Optimising Multi-feature Metric for Histology Image Retrieval

Qianni Zhang
http://www.elec.qmul.ac.uk/mmv/people/qianniz.htm
Ebroul Izquierdo
http://www.elec.qmul.ac.uk/mmv/people/ebroul/
School of Electronic Engineering and Computer Science
Queen Mary University of London
London, UK

Abstract

Feature combination for image classification and retrieval is an important design aspect in modern image retrieval systems. It is very valuable in medical applications and specially in histology applications in which different features are extracted to estimate tissue composition and architecture. This paper presents an approach to combine multiple textural features for histology image retrieval, following a late-fusion scheme. The multi-feature metric for feature combination is obtained using a multi-feature learning method. The experimental evaluation was carried out on a collection of histology images to evaluate the feature combination strategy. Experimental results show that it is possible to improve the system performance by appropriately defined multi-feature space, considering the structure and distribution of visual features.

1 Introduction

In biology and medicine, histology is a fundamental tool that provides information on architecture and composition of tissues at microscopic level. Nowadays, images of tissue slides are often digitized to document procedures and to support findings. However, these collections are often huge in size and thus hide a latent source of information that can be to be greatly exploited if suitable mechanisms are available for access the data [6].

There have been many content-based image retrieval (CBIR) systems designed for medical applications in recent years [3]. Histology image retrieval has been an active research topic for modeling visual similarity measures and retrieve tissue slides in some semantic categories. In [12], a CBIR system was proposed for retrieving histology images from prostate, liver and heart tissues, based on four different visual characteristics. The work in [10] described a system to index histology images of gastro-intestinal tract, by categorising image blocks into semantic classes based on local visual patterns. The approach proposed in [7] uses a boosting algorithm based on multiple distance measures computed on a fixed set of features to retrieve and classify breast histology slides. These works all intended achieving better performance in histology image understanding by employing multiple visual features. However, strategies for feature combination are commonly underestimated, whereby the use of simple feature vector concatenations is done.

Such approaches to feature combination shares a common problem. Different image features often have their own structure, distribution and metric space. Direct concatenation of

feature vectors could result in meaningless similarity measures. For instance, feature histograms are usually compared using a similarity measure for probability distributions while feature vectors should be matched using Euclidean metrics. In addition, even if two different features are being compared with the same metric, their scale, domain and distribution may be completely different due to the intrinsic descriptor nature [4].

In this paper a *late-fusion* strategy is followed to combine low-level features for histology image retrieval. A suitable feature combination metric is learned using a Multi-Objective Learning (MOL) method [11], which involves an multi-objective optimisation (MOO) process for learning [9]. The main advantage in the MOL method is that it is able to find a multi-feature metric that can simultaneously encapsulates different aspects of the most representative visual patterns for each concept, without however assigning fixed relevance factors to each one of the used visual features. Five different textural features are considered in our experiments, and the proposed approach has been tested in a collection of 2,828 images with various examples of the two types of fundamental tissues in biology: epithelial and muscular tissues.

This paper is organized as follows: Section 2 presents the visual analysis in histology databases, including feature extraction and normalisation. Section 3 introduces the MOL method for multi-feature based retrieval in histology image sets. The evaluation procedure and experimental results are presented in Section 4 and Section 5 discusses some concluding remarks.

2 Visual feature extraction and analysis

This paper focuses on developing an approach for automatically combining low-level visual features for retrieval of histology images according to their fundamental tissue. Therefore, the extraction and analysis of useful visual features in histology images is an essential step.

The proposed multi-feature combination approach is independent of used features. Textures features together with architectural features have been suggested as prominent characteristics for histology image analysis [1, 5]. Without losing generality, five textural features have been selected to describe histology image contents: Gabor textures, Tamura textures, Zernike moments, SIFT-based dictionary and DCT dictionary.

Among these five features, the first three are image feature vectors computed per block in order to obtain characteristic patterns in different regions [8]. The other two are histogram features constructed using a bag-of-features approach, that allows estimation of local patterns in images[2].

In the next step, each of these features are normalised to guarantee the appropriate comparison of different measurements that differ in scale and domain, while preserving the underlying characteristics of the data. In the proposed feature combination approach, distance metrics computed from different image features are normalised based on the fitted probability density functions for their corresponding feature spaces.

3 Multi-feature based histology image retrieval

The proposed approach to multi-feature based histology image retrieval relies on multi-objective learning (MOL) method, that is able to automatically learn a suitable multi-feature metric from a small sized training set.

