis of angiograms can act as a useful marker for trials investigating treatments, e.g., for diabetes and vessel occlusions; current studies are hampered by the lack of objective outcome measures. Secondly, as exact aetiologies are still being investigated, quantitative characterisation of retinal features together with accurate recording of patients may provide insight into the pathogenesis of potentially crippling disorders. Finally, in areas where expert ophthalmic opinion is not available, image analysis may help to identify patients at higher risk of developing neovascular complications requiring careful follow-up or laser treatment.

To our best knowledge, the literature of retinal image analysis includes only a few papers on detecting ischemic and hypoxic retinal regions. A possible reason is that the characterisation of ischemic regions in FA in terms of own image properties, e.g., texture, shape or intensity, proves elusive. In contrast, the appearance of several retinal elements and pathology signs is rather well defined, e.g., the optic disk is generally elliptical; vessels are elongated structures with predictable intensity properties; leakages cause wide, bright areas in FA. Jasiobedzki et al. [5,6] segmented angiograms into regions using morphology, then estimated the degree of perfusion as the local density of the microcapillary net. The authors report good agreements with clinician annotations on an unspecified number of angiograms. No characterisation of the data (e.g., image quality, presence of distracting signs or symptoms) was included. Conrath et al. [7] reported semi-automatic detection of avascular foveal zone (FAZ), which is enlarged in diabetic retinopathy, in angiograms. They selected angiograms (1280×1000 pixels) free of interfering artefacts (exudates, laser scars, haemorrhages etc.), and the commercially available ENVI geostatistics package. They concluded that FAZ detection was possible but improvements were necessary in image quality. Such improvements include instruments enabling observation of the retinal periphery, where several diabetic signs appear first. Our work is based on angiograms acquired with an Optos ultra-wide-field scanning laser ophthalmoscope (P200MA), capable of field of view of more than 200 degrees at 3900×3072 pixels with 256 grey levels (Figure 1). [8, 9] are examples of work using images from such sensors. Using such images, we propose a *detection process based on co-occurring pathology signs* to characterise ischemic regions reliably. The basic idea is to leverage the simultaneous occurrence of ischemia with a number of other signs, typical for the state of progress of the condition in a diabetic patient. Our clinical authors have identified a set of signs supporting the hypothesis that a candidate region is ischemic; the problem is then cast as one of combining evidence from multiple classifiers. Evidence combination is performed by AdaBoost [10], whereby a reliable

*e-mail: c.r.buchanan@hw.ac.uk

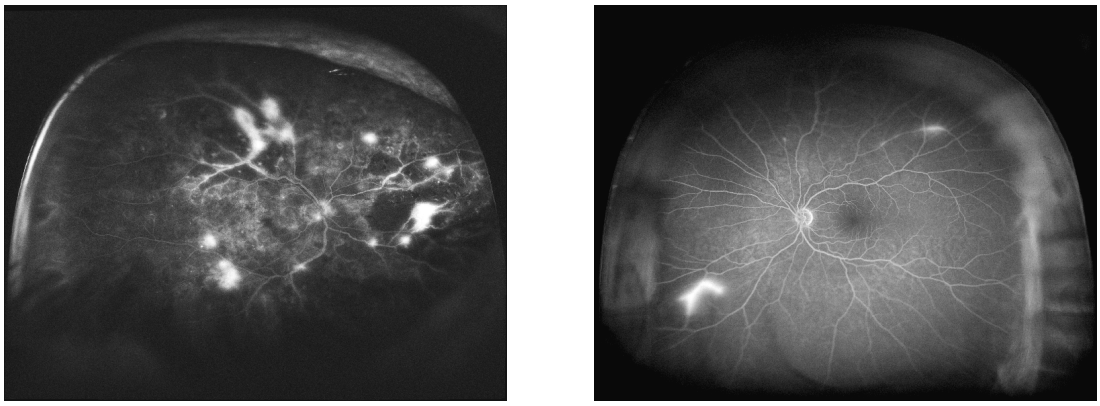


Figure 1. Two examples of UWFV angiograms with (left) and without ischemic regions.

committee classifier is built from a set of weak classifiers. We describe our preliminary prototype using time-intensity information and four contextual signs (distance from vessel bifurcations, density of vessels, presence of leakages and microaneurysms), and initial results on UWFV FA sequences indicating the reliability of the context-based scheme.

We are not aware of contextual detection of ischemic regions in retinal image analysis. Previous investigations on the use of context for reliable location of retinal elements, e.g., in the presence of seriously degraded images, has concentrated on *anatomical knowledge*, for example the relative position of optic disk, macula and blood vessels [8,11]. Our definition of context for ischemic regions extends this idea to the co-occurrence of pathological signs, based on *medical knowledge* about the progress of a condition (the set of co-occurring signs vary as the condition progresses). Context-based reasoning has also been reported for computer-assisted diagnosis STARE [12].

2 Methodology

The algorithm is visualised in Figure 2 and detailed in the following sections. The input is a sequence of FA frames. The outputs are the co-ordinates and approximate contours of regions classified as ischemic. Detection is performed on a pixel-by-pixel basis, where we consider the likelihood that a particular pixel is within an ischemic region. This is achieved by a supervised learning algorithm (AdaBoost) using appearance, temporal and contextual features. The algorithm is trained with examples of ischemic and non-ischemic regions.

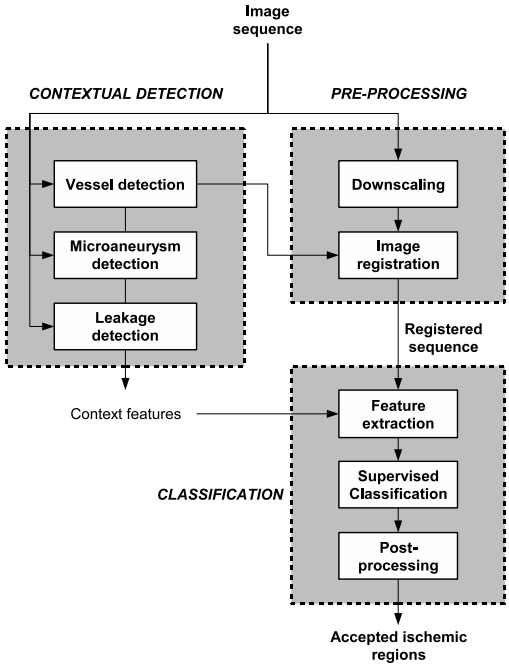


Figure 2. System architecture.

2.1 Pre-processing

Pre-processing consists of vessel network location, downscaling and alignment. A standard, simple illumination normalisation is also performed to compensate partially for effects due to patient motion or instrument adjustment.

Vessel network location. For each frame $(I_1,\ldots,I_n)$ vessels are located automatically and used for both image alignment and contextual detection. In brief, a vessel mask is constructed for each frame by applying matched filtering to enhance linear structures followed by grey-level thresholding. Bifurcations are located by skeletonisation and application of rules in an 8-connect pixel neighbourhood over the skeleton image [13].

Downscaling. To improve efficiency, frames are downscaled by a factor of four using bicubic interpolation. Fourfold downscaling does not destroy the intended targets (ischemic areas and context). In addition, the impact of sensor noise is reduced.

