

Deformable 2D Shape Matching Based on Shape Contexts and Dynamic Programming

Iasonas Oikonomidis and Antonis A. Argyros

Institute of Computer Science, Forth
and
Computer Science Department, University of Crete
{oikonom, argyros}@ics.forth.gr
<http://www.ics.forth.gr/cvrl/>

Abstract. This paper presents a method for matching closed, 2D shapes (2D object silhouettes) that are represented as an ordered collection of shape contexts [1]. Matching is performed using a recent method that computes the optimal alignment of two cyclic strings in sub-cubic runtime. Thus, the proposed method is suitable for efficient, near real-time matching of closed shapes. The method is qualitatively and quantitatively evaluated using several datasets. An application of the method for joint detection in human figures is also presented.

1 Introduction

Shape matching is an important problem of computer vision and pattern recognition which can be defined as the establishment of a similarity measure between shapes and its use for shape comparison. A byproduct of this task might also be a set of point correspondences between shapes. The problem has significant theoretical interest. Shape matching that is intuitively correct for humans is a demanding problem that remains unsolved in its full generality. Applications of shape matching include but are not limited to object detection and recognition, content based retrieval of images, and image registration.

A lot of research efforts have been devoted to solving the shape matching problem. Felzenszwalb et al. [2] propose the representation of each shape as a tree, with each level representing a different spatial scale of description. They also propose an iterative matching scheme that can be efficiently solved using Dynamic Programming. Ebrahim et al. [3] present a method that represents a shape based on the occurrence of shape points on a Hilbert curve. This 1D signal is then smoothed by keeping the largest coefficients of a wavelet transform, and the resulting profiles are matched by comparing selected key regions. Belongie et al. [1] approach the problem of shape matching introducing the shape context, a local shape descriptor that samples selected edge points of a figure in log-polar space. The resulting histograms are compared using the x^2 statistic. Matches between corresponding points are established by optimizing the sum of matching costs using weighted Bipartite Matching (BM). Finally, a Thin Plate Spline (TPS) transformation is estimated, that warps the points of the first shape to

the second, based on the identified correspondences. This process is repeated for a fixed number of iterations, using the resulting deformed shape of the previous step as input for the next step. A very interesting work that utilizes shape contexts is presented in [4]. The goal of this work is to exploit the articulated nature that many common shapes possess to improve shape matching. The authors suggest that the distances and angles to be sampled should be measured only inside the closed contour of a figure.

In this work we are interested in the particular problem of matching deformable object silhouettes. The proposed method is based on shape contexts and the work of Belongie [1]. It is assumed that a 2D shape can be represented as a single closed contour. This is very often the case when, for example, shapes are derived from binary foreground masks resulting from a background subtraction process or from some region-based segmentation process. In this context, shape matching can benefit from the knowledge of the ordering of silhouette points, a constraint that is not exploited by the approach of Belongie [1]. More specifically, in that case, two silhouettes can be matched in sub-cubic runtime using a recently published algorithm [5] that performs cyclic string matching employing dynamic programming. The representation power of shape contexts combined with the capability of the matching algorithm to exploit the order in which points appear on a certain contour, result in an effective and efficient shape matching method.

Several experiments have been carried out to assess the effectiveness and the performance of the proposed method on benchmark datasets. The method is quantitatively assessed through the bull’s-eye test applied to the MPEG7 CE-shape-1 part B dataset. More shape retrieval experiments have been carried out on the “gestures” and “marine” datasets. Additionally, the proposed shape matching method has been employed to detect the articulation points (joints) of a human figure in monocular image sequences. Specifically, 25 human postures have been annotated with human articulation points. Shape matching between a segmented figure and the prototype postures results in point correspondences between the human figure and its best matching prototype. Then TPS transfers known points from the model to the observed figure.

Overall, the experimental results demonstrate that the proposed method performs very satisfactory in diverse shape matching applications and that the performance of shape matching can be improved when the order of points on a contour is exploited. Additionally, its low computational complexity makes it a good candidate in shape matching applications requiring real-time performance.