3.1 Visual representation of tissue types

For a given type of tissue, let us denote a group of training samples, referred to as the representative group, as $R = \{r_i | i \in [1,m]\}, R \in \Gamma$, where m is the total number of training samples in R and Γ is the complete dataset to be classified. Having the representative group ready, the next step will be to calculate the centroid of training samples using each one of the different distance measures of the five considered feature spaces, as described in Section 2. A centriod in a feature space, $f_j, j = 1,2,...,5$, is calculated by finding the representative sample with the minimal sum of distances to all other representative samples in R. All the centroids across different feature spaces form a particular set of vectors $\bar{V} = \{\bar{v_1}, \bar{v_2}, ..., \bar{v_5}\}$, in which each $\bar{v_j}$ is the centroid vector according to feature space f_j. In general, $\bar{V}$ is referred to as the *generalised centroid* of its representative group R, since it does not necessarily represent a specific block of R. Taking $\bar{V}$ as an anchor, for each $v_{i,j}$ denoting the vector in feature space f_j extracted from an image r_i, the distance from r_i to the centroid $\bar{v_j}$ can be estimated by $d_{i,j} = d_j(\bar{v}_j, v_{i,j}), j = 1,2,...,5$. For the remaining of this paper, when these distance values are mentioned, they are all supposed to have been normalised according to Section 2. For the training group R of a particular tissue type, a distance matrix of size $m \times 5$ is constructed, in which each element is the distance from a training sample in one of the feature spaces. In this way, each class is represented by a distance matrix covering multiple feature spaces.

3.2 Multi-feature space learning and retrieval

Based on the distance matrix, a set of objective functions can be constructed for multi-feature metric optimisation. For a representative group R, a set of objective functions are formed as weighted linear combinations of feature-specific distances:

$$\left\{ \; D_i = \sum_{j=1}^{5} \alpha_j d_{i,j}, i = [1,2,...,m], j = [1,2,...,5] \; \right\} \tag{1}$$

The MOL process seeks to learn from the representative group an optimal set of coefficients $\{\alpha_j | j = 1,...,5\}$, subject to the constraint: $\sum_{j=1}^{5} \alpha_j = 1$. The problem of learning a multi-feature metric is now transformed to finding a solution that optimises each of these objective functions in (1). In this case, an optimum is defined as the minimum of $D_i, i = 1,2,...,5$. The difficulty in finding such an optimum is that different representative samples may display different visual characters, resulting in different interests of their corresponding objective functions. These differences need to be harmonised via simultaneous optimisation of multiple objectives in a global optimum. To solve this problem, the MOO strategy is employed in the MOL method for optimising the set of objectives and learning an 'optimal' multi-feature metric. The MOO strategy is able to find a general optimum across potentially conflicting objectives, and thus is widely used in real-life optimisation problems.

For a particular tissue type, an optimal multi-feature metric $\hat{D} = \sum_{j=1}^{5} \hat{\alpha_j} d_{i,j}$ can be obtained using the MOL method. Using $\hat{D}$, a multi-feature distance can be calculated for any image $I_i \in \Gamma$. According to these multi-feature distances, images can be retrieved based on their ranked multi-feature distances with respect to the query class.

4 Experiments

The histology image database used in this research contains 20,000 samples. Most samples in this dataset are unannotated, making these samples inaccessible using textual-based

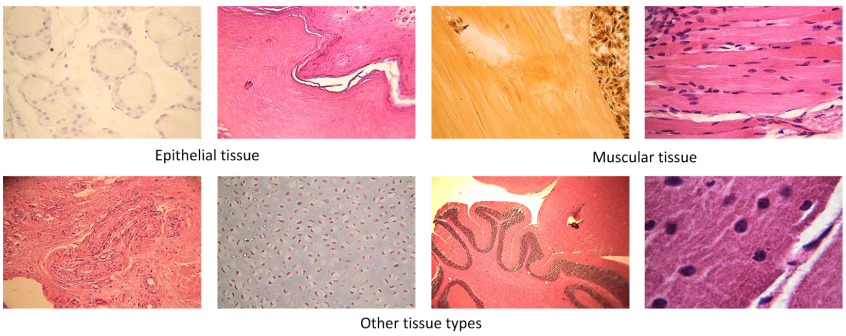

Figure 1: Image samples of tissue types.