Image registration. The frame alignment algorithm considers each pair (I_i, I_{i+1}) of consecutive frames and matches control points (vessel bifurcations) between image pairs. Bifurcations are detected as sketched above. A bifurcation matching algorithm then searches for corresponding pairs of control points. Correspondences are computed by contextual feature matching; attributes include branching angles, vessel widths, vessel cross-profiles and the displacement vectors to neighbouring bifurcations. Finally, a global affine transform identified by the correspondences is estimated by least squares and applied. Rigid registration provides an approximate solution sufficient for this initial study but which ought to be replaced by a more sophisticated scheme offering improved registration accuracy.

2.2 Feature definition

From the downscaled frame sequence we extract features of appearance, time and context associated with each pixel.

Average image. An average image is constructed by simply taking the pixel-wise average over all frames. The intention is to capture areas that remain consistently dark over an FA session.

The time-intensity profile. The *time-intensity profile* is the function of pixel intensities over the whole sequence. We expect true ischemic regions to show little increase in intensity in comparison to areas of normal perfusion. We consider averages of the time-intensity profile values in five adjacent temporal intervals (bins) spanning the sequence and the relative change between adjacent bins. We observed that capture times vary for each session, i.e., the k-th frame (same position in sequence) of different sessions may be captured at different times from injection. Consequently, before further analysis we must obtain a measure of the signal at certain absolute times for comparison between sessions, and for this reason we perform binning of the frames by time stamp, not by frame number. The chosen time windows are (in seconds): before 30; 30-59; 60-149; 150-299; after 300. Each bin may contain from 3 to 6 frames approximately, depending on the sequence. For each bin b, an average value $x(b)$ is calculated. The result for each pixel is a 5-dimensional feature vector, $x(1),\ldots,x(5)$. To capture the relative change in intensity between bins we simply calculate the differences between $x(b)$ and $x(b-1)$ for $b=2,\ldots,b=5$. As a result, the temporal feature is a 9-dimensional vector (5 bins, 4 differences).

Contextual features. We incorporate information of surrounding vasculature and pathologies; namely the vessel mask, bifurcation points, areas of hyper-fluorescent leakage and microaneurysms. The associated detectors cannot be detailed for reasons of space; in essence, vessels and bifurcations are detected as sketched in Section 2.1; leakages by spatio-temporal classification of intensity profiles [14], and microaneurysms by matched filtering, thresholding, region growing and analysis of size, shape and intensity, following [15]. For each of the contextual targets we obtain a binary image mask for each of the five time bins. Distance maps (the contextual features) are created for each mask by a Euclidean distance transform defining the distance between each pixel and the nearest nonzero pixel in the mask.

2.3 Classification

For each candidate pixel the final feature vector is a combination of average image, time-intensity profile and contextual features. We address this problem as a binary classification task, where a pixel (feature vector) is assigned to one of two classes, ischemic or non-ischemic.

Training the AdaBoost classifier. Ground truth data with ischemic regions were marked out for training. Accordingly, the training data consists of a set of feature vectors $\mathbf{x}_1,\ldots,\mathbf{x}_N$ with corresponding target labels $t_1,\ldots,t_N$ where $t_j \in$

$\{-1,+1\}$. AdaBoost [10], an adaptive boosting scheme, is used to combine a series of weak classifiers into a powerful committee classifier. For this study each weak classifier is a decision stump, equivalent to a simple threshold on one of the input variables. During training AdaBoost creates a series of weak classifiers over a number of rounds $m = 1,\ldots,M$. For each round a distribution of weights is updated indicating the importance of examples in the data set where the weights of incorrectly classified examples are increased. In this way the next classifier focuses more on those examples. The result is a series of weak classifiers $h_1,\ldots,h_M$ and corresponding weighting coefficients $\alpha_1,\ldots,\alpha_M$, with greater weights assigned to more accurate classifiers. Effectively we employ AdaBoost to select the most informative criteria with which to distinguish between the two classes given the supplied features.

Classifying pixels in unseen sequences. To classify pixels in an unseen test sequence, a set of feature vectors $\mathbf{y}_1,\ldots,\mathbf{y}_N$ is extracted, applying the aforementioned methods to detect vessels, microaneurysms and areas of fluorescein leakage. This data is supplied to the trained classifier, which queries each of the weak learners and combines their outputs. The result is a prediction $H(\mathbf{y_i})$ for each test example $\mathbf{y}_i$, where the sign denotes the class as positive or negative (Eq. (1)). Changing the parameter θ (normally set to zero) makes the prediction more conservative or more speculative.

$$H(\mathbf{y}) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(\mathbf{y}) - \theta\right) \tag{1}$$

Post-processing segmentation. Classification results are recorded in a final image (map), where regions of positive predictions are considered ischemic. Morphological closing is employed to minimise small areas considered to be unreliable. Accepted regions can be characterised by area and level of severity defined by the magnitude of their score from Eq. (1). Boundary tracing is then applied to the accepted regions. The final output is a list of region coordinates scaled to match the size of the input image. Notice that we concentrate on reliable detection at this stage, meaning the existence and approximate position of ischemic regions, not on their exact extent and shape. Exact characterisation is however already computed by the scheme components, but requires refinements to yield sufficient reliability.

3 Experimental results

For the purpose of this preliminary study we acquired five anonymous UWFV FA sequences (three with ischemia and other pathologies and two absent of ischemia). The sequences contain 20 to 40 frames captured over approximately six minutes from injection of the fluorescein dye. Ground truth data for all sessions was obtained by manually tracing ischemic regions on registered frames. Care was taken to avoid including vessels in the ischemic regions. This dataset was used to train and test the classifier using a *leave-n-out* procedure so that training was performed on four sequences and testing performed on the remaining held-out sequence. Training of the classifier proceeds as described in Section 2.3. We observed that good generalisation can be achieved by approximately 200 training rounds.

Two examples of final detections are presented (Figure 3) and compared against ground-truth data. On average above 80% of the pixels were classified correctly over all test rounds. More thorough evaluation is planned. Inspection of the results for the two non-ischemic retinal sequences reveals no false alarms even in the presence of leakages and microaneurysms. Overall predictions tend to be conservative. Higher sensitivity can be achieved by decreasing θ in Eq. (1) but this may lead to an unacceptable increase in false positives. In a clinical screening context it may be preferable to decrease the rate of false negatives, as it is better to raise a limited number of false alarms (to be checked by clinicians) than missing some pathology altogether. We believe some error and the tendency for conservative predictions is largely due to the limited training data in this initial study. We must also consider that ground truth boundaries are rather subjective. Other misdetections are due to registration error and image acquisition (e.g., eyelids).

4 Discussion and future work

We have presented a context-based algorithm detecting ischemic regions in fluorescein angiograms. Our main contributions are a context-based detection scheme for retinal ischemic regions, a definition of context based on a pathological (not just anatomical) knowledge, and the use of UWFV angiograms, not yet common in the literature but capable of showing large parts of the retinal periphery. Experimental results suggest that our scheme is a promising solution for the detection of ischemic regions, the image appearance of which, *per se*, is not easily characterised. Further, more extensive tests are planned to assess the power of our scheme, for which purpose a larger set of FA sequences are being collected. This work opens up a wealth of future investigations. These include improvements to registration, further investigation of contextual signs and meaningful comparison of medical and automatic judgements.

