Invariant Shape Matching for Detection of Semi-local Image Structures

Lech Szumilas, Horst Wildenauer, and Allan Hanbury

Vienna University of Technology
Gußhausstraße 27-29 / E376, A-1040 Wien, Austria
{szumilas,wildenauer}@acin.tuwien.ac.at,
hanbury@prip.tuwien.ac.at
http://www.acin.tuwien.ac.at

Abstract. Shape features applied to object recognition has been actively studied since the beginning of the field in 1950s and remain a viable alternative to appearance based methods e.g. local descriptors. This work address the problem of learning and detecting repeatable shape structures in images that may be incomplete, contain noise and/or clutter as well as vary in scale and orientation. A new approach is proposed where invariance to image transformations is obtained through invariant matching rather than typical invariant features. This philosophy is especially applicable to shape features such as open edges which do not have a specific scale or specific orientation until assembled into an object. Our primary contributions are: a new shape-based image descriptor that encodes a spatial configuration of edge parts, a technique for matching descriptors that is rotation and scale invariant and shape clustering that can extract frequently appearing image structures from training images without a supervision.

Keywords: Shape features, image descriptor, model extraction.

1 Introduction

Edges are an intuitive way to represent shape information, but the problems associated with the edge detection such as edge fragmentation, missing edges due to occlusions or low contrast as well as changes in object scale and orientation affect the final result based on edge matching or classification [1]. To overcome these issues we introduce a novel semi-local shape descriptor which represents the shape of an image structure by means of edges and their configurations. Our *Radial Edge Configuration*-descriptor (REC) encodes edges found in a neighborhood of an interest point (see Section 2) as a sequence of radial distances in a polar coordinate system (centered on the interest point). Thus, the similarity of shape is assessed by the comparison of local edge configurations. Here, our main contribution is the definition of a rotation and scale-invariant distance measure between edge configuration descriptors that is able to match multiple

[1] Our method utilizes Canny edge detector.

M. Kamel and A. Campilho (Eds.): ICIAR 2009, LNCS 5627, pp. 551–562, 2009.

edges, preserving their spatial relationships, and reject outlier edge pairs at the same time. This allows for a comparison of image structures across different scales, with only partially established correspondences. Another particularity of the chosen approach is that scale and orientation are not estimated during descriptor extraction. Instead they are established as relative entities between two REC descriptors during the distance calculation, which leads to more stable results.

We also introduce a method for weakly supervised learning of structure models that are represented by a set of REC descriptors with individual edges weighted accordingly to their repeatability and similarity within the same category of structures. The structure model learning is achieved through shape clustering presented in Section 5. The quality of extracted structure models is evaluated on database of MRI spine images described in Section 6.

The shape clustering is related to agglomerative hierarchical clustering but operates on variable length feature vectors, specifically Radial Edge Configurations. The result of shape clustering are "mean" edge fragment configurations (represented by REC descriptors) that can be used to locate similar structures in the image.

2 Symmetry Based Interest Points

The *Radial Symmetry Transform* (RST) attempts to find locations in the image where the intensity distribution attains locally maximal radial symmetry. The method tends to locate interest points approximately at the centers of round/isotropic structures or along the symmetry axis of elongated shapes. The symmetry measure $S_r(x, y)$ is calculated for each pixel (x, y) of the image separately and the interest points are aligned with local symmetry maxima.

$$S_r(x,y) = -\sum_{i=-r}^{r}\sum_{j=0}^{r} g(\sqrt{i^2+j^2}, \sigma_r = 0.5r)\|\boldsymbol{I}(x+i,y+j) - \boldsymbol{I}(x-i,y-j)\| \quad (1)$$

where $\boldsymbol{I}(x+i, y+j)$ is an image pixel intensity or color at coordinates $(x+i, y+j)$ and r defines the image window size used for the symmetry measure calculation to be a $(2r + 1) \times (2r + 1)$ rectangle. Each contribution of the pixel pair at $(x + i, y + j)$ and $(x - i, y - j)$ is weighted by the Gaussian $g(\sqrt{i^2 + j^2}, r)$ which decreases the influence of pixel pairs at increasing distance from (x, y) and normalizes the transform with respect to the chosen scale R.

