The Multi Angular Descriptor: new binary and gray images descriptor for shape recognition

Raid Saabni

Faculty of Engineering, Tel-Aviv University, Tel-Aviv, Israel
Triangle R&D Center, Kafr Qarea, Israel
saabni@cs.bgu.ac.il

Alex Bronstein

Faculty of Engineering, Tel-Aviv University, Tel-Aviv, Israel
bron@eng.tau.ac.il

Abstract

In this paper we present a new shape descriptor, The Multi Angular Descriptor (MAD), for shape based object recognition. In the binary case, from each contour point, the Multi Angular Descriptor captures the angular view to multi resolution rings in different heights. In the gray level case, it captures the weighted distribution over relative positions of the shape points to multi resolution rings around the centroid. The multi angular descriptor is robust to noise and small deformations and have very flexible variables which can be tuned for different tasks. The extension of the (MAD) descriptor to the gray level case can be seen as an extension of the shape context to gray level images which enables dealing with low quality images. Testing the proposed descriptor on the MNIST dataset [1] and a private dataset using two matching techniques gave better results comparing to the Shapes Context and the Histogram of Oriented Gradients (HOG).

Key words: Shape Descriptor; Shape Context; Histogram of Oriented

Gradients (HOG);Handwriting Recognition; Shape Matching; Angular Radial Transform.

1. Introduction

Comparing and matching Objects is one of the common problems frequently targeted in most computer vision applications. Generally, image recognition and retrieval systems aim to match shapes of objects in the real world to each other, using object models which are mostly represented in some feature space. It is difficult to specify what is a proper description of a shape, since shapes many times are considered to be similar despite variation in size, orientation and boundary. Many times simple geometric features are used to describe shapes for matching, but mostly such features can only discriminate shapes with large differences. This kind of simple features are usually used as filters to eliminate false hits or as a part of features combinations. Therefore, richer descriptors such as SIFT, Shape Context, Moments and others were developed and presented to enable better matching and similarity measurement for shape matching and retrieval. In binary images, contour-based methods based on the use of shape boundary points are the most common and general features for recognition and classification of shapes. Tangent Angles, Contour Curvature, Chain Codes and Shape Context are part of the existing descriptors extracted from the shape's contours. In gray level Images, descriptors based on gradients such as the Scale Invariants Feature Transform (SIFT) and the Histogram of Oriented Gradients (HOG) are commonly used.

The 'Shape contexts' descriptor presented by Belongie and Malik *et al.* [2], is a boundary based descriptor which describes a distribution of all boundary points with respect to each point on the boundary. The Shape Context descriptor computes the histogram of relative polar coordinates of any single boundary point and have been proved to be an efficient feature for matching binary images. There is no direct extension of the Shape Context to gray level images considering it's dependency with a clear and ordered boundaries of the matched components. In [3, 4], a novel approach for image representation based on geometric distribution of edge pixels was presented, where the edge map have been divided into MXN Angular Radial partitions and extracted local features for these partitions. The entire image is then described as a set of spatially distributed invariant feature descriptors using the magnitude of the Fourier transform. The approach is scale and rotation invariant and tolerates small translations and erosions.

In this paper, we present a shape descriptor based on angles taken from different view points from rings around the shape centroid with different sizes and heights. The independence of the rings with boundary points, enables extending the descriptor to gray level images and multiple disconnected components. In the presented approach, the shape is treated as a two dimensional set of points and the different rings are upper view points from different sizes and heights. Sizes and heights of these rings are calculated using the diameter and centroid of the shape to enables scale and translation invariance. The presented descriptor can be applied to edge maps images and gray level images, and not only to contours as in the shape context. Generally, the presented descriptor can be thought of, as merging the two descriptors (Shape Context and Angular Radial Transform) and extending both of them.

The rest of this paper is organized as follows: in Section 2 we briefly overview some of the closely related work of shape descriptors used in handwriting word matching and image retrieval. In Section 3, we describe our shape descriptor in details. Experimental results are presented in Sections 4.

