TESPAR-BASED 2D SHAPE RECOGNITION

Corneliu FLOREA[1], Laura FLOREA[2], Constantin VERTAN[3]

In this paper we introduce a novel shape descriptor based on a three step procedure. The first step aims to offer invariance to rotation. The second step decreases the dimensionality from 2D to a pair of 1D waveforms by using either image projections, or contour tracing, which in the third step are encoded with the TESPAR (Time-Encoded Signal Processing and Recognition) procedure. The proposed shape encoding preserves the most important features of shapes, being suitable for posterior clustering as a filtering step. The method is tested on the MPEG-7 database, providing promising results in both accuracy and execution time.

Keywords: shape recognition; TESPAR; rotation invariance; principal component analysis

1. Introduction

The analysis and understanding of visual shapes is one of the most important topics in pattern recognition. A shape is an object's silhouette that is typically retrieved after a segmentation step, has a significance in the context of the given problem to be solved by image processing and analysis and, thus, is the typical result of turning the multi-planar/grayscale image into a binary image. Notable examples of this model are graphical signs, letters or musical signs in image based document analysis and recognition [1].

Beyond the recognition of printed or handwritten signs, letters or musical signs, the shape identification, recognition, and classification is necessary in a plethora of problems. For instance, in high dynamic range, the scene may be decomposed in frameworks (i.e. shapes) [2], which may be further individually described, being a reliable tool for scene understanding and thus adaptive tone compression. Another example is the face analysis, where the segmented mouth or the eye silhouettes are indicators of the expression dynamics [3], [4].

Since the 2D objects are projections of 3D objects, their 2D silhouettes may change [5] due to:

[1] Associate Professor, The Image Processing and Analysis Laboratory, LAPI, University POLITEHNICA of Bucharest, Romania, e-mail: corneliu.florea@upb.ro

[2] Lecturer, The Image Processing and Analysis Laboratory, LAPI, University POLITEHNICA of Bucharest, Romania, e-mail: laura.florea@upb.ro

[3] Professor, The Image Processing and Analysis Laboratory, LAPI, University POLITEHNICA of Bucharest, Romania, e-mail: constantin.vertan@upb.ro

- The variability of the view point with respect to objects. This may lead to object scaling, rotation, warping.
- Non-rigid object motion (e.g., people walking or fish swimming). In such a case the object is articulated.
- Noise due to digitization or flaws in the segmentation process.

Traditionally, descriptors of shape silhouette may fall in one of the following categories: *contour-based* methods and *region-based* methods [6]. Each class may be further divided into structural approaches (segments/sections primitives) and global approaches (i.e. the shape is represented as a whole). Taking into account that shapes may be represented into the original spatial domain or in the transform domain, based on whether the shape features are derived from the spatial domain or from the transformed domain, this provides a supplemental degree for classification.

Common, simple, global descriptors are the area, the compacticity ratio (perimeter^2/area), the eccentricity (ratio of the length of major axis to the length of minor axis), major axis orientation, and bending energy [7]. These simple global descriptors are able to discriminate shapes within large categories, thus, they are pre-filters that eliminate false positives and can be used as an initial separation procedure.

The more recent shape descriptors approaches aim at robustness, employing representations that are invariant (or insensitive) to certain groups of deformations. For example, the popular scale invariant feature transform (SIFT) [8] selects scale invariant blob regions for reliable key point identification. The use of the statistical moments works toward invariance to similarity transforms. Next, Ling and Jacobs [9], showed that robustness to object articulation is achieved by using the inner-distance.

Yet, we must point out that the more recent methods, such as the *aspect space* [10] while being highly accurate are also very complex, going up to $O(N^3)$ operations (where N is the number of pixels forming the shape). Under such circumstances, there is need for simpler and computationally efficient methods, especially since silhouette recognition acts as an intermediate step of more complex systems. In this paper we propose a simple binary shape / silhouette descriptor extracted in three steps: rotation normalization, dimensionality reduction to a pair of 1D waveforms and TESPAR encoding. The main contribution is the introduction of TESPAR-based shape descriptors. The proposed system is tested on the MPEG-7 silhouette database.