In the basic version, the interest point locations $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ correspond to the maxima of the S_r transform:

$$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \underset{x,y}{\texttt{argmax}}\,(S_r) \quad (2)$$

It is also possible to obtain a scale adapted set of interest points using a similar iterative approach as for the scale adapted Harris detector [4]. In this

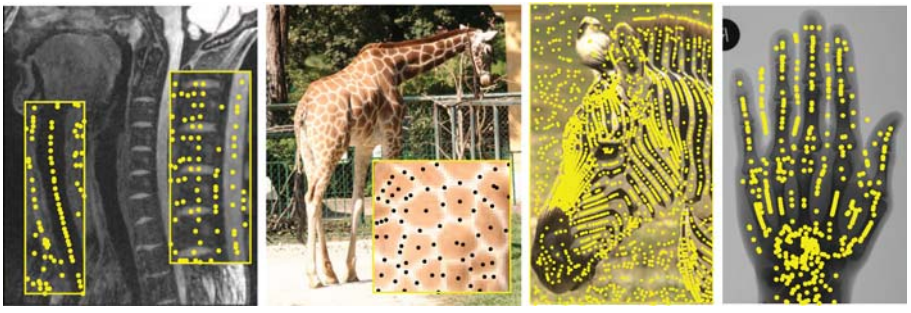

Fig. 1. Examples of RST based interest points computed at a single scale ($r = \varsigma/50$, where ς is a lower value out of horizontal and vertical image size in pixels)

case the interest point locations are detected using the symmetry transform and the related scale is detected using the Laplacian operator. Alternatively, an approximation of the scale adapted symmetry measure is a sum of S_r over a sparse set of radii R:

$$S = \sum_{r \in R} S_r \tag{3}$$

Examples of interest point detection are presented in Figure 1.

3 Edge Matching in Polar Coordinates

The complexity of edge matching is primarily associated with the difficulty in assigning a scale to the edge – a part of one edge may be matched to another edge or to itself at a larger scale (e.g. straight edges or fractal like structures). Polar coordinates allow the definition of an edge scale locally, based on the relative position to the origin of a coordinate system. However, the matching of a part of an edge to a part or whole of another edge is still admissible.

The origin of the coordinate system is associated with the interest point location.

The REC descriptor consists of a variable number of K continuous edges. The k-th edge Γ_k is encoded as an ordered list of radial boundary points, each representing the distance $r_{k,i}$ along the i-th ray from the origin of the polar coordinate system:

$$\Gamma_k = \{r_{k,i} : i \subset \mathbb{N}_0^+; i = (b_k ... b_k + n_k) \bmod N\} \tag{4}$$

where b_k denotes the index of the first ray and n_k is the number of rays the edge occupies. The modulo operation is used to ensure that index $i < N$, where N describes the total number of rays (polar resolution) and in all our experiments is set to 64, which we found to offer a good compromise between accuracy and computational cost.

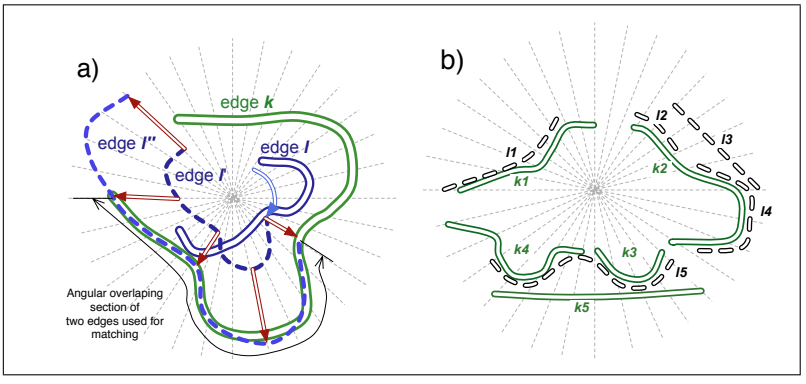

Fig. 2. a) example of matching edge k and l in polar coordinates. Edge l' is a rotated version of l and l'' is scaled version of l' relative to the origin of the coordinate system. b) example of edge correspondences in two descriptors (edges k and l).