The rest of the paper is organized as follows. The proposed method is presented in Sec. 2. Experimental results are presented in Sec. 3. Finally, Sec. 4 summarizes the main conclusions from this work.

2 The Proposed Shape Matching Method

The proposed method utilizes shape contexts to describe selected points on a given shape. A fixed number of n points are sampled equidistantly on the contour

of each shape. For each of these points, a shape context descriptor is computed. To compare two shapes, each descriptor of the 1st shape is compared using the x^2 statistic to all the descriptors of the 2nd, giving rise to pairwise matching costs. These costs form the input to the cyclic string matching, and correspondences between the shapes are established. These correspondences are used to calculate a Thin Plate Splines based alignment of the two shapes. A weighted sum of the cyclic matching cost and the TPS transformation energy forms the final distance measure of the two shapes. The rest of this section describes the above algorithmic steps in more detail.

2.1 Scale Estimation and Point Order

The first step of the method is to perform a rough scale estimation of the input shape. As in [1], the mean distance between all the point pairs is evaluated and the shape is scaled accordingly. Denoting the i th input point as pt_i , the scale a is estimated as

$$a = \sum_{i=1}^n \sum_{j=i+1}^n \frac{2 \|pt_i - pt_j\|}{n(n-1)}. \quad (1)$$

Then, every input point is scaled by $1/a$.

The order (clockwise/counterclockwise) in which silhouette points are visited may affect the process of shape matching. Therefore, we adopt the convention that all shapes are represented using a counter-clockwise order of points. To achieve this, the sign of the area of the polygon is calculated as

$$A = \frac{1}{2} \sum_{i=1}^n x_i y_{i+1} - x_{i+1} y_i, \quad (2)$$

with $x_{n+1} = x_1$ and $y_{n+1} = y_1$. If A is negative, the order of the input points is reversed.

2.2 Rotation Invariant Shape Contexts

For the purposes of this work, rotation invariance is a desirable property of shape matching. As mentioned in [1], since each shape context histogram is calculated in a log-polar space, rotation invariance can be achieved by adjusting the angular reference frame to an appropriately selected direction. A direction one can use for imposing rotation invariance in shape contexts, is the local tangent of the contour. In this work this direction is estimated using cubic spline interpolation. First, the 2D curve is fitted by a cubic spline model. Cubic splines inherently interpolate functions of the form $f : \mathbb{R} \rightarrow \mathbb{R}$. It is easy to extend this to interpolate parametric curves on the plane (functions of the form $\gamma : \mathbb{R} \rightarrow \mathbb{R}^2$), by concatenating two such models. The next step is to compute the derivatives of the two cubic spline models at each point of interest. For each such pair of derivatives, the local tangent is computed by taking the generalized arc tangent function with two arguments. This method has the advantage that the computed

angles are consistently aligned not only to a good estimate of the local derivative, but also to a consistent direction. The estimated local contour orientation is then used as the reference direction of shape contexts towards achieving descriptions that are rotationally invariant.

2.3 Cyclic Matching

The comparison of a pair of shape contexts can be performed with a number of different histogram comparison methods. In this work, the x^2 statistic is selected as in [1]:

$$\chi^2(h_1, h_2) = \frac{1}{2} \sum_{k=1}^K \frac{[h_1(k) - h_2(k)]^2}{h_1(k) + h_2(k)}, \quad (3)$$

where h_1 and h_2 are the compared histograms, each having K bins. The comparison of two shapes is performed by considering a 2D matrix C . The element (i, j) of this matrix is the x^2 statistic between the i th shape context of the first shape and the j th shape context of the second shape. Any such pair is a potential correspondence. Belongie et al. [1] use Bipartite Matching to establish a set of 1-to-1 point correspondences between the shapes. However, by exploiting the order that is naturally imposed by the contour, the search space can be significantly reduced.