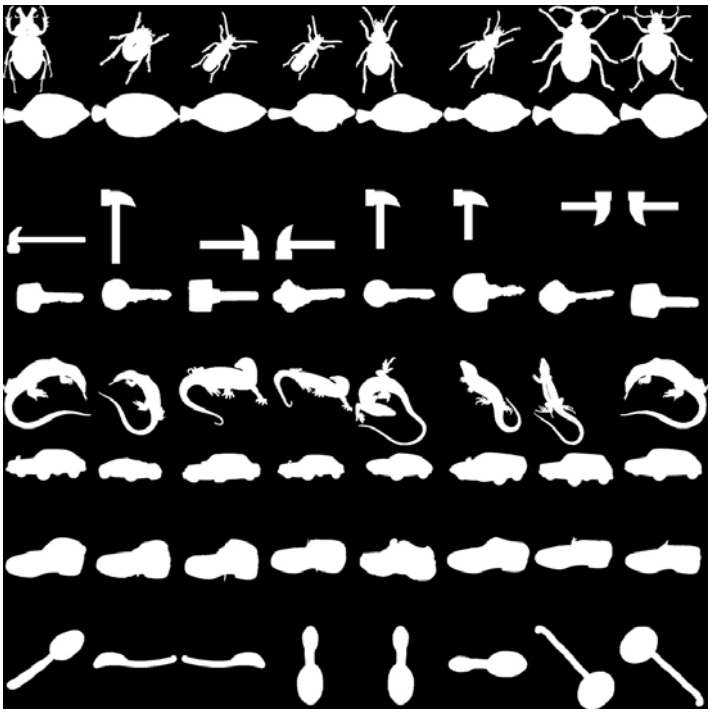

Fig 1. An example showing 8 typical classes of objects from the MPEG-7 database. Silhouettes from the same class are one the same line; note the within- and the inter-class variability.

2. MPEG7-CE-Shape-1 Database

The proposed framework and shape descriptors are tested for the shape recognition task on the widely used MPEG-7 CE-Shape-1 part B database [5]. The database contains 1400 silhouette images grouped within 70 classes, where each class contains 20 images different shapes (some examples are shown in Fig. 1). The shapes varies by articulation, rotation, perspective, etc.

The performance of different solutions is measured by the so called Bull's Eye score: every shape in the database is submitted as a query and the number of shapes from the same class from the most similar 40 returned results is counted. The Bull's Eye score is then defined as the ratio of the number of correct hits to the best possible number of hits (which is 20 x 1400). The Bull's Eye metric evaluates the capacity of a descriptor to preserves objects from the same class as compact clusters in larger sets of data.

3. Shape description

The proposed object descriptor consists in three steps: rotation normalization, shape regression and TESPAR encoding. The step of shape regression refers to the reduction of the two dimensional form to a set of 2 linear vectors. Two methods are envisages for this case, methods that are based on integral projections or methods that are based on contour tracing. The resulting (one-dimensional) waveforms are encoded with TESPAR for constructing the shape final descriptor.

3.1. Rotation normalization

As one may notice in fig. 1, the data from each class is originally presented with various rotations. Taking into account that the proposed descriptors are not invariant to rotation, to gain robustness with respect to this category of variation, a dedicated strategy must be employed.

The main idea is to separate shapes into elongated and compact categories. In the elongated shape case, the main axis gives the shape inclination. For the compact shapes, the angle of the major axis is subject to interference from articulation and a different measure should be provided. Here the relative angular position of the weighting center of mass of the shape with respect to the geometrical center is considered.

The actual algorithm that is applied to any shape, consists of the following steps :

1. Consider the shape as a set of pixels. Each pixel is a point into a 2D space, represented as a pair of *(row, column)* coordinates. Thus, the shape is the particular output of a 2D random variable.
2. Given the shape (random variable), compute the covariance matrix. Next, determine the eigenvectors (major axis and minor axis) and eigenvalues (the variance on each eigenaxis).
3. Compute the *eccentricity* as the eigenvalues ratio. If this is larger than a threshold (i.e. 2), then the shape is elongated (e.g. the spoons or the hammers from fig. 1). For elongated shapes, we rotate the shape such that the major axis is horizontal.
4. If the shape is not elongated (the eigenvalue ratio is smaller than the threshold), it is compact (e.g. fish in fig. 1) and the shape is rotated such that the weighted center of mass of the shape is directly above the shape geometrical center.

