Affine Invariant Image Segmentation

Abhir Bhalerao and Roland Wilson
Department of Computer Science
University of Warwick
Coventry, UK
{abhir|rgw}@dcs.warwick.ac.uk

Abstract

This paper introduces a method of image segmentation that partitions an image into regions exhibiting self-similarity. The new method is well suited to the description and segmentation of images containing textures, such as fingerprints and medical images. An affine texture description which models an image as a collection of self-similar image blocks is developed and it is shown how textural regions can be approximated by a single prototype block together with a set of affine transformation coefficients. The transformations are estimated by Gaussian mixture modelling of the prototype block spectrum and a least-squares estimation to fit the model to the target block. An appropriate set of prototype blocks is determined for a given an image by clustering in a feature subspace of discrete samples from an affine symmetry group. We consider how the texture synthesis outlined in the first stage can be combined with the prototype clustering in the second, to build a novel image segmentation method. Experimental results demonstrate the potential of the method on widely varying images.

1 Introduction

Images exhibit self-similarity not only because there is a great deal of redundancy in image regions that are flat or change in colour smoothly, but also in the obvious rotational symmetry of small sections of image contours: edges and lines. As well as rotational symmetry, there are changes of scale such as 'thin'-edges and 'fat'-edges, small structures and larger versions of the same. Natural scenes, biological images, or images containing repetitions, often have a tiling of some image 'patch' or texton. Notable image models that have exploited image self-similarity for image analysis are the fractal image coding methods of Jacquin and others [1], and warping methods first reported by Volet [2] for texture. Syntactical or structural approaches attempt to capture the repetition and self-similarity of a texture by picking a texton and defining rules for its placement to synthesise the entire texture. For some types of texture, a structural model alone is unsuitable, so statistical texture models, such as MRFs, have been used to parameterise the randomness of the texture. With combined structural and random models [3] a stochastic element is 'added' to a purely deterministic local texton model to account for the effects of texture variation due to surface properties, lighting or noise in the imaging process. Natural textures, however, differ fundamentally from man-made textures by a degree of randomness, syntactical errors or faults, that is inherent at the structural level. This is why, for example, a simple tiling of a texture patch from a natural texture does not have a natural appearance [4].

BMVC 2004 doi:10.5244/C.18.18

In this work using Gaussian Mixture Modelling (GMM), we build on the work of Hsu, Calway and Wilson [5, 6], whose two-component affine texture model combined a structural texton plus stochastic residual that enabled the synthesis of a range of textures from periodic to random. Moreover, this affine placement model itself can accurately model small structural variations which seem important in natural texture.

Two questions are tackled:

1. Given two self-similar image patches, is it possible to accurately and efficiently estimate the symmetry transformation (rotation, scaling and shear)?
2. Given an image containing self-similarities, can a small set of prototype image patches which can be used to model the entire image, to some level of distortion, be determined?

We use a Gaussian mixture model to estimate the affine transformation (1), and apply feature clustering to achieve (2). We thus obtain a concise way to describe the image and perform image segmentation where image features are small representative patches from the image, the feature clustering is invariant to local affine transformations of these patches, and the accuracy of the segmentation can be measured by image reconstruction.

2 A Model for Affine Self-Similarity

Image regions are modelled as sets of repeating texture patches, assuming that each region can be adequately represented by affine coordinate transformations of some prototypical patch (the source texton) and an intensity scaling. Denoting the coordinates of a target patch, f_j, by $\mathbf{y} = (p,q)^T, 0 \leq p,q, \leq N$, the target patch estimate is

$$\hat{f}_j(\mathbf{y}) = \alpha_{ij} f_i(T_{ij}(\mathbf{x})), \qquad T_{ij}(\mathbf{y}) = A_{ij}\mathbf{x} + \mathbf{t}_{ij} \tag{1}$$

$$\alpha_{ij}^2 = \sum_{\mathbf{x}} f_i^2(\mathbf{x}) / \sum_{\mathbf{x}} f_j^2(\mathbf{x}) \tag{2}$$

where T_{ij} is a 2D affine transformation of the coordinates of some source patch f_i and α_{ij} is a scaling for image intensity changes. The transformation consists of a linear part, A_{ij}, and a translation, $\mathbf{t}_{ij}$. It is convenient to use overlapping square image patches (blocks) as textons of size B, such as 16×16, 32×32 etc. By using a squared cosine window function centred on the block, $w(\mathbf{y}) = \cos^2[\pi p/B]\cos^2[\pi q/B]$ and having the blocks overlap by 50%, it is possible to sum the transformed patches, $w(\mathbf{y}) f_j(\mathbf{y})$, without blocking artifacts being introduced.