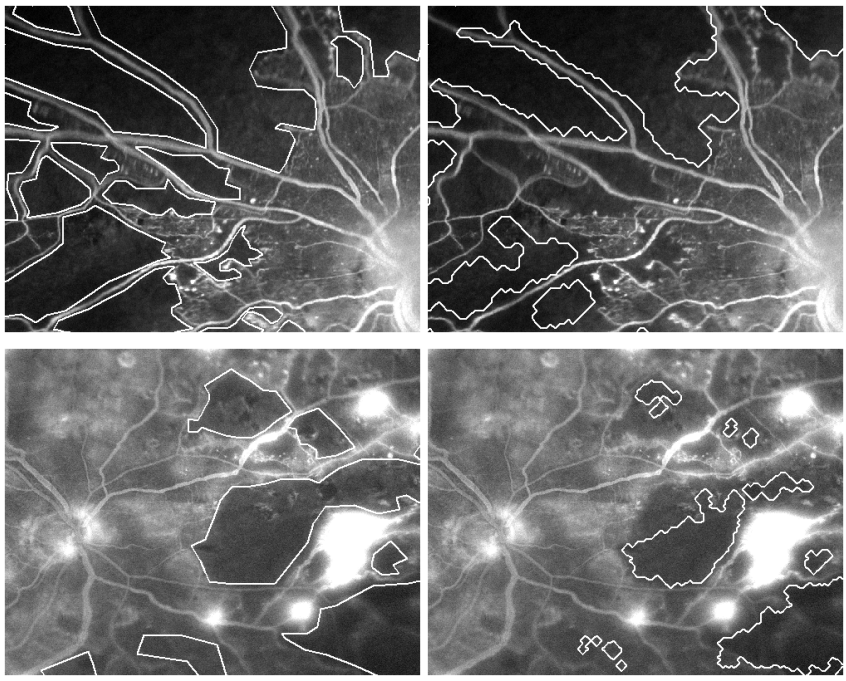


Figure 3. Two ischemic examples showing manual annotations (left) and overlay of final predictions (right)

Acknowledgements

This work was supported by Optos plc and a Knowledge Transfer Partnership grant (KTP000684) of the UK Department of Trade and Industry.

References

1. D. Verma. "Pathogenesis of diabetic retinopathy: the missing link?" *Medical Hypotheses* **41**, pp. 205–210, 1993.
2. E. T. D. R. S. R. Group. "Fluorescein angiographic risk factors for progression of diabetic retinopathy." *Ophthalmology* **98**, pp. 834–840, 1991.
3. T. D. R. S. R. Group. "Preliminary report on effects of photocoagulation therapy." *Am. Journ. Ophthalm.* **81**, pp. 383–396, 1976.
4. T. C. V. O. S. G. N. report. "A randomized clinical trial of early panretinal photocoagulation for ischemic central vein occlusion." *Ophthalmology* **102**, pp. 1434–1444, 1995.
5. P. Jasiobedzki, D. McLeod & C. Taylor. "Detection of non-perfused zones in retinal images." In *Proc. Fourth Annual IEEE Symposium Computer-Based Medical Systems*, pp. 162–169. 1991.
6. P. Jasiobedzki, C. Taylor & J. Brunt. "Automated analysis of retinal images." *Image and Vision Computing* **11(3)**, pp. 139–144, 1993.
7. J. Conrath, O. Valat, R. Giorgi et al. "Semi-automated detection of the foveal avascular zone in fluorescein angiograms in diabetes mellitus." *Clinical and Experimental Ophthalmology* **34**, pp. 119–123, 2006.
8. E. Trucco & P. Kamat. "Locating the optic disk in retinal images via plausible detection and constraint satisfaction." In *Proc. IEEE Int Conf on Image Processing (ICIP'04)*, volume CD 1. IEEE, 2004.
9. A. Manivannan, J. Plskova, A. Farrow et al. "Ultra-wide-field-of-view fluorescein angiography of the ocular fundus." *American Journ of Ophthalmology* **140**, pp. 525–527, 2006.
10. R. Schapire. "The boosting approach to machine learning: An overview." *MSRI Workshop on Nonlinear Estimation and Classification* 2002.
11. A. Hoover & M. Goldbaum. "Locating the optic disk in a retinal image using the fuzzy convergence of blood vessels." *IEEE Trans. on Medical Imaging* **22(8)**, pp. 951–8, 2003.
12. R. Pai, A. Hoover & M. Goldbaum. "Automated diagnosis of retinal images using evidential reasoning." In *15th Int Conf on Systems Engineering*. 2002.
13. H. Azegrouz & E. Trucco. "Max-min central vein detection in retinal fundus images." *Proc. IEEE Int Conf on Image Processing (ICIP'06)* pp. 1925–1928, 2006.
14. C. Buchanan & E. Trucco. "Automated detection of hyper-fluorescent leakage in retinal angiographic sequences by temporal analysis." In *Proc. MIUA 2007, to appear*. 2007.
15. T. Spencer, J. Olson, K. McHardy et al. "An Image-Processing Strategy for the Segmentation and Quantification of Microaneurysms in Fluorescein Angiograms of the Ocular Fundus." *Computers and Biomedical Research* **29(4)**, pp. 284–302, 1996.

Classifying mammograms using texture information

Arnau Oliver[a]*, Xavier Lladó[a], Robert Martí[a], Jordi Freixenet[a], and Reyer Zwiggelaar[b]

[a]Institute of Informatics and Applications, University of Girona, 17071, Girona, Spain
[b]Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, UK

Abstract. In an ongoing effort to assist radiologists in detecting breast cancer early, this paper focuses on breast characterisation according to internal tissue characteristics. This is an important feature because it has been demonstrated that women with dense breasts are more likely to suffer breast cancer, and also, the performance of automatic mass detection methods decreases in dense breasts. The strategy of our proposal firstly identifies regions with similar grey-level by using a clustering strategy. Subsequently, texture descriptors are extracted from each cluster by using Local Binary Patterns and Co-occurrence Matrices, and finally used to train a classifier. Results obtained from the complete MIAS database and using a leave-one-out strategy show a correct classification of 78% when compared to expert assessment.

1 Introduction

Mammographic Computer Aided Diagnosis (CAD) systems are being developed to assist radiologists in the evaluation of mammographic images [1]. However, recent studies have shown that CAD performance decreases as the density of the breast increases, either decreasing the sensitivity [2] (the accuracy in which the lesions are detected) or decreasing the specificity [3] (increasing the number of regions being normal tissue but marked by the automatic systems as suspicious lesions). In addition, it is well-known that there is a strong positive correlation between breast parenchymal density in mammograms and the risk to develop breast cancer [4]. Therefore, as Taylor [5] suggested, the development of automatic methods for classification of breast tissue are justified by, at least, two factors:

- To permit a better use of the time and skills of expert radiologists by allowing difficult mammograms to be examined by the most experienced readers.
- To increase the scope for CAD of abnormalities by concentrating on the easier (fatty) mammograms.

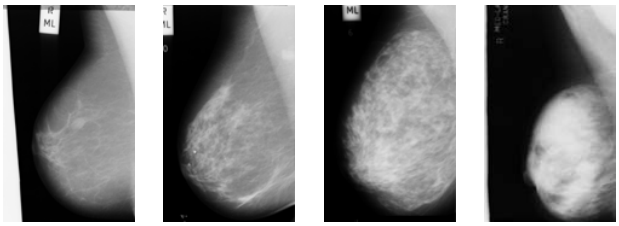
Classification in mammographic risk assessment can be based on a number of categories which might not describe the same mammographic features [6]. However, the American College of Radiology Breast Imaging Reporting and Data System (BIRADS) [7] is becoming a standard on the assessment of mammographic images and uses four categories for density evaluation:

- BIRADS I: the breast is almost entirely fatty,
- BIRADS II: there is some fibro-glandular tissue,
- BIRADS III: the breast is heterogeneously dense,
- BIRADS IV: the breast is extremely dense.