3.2 Shape regression to a pair of 1D waveforms

Integral projections. For an image I(x,y), x=1,…M, and y=1,…,N, the horizontal and vertical integral projections functions are defined as:

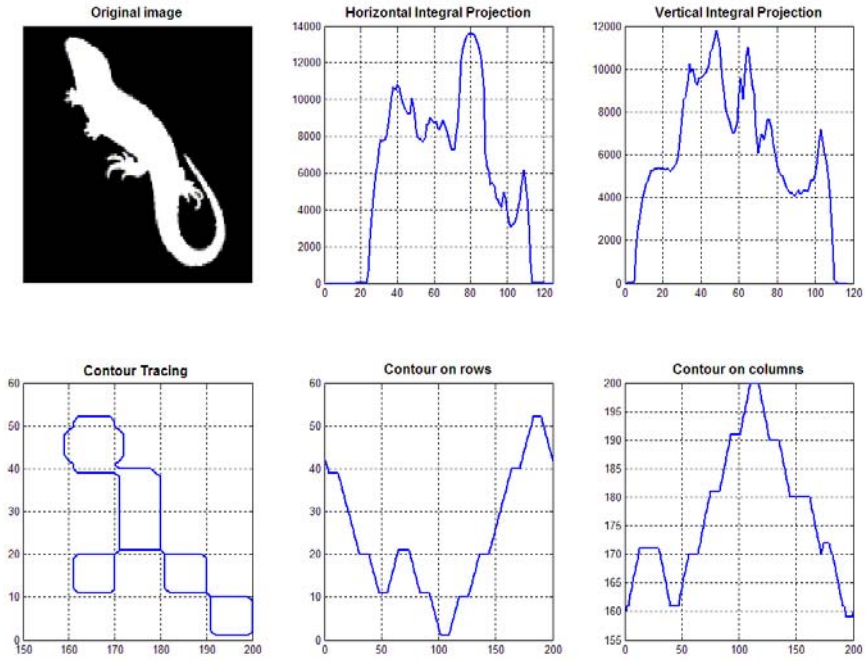

Fig. 3 Original image and the two resulting integral projection functions (top row) and respectively the contour waves as computed by Moore neighbor tracing algorithm

$$I_H(y) = \sum_{y=1}^{N} I(x,y) \tag{1}$$

and respectively, the vertical one:

$$I_V(y) = \sum_{x=1}^{M} I(x,y) \tag{2}$$

The dimensionality reduction using integral projections functions is illustrated in Fig. 3, top row. The integral projection shows the distribution of the shape over rows and respectively columns.

Contour tracing. The alternative choice to projections is to use the contour tracing, technique that is also known as *border following* or *boundary following*; it

is applied to objects within digital images in order to extract their boundary. In fact, given a 2D shape the method will return a pair of waveforms: the row coordinates of the contour and respectively the column coordinates. We used the Moore-Neighbor Tracing algorithm [11] and an illustration of it is presented in Fig. 3, bottom row.

The Moore neighborhood of a pixel, P, is the set of 8 pixels which share a vertex or edge with that pixel. The main contour tracing algorithm steps are:

- The starting point is a white pixel (in shape), having in its neighborhood black pixels.
- Extract the contour by going around the pattern in a clockwise direction.
- Every time we hit a white pixel, P, backtrack (i.e. go back to the black pixel on which we stand), then, go around the pixel P in a clockwise direction, visiting each pixel in its Moore neighborhood, until we hit a white pixel.
- The algorithm terminates when the start pixel is visited for a second time (i.e. the shape was surrounded).

In this case the two waveforms are the contour rows and columns with respect to contour perimeter.

3.3. TESPAR encoding

TESPAR (Time-Encoded Signal Processing and Recognition) encoding was introduced by King and Phipps [12] as a technique for representation and recognition of 1D, band-limited speech signals. It has been used, for instance, in speaker recognition [13] and was used in image analysis as part of a system for for iris center localization [14]. The TESPAR encoding is based on the determination of zero-crossings of the signal. The interval between two consecutive zero-crossings is called an *epoch*. While in the original TESPAR paper [11], an epoch was described by two parameters (the duration and the shape), inspired by [14], we consider here three parameters (as presented in Fig. 4) and we replace the zero-crossing by the intersection with a key-level:

- Duration - the number of samples in the epoch;
- Amplitude - the maximal signed deviation of the signal with respect to the key-level;
- Shape - the number of local extremes in the epoch.