2. Related Work

Matching shapes is very important for shape classification and image retrieval, therefore, shape descriptors play a major rule in Document Image Analyses such as in character and handwriting recognition, symbol and logo recognition, or generally speaking shape recognition and matching. In the literature, we can find several surveys summarizing advances in shape descriptors either in the context of shape analysis [5] or in the more general context of computer vision and pattern recognition [5–8]. Different taxonomies of shape descriptors according to different points of view have been presented trying to make some order in this wide field. T.Pavlidis[7], divides the shape descriptors in several binary classes: external and internal algorithms; scalar and domain transforms; and information preserving and information non-preserving methods. Mehtre *et al.*[6], classified shape descriptors as boundary based methods and region based methods. Zhang *et al.*[8], differentiate between contour and region based descriptors but they simplify the classification by only differentiating between structural and global descriptors. Finally, Trier *et al.*[5] introduced another point of view distinguishing among features extracted from binary images and gray-scale images. Another taxonomy divides them to appearance-based models, where gray or color values of images are directly used to measure similarity, and feature based methods which use characteristics and descriptors of the target objects.

In general, successful description of a shape should contains sufficient information to gather similar and distinguish between different target objects. These methods can be divided into two categories, the area-based methods and the boundary-based methods. Simple descriptors, for example perimeter length, curvature, and bending energy, have been applied widely but proved to be efficient only as part of a feature set or for eliminating far candidates. Shapes of the same object can be defined as the equivalence class under a group of transformations mostly including scale, translation and small distortions. Shape classifying in such case is belonging a given shape to it's equivalence class using shape distance measurement. Appearance based method makes a direct use of gray values within the visible portion of the objects, where feature based methods focus on the shape geometry. The appearance information is used to find the correspondences and to align the gray scale values to compare brightness of two different shapes.

The first group of work, on appearance based recognition, makes direct use of pixel brightness values as presented in [9]. Several other approaches in this vein [10, 11], first attempt to find correspondences between the two images before doing the comparison. As an alternative, there are a number of methods that build classifiers without explicitly finding correspondences. In such approaches, one relies on a learning algorithm having enough examples to acquire the appropriate invariance. These approaches have been used for handwritten digit recognition [12, 13], face recognition [14], and isolated 3D object recognition [15]. In contrast, techniques that perform recognition based on shape information attempt to capture global structure of extracted edge or silhouette features. Silhouettes have been described (and compared) using Fourier descriptors [16], skeletons derived using Blums medial axis transform [17], or directly matched using dynamic programming. Other approaches [18–20] treat the shape as a set of points in the 2D image, extracted using, say, an edge detector. Another set of methods compute correspondences between edge points, such as the work of Carlsson [21], which uses order structure, and the work of Johnson and Hebert [22] and Chui and Rangarajan [23]. Recent years have seen the emergence of hybrid approaches [24–27] that capture appearance information through a collection of local image patches. Shape information is encoded via spatial relationships between the local patches. The locations for the local patches are selected with various interest point operators, and are represented either as raw pixel values [25] or histograms of image gradients [24, 26], termed SIFT descriptors (Scale Invariant Feature Transform).

Belongie and Malik [2], presented the "'Shape context"' feature descriptor describing shapes in a way that allows for measuring shape similarity and the recovering of point correspondences. The basic idea of shape contexts is illustrated in Figure 1. A shape is represented by a discrete set $Ps = p_1, ..., p_n$ of n points sampled from the internal and external contours on the shape. the $n-1$ vectors which are the set of vectors originating from a point to all other sample is a rich but too detailed description of the shape. Therefore, the authoress suggested to capture this information as the distribution of the relative positions of the remaining $n-1$ points in a spatial histogram, See Equation 1. This histogram is defined to be the shape context of the origin point p_i.

$$h_i(k)) = \{q \neq p_i : q - p_i \in bin(k)\} \quad (1)$$

They have used bins that are uniform in log-polar space, making the descriptor more sensitive to positions of nearby sample points than to those of points farther away. The shape context of a point on a shape are made invariant under uniform scaling of the shape as a whole by normalizing all radial distances by the mean distance of all different pairs of points. Shape contexts are empirically demonstrated to be robust to deformations, noise, and outliers.