For the purpose of matching, we adopt the method presented in [5]. The matrix C of x^2 shape context comparisons forms the matching costs matrix needed for the cyclic matching. Along with the matching pairs, a matching cost c_m is calculated as the sum of costs of all the aligning operations that were used. Thus, c_m can be used as a measure of the distance between the two shapes.

2.4 Thin Plate Spline Computation

The final step of the presented shape matching method is the computation of the planar deformation that aligns two shapes. The alignment is performed using Thin Plate Splines. The input to this stage is the result of the previous step, i.e. a set of pairs of correspondences between two 2D shapes. The output is a deformation of the plane, as well as a deformation cost. This cost can be properly weighted along with the cost of the previous step to form the final matching cost or distance between the shapes.

The regularized version of the TPS model is used, with a parameter λ that acts as a smoothness factor. The model tolerates higher noise levels for higher values of λ and vice versa. Since the scale of all shapes is roughly estimated at the first step of the method, the value of λ can be uniformly set to compensate for a fixed amount of noise. For all experiments, λ was fixed to 1, as in [1].

Besides the warping between the compared shapes, a total matching cost \mathcal{D} is computed as

$$\mathcal{D} = l_1 c_m + l_2 c_b. \quad (4)$$

\mathcal{D} is a weighted sum of the cyclic matching cost c_m and the TPS bending cost c_b . While c_b has the potential to contribute information not already captured by c_m ,

in practice it proved sufficient to ignore the c_b cost, and use only the c_m cost as the distance \mathcal{D} between shapes (i.e. $l_1 = 1$ and $l_2 = 0$). For all the following, this convention is kept. It should be also noted that the TPS might be needed for the alignment of matched shapes, regardless of whether the c_b cost contributes to the matching cost \mathcal{D} . Such a situation arises in the joints detection application described in Sec. 3.2.

3 Experimental Results

Several experiments have been carried out to evaluate the proposed method. The qualitative and quantitative assessment of the proposed method was based on well-established benchmark datasets. An application of the method for the localization of joints in human figures is also presented. Throughout all experiments $n = 100$ points were used to equidistantly sample each shape. For the MPEG7 experiment (see Sec. 3.1) this results in an average subsampling rate of 13 contour pixels with a standard deviation of 828 pixels. This large deviation is due to the long right tail of the distribution of shape lengths. Shape contexts were defined having 12 bins in the angular and 5 bins in the radial dimension. Their small and large radius was equal to 0.125 and 2, respectively (after scale normalization). The TPS regularization parameter λ was set equal to 1 and the insertion/deletion cost for the cyclic matching to 0.75 (the χ^2 statistic yields values between 0 and 1).

3.1 Benchmark Datasets

The proposed shape matching method has been evaluated on the “SQUID” [6] and the “gestures” [7] datasets. In all the experiments related to these datasets, each shape was used as a query shape and the proposed method was employed to rank all the rest images of the dataset in the order of increasing cost \mathcal{D} . Figures 1(a) and 1(b), show matching results for the “SQUID” and the “gestures” datasets, respectively. In each of these figures, the first column depicts the query shape. The rest of each row shows the first twenty matching results in order of increasing cost \mathcal{D} . The retrieved shapes are, in most of the cases, very similar to the query.

The quantitative assessment of the proposed method was performed by running the bull’s-eye test on the MPEG7 CE-shape-1 part B dataset [8]. This dataset consists of 70 shape classes with 20 shapes each, resulting in a total of 1400 shapes. There are many types of shapes including faces, household objects, other human-made objects, animals, and some more abstract shapes. Given a query shape, the bull’s-eye score is the ratio of correct shape retrievals in the top 40 shapes as those are ranked by the matching algorithm, divided by the theoretic maximum of correct retrievals, which for the specific dataset is equal to 20. The bull’s eye score of the proposed method on the MPEG7 dataset is 72.35%. The presented method does not natively handle mirroring, so the minimum of the costs to the original and mirrored shape is used in shape similarity comparisons. By post-processing the results using the graph transduction method [9]