To estimate the affine transformations T_{ij}, the Fourier transform of the source patch, $F_i(\mathbf{u})$, is used to separate the transformation,

$$F_j(\mathbf{u}) = \frac{1}{|detA_{ij}|} \exp(i\mathbf{u}^T \mathbf{t}_{ij}) |F_i(|A_{ij}^T|^{-1}\mathbf{u})| \tag{3}$$

such that the linear part, A_{ij}, affects only the amplitude spectrum and the the translation $\mathbf{t}_{ij}$ is exhibited as a phase gradient. To estimate the affine transformation a three stage method is applied: (1) the amplitude spectrum of the source block is modelled by a Gaussian mixture; (2) a search is made for the linear transformation of this mixture model to fit the spectrum of the target block; (3) the best translation between the linear transformed source block and the target block is recovered by cross-correlation. These steps are shown in the flow chart of figure 1.

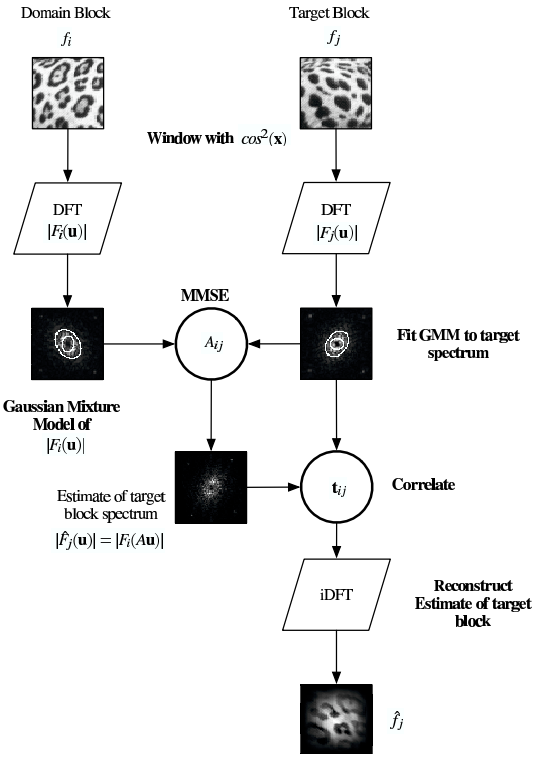

Figure 1: **Affine transformation estimation stages**. Source (domain) image block is i, and a target image block is j. The amplitude spectrum of the DFT of the source block is modelled by a low-order Gaussian mixture model centred on DC. A gradient descent search is made to determine the linear part of the transformation, A_{ij}, with the the target block amplitude spectrum. Then, the DFT spectrum of $F_i(A_{ij}\mathbf{u})$ and $F_j(\mathbf{u})$ are correlated to estimate the translation $\mathbf{t}_{ij}$.

2.1 A Gaussian Mixture Model of Spectral Amplitude

The amplitude spectrum, $|F_i(\mathbf{u})|$, is modelled as a two-dimensional, M component Gaussian mixture

$$G_i(\mathbf{u}) = \sum_m^M a_m \exp(-\mathbf{u}^T C_m^{-1}\mathbf{u}/2), \tag{4}$$

having centroids fixed on the origin of the spectrum and coordinates $\mathbf{u} = (r,s)^T$. The Gaussian parameters, $\{a_m, C_m\}$ are then estimated by minimising the residual error,

$$\sum_{\mathbf{u}} (|F_i(\mathbf{u})| - G_i(\mathbf{u}; a_m, C_m))^2,$$

using a standard non-linear optimisation method, Levenberg-Marquardt (LM) [7]. LM is a conjugate gradient descent method which requires the gradient with respect to the fitting

parameters to be known at each data point. It is not difficult to show that

$$\frac{dG_i(\mathbf{u})}{da} = G_i(\mathbf{u})/a, \qquad \frac{dG_i(\mathbf{u})}{dc_{rr}} = -G_i(\mathbf{u})r^2/2, \tag{5}$$
$$\frac{dG_i(\mathbf{u})}{dc_{rs}} = -G_i(\mathbf{u})rs, \qquad \frac{dG_i(\mathbf{u})}{dc_{rr}} = -G_i(\mathbf{u})s^2/2,$$