Calculating the distance between two REC descriptors involves finding correspondences between multiple edges. We describe a method to find the best fit between two edges, assuming one of the edges can be rotated and scaled relative to the origin of the polar coordinate system associated with the interest point (as shown in Figure 2). This operation is a prerequisite for the estimation of distance between two REC descriptors.

Fitting one edge to another corresponds to finding a transformation (rotation and scaling) which globally minimizes the spatial distance between corresponding boundary points of the two edges. It is important to note that while the scaling of an edge is performed in the continuous domain, the relative rotation is quantized into N rays. The relative scale $\varsigma_{k,l}^{a,b}$ between edge k belonging to the descriptor a and edge l belonging to the descriptor b, rotated by α rays, is calculated as follows:

$$\varsigma_{k,l}^{a,b}(\alpha) = \left(\sum_{i=b_{kl}}^{b_{kl}+n_{kl}} r_{k,i}^{a} r_{l,\bar{i}}^{b} \right) \Big/ \left(\sum_{i=b_{kl}}^{b_{kl}+n_{kl}} (r_{l,\bar{i}}^{b})^2 \right) \tag{5}$$

where b_{kl} is the first ray containing boundary points of both edges, n_{kl} is the number of consecutive rays containing boundary points from both edges for a given rotation α and $\bar{i} = (i - \alpha) \bmod N$. It is important to note that this scheme allows for partial edge matching, which means that only the overlapping section of the two edges is matched (as shown in Figure 2). However, only combinations of α for which $n_{kl} \geqslant \tau$ (in our experiments τ=5) are used, due to the fact that extremely short sections of an edge usually carry less information, which is made worse by the quantization process. It can be easily proven that the spatial distance between corresponding boundary points of the edges k and l, for a given rotation α, is minimized when edge l is scaled (multiplied) by $\varsigma_{k,l}^{a,b}(\alpha)$.

One way of estimating how well two edges fit together is to calculate the variation of relative scale between the corresponding boundary points:

$$\epsilon_{k,l}^{a,b}(\alpha) = \frac{1}{n_{kl}} \sum_{i=b_{kl}}^{b_{kl}+n_{kl}} \left| \log^2\left(\frac{r_{k,i}^{a}}{r_{l,\bar{i}}^{b}}\right) - \log^2\left(\varsigma_{k,l}^{a,b}(\alpha)\right) \right| \tag{6}$$

This equation is a scale independent fitting distance between two edges for a given relative rotation α. The $\log^2()$ operation is used to avoid impairment associated with the $\frac{r_{k,i}^{a}}{r_{l,\bar{i}}^{b}}$ measure. The relative rotation giving the best fit of the two edges is the one which minimizes the distance $\epsilon_{k,l}^{a,b}$:

$$\epsilon_{k,l}^{a,b} = \min_{\alpha}\left(\epsilon_{k,l}^{a,b}(\alpha) : n_{kl} \geqslant \tau\right) \tag{7}$$

Finding the transformation resulting in the best fit between two edges requires $\epsilon_{k,l}^{a,b}(\alpha)$ to be evaluated for all α (for which $n_{kl} \geqslant \tau$ holds).

4 Descriptor Distance

The REC descriptor contains a set of edges that are the result of edge detection around the corresponding interest point. In reality we should expect that some perceptible edges may be missing or fragmented due to weak gradients and noise. An additional problem is related to the fact that only a subset of edges in the two descriptors may correspond well, while others are related to non-similar image structures. For example we can find patches on a giraffe skin with a high shape similarity at a local scale, but the random distribution of the patches makes shape comparison irrelevant on a large scale. Thus we have to search for a subset of edges in both descriptors, which together give a low fitting error, while other edges are rejected as outliers.