Depending on the problem specifics, additional parameters of the epochs may be considered, such as the difference between the highest and the lowest

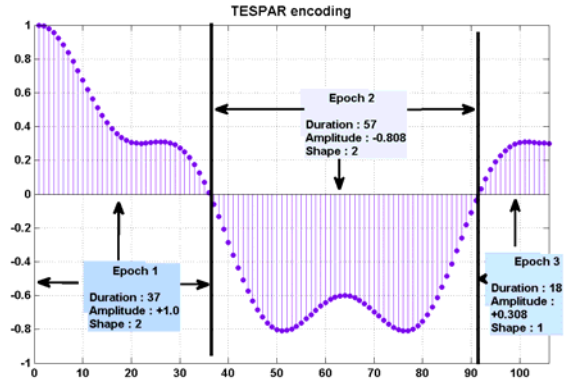

Fig 4. TESPAR Encoding of a given signal.
The associated code is: [37;+1;2; 57;-0.808;2; 18;0.308,1].

mode from the given epoch. Further extensions are at hand if an epoch is considered an approximation of a probability density function and the extracted parameters are the statistical moments of the said distribution.

TESPAR was developed for 1D signal encoding and its authors [12] recommend it as a reliable encoding technique for low, limited band signals. Indeed, the projections functions and the contour waveforms are low bandwidth signals. In both cases epoch select specific part of the shape.

Features. Given a rotation normalized silhouette, its descriptor is formed by the concatenating the TESPAR encoding of the two waveforms: horizontal and vertical projections functions and contour rows and respectively columns in the case of tracing.

4. Results

As said, the proposed descriptor is tested on MPEG-7 database with Bull's Eye objective score as reference for the accuracy.

4.1 Metrics used

Given a shape, in the process of determining the Bull's Eye score distance to all other shapes must be found. There exist multiple possibilities to compute such distance. Considering that two descriptors are $X=[x_1, x_2,\ldots x_P]$ and respectively $Y=[y_1, y_2,\ldots y_P]$, we considered the following metrics: the Manhattan - L1 distance, the Euclidean distance, the Hellinger coefficient and the cosine distance. We recall that the Hellinger Coefficient is defined as

$$H(X,Y) = \sqrt{\sum_{i=1}^{P}\left(\sqrt{x_i} - \sqrt{y_i}\right)^2} \tag{3}$$

while the Cosine distance is defined as:

$$C(X,Y) = \frac{\sum_{i=1}^{P} x_i y_i}{\sqrt{\sum_{i=1}^{P} x_i^{\,2}} \cdot \sqrt{\sum_{i=1}^{P} y_i^{\,2}}} \tag{4}$$

Table 1

Bull's Eye Score for the full MPEG-7 database, when various distances were used as similarity measures and the shape description was done with various methods. We marked with gray the best achieved result

Method \ Distance	L1	Euclidean	Hellinger	Cosine
Contour tracing (original image)	35.37%	32.62%	34.12%	28.71%
Contour tracing (image upscaled 2x)	33.95%	32.25%	32.89%	25.71%
Contour tracing (image upscaled 5x)	23.6%	21.44%	22.94%	19.36%
Image projections	27.95%	28.23%	24.56%	22.35%

4.2. Recognition accuracy

The results of the proposed system for various combinations of parameters, when testing on the full database (i.e. on 70 classes) are presented in *Table 1*. To give an idea of the complexity of data, a random chance classifier produces 2.8% accuracy.

While the shapes description with image projections is insensitive to image relative size, the latter impacts the contour tracing: larger the shape is, more possibilities for the Moore neighborhood, thus, more detailed view of the shape is retrievable. Consequently, we take into account increasing the shape size by various factors. Yet, the results show that the impact is negative and the best performance is retrieved at the original scale.

We note that even simpler and computationally faster, the Manhattan distance leads to best results, followed by the Hellinger coefficient. When

compared with Euclidean distance, the latter imposes a heavier penalty on differences on larger values and tends to ignore the ones on the smaller values.

Table 2

Bull's Eye Score for the full MPEG-7 database: comparative results.

Method	*Proposed*	*Random chance*	*Aspect space [10]*
Bull's eye score	35.37%	2.8 %	95,96%

As one may see, the proposed results are considerably higher than random chance, yet there is space for improvement as the aspect space method [10] reaching 96% retrieval accuracy as showed in table 2. The main flaw of the proposed method is the discrimination for articulated objects. In such a case, the contour tracing method is more robust, yet it also, sometimes fail.

However, if from the full 70-class database we retain only the 20 classes that contain least articulated objects, the Bull's Eye score increases to **92.75%** for contour tracing with L1 distance and to **90.50%** for image projections with Hellinger coefficient. In such a case the random chance leads to 9.48% accuracy.

Complexity and duration. The complexity order of all methods is linear with respect to the number of pixels of the image, as the shape signature is computed in linear time and rotation normalization takes most of it. Given a shape, its description takes 0.25 msec while using TESPAR encoded image projections and 0.4 msec with contour tracing. A Bull's Eye query on the described database takes 20 msec. The tests were performed with Matlab code on an Intel Xeon at 3.3 Ghz CPU.