where c_{rs} are the components of the inverse of the covariance matrix,

$$\begin{pmatrix} c_{rr} & c_{rs} \\ c_{rs} & c_{ss} \end{pmatrix} = C_m^{-1}.$$

In the experiments presented below, we have found that a mixture with 2 components is sufficient to model the general shape and amplitude of oriented and periodic patches. The purpose of this modelling is not to accurately fit the amplitude spectrum, but to capture its principal features such that a linear transformation of the model can *uniquely* fit the amplitude spectrum of a target image block.

2.2 Linear Transformation Estimation

The second step of the affine transformation estimation uses the mixture model of the source block spectrum and searches for a linear transformation that minimises the squared residuals, $\sum_{\mathbf{u}}(G_i(A_{ij}(\mathbf{u})) - |F_j(\mathbf{u})|)^2$. This search can be again performed by the LM method. This time the gradients of $H_i(\mathbf{u}) = G_i(A(\mathbf{u}))$ w.r.t. the parameters of the linear transformation matrix,

$$\begin{pmatrix} a_{rr} & a_{rs} \\ a_{sr} & a_{ss} \end{pmatrix} = A_{ij}$$

are needed:

$$\begin{aligned}
\frac{dH_i(\mathbf{u})}{da_{rr}} &= -H_i(\mathbf{u})(1+r)(c_{rr}, c_{rs})A\mathbf{u}/2, \\
\frac{dH_i(\mathbf{u})}{da_{rs}} &= -H_i(\mathbf{u})(1+s)(c_{rr}, c_{rs})A\mathbf{u}/2, \\
\frac{dH_i(\mathbf{u})}{da_{sr}} &= -H_i(\mathbf{u})(1+r)(c_{sr}, c_{ss})A\mathbf{u}/2, \\
\frac{dH_i(\mathbf{u})}{da_{ss}} &= -H_i(\mathbf{u})(1+s)(c_{sr}, c_{ss})A\mathbf{u}/2.
\end{aligned} \tag{6}$$

Note LM is prone to become trapped in local minima and more reliable performance is achieved by initialising the searches at several different start points. A is initially restricted to be Cartesian for these searches, so that $a_{rr} = a_{ss} = \gamma\cos(\phi)$ and $a_{rs} = -a_{sr} = \gamma\sin(\phi)$. When this converges, typically in under 10 iterations, a shearing of the model is permitted by letting the off-diagonal elements have different scaling values: $a_{rs} = \gamma_1\sin(\phi), a_{sr} = \gamma_2\sin(\phi)$.

2.3 Translation Estimation

The final step is estimating the translation, t_{ij}, which is exhibited in the phase spectrum of the block DFT. An estimated transformed block, $\hat{f}_j(\mathbf{y})$, is synthesised by the applying

the linear transformation: $\mathbf{y} = A_{ij}\mathbf{x}$. The translation, $\mathbf{t}_{ij}$, is then taken as the peak location of the cross-correlation i.e.

$$\mathbf{t}_{ij} = \arg\max_{\mathbf{t}}[\hat{f}_j \star f_j].$$

3 Self-Similarity Clustering

Image patches are related through the operation of the affine group $(A,\mathbf{t})f_j = f_i$; the process of image modelling described above produces a set of samples from this group A_{ij} that minimise the residuals $((A_{ij},\mathbf{t}_{ij})f_i - f_j)^2$. If the image patches i, j do not belong to the same image region, then the residual will be large and A_{ij} will fail to model a self-similarity. In these cases, to minimise the reconstruction error, a different source image patch f_l can be used, which is representative of the region to which f_j belongs. To discover a set of prototypical textons, ideally one for each image region that appropriately 'span' the image, a clustering approach is used.