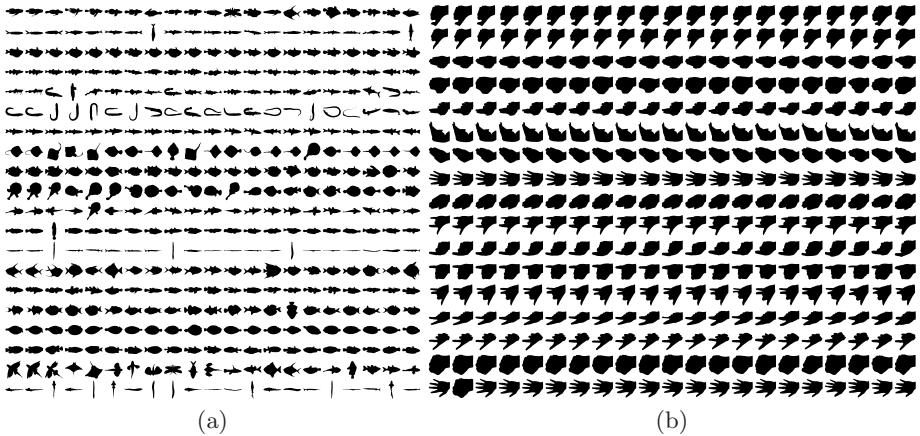


Fig. 1. Matching results for (a) the “SQUID” and (b) the “gestures” datasets

with the parameter values suggested therein, the score is increased to 75.42%. For comparison, the state of the art reported scores on this dataset are 88.3% for the Hilbert curve method [3] and 87.7% for the hierarchical matching method [2] (for more details, see Table 2 in [3]).

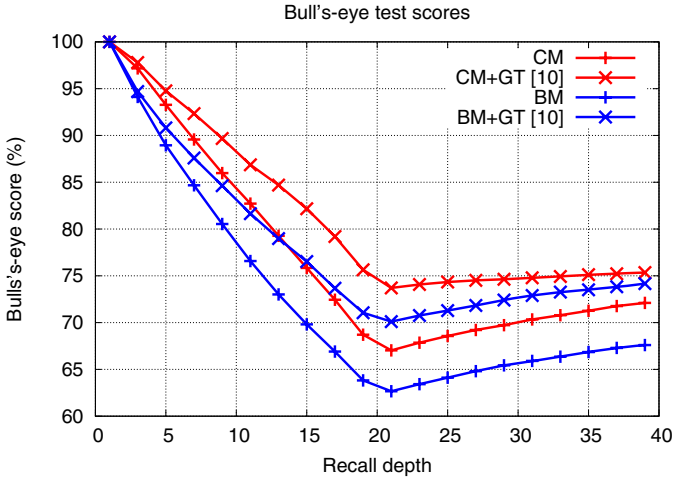
An extended investigation of the results of the bull’s-eye test is graphically illustrated in Fig.2(a). This graph essentially turns the rather arbitrary choice of the forty best results into a variable. The horizontal axis of the graph is this recall length variable, and the vertical axis is the percentage of correct results among the examined ones. The experimental results demonstrate that the cyclic string matching performs better than Bipartite Matching. Additionally, graph transduction improves both methods but does not affect the superiority of the cyclic matching compared to Bipartite Matching.

The essential advantage of cyclic matching over Bipartite Matching is the reduction of the search space: while Bipartite Matching searches among all possible permutations between two shapes, cyclic matching only considers the matchings that obey the ordering restrictions imposed by both shape contours. This effectively speeds up the matching process while yielding intuitive results. Sample¹ shape retrieval results on the MPEG7 dataset are shown in Fig.2(b).

3.2 Detecting Joints in Human Figures

Due to its robustness and computational efficiency, the proposed method has been used for the recovery of the joints of a human figure. For this purpose, a set of synthetic human model figures were generated. Two model parameters control the shoulder and elbow of each arm. Several points (joints and other points of interest) are automatically generated on each model figure. Figure 3

¹ The full set of results for the reported experiments is available online at <http://www.ics.forth.gr/~argyros/research/shapematching.htm>



(a)



(b)

Fig. 2. Results on the MPEG7 data set. (a) The bull's-eye test scores on the MPEG7 dataset as a function of the recall depth, (b) sample shape retrieval results.