Figure 1 shows example mammograms of each class (the mammograms are extracted from the MIAS database [8]). Note how the internal density of the breasts increases from BIRADS I (left) to BIRADS IV (right). It should be noted that besides density these BIRADS classes also included patterns that can be described as various textures. As such, it seems appropriate to include both aspects in an automatic classification approach.

Although BIRADS is becoming the radiologic standard, it has not been commonly used in the evaluation of automatic breast density classification approaches. Exceptions to this are the works of Bovis and Singh [9], Petroudi et al. [10], and our previous work [11–13]. Bovis and Singh [9] extracted features from the global breast and used a combination of classifiers for training and testing the system. Petroudi et al. [10] used textons to obtain a visual dictionary for breast classification. This work was extended in [12] adding also SIFT features, although we found that textons outperformed them. The strategy we used here is similar to the work presented in [11, 13], which in turn follows ideas from Bovis and Singh [9] and more straightforward compared to Petroudi et al. [10] and Bosch et al. [12]. We used the Fuzzy C-Means [14] for clustering the pixels with similar grey-level values into two classes and, subsequently, a set of features derived from co-occurrence matrices was extracted from each cluster. In [11] the used classifier was a Bayesian combination of kNN and a decision tree. Here, we use two more stable clustering algorithms for grouping the tissue with similar grey-level. In addition to using texture features extracted from the co-occurrence matrices, we also propose to use the Local Binary Patterns [15] for extracting the texture information. To our knowledge this is the first attempt to use LBP in the mammographic field analysis.

*Corresponding author, e-mail: aoliver@eia.udg.es



(a) (b) (c) (d)

Figure 1. Example mammograms, where the breast density increases from (a) BIRADS I to (d) BIRADS IV.

2 Clustering the tissue with similar appearance

As a first step, all mammograms are pre-processed to identify the breast region. Hence, the background, labels, and pectoral muscle are removed from the image. We used a previous developed approach [16] which results in a minor loss of skin-line pixels in the breast area, but those pixels are deemed not to be relevant for tissue estimation and the relative number of potentially affected pixels is small.

Subsequently, using a clustering algorithm, the pixels with similar grey-level values are grouped together. We use a clustering algorithm in order to obtain only two clusters of pixels, representing fatty tissue (the cluster with lower mean intensity) and dense tissue (the cluster with higher mean intensity). In previous works we showed that the Fuzzy C-Means algorithm outperforms the results obtained using other segmentation strategies [13]. In this work we want to compare it with two alternative clustering algorithms, which are the Normalised Cuts [17] and the Mean Shift [18] algorithms. All these approaches are summarised below:

- Fuzzy C-Means clustering allows each pattern of the image to be associated with every cluster using a fuzzy membership function. In our implementation, the function minimised by the algorithm is defined by:

$$e^2(\Xi, U) = \sum_{n=1}^{N} \sum_{t=1}^{T} u_{nt} ||p_n - c_t||^2 \tag{1}$$

  where Ξ is the partition of the image, U is the membership matrix: u_{nt} represents the membership of pattern p_n to belong to cluster t, which is centred at $c_t = \frac{\sum_{n=1}^{N} u_{nt} p_n}{\sum_{n=1}^{N} u_{nt}}$, N is the number of patterns in the whole image (i.e. the number of pixels), and T the number of clusters, which has to be known a priori. Instead of randomly initialising the algorithm, the two cluster centres were initialised with the grey-level values that represent 15% and 85% of the accumulative histogram of the mammogram pixels, respectively representing fatty and dense tissue.

- The Normalised Cuts algorithm [17] deals with segmentation as a graph partitioning problem, where each node represents a pixel and the edges the pixel neighbourhood. Given a similarity measure between each pair of points in the set, it tries to group together points having large affinity between them. Namely, the algorithm consists in four steps: 1. Given an image build a weighted graph and set the weight (w) on the edge connecting two nodes to be a measure of their similarity. 2. Solve $(D - W)x = \lambda Dx$ as an eigen decomposition problem, where D is a $N \times N$ diagonal matrix with $d(i) = \sum_j w(i, j)$. 3. Use the eigenvector with the second smallest eigenvalue to bipartition the graph. 4. Decide if the current partition should be subdivided and recursively divide the segmented parts if necessary. Note that as we only need two clusters, this last step is not necessary.

- The Mean Shift algorithm [18] is a robust, non-parametric density estimation based clustering approach for image segmentation. It assumes that the feature space can be regarded as an empirical probability density function where dense regions correspond to the modes of the unknown density. A two-step kernel-based iterative procedure is used for the localisation of the modes. First, the mean shift vector m is computed as:

$$m(x) = \frac{\sum_{i=1}^{n} x_i g(||\frac{x - x_i}{h}||^2)}{\sum_{i=1}^{n} g(||\frac{x - x_i}{h}||^2)} - x \tag{2}$$

  where x is the actual mean, x_i all the points in the kernel, h the bandwidth (the size of the kernel), and g is the profile of the used kernel (we used a flat kernel). And secondly, the centre of the kernel is translated by $m(x)$. Both steps are repeated until convergence.

3 Characterising the breast tissue

Once the pixels with the same grey-level have been grouped in the fatty and dense clusters, we extract a set of features from both regions, and a feature vector is constructed by concatenating the fatty and dense features. In particular, we test in this paper the use of Local Binary Patterns (LBP) [15] and the use of features extracted from co-occurrence matrices [19].

The LBP operator labels the pixels of an image by thresholding the neighbourhood of each pixel with the centre value and considering the result of this thresholding as a binary number. When all the image pixels have been labeled with the corresponding LBP codes, the histogram of the labels is computed and used as a texture descriptor. However, this representation is not directly applicable to our work, because we need to characterise the clusters instead of the full image. Thus, we slightly modify the LBP in two different ways. Firstly, those pixels with a neighbourhood belonging to both clusters are not considered when extracting the texture. And secondly, instead of extracting one histogram from each mammogram, we extract two histograms per mammogram: one for the fatty cluster and the other one for the dense cluster. Hence, the concatenation of both histograms is taken as the full description of a mammogram.

Co-occurrence matrices are essentially two-dimensional histograms of the occurrence of pairs of gray-levels for a given displacement vector. Formally, the co-occurrence of gray-levels can be specified as a matrix of relative frequencies P_{ij}, in which two pixels separated by a distance d and angle θ have gray-levels i and j. The full co-occurrence matrices are not generally used as features (mainly due to their high dimensionality and potential sparseness), but instead a large number of features derived from such matrices have been proposed [19].

4 Classifying the features

Once the features are extracted, a classifier is trained and used in a posterior step to classify new mammograms. In this paper we test the use of four different classifiers [20]: kNN in combination with SFS feature selection, Fisher discriminant analysis using a linear discriminant feature selection algorithm, C4.5 decision tree classifier, and support vector machine.

5 Results

We test our approach using all 322 mammograms of the MIAS database [8], classified in BIRADS categories using the majority opinion of three experts (in the 12 cases where there was complete disagreement, the intermediate class was used). The results are obtained using a leave-one-woman-out strategy, where a query mammogram is tested by a classifier trained on the rest of the mammograms belonging to different women, and this procedure is repeated until all the mammograms have been used as a query image.