In the supervised case, we cluster not just to the representative (given) cluster centroids, but to an *extended family* of cluster centres which are created by taking a finite sub-group from the group $(A_{ij},\mathbf{t}_{ij})f_l$, where l is a given prototypical block. As with the GMM model estimation described above, by extroplating the sub-group of rotations and scaling from the amplitude spectrum prototype block, makes the family translation invariant. In our experiements we have restricted these prototype transformations to a sub-group of rotations and scalings of f_l and ignored shears. We have used 8 rotations and 8 scalings to create a cluster family $f_{lk}, 0 \leq k < 64$. Having applied each transformation, k, these prototype blocks give B^2 dimensional vectors: $f_{lk}(\mathbf{x}) \rightarrow \mathbf{f}_{lk}$. The remaining image blocks are then clustered using a nearest neighbour criterion. We can imagine that the data feature vectors cluster closest points on manifolds represented by the k rotation/scaling family of each prototype. Thus each data block is labelled as the closest family member which have already taken into account the affine group invariance. Schölkopf has suggested this same idea but for extending SVM vectors [8].

In the unsupervised case, each image data block has to have a set of likely affine-variants created before clustering can take place. This is clearly computationally arduous and requires a more sophisticated clustering method. Techniques for affine invariant clustering are discussed in detail by Fitzgibbon and Zisserman [9] who describe a distance metric incorporating a prior on the transformation parameters. These methods may be more suited to the unsupervised problem.

4 Experimental Results

Figure 2 shows results of self-similarity image modelling and reconstruction on a reptile skin image taken from the Brodatz database (top row), a fingerprint (second row) and two wild-life images with zebras and a jaguar (third and fourth rows). Columns (b)-(d) show synthesis results using a single domain block (highlighted in white on the original (a)) at overlapping block sizes $B = 16, 32$ and 64 pixels. The mixture modelling analysis was performed on whitened (Laplacian pyramid) version of the original (see [5] for details) and coarse version of the low-pass information was added back at the end. The number of parameters per pixel for these images is related to the block size, B and the image size N and can be estimated by: $P(B) = \frac{11}{(B/2)^2} + (B/N)^2$. This assumes that each affine

transformation requires 6 parameters, with 1 for the intensity scaling, α, and the low-pass information encoded at twice the spatial resolution as the number of blocks in each dimension i.e. 4 extra parameters per block: the total being $6+4+1=11$ per block. In addition, each prototype itself has B^2 pixels per image:

Params/pixel	$B=8$	$B=16$	$B=32$	$B=64$
$P(B)$	0.688	0.173	0.047	0.026

The results demonstrate that the chosen texton is able to describe and reconstruct well the original data including the imperfections in the repetitive cellular pattern of the reptile scales, the cuts and minutiae of the fingerprint and the multi-directional and bifurcating stripes of the zebra. Even the reconstructions at the largest texton size (the lowest 'rate', at $P(64)$), exhibit the essential structural features of the image, yielding a 'painted' effect. In the zebra reconstructions, the background has no obvious deterministic component and is modelled largely by the amplitude scaling α_{ij} of (2). Where there is a structural texture present such as in the jaguar image the prototype, taken from the back of the jaguar containing the jaguar 'print', can be transformed adequately to become the leaves of the forest behind the animal as well as the bark of the log.

	SNR dB					
Block size	16		32		64	
Params/pixel	0.688		0.173		0.047	
`reptile`	13.3	-	8.6	-	5.3	-
`finger`	15.4	15.1	10.9	11.1	6.9	7.2
`zebra`	14.4	14.8	10.4	(8.1)	6.9	7.6
`jaguar`	16.0	16.3	11.5	11.8	7.8	8.5
`girl`	14.8	18.4	11.0	14.7	8.8	11.1

Table 1: **Estimates of reconstruction quality**. Results based on 1 or more chosen prototypes using affine invariant block based texture model. The SNR in first column for each given block size (e.g. 16×16) is for single prototype reconstructions shown in figure 2. The second column shows the errors (again for each block size), for multi-prototype reconstruction using the self-similarity clustering as illustrated in figure 3