The primary idea behind the matching of multiple edges in the descriptors a and b is summarized below:

1. Perform edge fitting for admissible edge pair combination k and l, resulting in P putative transformations.
2. Repeat multiple edge fitting for P transformations. Choose the one which gives the lowest overall fitting error for the descriptor.
   (a) Rotate and scale all edges in descriptor b according to the current transformation and find the edge correspondences between two descriptors.
   (b) Remove outliers and calculate the final distance from all corresponding edge pairs.

The most computationally demanding task is finding edge correspondences for a given relative scale and rotation. The difficulty is associated with the possibility that a single edge in one descriptor may correspond to more than one non-overlapping edges in the other descriptor. An example of such multi-correspondences is shown in the Figure 2-b – edge $k2$ corresponds to edges $l2$ and $l4$, while edges $k4$ and $k3$ correspond to edge $l5$. Note that edge $l3$ could be

also matched to the edge $k2$, but it overlaps with edges $l2$ and $l4$, which produce a better fit with edge $k2$. The process of finding edge correspondences can be divided into several steps:

1. Find overlapping edge pairs in a: $\phi^a_{k1,k2} = \begin{cases} 1, \text{ if } k1 \text{ and } k2 \text{ overlap } \geqslant \tau \\ 0, \qquad \text{otherwise} \end{cases}$
2. Find overlapping edge pairs in b: $\phi^b_{l1,l2} = \begin{cases} 1, \text{if } l1 \text{ and } l2 \text{ overlap } \geqslant \tau \\ 0, \qquad \text{otherwise} \end{cases}$
3. Find overlapping edge pairs between a and b: $\phi^{ab}_{k,l} = \begin{cases} 1, \text{ if } k \text{ and } l \text{overlap } \geqslant \tau \\ 0, \qquad \text{otherwise} \end{cases}$
4. Find edge correspondence. The edge l is correspondent to edge k if:

$$\epsilon^{a,b}_{k,l} = \min_{f,g}\Big(\epsilon^{a,b}_{f,g} : f \in \{\phi^{ab}_{f,l} = 1 \wedge \phi^a_{f,k} = 1\}; g \in \{\phi^{ab}_{k,g} = 1 \wedge \phi^b_{l,g} = 1\}\Big) \quad (8)$$

which means that edges k and l correspond when the distance $\epsilon^{a,b}_{k,l}$ is the minimum among all combinations of edges f and g which overlap with k and l. This condition allows the association of multiple non-overlapping edges in one descriptor with a single edge in another descriptor.

The final distance between two descriptors a and b is a weighted sum of individual edge-pair (k, l) distances:

$$\epsilon^{a,b} = \frac{1}{\sum_{k,l} \upsilon^a_k \upsilon^b_l} \sum_{k,l} \upsilon^a_k \upsilon^b_l \epsilon^{a,b}_{k,l} \quad (9)$$

where the weights υ_k and υ_l describe the confidence of edge match:

$$\upsilon_k = \frac{\widehat{s}^a_k}{s^a_k} \quad (10)$$

where s^a_k is the total length of edge k in descriptor a and $\widehat{s}^a_k$ is the length of all edge fragments that were matched to edges in the descriptor b. The edge match confidence reaches 1 if it was completely matched to other edge or edges and is 0 if it was not matched to any edges.

During our matching tests we found that a simple outlier removal scheme helped to improve results when only a part of the structure in the two descriptors was found to correspond.

Examples of finding similar image structures through the edge matching are presented in Figures 3, 4 and 5. Majority of descriptors are matched to similar structures despite differences in scale, orientation and shape deformations.