In these experiments we tested different parameters for LBP (we refer to the work of Ojala et al. [15] for details), including the mapping (uniform, rotation-invariant, and uniform rotation-invariant), neighbourhood (4, 8, and 16), and distance (1,2,4). In what follows we used the best results experimentally obtained, which were the ones using the uniform rotation-invariant mapping and 8-neighbourhood with distances 1 and 2, and 16-neighbourhood with distances 1 and 4. The complete size of the used descriptor was 112. We also tested different parameters for the co-occurrence matrices. In this paper we use four different directions: $0°$, $45°$, $90°$, and $135°$, and three distances equal to 1, 5, and 9 pixels. Note that these values were empirically determined in our experiments and are related to the scale of the texture features found in mammographic images. For each co-occurrence matrix the following features were used: contrast, energy, entropy, correlation, sum average, sum entropy, difference average, difference entropy, and homogeneity features. As each of these features is extracted from each class, we deal with 216 co-occurrence features in total.

5.1 Segmentation comparison

In this section we used the same parameters for LBP and the same classifier (kNN+SFS) while varying the segmentation algorithm used for segmenting the fatty and dense regions. This allows us to compare how the segmentation strategy modifies the final results.

Table 1 shows the results obtained in each case, obtaining 65%, 78%, 73%, and 77% correct classification when no segmentation, Fuzzy C-Means, Normalised Cuts, and Mean Shift segmentation algorithms were used respectively.

Note that the best performances are obtained when using the Fuzzy C-Means algorithm and the Mean Shift, while Normalised Cuts obtained intermediate results. The no segmentation strategy provided the worst results. This shows the benefits of the proposed segmentation step for clustering the tissue with similar grey-level. On the other hand, the low performance of the Normalised Cuts algorithm is due to the fact that the affinity function of this algorithm takes the distance between pixels into account, which results in filled-in regions in the images. However, in this not the aim of our work, because usually there are some small clusters of dense tissue inside large regions of fatty tissue, and viceversa. These regions are effectively detected using the Fuzzy C-Means and the Mean Shift algorithms.

(a)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	71	13	1	2
B-II	14	70	17	2
B-III	4	24	57	10
B-IV	1	7	17	12

(b)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	73	13	0	1
B-II	7	77	17	2
B-III	0	14	75	6
B-IV	1	1	9	26

(c)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	70	13	4	0
B-II	9	70	21	3
B-III	2	16	69	8
B-IV	1	4	6	26

(d)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	75	6	4	2
B-II	8	83	10	2
B-III	0	18	70	7
B-IV	1	2	14	20

Table 1. Confusion matrices obtained when (a) no segmentation, (b) Fuzzy C-Means, (c) Normalised Cuts, and (d) Mean Shift were used in the segmentation of fatty and dense tissue. B-I stands for BIRADS I, etc.

5.2 Features comparison

The aim of this section is to demonstrate the usefulness of computing texture information. Thus, the Fuzzy C-Means is used as a segmentation strategy, kNN+SFS as the classifier, and we compare the results obtained using LBP with those obtained using co-occurrence matrices. Table 2 shows the confusion matrices for both cases, obtaining 79% correct classification when using features derived from co-occurrence matrices and 78% for LBP. Note that co-occurrence matrices discriminate better low dense breasts, while Local Binary Patterns performs better the discrimination in dense breasts. The difference between these two results is not significant and, in fact, these results can well be influenced by the different dimensionality of the two set of features and to evaluate this is part of our future research.

(a)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	75	8	1	3
B-II	7	77	15	4
B-III	1	10	79	5
B-IV	3	3	7	24

(b)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	73	13	0	1
B-II	7	77	17	2
B-III	0	14	75	6
B-IV	1	1	9	26

Table 2. Confusion matrices obtained when (a) using co-occurrence matrices and (b) using Local Binary Patterns for texture extraction.

5.3 Classifiers comparison

Finally, we tested here the use of different classifiers. Hence, the same segmentation algorithm is used (the Fuzzy C-Means) and the same set of LBP features as it allows obtaining similar results with lower computational cost.

Table 3 shows the confusion matrices for kNN with SFS, linear discriminant analysis, C4.5 decision tree, and support vector machine with a polynomial kernel, obtaining respectively, 78%, 65%, 68%, and 65% correct classification. The high performance of the kNN+SFS algorithm compared to the other algorithms shows the difficulty of the problem, being hard to find a real boundary separating the four classes (for the SVM we used a 10-fold cross-validation instead of the leave-one-out to avoid overfitting). The unexpected poor performance of SVM is an aspect which will need further investigation.

(a)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	73	13	0	1
B-II	7	77	17	2
B-III	0	14	75	6
B-IV	1	1	9	26

(b)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	78	9	0	0
B-II	15	59	28	1
B-III	0	31	54	10
B-IV	2	2	15	18

(c)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	74	10	1	2
B-II	14	65	22	2
B-III	2	29	56	8
B-IV	2	2	8	25

(d)

Manual \ Automatic	B-I	B-II	B-III	B-IV
B-I	77	9	0	1
B-II	16	55	31	1
B-III	0	35	50	10
B-IV	1	4	14	18

Table 3. Confusion matrices obtained when using (a) kNN classifier (k=3), (b) Fisher Discriminant Analysis, (c) C4.5 Decision Tree, and (d) Support Vector Machine.

6 Conclusions

In this paper we have quantitatively compared the use of Fuzzy C-Means, Normalised Cuts, and Mean Shift for grouping the tissue with similar grey-level, as a first step of a full strategy for classifying the breasts according their internal density. We have found that, due to the internal nature of the algorithm (the use of the distance between pixels into the affinity function), the performance of the Normalised Cuts is lower compared to the use of both other algorithms. On the other hand, Mean Shift and Fuzzy C-Means obtained similar performance, around 78% correct classification. We consider these results in-line with the expected results, because when comparing the agreement between each individual expert annotations and the consensus opinion we obtained 78%, 89%, and 72% agreement.

In addition, we effectively tested the use of Local Binary Patterns and co-occurrence matrices for describing the breast tissue textural information. Results obtained from the complete MIAS database and using a leave-one-woman-out strategy show that LBP and co-occurrence matrices features provide similar overall results, although LBP performs better in dense breasts while co-occurrence matrices in fatty breasts. Future work will concentrate on further evaluation of variation in these features.

Acknowledgements

This work was partially supported by MEC grant nb. TIN2006-08035 and grant IdIBGi-UdG 91060080. We also thank Dr. Josep Pont and Dr. Elsa Pérez from the Hospital Josep Trueta of Girona (Spain) and Dr. Erika R.E. Denton from the Norfolk and Norwich University Hospital for providing the ground-truth data used in this paper.