Preliminary results of self-similarity clustering are presented in figure 3 and 4. Different blocks sizes were used to perform the clustering and subsequent reconstructions: $B=16,64$ results are shown in figure 3 for the fingerprint, zebra, jaguar and girl images. In all cases, 3 texton blocks were used to seed the cluster families (shown on the original images in figure 2, (a)). For the girl image, blocks in the background, on the edges of the mirror and hair, and in the feathers were used. The clustering results on the fingerprint show a grouping of region blocks at orientations radially around the central whorl feature. This is to be expected, as the cluster manifold is expanded to include 8 rotations around the circle. Since textons with horizontal, vertical and diagonal ridges were chosen, the labels roughly alternate between these three classes. In the case of the zebra, the background is clearly labelled as one group and the vertical and horizontal stripe labels together identify the two zebras. For the jaguar image, the animal is fairly well segmented from the log (foreground) and the background leaves and trees. The final two columns are the corresponding reconstruction results using all three texton blocks in both images. The fingerprint results are marginally better than the single source block reconstructions

shown in figure 2. For the zebras, the background is better reconstructed by the texton representing grass but the reconstructions show some amplitude modelling errors. This is partly caused by the user specification of the the texton on the back of the right-hand zebra falling on a white stripe: it does not contain any black to properly reconstruct the belly of the right-hand zebra. However, for the zebras, the multi-block reconstructions are overall better than those in figure 2. With the girl image, the visual results are hard to interpret as it is difficult to manually select a representative compact set of prototypes. We believe that an unsupervised texton selector, as outlined in the previous section, would alleviate such problems and produce better reconstructions. Table 1 presents a quantitative comparison of the single and multiple prototype reconstructions (guided by the self-similarity clustering). These results confirm the qualitative findings that the reconstructions based on the clustering are marginally better.

Finally, we ran a self-similarity clustering experiment on a slice of CT data, figure 4. What is notable about these results is that we were able to specify prototype blocks in textural or piece-wise constant regions, such as the background and brain, as well as *edge regions*, such as the skull and internal bones. The clustering results show the method is easily able to highlight the similar regions and could form the basis of a feature set for a contextual classifer, like MRF model, or a global approach using deformable templates or contours.

5 Conclusions

A self-similarity based image model has been presented which can be used to synthesise textural image regions from representative image patches with good accuracy. We have shown how a clustering approach can be used to discover, i.e. segment, these prototypical image patches. The crux of the method is the underlying affine-invariant group model of the patch that allows us to extend the given cluster centre to its immediate affine-group family simplifying the clustering process. In the supervised case, the reconstruction results are indifferent and the choice of representative patch appears critical. We hope to use a hierarchical, unsupervised clustering to improve the results further. Despite these problems, the affine-group mixture modelling and self-similarity reconstruction may provide a basis of other applications such as content based image retrieval [10].

References

[1] A. Jacquin. Image Coding Based on a Fractal Theory of Iterated Contractive Image Transformations. *IEEE Trans. Image Proc.*, 1:18–30, 1991.

[2] P. Volet and M. Kunt. Synthesis of Natural Structured Textures. In *Signal Processing III: Theories and Applications*, pages 913–916. 1986.

[3] F. Liu and R. W. Picard. Periodicity, Directionality, and Randomness: Wold Features for Image Modeling and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):722–733, 1996.

[4] A. A. Efros and W. T. Freeman. Image Quilting for Texture Synthesis and Transfer. In Eugene Fiume, editor, *SIGGRAPH 2001, Computer Graphics Proceedings*, pages 341–346. ACM Press / ACM SIGGRAPH, 2001.