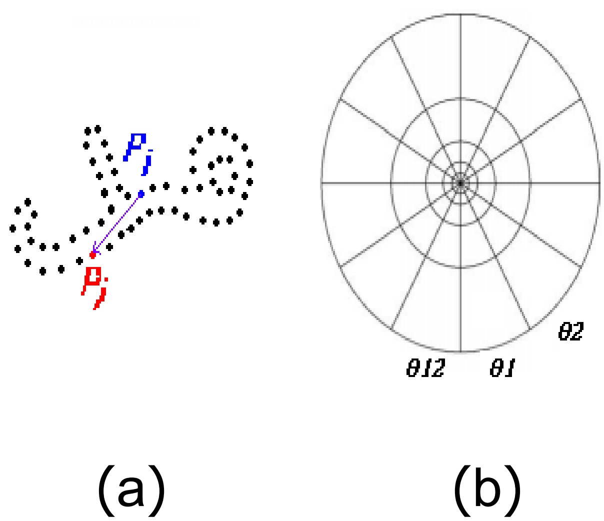

Figure 1: (a) The outer contour of the main component of an Arabic word-part. (b), five bins are used for $\log r$ and 12 bins for the angle θ.

The Angular Radial Transform (ART) *et al.*[3, 4], is a novel approach for image representation based on geometric distribution of edge pixels. Not like shape context, input image may consist of several complex objects. It is a moment-based image description method adopted in MPEG-7 as a region-based shape descriptor. The ART is a complex orthogonal unitary transform defined on a unit disk that consists of the complete orthogonal sinusoidal basis functions in polar coordinates. It starts by converting the color image to a gray intensity image by eliminating the hue and saturation while retaining the luminance. IN the next step the algorithm normalizes the gray image to the size $201X201$ followed by applying the Canny edge operator for edge detection. The normalized edge image I, is employed for feature extraction where edge pixels are considered as '1' and '0' otherwise. In the following, we consider pixels $I(\rho,\theta)$ to be either equal to 1 for edge pixels or 0 for non-edge pixels. The algorithm uses the surrounding circle of I for partitioning it to MXN sectors, where M is the number of radial partitions and N is the number of angular partitions. (see Fig. 1). The number of edge points in each sector of I is chosen to represent the sector feature. The scale invariant image feature $f(k,i)$ is then defined to be:

$$f(k,i) = \sum_{\rho=KR/M}^{(K+1)R/M} \sum_{\theta=i2\pi}^{((i+1)2\pi)/N} I(\rho,\theta) \tag{2}$$

3. Our Approach

In this paper, we present a translation and scale invariant descriptor. Orientation invariance can be also achieved by using the Fourier Transform as in (ART) or measuring angles at each point relative to the direction of the tangent at that point as in Shapes Context. In our case we prefer to avoid significant orientation invariance since word matching and recognition does not tolerate high variance in the orientation of compared shapes, see the case of digits $'6'$ and $'9'$.

Given a binary image I of a connected component(CC), We start by calculating the centroid C and the diameter D of the image I. The calculated values of C and D are used to determine and draw a set of rings centered by C with different radius values which are derived from the diameter D. We treat the rings as lying on different heights above the given shape where larger rings overlay closer to the shape. In the next step, we treat each ring as a set of k points taken uniformly distant from each other. Each point in each ring serves as an upper view point watching each pixel (contour point) in the shape.

The main idea of the presented descriptor is to generate a sequential concatenation of upper view points from different heights and resolution to the 2-D shape. The multi resolution and heights will enable capturing more information in different resolutions and by that enabling a local and semi-global description of the given shape. In the following subsections we will give in more detail a full description of the proposed descriptor for binary images followed by the extension to gray level images.

3.1. Upper View Points and Angle Descriptors

Let I be a binary image with the size nXm including one Connected Component(CC), and Let C and D be the centroid and the diameter of the shape respectively. Let $P = \{P_i\}_{i=1}^{l}$ a set of l point taken uniformly from the extracted contour of the CC. Given a view point V_j from a given ring with height h over the shape, the $l-coordinate$ vector of angles, obtained by connecting the point V_j with each point $P_i \in P$ and the plain of the shape is a rich description of the shape from this view point. As expected, one view point have some limitations, therefore, the key idea is to give additional view points from different directions, therefore, additional points of view will enable richer and more accurate description of the shape. Lifting the points of view up to different heights from the $2-D$ shape gives additional views to the shape, avoids intersection with segments of the shape, and integrates the distances into the angle value. In our case we pick the different view points with the same height to lay on the same ring on a plane which is parallel to the plane of the $2-D$ shape. The view points are taken uniformly from the given ring and their number can be tuned as a derivative of the recognition task and the given shape. Depending on the classification algorithm and the recognition task, we can determine the number of layers and view points to enable different options of multi resolution look at the shape.