5 Clustering of Radial Edge Configurations

Clustering of local image descriptors (e.g. SIFT) is the basis of object recognition techniques such as "bag of keypoints" [5] as well as part based models [3]. In these cases clustering allows for a compact (data reduction) representation

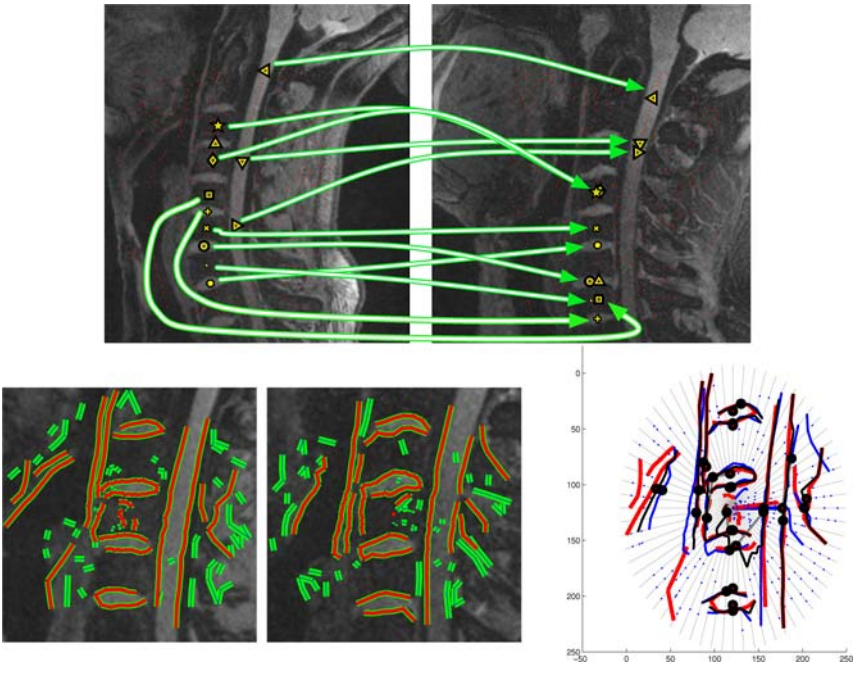

Fig. 3. Top row: example of descriptor matching between different MRI images. Only a representative subset of interest point matches is shown to avoid clutter. Bottom row: example of two similar image structures matched. The first two images show corresponding image patches and the extracted edges (edges which length falls below configurable threshold are not used for matching and marked with a green color). The third image shows correspondence of edges from two descriptors (red and blue respectively) and the resulting mean edges after descriptor merging (black). Note that not all edges have been matched. We strongly advise to view all images in color.

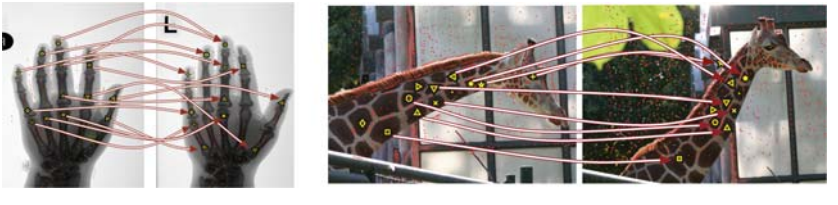

Fig. 4. Examples of descriptor matching. Corresponding descriptor locations are connected with arrows and marked with a unique symbol.

of distinctive image structures. Among the most popular clustering methods are hierarchical, k-means and kd-tree clustering. The first difference between clustering of typical image descriptors and clustering of the REC descriptor is that the later produces a variable length feature vector (the number of edges can vary significantly). This prevents the use of k-means and kd-tree clustering which