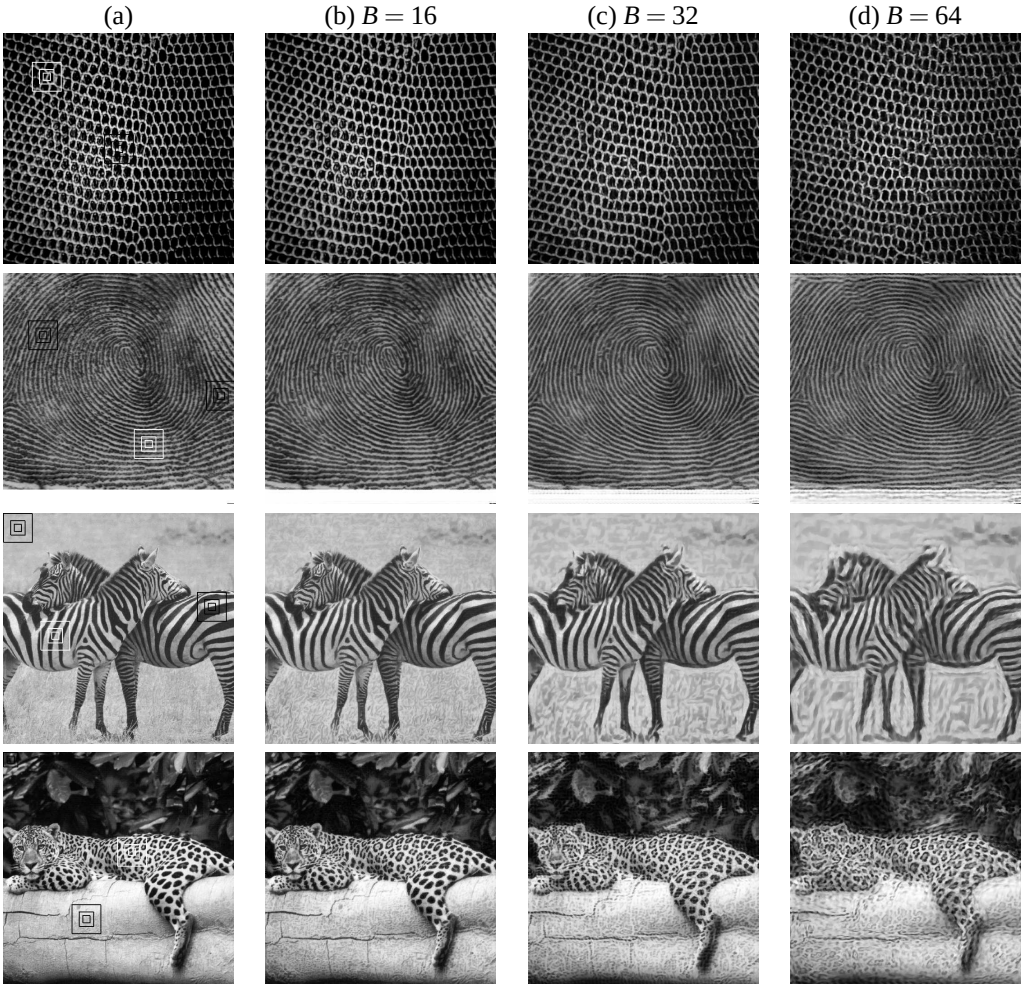

Figure 2: **Affine self-similarity image model reconstructions**: (a) using the texton blocks marked in white at block sizes $B = 16, 32, 64$ (b)-(d).