Formally, Let R be a ring with the radius r and the center C positioned above the shape S with the height h. let $V = \{V_i\}_{i=1}^{n}$ be a set of n view points lying uniformly on the ring R. We define $\alpha(V_{ij})$ to be the angle between the segment $\overline{V_i, p_j}$ and the plain contains the shape S, see Figure 2. We define the vector Vec_i to be $Vec_i = \{\alpha(V_{ij})\}_{j=1}^{l}$. The vector Vec_i can be seen as watching the shape S from one upper view point, the point V_i.

We define L, a ring layer description of the shape S as a triple (r, n, h) representing a ring R with radius r and overlaid parallel to the shape plain with the height h. The ring will be represented by n points taken uniformly distant from each other on R. This layer as a feature set will be representing

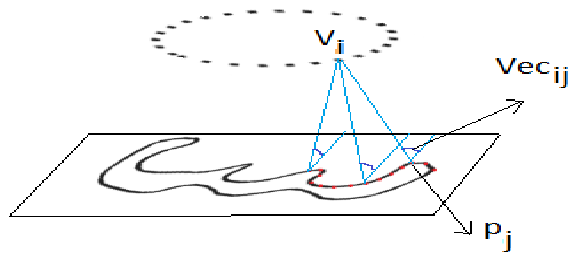

Figure 2: In this figure we can see an example of three line segments drawn from the same view point V_i, generating the three angles Vec_{ij} with the plane of the shape. When the parameter j goes over all contour points we get the vector Vec_i describing the shape from the view point V_i. With the parameter i goes over all view points.

the set of n feature vectors, Vec_i, each represents the angles from one specific view point V_i in the the ring to all points of the shape contour having the same order as the given contour.

A formal definition $FV(S)$ of the layer L as a feature vector describing the shape S will be the vector generated by concatenating all the vectors $\{Vec_i\}_{i=1}^{n}$ taken from all view points of the layer L.

In some cases, a richer shape feature set may be needed. In such cases we may use the descriptor generated using different layers with different parameters representing different resolutions enabled by varying height and radius values of the rings. Experimental results show that the number of layers and parameters values are derived from the task and the classifier used for matching shapes.

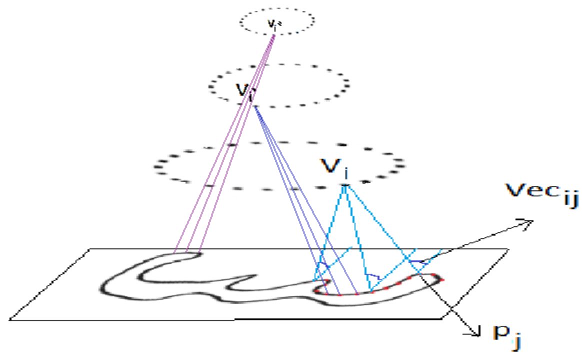

Figure 3: In this figure we can see a multi resolutin view to the shape. Three different view point from three different rings are drawn to calculate angles to the contour points of the shape.

3.2. Extension to Gray Level Images and Multi Components Shapes

Gray level images with low quality where binarization results does not guarantee consistently the number and the quality of resulted components poses real limitation for using the Shape Context descriptor. The presented descriptor uses view points taken from rings out of the shape and therefore not significantly affected by the results of the binarization step. In such cases, limited results of the binarization process can still serve to determine the values of D,H and C as in the binary case. From the other hand, since the contour results are not guaranteed, we modify the presented descriptor to captures the weighted distribution over relative positions of the shape pixels. Multi resolution rings around the centroid can still be used for richer descriptions. In this case each feature descriptor from a view point is a log-polar weighted histogram of the coordinates of each pixel in the shape. The gray level value of pixels are used as weights, and the reference points are the view points. In low quality images, one may emphasize edges using *Sobel* or *Gaussian* low pass edge-emphasizing filters for better performance. In our case we have used the *Sobel* edge-emphasizing filter on the gray level image. A predefined low value have been used to binarize only the back ground of the image and later on calculate the values C and D for the centroid and Diameter respectively. The ring centered on C with the diameter D have been used to generate n view points taken uniformly distant from the ring.