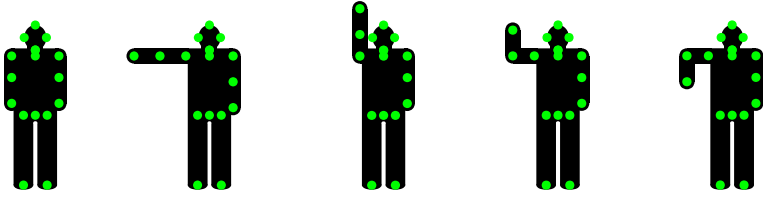


Fig. 3. The five configurations for the right arm. The contour of each figure is used as the shape model; Marked points are the labeled joints.

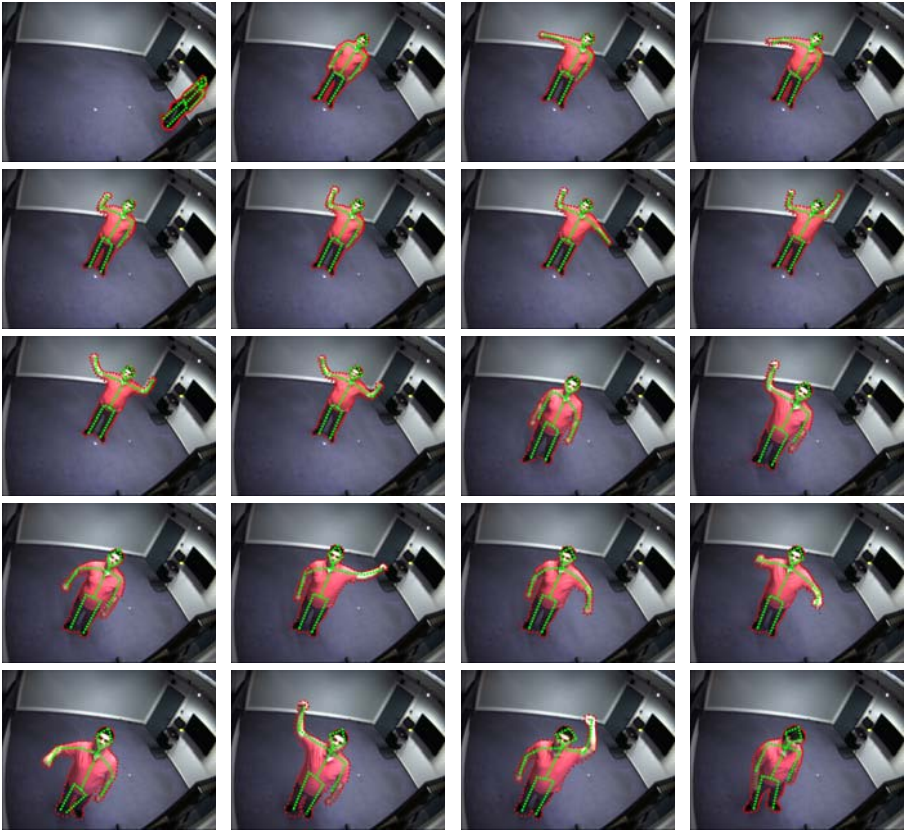


Fig. 4. Characteristic snapshots from the joints detection experiment

shows five such model figures for various postures of the right arm. A total of 25 models were created, depicting all possible combinations of articulations of the right (as shown in Fig.3) and the left arm.

In the reported experiments, the background subtraction method of [10] has been employed to detect foreground figures. Connected components of the resulting foreground mask image are then considered. If there exist more than one

connected components on the foreground image, only the one with the largest area is maintained for further processing. Its silhouette is then extracted and a fixed number n of roughly equidistant points are selected on it. This list of points constitutes the actual input to the proposed matching method. Each figure is compared to all model figures. The model with the lowest cost \mathcal{D} is picked as corresponding to the input. The TPS transformation between the model and the input is subsequently used to warp the labeled points of interest on the input image.