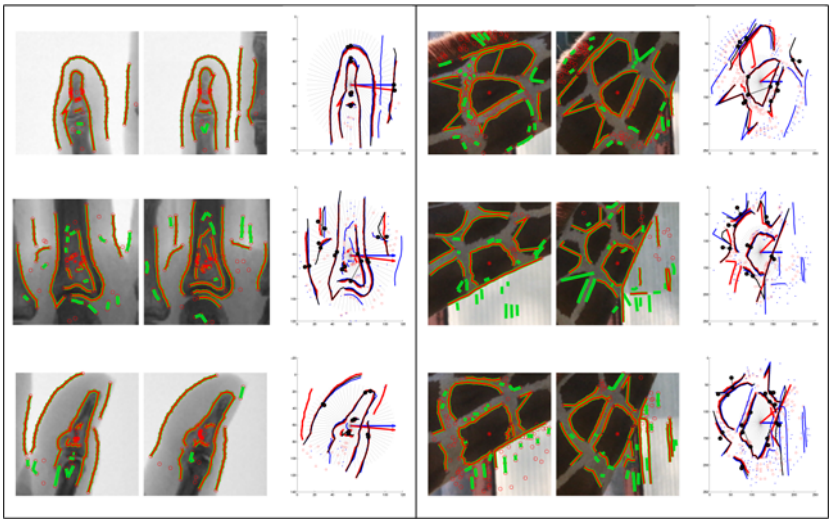

Fig. 5. Examples of REC matching in X-Ray images of hands (left) and the giraffe skin (right). The first two columns contain corresponding image regions and third column shows edge correspondence (black lines depict mean edges).

require constant dimensionality of the feature vectors. The second difference is that the clustering of REC descriptors assigns weights to edges and individual boundary points along the edges that depend on the edge repeatability across training instances of the same structure type and the amount of variability an edge exhibits across the training instances.

The REC descriptor is clustered using agglomerative hierarchical clustering [2] based on the REC distance defined in Section 4. Clustering starts with finding the closest pairs between a set of descriptors extracted from the training data set labelled as clustering level $t = 0$. The closest pairs are merged into nodes at the next clustering level and the same procedure is repeated on these nodes. The closest descriptor pairs are merged only if the matching distance between them does not exceed the threshold τ. Therefore clustering is performed until no more pairs can be merged. Parameter $\tau = 0.4$ was experimentally chosen and used in all tests presented in this chapter. The merging of two descriptors is an operation which generates a single edge for each set of corresponding edges in two descriptors as described in Section 4. Recall that a single edge in one descriptor can correspond to several edges in another descriptor and that some edges do not have any correspondences and are down-weighted in the merged descriptor. The edge kl, which is a result of merging of edges k and l, is obtained by averaging the boundary point positions from both edges:

$$\Gamma_{kl} = \{0.5(r_{k,i} + r_{l,i-\alpha \bmod N_0^+}) : i \subset \mathbb{N}; i = (b_{kl}...b_{kl} + n_{kl}) \bmod N\} \qquad (11)$$

In addition, each boundary point is assigned the weight that is corresponding to the distance between two merged boundary points and includes the boundary point weights from the previous clustering level. This way edges are prioritized according to their similarity across the clustering levels.

$$w_{kl}^{t}(i) = \omega_p(w_k^{t-1} + w_l^{t-1}) + \omega_d \exp\left(- \left(1 - \frac{\max\left(r_{k,i}^{a}, \varsigma_{k,l}^{a,b} r_{l,\bar{i}}^{b}\right)}{\min\left(r_{k,i}^{a}, \varsigma_{k,l}^{a,b} r_{l,\bar{i}}^{b}\right)} \right)^2 / \sigma^2 \right) \quad (12)$$

where σ was set to 0.25 in all experiments and regulates the down-weighting depending on the local edge deformation – the difference between relative boundary point scale and the relative descriptor scale. The parameters ω_p and ω_d regulate the influence of edge weights from previous cluster level $t-1$ (history) and the differences between merged edges (deformation) respectively onto the final weight $w_{kl}^{t}(i)$. These were set to $\omega_p = 0.25$ and $\omega_d = 0.75$ in all experiments which prioritizes the influence of "deformation" over the "history". The edges without correspondences are copied into the merged descriptor and the corresponding weights are divided by two – if such an edge consequently has no correspondences at multiple clustering levels its weight is reduced to approximately 0.