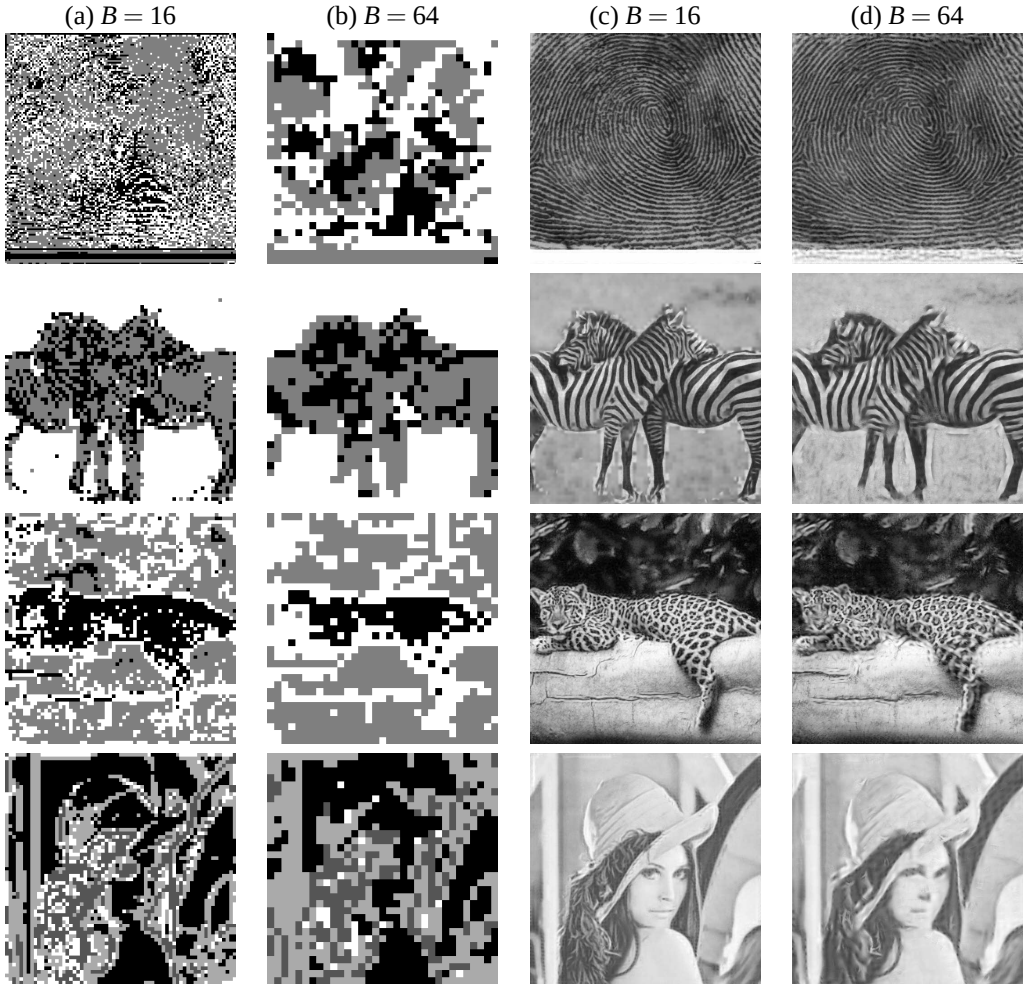

Figure 3: **Supervised self-similarity clustering results**: (a)-(b): results on fingerprint, zebras, jaguar and girl images grouping image blocks ($B = 16, 64$) that are related by an affine-group transformation. (c)-(d) Image reconstructions using 3 prototype texton blocks selected by self-similarity clustering results using (a),(b).

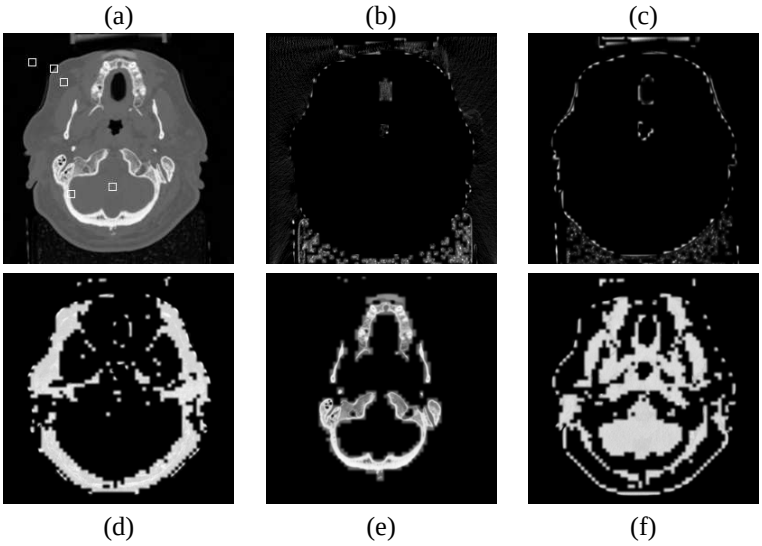

Figure 4: **Self-similarity clustering results on medical data**. An image slice from a CT data set size 256×256 was used (a). The user selected prototype blocks are marked in white on the original. (b)-(f) images show labelled blocks falling into the image classes: (b) background; (c) skull; (d) tissue 1; (e) tissue/bone; (f) tissue 2. Note that although (c) and (e) represent 'edge' type image features, the prototype selection is the same as for region feature classes such as 'tissue' or 'background'.

[5] T-I Hsu and R. Wilson. A Two-Component Model of Texture for Analysis and Synthesis. *IEEE Trans. on Image Processsing*, 7(10):1466–1476, October 1998.

[6] A. D. Calway. Image Representation Based on the Affine Symmetry Group. In *Proc. ICIP 1996*, pages 189–192, 1996.

[7] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, second edition, 1992.

[8] B. Schölkopf and C. Burges and V. Vapnik. Incorporating Invariances in Support Vector Learning Machines. In *"Artificial Neural Networks, ICANN'96"*, volume 1112, pages 47–52, 1996.

[9] A. Fitzgibbon and A. Zisserman. On Affine Invariant Clustering and Automatic Cast Listing in Movies. In *"Proc. 7th ECCV 2002"*, 2002.

[10] S. Lazebnik, C. Schmid, and J. Ponce. A Sparse Texture Representation Using Affine-Invariant Regions. In *Proc. IEEE CVPR*, pages 272–277, 2003.