References

1. T. W. Freer & M. J. Ulissey. "Screening mammography with computer-aided detection: Prospective study of 12860 patients in a community breast center." *Radiology* **220**, pp. 781–786, 2001.
2. S. Obenauer, C. Sohns, C. Werner et al. "Impact of breast density on computer-aided detection in full-field digital mammography." *J. Digit. Imaging* **19(3)**, pp. 258–263, 2006.
3. R. F. Brem, J. W. Hoffmeister, J. A. Rapelyea et al. "Impact of breast density on computer-aided detection for breast cancer." *Am. J. Roentgenol.* **184(2)**, pp. 439–444, 2005.
4. J. N. Wolfe. "Risk for breast cancer development determined by mammographic parenchymal pattern." *Cancer* **37**, pp. 2486–2492, 1976.
5. P. Taylor, S. Hajnal, M. H. Dilhuydy et al. "Measuring image texture to separate "difficult" from "easy" mammograms." *Brit. J. Radiol.* **67**, pp. 456–463, 1994.
6. I. Muhimmah, A. Oliver, E. R. E. Denton et al. "Comparison between Wolfe, Boyd, BI-RADS and Tabár based mammographic risk assessment." In *Int. Work. Dig. Mammography*, pp. 407–415. 2006.
7. American College of Radiology. *Illustrated Breast Imaging Reporting and Data System BIRADS*. American College of Radiology, third edition, 1998.
8. J. Suckling, J. Parker, D. R. Dance et al. "The Mammographic Image Analysis Society digital mammogram database." In *Int. Work. Dig. Mammography*, pp. 211–221. 1994.
9. K. Bovis & S. Singh. "Classification of mammographic breast density using a combined classifier paradigm." In *Med. Image Underst. Anal.*, pp. 177–180. 2002.
10. S. Petroudi, T. Kadir & M. Brady. "Automatic classification of mammographic parenchymal patterns: A statistical approach." In *IEEE Conf. Eng. Med. Biol. Soc.*, volume 1, pp. 798–801. 2003.
11. A. Oliver, J. Freixenet & R. Zwiggelaar. "Automatic classification of breast density." In *IEEE Int. Conf. Image Proc.*, volume 2, pp. 1258–1261. 2005.
12. A. Bosch, X. Muñoz, A. Oliver et al. "Modeling and classifying breast tissue density in mammograms." In *IEEE Conf. Comput. Vision Patt. Rec.*, volume 2, pp. 1552–1558. 2006.
13. A. Oliver, J. Freixenet, R. Martí et al. "A comparison of breast tissue classification techniques." In *Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, volume 2, pp. 872–879. 2006.
14. J. C. Bezdek. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
15. T. Ojala, M. Pietikäinen & T. Mäenpää. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns." *IEEE Trans. Pattern Anal. Machine Intell.* **24(7)**, pp. 971–987, 2002.
16. D. Raba, A. Oliver, J. Martí et al. "Breast segmentation with pectoral muscle suppression on digital mammograms." In *Lect. Not. Comp. Sc.*, volume 3523, pp. 471–478. 2005.
17. J. Shi & J. Malik. "Normalized cuts and image segmentation." *IEEE Trans. Pattern Anal. Machine Intell.* **22(8)**, pp. 888–905, 2000.
18. D. Comaniciu & P. Meer. "Mean sift: a robust approach toward feature space analysis." *IEEE Trans. Pattern Anal. Machine Intell.* **24(5)**, pp. 603–619, 2002.
19. R. M. Haralick, K. S. Shanmugan & I. Dunstein. "Textural features for image classification." *IEEE Trans. Syst., Man, Cybern.* **3(6)**, pp. 610–621, 1973.
20. R. O. Duda, P. E. Hart & D. G. Stork. *Pattern Classification*. John Wiley & Sons, New York, second edition, 2001.

Segmentation based on Textons and Mammographic Building Blocks

Izzati Muhimmah[a], Wenda He[a], Erika RE Denton[b] and Reyer Zwiggelaar[a*]

[a]Department of Computer Science, University of Wales Aberystwyth, SY23 3DB, UK
[b]Radiology Department, Norfolk & Norwich University Hospital, Norwich NR4 7UY, England

Abstract. The objective of this paper is to investigate a potential approach of mammographic image segmentation based on textons and mammographic building blocks (i.e. nodular, linear, homogeneous, and radiolucent) as described in Tabár's tissue model. The texton approach is based on clustering filter responses in a high dimension for a particular building block. The texton selection process is based on a combination of visual assessment (probability maps) and minimum spanning tree topological information. The initial segmentation results are promising, which may lead us to an automatic Tabár's mammographic risk assessment.

1 Introduction

In mammographic imaging, anatomical breast structures are enhanced by the attenuation and tissue composition of those structures [1]. According to Tabár *et al.* [2], there are four building blocks in mammographic images which compose the normal breast anatomy as follows: *nodular densities* mainly corresponds to Terminal Ductal Lobular Units (TDLU), *linear densities* could be corresponded to either ducts, fibrous strand or blood vessels, *homogeneous-structureless densities* correspond to fibrous tissues which appearance could hide the underlying normal TDLU and ducts as well as their alterations due to hyperplastic breast changes, and *radiolucent areas* are related to adipose fatty tissues, which appears as dark areas. Some image patches, which contain linear, nodular, and homogeneous densities are shown in Figure 1. From an image processing point of view, these building blocks can be seen as different textures and hence could drive a tissue segmentation process.

In addition, Tabár *et al.* have proposed a mammographic risk assessment based on quantitative approach using the relative proportion of the mammographic building blocks. This is known as the Five Mammographic Parenchymal Patterns classification approach [2]. Patterns I-III correspond to lower breast cancer risk, whilst Patterns IV-V relate to higher risk. Relative composition of the four building blocks (in the order [nodular, linear, homogeneous, radiolucent]) provides the feature space for risk classification. **Pattern I** has composition [25%, 15%, 35%, 25%]. **Pattern II** is composed of [2%, 14%, 2%, 82%]. **Pattern III** is similar in composition to Pattern II, except that the retroareolar prominent ducts are often associated with periductal fibrosis. The composition of **Pattern IV** is: [49%, 19%, 15%, 17%]. **Pattern V** is composed as [2%, 2%, 89%, 7%]. Mammographic images representing these five risk patterns can be found in Figure 2. A comparison between this and other mammographic risk classification schemes can be found in [4].

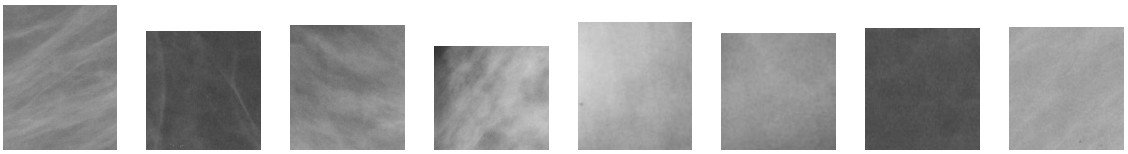
To develop an automated mammographic risk assessment approach, according to the above description, the challenge is to segment a given mammographic image with respect to Tabar's building blocks, estimate their relative proportions and link these to mammographic risk. This paper is focused on mammographic texture segmentation based on Tabár's building blocks and uses texton which have been shown successful for the recognition of natural textures [5, 6].

2 Texton

Julesz firstly hypothesized that iso-second-order textures are globally preattentive indistinguishable [7]. However, he refuted this by showing cases where iso-second-order textures are easily discriminated because they are composed of local conspicuous features, such as *'closure'*, *'corner'*, *'connectivity'* and *'granularity'*. These local features are known as 'textons'. He concluded that two textures having different density changes in textons were pre-attentively distinguishable [8]. Subsequently, he postulated a link between textons and filter banks [9]. It should be noted that this concept was designed for binary textures.

For grey-scale images, Leung and Malik [5] defined texture as entities which have spatially repeated properties. When a texture is filtered, one could expect that some pixels have similar filter responses. They suggested that there should be several distinct filter response vectors to represent a texture. They proposed the cluster centers of the filter responses as the textons.