At clustering level $t = 0$ all boundary point weights are set to 1 which means that all edges in every descriptor have identical priority.

The result of clustering is a set of REC descriptors, which contain edges resulting from edge merging across a number of clustering levels. The weights assigned to the edges are then used during matching cluster nodes (structure models) to descriptors in the test data set. The edge distance (6) is then replaced with:

$$\epsilon_{k,l}^{a,b}(\alpha) = \frac{\sum_{i=b_{kl}}^{b_{kl}+n_{kl}} w_{k,i}^{a} \left| \log^2\left(\frac{r_{k,i}^{a}}{r_{l,\bar{i}}^{b}}\right) - \log^2\left(\varsigma_{k,l}^{a,b}(\alpha)\right) \right|}{\sum_{i=b_{kl}}^{b_{kl}+n_{kl}} w_{k,i}^{a}} \quad (13)$$

where descriptor a corresponds to the cluster node and weights for descriptor b corresponding to the detected structure are set to 1.

6 Weakly Supervised Model Extraction in MRI Spine Images

The intention of this test scenario is to show the discriminative capabilities of structure models obtained from shape clustering. The evaluation is performed on MRI spine images, that contain characteristic structures such as vertebrae, disks and the spinal cord. Figure 6 shows examples of MRI images used in this evaluation as well as examples of the manual structure annotation that assigns structure type labels to the symmetry based interest points[2] The annotation of

[2] The MRI image database consists of 30 images.

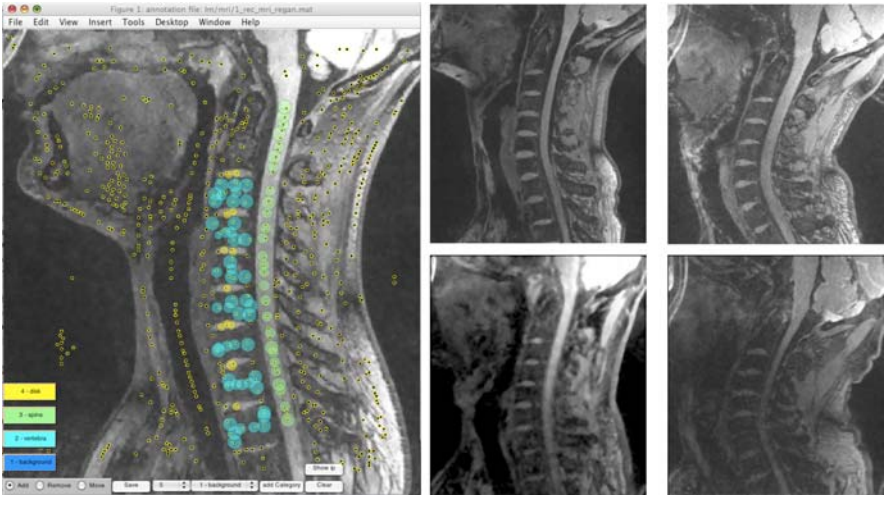

Fig. 6. Left: Example of MRI annotation. The categories represent 3 characteristic structures (visible as color disks covering corresponding interest points) and the background (interest points that were not annotated). Right: Examples of test images.

a single image can be performed in less than one minute – the annotation of structure boundaries is not needed.

The localization of vertebrae, disks and spine has a medical application of providing landmarks for image segmentation and global structure localization [1].

The structure model extraction is performed using shape clustering described in Section 5. The training descriptor database contains approximately 10% of all images. Every category is separately clustered which produces cluster trees containing structure models (see Figure 7). The structure models are then matched to the test images and the classification accuracy based on minimum matching distance is estimated. Table 1 shows classification accuracy (true positives) at equal error rate.