5. Conclusions and continuations

In this paper we have proposed two new simple binary shape descriptors. The descriptors relies on rotation normalization, data reduction to a pair of 1D waveforms and TESPAR encoding. The descriptors shown promising results on the very challenging MPEG-7 database, especially if only classe with non articulated objects are considered. Specifically they achieve high accuracy when described no-articulated compact objects. While the achieved retrieval rate (bull's eye score) is inferior to state of the art methods, the proposed method is significantly faster; thus in the current shape the proposed descriptor is suitable to discriminate among silhouettes of non-articulated objects such as the mouth in various expressive positions. Due to its simplicity and speed, the proposed descriptor is a viable solution as pre/post-processing in other systems and not as a stand-alone box.

Taking into account that the inner distance introduced in [9] was showed to be robust with respect to object's articulation the next step of the research is to subtract from such descriptors those features that would increase its robustness to

articlation. In such a way we hope to significantly increases the overall performance on the full database.

Acknowledgements

This work was partially supported by the Romanian Sectoral Operational Programme Human Resources Development 2007-2013 through the European Social Fund Financial Agreements POSDRU/159/1.5/S/134398 (Knowledge).

R E F E R E N C E S

[1] *Th. Plötz, G. A. Fink* "Markov models for offline handwriting recognition: a survey", International Journal on Document Analysis and Recognition (IJDAR) 2009, **Vol**. 12(4), pp 269-298.

[2] *G. Krawczyk, K. Myszkowski, H.-P. Seidel,* "Lightness perception in tone reproduction for high dynamic range images", in: Proc. of EUROGRAPHICS, Computer Graphics Forum, **vol**. 24, 2005

[3] *S. Asteriadis, N. Nikolaidis, I. Pitas,* "Facial feature detection using distance vector fields". Pattern Recognition, **Vol.** 42(7), pp. 1388 – 1398, 2009

[4] *R. Vranceanu, C. Florea, L. Florea, C. Vertan,*: "Gaze direction estimation by component separation for recognition of eye accessing cues". Machine Vision and Application, **Vol.** 26(2-3), pp. 267–278, 2015.

[5] *L. J. Latecki, R. Lakamper, U. Eckhardt. "*Shape descriptors for non-rigid shapes with a single closed contour". In Proc. of Computer Vision and Pattern Recognition, **Vol. 1**, *pp.* 424–429, 2000.

[6] *D. Zhang, G. Lu* "Review of shape representation and description techniques", Pattern Recognition, **Vol.** 37, 2004, pp. 1 – 19.

[7] *I. Yong, J. Walker, J. Bowie*, "An analysis technique for biological shape", Comput. Graphics Image Process. **Vol.** 25, pp.357–370, 1974.

[8] *D.G. Lowe*. "Distinctive image features from scale-invariant keypoints". Int. Journal of Computer Vision, **Vol.** 60(2), pp. 91–110, 2004.

[9] *H. Ling, D. W. Jacobs*. "Shape classification using the inner-distance". In IEEE Trans. On Pattern Analysis and Machine Intelligence, **Vol.** 29(2), pp.286–299, 2007.

[10] *H. Ling, X. Yang, L. J. Latecki* "Balancing Deformability and Discriminability for Shape Matching" In Proc. of European Conference on Computer Vision, **Vol.** 6313 LNCS, pp. 411-424, 2010,

[11] *G. Toussaint*, "Grids, Connectivity And Contour Tracing", http://jeff.cs.mcgill.ca /~godfried/teaching/mir-reading-assignments/Chapter-2-Grids-Connectivity-Contour-Tracing.pdf, Retrieved, March, 2015

[12] *R.A. Kings, T.C. Phipps*, "Shannon, TESPAR and approximation strategies". Computers & Security **Vol. 18,** pp**.** 445 – 453, 1999.

[13] *A. Sheela, S. Prasad*, "Linear Discriminant Analysis F-Ratio for Optimization of TESPAR&MFCC Features for Speaker Recognition" , Journal of Multimedia, **Vol. 2(6),** pp. 34-43, 2007.

[14] *L. Florea, C. Florea, R. Vranceanu, C. Vertan* "Zero-crossing Based Image Projections Encoding for Eye Localization", EUSIPCO 2012, pp. 150-154, 2012, Bucuresti, Romania.