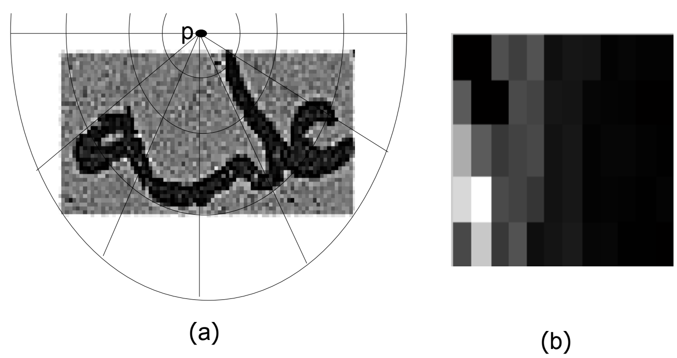

Figure 4: In the left side, we see the different bins from one view point to a gray level image. in the right the histogram from the reference point p.

Let I be a gray level image with the size n X m. We start with a pre-processing step where edge emphasizing filter is used and all high probability back ground pixels are turned to zero. In the next step we calculate the centroid and the diameter of the preprocessed shape. We use the values D and C to draw a circle with the center C and diameter D. n points taken uniformly distant on the ring are used as view points to generate a vector of weighted distribution of pixels on relative area. The number of pixels in each area weighted by the gray level of each pixel is used to generate a log-polar histogram of the shape gray value pixels measured using the view points on the ring as reference points. In our case we have used 5 and 12 bins for $logr$ and θ respectively. This histogram is defined to be the histogram of the view point V_i on the given ring. The vector with size n where each coordinate is the histogram $H(V_i)$ of the view point V_i is a feature vector describing the given shape I. This definition can be seen as an extension of the known

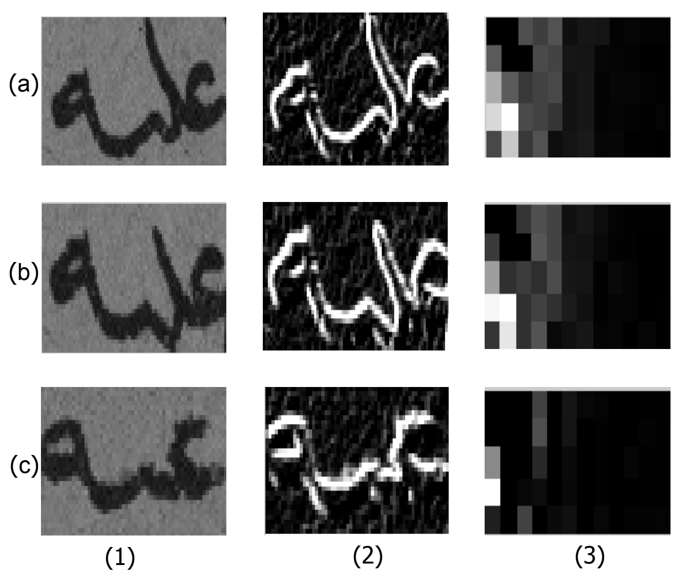

Figure 5: Three gray level images of Arabic word-parts are presented in three rows. In column (1), we can see the original image, in (2), the image manipulated with edge emphasizing filter and in the third (3), we can see the Histogram of the shape using our descriptor from the same point of view. as can be seen in the the third column, Histograms of similar shapes are similar from the same view point

shape context. In this definition we use the circle with it's view points to replace contour points which is a robust representation of low quality shapes. To get the maximum benefit of the gray level information we use the gray level values of each pixel to calculate a weighted distribution of the pixels. As expected, high values which are foreground pixels with high probability, contributes more value to the histogram in the related area.