*Email: [iim04,rrz]@aber.ac.uk.



(a) (a) (b) (b) (c) (c) (d) (d)

Figure 1. Examples of Tabár's building blocks: (a) linear, (b) nodular, (c) homogeneous, and (d) radiolucent areas. These patches are subsampled from the MIAS database [3] by an expert radiologist. It should be noted that there is large variation in density and texture patterns within a building-block.

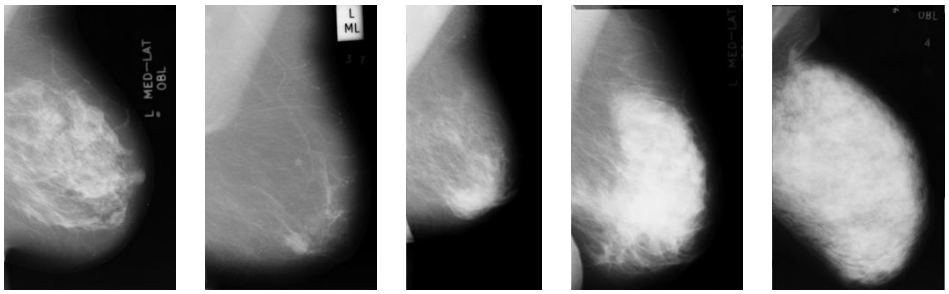


Figure 2. Examples of the Five Mammographic Risk Patterns. From left to right: Patterns I−V.

3 Data

A breast screening expert radiologist subsampled images to provide patches of linear, nodular, homogeneous, and radiolucent textures of the MIAS database [3]. The collection of 253 linear, 199 nodular, 70 homogeneous, and 121 radiolucent patches were obtained (see Figure 1 for examples). From previous experiments, the breast edge area needs to be taken into account in the modelling [10]. Additional collection of 93 linear, 8 nodular, 30 homogeneous, and 245 radiolucent patches of the breast edge areas were obtained. It should be noted that the patches cover various breast density groups, so we normalised them to zero mean and unit variance.

4 Method

As the ideal filters bank for mammograms is currently unknown, 56 filters were used, which consisted of iso-tropic Gaussian and Laplace of Gaussian filters at four scales (σ = $\{$ 1, 2, 4, 8$\}$), and of first- and second-order derivatives of Gaussian-anisotropic filters (edges and lines) [11] at four scales ((σ_x, σ_y) = $\{$(1,2), (2,4), (4, 8), (8,16)$\}$) and six orientations (θ = $\{$0, 30, 60, 90, 120, 150$\}$). These filters had a 63×63 window size and L_1 normalised. In summary, each pixel will be represented as a 56 dimensional feature vector.

In the first instance, the aim is to obtain textons which represent the mammographic building blocks, which are linear, nodular, homogeneous, and radiolucent patterns. Learning stage I of the Varma and Zisserman (VZ) algorithm [6] was followed to achieve this: for each group of texture images (in our case the mammographic nodular, linear, homogeneous, or radiolucent patches) every normalised image was convolved with 56 normalised filters; the resultant filter responses were (contrast) normalised using Equation 1 [5, 6]; the normalised filter responses are aggregated and clustered into textons using the k-means algorithms. It should be noted that a resulting texton will also be 56 dimensional vector. We used 25 cluster centers as the number of initial textons (as subsequently the number of textons are reduced, the randomly chosen initial number of textons is not critical). Schematic diagram of this methodology can be seen in Figure 3.

$$F(x) \leftarrow F(x) \left[log \left(1 + \left\| F(x) \right\|_2 / 0.03 \right) \right] / \left\| F(x) \right\|_2 \tag{1}$$

Once the textons have been estimated, it becomes possible to obtain a probability map for a specific texton, T_i, and this is given by :

$$V_i(x,y) = 1 - \left(\frac{\sqrt{\left[I(x,y) - T_i \right]^2}}{\sum_{k=1}^{25} \sqrt{\left[I(x,y) - T_k \right]^2}} \right) \tag{2}$$

here, $I(x,y)$ represents the 56 dimensional filter responses of an image pixel at position (x,y).

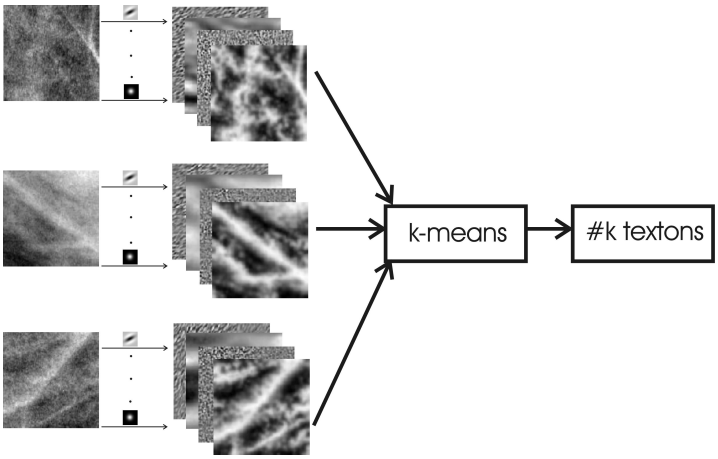


Figure 3. Learning stage I of the Varma and Zisserman (VZ) algorithm. On the left a coloumn of image patches is shown, which are filtered by 56 filters (second coloumn) resulting in a set of filter-response images (third coloumn). The 56 filter-response images are used to populated a 56 dimensional feature space and k-means is used to obtain cluster centres which are the resulting textons.

It is plausible to assume that textons will capture either the meaningful texture primitives or noisy pixels. So, texton reduction/selection can be seen as an essential and potentially beneficial step. We expect that the selected textons will be robust representations of the mammographic building blocks.

The approach used for texton selection has two distinct components which have been combined. In the first instance, an Euclidean minimum spanning tree is obtained for a set of textons. Such a tree indicates likely topological correct connectivity in high dimensional spaces (as long as the space is not sparsely populated). Subsequently, probability maps (using Equation 2) of representative image patches are used as the nodes of the tree. See Figure 4 for an example of such a minimum spanning tree for all textons. It is possible to obtain such a tree representation for various sets of textons. For the overall texton selection the minimum spanning trees for linear, nodular, homogeneous, and radiolucent textons were compared as individual trees, which indicated outliers and common textons. In addition, a tree was generated from combined sets of textons and clustering within this tree was used to ensure a distinction between textons representing the various Tabár's building blocks.

Distinct textons tend to be situated towards the outer edges of the tree, while common texture/intensity aspects tend to be modelled by textons in the central part of the tree. It should be noted that the currently used display approach (see Figure 4) does use a fixed distance between the nodes, but in practise the distance can vary by almost an order of magnitude (especially for the trees that combine various texton sets). The visual aspects and distinctiveness of the textons are taken into account to determine a set of textons that will be distinctive in describing the various Tabar mammographic building blocks. Visual selection was based on the ability of a texton to reproduce Tabár's building blocks in clear examples of these.

5 Results

The minimum spanning tree produced from the combined set of all textons, 200 in total, can be found in Figure 4. This result provides visual confirmation of our expectations. Selective textons map specific mammographic structures at various scales, while other textons seem to respresent background texture information. Based on this it is plausible that such textons can be used for segmentation purposes. Based on distinctiveness and visual evaluation of the minimum spanning tree results eight textons were selected to represent Tabar's mammographic building blocks. With respect to Figure 4, these are textons L7 and BEL23 (linear), N6 and BEN6 (nodular), H13 and L9 (homogeneous), and F12 and BEL4 (radiolucent).