Figure 4 shows characteristic snapshots from an extensive experiment where a human moves in a room in front of a camera while taking several different postures. The input image sequence contains approximately 1200 frames acquired at 20 fps. Having identified the joints, a skeleton model of each figure is obtained. Interestingly, the method performs well even under considerable scale and perspective distortions introduced because of the human motion that result in considerable differences between the actual foreground silhouettes and the considered prototypes.

The results presented in Fig.4 have been obtained without any exploitation of temporal continuity. This may improve results based on the fact that the estimation of the human configuration in the previous frame is a good starting point for the approximation in the current frame. To exploit this idea, at each moment in time, a synthetic figure like the ones shown in Fig.3 is custom rendered using the joint angles of the estimated skeleton. Thus, the result of the previous frame is used as a single model figure for estimating the human body configuration in the current frame. In case that the estimated distance between the synthetic model and the observed figure exceeds a specified threshold, the system is initialized by comparing the observed figure with the 25 prototype figures, as in the previous experiment. The exploitation of temporal continuity improves significantly the performance of the method.

4 Discussion

This paper proposed a rotation, translation and scale invariant method for matching 2D shapes that can be represented as single, closed contours. Affine transformations can be tolerated since the shape contexts are robust (but not strictly invariant) descriptors under this type of distortion. The performance of the method deteriorates gradually as the amount of noise increases. In this context, noise refers to either shape deformations due to errors in the observation process (e.g. foreground/background segmentation errors, sampling artifacts etc) or natural shape deformations (e.g. articulations, perspective distortions, etc).

The time complexity of the method is $\mathcal{O}(n^2 \log(n))$ for n input points, an improvement over the respective performance of [1], which is $\mathcal{O}(n^3)$. In the application of Sec. 3.2, the employed unoptimized implementation performs 25 shape comparisons per second, including all computations except background subtraction. By exploiting temporal continuity, most of the time the method needs to compare the current shape with a single prototype, leading to real time

performance. Overall, the experimental results demonstrate qualitatively and quantitatively that the proposed method is competent in matching deformable shapes and that the exploitation of the order of contour points besides improving matching performance, also improves shape matching quality.

Acknowledgments

This work was partially supported by the IST-FP7-IP-215821 project GRASP. The contributions of Michel Damien and Thomas Sarmis (members of the CVRL laboratory of FORTH) to the implementation and testing of the proposed method are gratefully acknowledged.

References

1. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *IEEE Transactions on PAMI* 24, 509–522 (2002)
2. Felzenszwalb, P., Schwartz, J.: Hierarchical matching of deformable shapes. In: *CVPR 2007*, pp. 1–8 (2007)
3. Ebrahim, Y., Ahmed, M., Abdelsalam, W., Chau, S.C.: Shape representation and description using the hilbert curve. *Pat. Rec. Let.* 30, 348–358 (2009)
4. Ling, H., Jacobs, D.: Shape classification using the inner-distance. *IEEE Transactions on PAMI* 29, 286–299 (2007)
5. Schmidt, F., Farin, D., Cremers, D.: Fast matching of planar shapes in sub-cubic runtime. In: *ICCV 2007*, pp. 1–6 (2007)
6. Mokhtarian, F., Abbasi, S., Kittler, J.: Robust and efficient shape indexing through curvature scale space. In: *BMVC 1996*, pp. 53–62 (1996)
7. Petrakis, E.: Shape Datasets and Evaluation of Shape Matching Methods for Image Retrieval (2009), <http://www.intelligence.tuc.gr/petrakis/>
8. Jeannin, S., Bober, M.: Description of core experiments for mpeg-7 motion/shape (1999)
9. Yang, X., Bai, X., Latecki, L.J., Tu, Z.: Improving shape retrieval by learning graph transduction. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part IV*. LNCS, vol. 5305, pp. 788–801. Springer, Heidelberg (2008)
10. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 2, pp. 28–31 (2004)