4. Experimental results

We have used two known descriptors to compare results with the presented one. The Shape context descriptor was used for the binary images and the Histogram of Oriented Gradients (HOG) for gray Level images. We have used handwriting digits and Arabic word-parts recognition tasks to test and compare our proposed descriptor. The well known MNIST database have been used for digit recognition while a private database including $20,000$ word-parts, merged with a modified version of part of the off-line Arabic database IFN/ENIT [28] was used to train and evaluate an Arabic handwriting word-parts recognition systems. The MNIST dataset, is derived from the NIST dataset, and has been created by Yann LeCun [1]. The MNIST dataset consists of handwritten digit images. The examples for training include $60,000$ examples for all digits and $10,000$ examples for testing. All digit images in this dataset have been size-normalized and centered in a fixed size image of $28x28$ pixels. In the original dataset each pixel of the image is represented by a value between $0and255$, where 0 is black, 255 is white and anything in between is a different shade of gray. The $IFN/ENIT$ database has been slightly modified to work with the main part of word-parts as connected components, i.e, split components for single word-parts have been rejoined to single one. Touching components have been split to the word-parts they represent. The database has been reorganized as a list of sets each containing multiple shape of a word-part. The manually modified version of the $IFN/ENIT$ database includes $10,000$ different shapes of 400 word-parts. Additionally, a private set of $20,000$ images of the same 400 word-parts have been collected using 50 students. To test the (MAD) descriptor with gray level Images we have used 40 pages of historical documents taken from from the Juma'a Al-majid Center in Dubai. These pages were segmented to words and to word part and labeled. A Hierarchical clustering process using the DTW with each descriptor as the metric distance between each two images. resulted clusters where used to measure precision and recall where right clustering of a word/word part was considered as right positive and false negative other wise.

To compare results, we have used the basic DTW algorithm and the Hungarian algorithm. The Hungarian method was used in [2] to find the correspondence between the two sets of points in the matched shape. The shape context descriptor of each point was used to measure there similarity between each pair of points. While the Hungarian method is used to find the correspondence, the DTW is used to measure the similarity between the two shapes without finding the correspondence between the two points.

Results in Tabel 1, show that the presented descriptor outperforms the results using the same system with the Shape Context for binary images and the HOG for gray level Images.

Using the fact that the MNIST dataset includes close to 6000 different

MNIST DATA SET				
	Shape Context Descriptor		MAD Descriptor	
	precision	Recall	precision	Recall
DTW	93.6%	91.8%	95.8%	91.1%
HUM	96.15%	95.17%	96.81%	95.6%
Private DATA SET				
	Shape Context Descriptor		MAD Descriptor	
	precision	Recall	precision	Recall
DTW	81.2%	82.8%	84.2%	82.1%
HUM	78.9%	81.8%	78.8%	79.1%
Gray Level Dataset				
	HOG Descriptor		MAD Descriptor	
	precision	Recall	precision	Recall
DTW	71.2%	73.8%	80.3%	82.4%
HUC	70.1%	71.1%	76.5%	77.2%

Table 1: Precision and Recall results of the three compared systems (MNIST, Private dataset and Gray Level Dataset), in terms of counting false positives and false negatives. In the second system 20% the shapes of each word-part were extracted out to be searched for. In the last system, we have used the results of heirarchial clustering to evaluate the system performance.

shapes of each digit we have tested the recognition ability of the compared descriptors with different sizes of the sets of template. We have generated eight sets with different sizes. The first set included 100 samples, the second 500 and the rest are 1000,2000, 3000,4000,5000 and the 8'th included all the samples. We have used the DTW method to test the recognition rates of the two descriptors with the different sets of samples. Results in Figure 6, show that the presented descriptor outperforms the Shape Context when the number of templates is small, and becomes closer to it when the number of the presented samples for comparing is large.

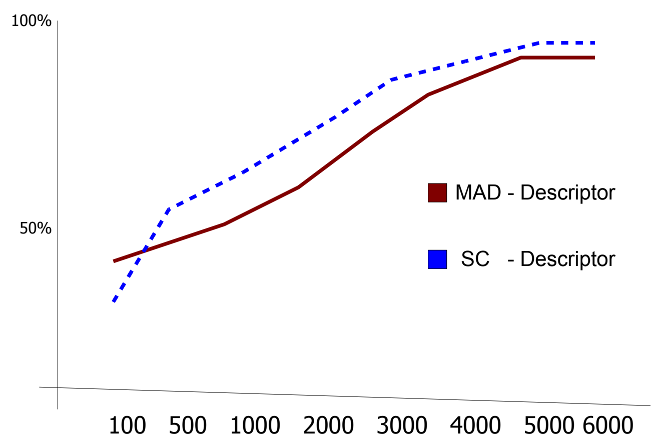

Figure 6: As seen in this figure, Results show that the MAD descriptors recognition ability outperforms the Shape context descriptor especially when the number of samples for templates and training are small, but continues to slightly outperform it even with large number of samples.

5. Acknowledgments

We would like to thank Prof’ D. Ron from Tel-Aviv University for her insightful comments which led to several improvements in the presentation of this paper. This work has been supported in part by the ’Vatat and Pikovsky Valazzi Post Doc scholarship’ and the DFG-Trilateral Grant No.8716.