Resulting segmentation on a mammographic image can be found in Figure 5. It should be clear that the most distinct linear structures (largest width) have been segmented as linear structures. However, this does not seem to be the case for smaller linear structures. Nodular and homogeneous structures have been realistically modelled. It should be noted that the breast-edge area seems to be segmented into radiolucent classes and the smaller linear structures within this area has not been segmented although they had been modelled.

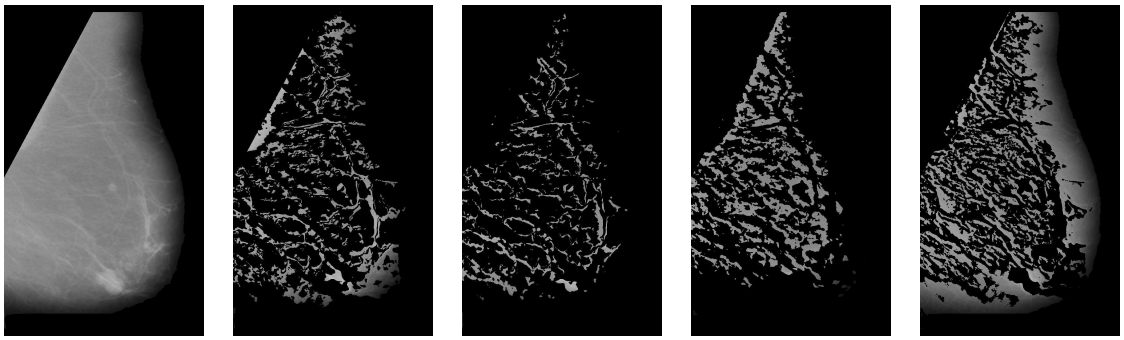
Figure 4. Probability maps of 200 textons of all trained textures. Bold labels indicate textons which were selected for segmentation (please refer to the electronic version for better view of this figure). The probability maps are clustered according to a minimum spanning tree.

These are very initial results and full evaluation based on the MIAS database is currently being investigated. This evaluation will cover a comparison between the automatically obtained Tabár building blocks distribution for the five Tabár classes (and potentially the four BIRADS classes [12]). It will also cover automatic mammographic risk classification using the relative percentage of occurrence of the Tabár building blocks as a feature space.

6 Discussion

Clearly, the presented results can be dependent by the choice of filter-bank that was used. One aspect of our future work will concentrate on the effect of the choice of filters. Secondly, the segmentation approach will be applied to a large mammographic dataset and the relative proportion of the four classes will provide the input for a Tabar mammographic risk assessment model.

It should be noted that this is, to our knowledge, the first attempt at providing automatic segmentation of mammographic images according to the Tabár mammographic tissue building block model. There have been several approaches to the segmentation of dense regions in mammographic images [13,14] and some of these are based on texton information [15, 16]. Petroudi and Brady have used the textons in Hidden Markov Random Fields modelling for three category density classification [15]. They pointed out that textons based on a set annotations by radiologists are expected to lead to concise segmentations [15], which is the approach used here. An attempt to link texton classification and Wolfe's risk assessment has also been reported by Gong *et al.* [16]. Whilst, Bosch *et al.* reported patch-based textons as a descriptors for mammographic density classification using BIRADS categories [12].



(a) (b) (c) (d)

Figure 5. Left image shows the original image, with background and pectoral muscles removed. Segmentation results: (a) linear, (b) nodular, (c) homogeneous, (d) radiolucent.

7 Conclusions

We have presented initial results of texton driven segmentation of mammographic images. The texton selection step follows Tabár's mammographic tissue building block model and uses visual and minimum spanning tree topological information. The initial segmentation results are promising, but at the same time have indicated areas of future work. Segmentation results as described can form the basis for an automatic Tabár mammographic risk assessment system.

References

1. H. H. Barrett & W. Swindell. *Radiological Imaging: The Theory of Image Formation, Detection, and Processing*, volume 1. Academic Press, 1981.
2. L. Tabár, T. Tot & P. B. Dean. *Breast Cancer: The Art and Science of Early Detection with Mammography*. Georg Thieme Verlag, 2005.
3. J. Suckling, J. Parker, D. Dance et al. "The mammographic image analysis society digital mammogram database." *Excerpta Medica. International Congress Series* **1069**, pp. 375–378, 1994.
4. I. Muhimmah, A. Oliver, E. Denton et al. "Comparison between Wolfe, Boyd, BI-RADS and Tabár based mammographic risk assessment." In S. Astley, M. Brady, C. Rose & R. Zwiggelaar (editors), *LNCS 4046 Digital Mammography*, pp. 407–415. Springer-Verlag Berlin Heidelberg, 2006.
5. T. Leung & J. Malik. "Recognizing surfaces using three-dimensional textons." *Proceeding of International Conference Computer Vision* pp. 1010–1017, 1999.
6. M. Varma & A. Zisserman. "A statistical approach to texture classification from single images." *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis* **62(1–2)**, pp. 61–81, April 2005.
7. B. Julesz, E. Gilbert, L. Shepp et al. "Inability of humans to discriminate between visual textures that agree in second-order statistics – revisited." *Perception* **2(4)**, pp. 391–405, 1973.
8. B. Julesz. "Textons, the elements of texture perception, and their interactions." *Nature* **290(12)**, pp. 91–97, 1981.
9. B. Julesz & J. Bergen. "Textons, the fundamental elements in preattentive vision and perception of textures." *Bell Systems Technical Journal* **62(6)**, pp. 1619–1645, 1983.
10. I. Muhimmah, W. He, E. Denton et al. "Mammographic texton representation based on the Tabár tissue model." *BMVC The Inagural Student Papers Meeting* 28 March 2007.
11. J.-M. Geusebroek, A. W. M. Smeulders & J. van de Weijer. "Fast anisotropic gauss filtering." *IEEE Transactions on Image Processing* **12(8)**, pp. 938–943, August 2003.
12. A. Bosch, X. Muñoz, A. Oliver et al. "Modeling and classifying breast tissue density in mammograms." *Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2**, pp. 1552–1558, 2006.
13. A. Oliver, J. Freixenet, R. Martí et al. "A comparison of breast tissue classification techniques." *International Conference on Medical Image Computing and Computer Assisted Intervention LNCS 4191* pp. 872–879, October 2006.
14. R. Ferrari, R. Rangayyan, R. Borges et al. "Segmentation of the fibro-glandular disc in mammograms via gaussian mixture modelling." *Med. and Biol. Eng. and Comp.* **42**, pp. 378–387, 2004.
15. S. Petroudi & M. Brady. "Breast density segmentation using texture." In S. Astley, M. Brady, C. Rose & R. Zwiggelaar (editors), *LNCS 4046 Digital Mammography*, pp. 609–615. Springer-Verlag Berlin Heidelberg, 2006.
16. Y. C. Gong, M. Brady & S. Petroudi. "Texture based mammogram classification and segmentation." In S. Astley, M. Brady, C. Rose & R. Zwiggelaar (editors), *LNCS 4046 Digital Mammography*, pp. 616–625. Springer-Verlag Berlin Heidelberg, 2006.

Author Index