The results in Table 1 show that clustering improves detection accuracy (vs. unclustered models) for all categories except the background e.g. the models of vertebrae obtained from descriptor clustering are correctly matched to 87% vertebrae related descriptors in the evaluation data set while without clustering only 68% of descriptors are correctly matched. The clustered descriptors have

Table 1. Detection accuracy in test images using structure models obtained from descriptor clustering. The numbers represent true positives at equal error rate.

	Vertebrae	Disk	Sine	Background
Unclustered	0.6809	0.8875	0.8511	0.8596
Clustered $\tau = 0.4$	0.8723	0.8625	0.9149	0.6555

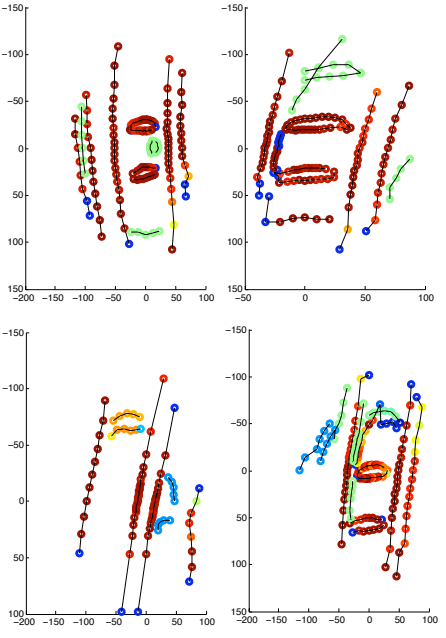

Fig. 7. Example of weights assigned to boundary points in the process of shape clustering (top: vertebrae models, bottom: disk and spine models)

weights assigned to the encoded edges that describe repeatability of them among examples in the training data set while these weights are set to 1 in unclustered descriptors. This explains why repeatable structures such as vertebrae, disc and spine are better detected by structure models obtained from descriptor clustering. Background detection however shows the opposite trend due to higher variability of background related structures than in the case of other categories e.g. compare the structures behind the spine in examples in Figure 7. The improvement of background matching is possible either by using training data set that contains majority of structures occurring in the test data set or by learning and detecting spatial relationship between detected structures (e.g. [1]).

7 Conclusions

We have presented a method for clustering shapes that uses an edge based semi-local shape descriptor (REC) together with a robust scale and rotation invariant distance measure. This allows us to perform clustering of the descriptors in order to obtain a consistent representation of similar image structures.

The presented test scenario shows the applicability of the REC descriptor to detection of image structures in medical images. The MRI images used for supervised learning of characteristic anatomical structures contain structures that

differ in scale and orientation while edge detection performed on these images produces fragmented structure boundaries due to low image contrast and noise. Despite these problems and the inconsistency of interest point detection the supervised learning of anatomical structures in MRI images produced structure models that resulted in correct detection of more than 80% of corresponding structures in the validation data set.

Future research will concentrate on the replacement of symmetry based interest points with edge key points corresponding to high curvature locations along detected edges. These key points are significantly less exposed to the interest point drift affecting symmetry interest points and blob detectors. An additional advantage of using these key-points is their ability to estimate the descriptor orientation from a local edge orientation, thereby reducing the search for relative orientation between two descriptors and overall computational complexity. Finally the descriptor distance will be altered to make it affine invariant with the ability to control the amount of affine transformation allowed.

Acknowledgements

This work was partly supported by the European Union Network of Excellence MUSCLE (FP6-507752) and the European Union project GRASP (FP7-215821).

References

1. Donner, R., Micusik, B., Langs, G., Szumilas, L., Peloschek, P., Friedrich, K., Bischof, H.: Object localization based on markov random fields and symmetry interest points. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part II. LNCS, vol. 4792, pp. 460–468. Springer, Heidelberg (2007)
2. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley Interscience, Hoboken (2000)
3. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: ECCV 2004 Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, pp. 17–32 (2004)
4. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60(1), 63–86 (2004)
5. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)