References

[1] Y. LeCun., The mnist database of handwritten digits., http://www.research.att.com/ yann/exdb/mnist, AT&T Labs (1994).

[2] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Analysis and Machine Intelligence 24 (2002) 509–522.

[3] M. Bober, Mpeg-7 visual shape descriptors, IEEE Trans. Cercuit Sys. 1(6).

[4] W. Kim., Y. S. Kim., A net region based shape descriptor, Tech. rep., Pisa (1999).

[5] ivind Due Trier, A. K. Jain, T. Taxt, Feature extraction methods for character recognition - a survey, Pattern Recognition 29(4) (1996) 641662.

[6] B. M. Mehtre, M. S. Kankanhalli, . W. F. Lee., 33(3):319337, Shape measures for content based image retrieval: a comparison, Information Processing and Management 33(3) (1997) 319–337.

[7] T. Pavlidis., Survey: A review of algorithms for shape analysis., Computer Graphics and Image Processing 7(7) (1978) 243–258.

[8] D. Zhang, G. Lu., Review of shape representation and description techniques, Pattern Recognition 37 (2004.) 119.

[9] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognitive Neuroscience 3 (1991) 71–96.

[10] M. Lades, C. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, W. Konen, Distortion invariant object recognition in the dynamic link architecture, IEEE Trans. Computers 42 (1993) 300311.

[11] T. Cootes, D. Cooper, C. Taylor, J. Graham, Active shape models - their training and application, Computer Vision and Image Understanding 61 (1995) 38–95.

[12] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, vol. 86, no. 11, pp. 22782324, 1998.

[13] C. Burges, B. Scholkopf, Improving the accuracy and speed of support vector machines, in nips, 1997, pp. 375381.

[14] B. Moghaddam, T. Jebara, A. Pentland, bayesian face recognition, pattern recognition, vol. 33, no. 11, pp. 17711782, november 2000.

[15] H. Murase, S. Nayar, Visual learning and recognition of 3-d objects from appearance, int. journal of computer vision, vol. 14, no. 1, pp. 524, jan. 1995.

[16] C. Zahn, R. Roskies, Fourier descriptors for plane closed curves, ieee trans. computers, vol. 21, no. 3, pp. 269281,march 1972.

[17] D. Sharvit, J. Chan, H. Tek, B. Kimia, Symmetry-based indexing of image databases, j. visual communication and image representation, 1998.

[18] G. Borgefors, Hierarchical chamfer matching: A parametric edge matching algorithm, ieee trans. pami, vol. 10, no. 6, pp. 849865, 1988.

[19] D. Huttenlocher, R. Lilien, C. Olson, View-based recognition using an eigenspace approximation to the hausdorff measure, pami, vol. 21, no. 9, pp. 951955, sept. 1999.

[20] D. Gavrila, V. Philomin, Real-time object detection for smart vehicles, in proc. 7th iccv, 1999, pp. 8793.

[21] S. Carlsson, Order structure, correspondence and shape based categories, in shape contour and grouping in computer vision. springer lncs 1681, 1999, pp. 5871.

[22] A. E. Johnson, M. Hebert, Recognizing objects by matching oriented points, in cvpr, 1997, pp. 684689.

[23] H. Chui, A. Rangarajan, A new algorithm for non-rigid point matching, in cvpr, vol. 2, june 2000, pp. 4451.

[24] D. G. Lowe, Distinctive image features from scale-invariant keypoints, int. journal of computer vision, vol. 60, no. 2, pp. 91110, 2004.

[25] R. Fergus, P. Perona, A. Zisserman, Object class recognition by unsupervised scale-invariant learning, in proceedings of the ieee conference on computer vision and pattern recognition, vol. 2, 2003, pp. 264271.

[26] G. Dorko, C. Schmid, Selection of scale invariant neighborhoods for object class recognition, in proc. 9th int. conf. computer vision, 2003, pp. 634640.

[27] Y. Amit, D. Geman, K. Wilder, Joint induction of shape features and

tree classiers, ieee trans. pami, vol. 19, no. 11, pp. 13001305, november 1997.

[28] M. Pechwitz, S. S. Maddouri, V. Mrgner, N. Ellouze, H. Amiri, Ifn/enit - database of handwritten arabic words, in: In Proc. of CIFED 2002, 2002, pp. 129–136.