Table 1: Retrieval evaluation of two tissue types

	Epithelial			*Muscular*		
Approach	MAP	R-prec	Prec 20	MAP	R-prec	Prec 20
MOL	**0.429**	**0.367**	**0.850**	0.303	0.298	**0.750**
Linear	0.318	0.323	0.650	**0.307**	**0.325**	0.600

search methods. To evaluate the performance of the proposed multi-feature based retrieval approach, we selected a subset containing 2,828 histology images, which are manually labeled. The subset contains 804 samples for epithelial tissue and 514 for muscular tissue. The rest of images in the subset are for other types of tissue and are labeled as 'other' in ground-truth. The training set of each of the two types of tissue contains ten or less samples. The performance measures presented include Mean Average Precision (MAP); R-Precision, which is obtained at the point where precision and recall get the same value; and Precision after the first 20 retrieved samples, as shown in Table 1.

In Tables 1, we also show another retrieval result based on a direct linear combination of distances from different feature spaces with the same weight. The proposed multi-feature retrieval using MOL metric performed better for all evaluation criteria for the Epithelial type. Direct linear combination metric resulted in better MAP and R-precisions for Muscular type, but MOL metric lead to much better precision in the first 20 retrieved samples in this case.

5 Conclusions and future work

This paper proposes a strategy for multi-feature based retrieval in histology image databases. The multi-feature metric is obtained using a MOL method, which automatically derives suitable metric for feature combination based on a small set of training samples. Experimental performance of the proposed approach is presented and analysed. The evaluation of results shows that, in the used experimental set-up, the proposed strategy was able to provide more precise retrieve results with small size of training sets compared to single features compared to using single features or linear combination of feature metrics. Future work could include to test the proposed strategy with more query types and bigger dataset.

References

[1] N. Bonnet. Some trends in microscope image processing. *Micron*, 35(8):635–653, December 2004. ISSN 09684328. doi: 10.1016/j.micron.2004.04.006. URL http:

//dx.doi.org/10.1016/j.micron.2004.04.006.

[2] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, 2004.

[3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, April 2008. ISSN 0360-0300. doi: 10.1145/1348246.1348248.

[4] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision - ECCV 2002*, Lecture Notes in Computer Science, chapter 7, pages 349–354. 2002.

[5] C. Loukas. A survey on histological image analysis-based assessment of three major biological factors influencing radiotherapy: proliferation, hypoxia and vasculature. *Computer Methods and Programs in Biomedicine*, 74(3):183–199, June 2004. ISSN 01692607. doi: 10.1016/j.cmpb.2003.07.001. URL http://dx.doi.org/10.1016/j.cmpb.2003.07.001.

[6] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1): 1–23, February 2004. doi: http://dx.doi.org/10.1016/j.ijmedinf.2003.11.024. URL http://dx.doi.org/10.1016/j.ijmedinf.2003.11.024.

[7] Jay Naik, Scott Doyle, Ajay Basavanhally, Shridar Ganesan, Michael D. Feldman, John E. Tomaszewski, and Anant Madabhushi. A boosted distance metric: application to content based image retrieval and classification of digitized histopathology. volume 7260, pages 72603F+. SPIE, 2009.

[8] Nikita Orlov, Lior Shamir, Tomasz Macura, Josiah Johnston, D. Mark Eckley, and Ilya G. Goldberg. Wnd-charm: Multi-purpose image classification using compound image transforms. *Pattern Recognition Letters*, 29(11):1684–1693, August 2008. ISSN 01678655. doi: 10.1016/j.patrec.2008.04.013. URL http://dx.doi.org/10.1016/j.patrec.2008.04.013.

[9] R.E. Steuer and RE Steuer. *Multiple criteria optimization: Theory, computation, and application*, volume 233. Wiley, 1986.

[10] H. L. Tang, R. Hanka, and H. H. S. Ip. Histological image retrieval based on semantic content analysis. *Information Technology in Biomedicine, IEEE Transactions on*, 7(1): 26–36, 2003. doi: http://dx.doi.org/10.1109/TITB.2003.808500. URL http://dx.doi.org/10.1109/TITB.2003.808500.

[11] Q. Zhang and E. Izquierdo. Combining low-level features for semantic inference in image retrieval. *Journal on Advances in Signal Processing*, 2007.

[12] Lei Zheng, A. W. Wetzel, J. Gilbertson, and M. J. Becich. Design and analysis of a content-based pathology image retrieval system. *Information Technology in Biomedicine, IEEE Transactions on*, 7(4):249–255, 2003. doi: 10.1109/TITB.2003.822952. URL http://dx.doi.org/10.1109/TITB.2003.822952.