Shape Context: A new descriptor for shape matching and object recognition

Serge Belongie, Jitendra Malik and Jan Puzicha
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley, CA 94720, USA
{sjb,malik,puzicha}@cs.berkeley.edu

Abstract

We introduce a new shape descriptor, the *shape context*, for correspondence recovery and shape-based object recognition. The shape context at a point captures the distribution over relative positions of other shape points and thus summarizes global shape in a rich, local descriptor. Shape contexts greatly simplify recovery of correspondences between points of two given shapes. Moreover, the shape context leads to a robust score for measuring shape similarity, once shapes are aligned.

The shape context descriptor is tolerant to all common shape deformations. As a key advantage no special landmarks or key-points are necessary. It is thus a generic method with applications in object recognition, image registration and point set matching. Using examples involving both handwritten digits and 3D objects, we illustrate its power for object recognition.

1 Introduction

The last decade has seen increased application of statistical pattern recognition techniques to the problem of object recognition from images [8, 7, 6]. Typically, an image with n pixels is regarded as an n dimensional feature vector formed by concatenating the brightness values of the pixels. Given this representation, a number of different strategies have been tried, e.g. nearest-neighbor techniques after extracting principal components [8, 7], or training a discriminative convolutional neural network classifier [6]. Impressive performance has been demonstrated on datasets such as digits and faces.

In our opinion, a vector of pixel brightness values is a somewhat unsatisfactory representation of an object. Basic invariances e.g. to translation, scale and small amount of rotation must be obtained by suitable pre-processing or by the use of enormous amounts of training data [6]. This has motivated alternative approaches such as [1] who find key points or landmarks, and recognize objects using the spatial arrangements of point sets. However not all objects have distinguished key points (think of a circle for instance), and using key points alone sacrifices the shape information available in smooth portions of object contours.

Our approach therefore uses a general representation of shape – a set of points sampled from the contours on the object. Each point is associated with a novel descriptor, the *shape context*, which describes the coarse arrangement of the rest of the shape with respect to the point. This descriptor will be different for different points on a single shape S; however corresponding (homologous) points on similar shapes S and S'will tend to have similar shape contexts. Shape contexts are distributions and can be compared using the χ^2 statistic. Correspondences between the point sets of S and S' can be found by solving a bipartite weighted graph matching problem with edge weights C_{ij} defined by the χ^2 distances of the shape contexts of points i and j. Given correspondences, we can calculate a similarity measure between the shapes S and S'. This similarity measure can be used in a nearest-neighbor classifier for object recognition.

Appealing features of the approach are that it is very simple and robust, the standard invariances are built in for free, and as a consequence we develop a classifier which is effective when only a small number of training examples are available.

This paper is organized as follows. We first discuss related work on shape matching in Sect. 2. Next, we introduce the shape context and our method for establishing correspondences in Sect. 3. We present experiments which show that shape matching using this approach is robust and accurate. Recognition results on the MNIST digit dataset and the Columbia COIL dataset are in Sect. 4. We conclude in Sect. 5.

2 Related Work on Shape Matching

In the context of image retrieval and shape similarity, several shape descriptors have been proposed, ranging from moments and Fourier descriptors to Hausdorff distance and the medial axis transform. For an overview and a detailed discussion of shape matching techniques, the reader is referred to [9]. It should be emphasized that our approach is generically applicable as opposed to most shape matching techniques that are restricted to silhouettes and closed curves. In our framework shape refers to any type of boundary information, and in consequence, our algorithm is applicable for a large variety of recognition problems.

At its core, shape contexts can be understood as a point set matching technique. Most closely related is the work of [3] which proposes an iterative optimization algorithm to jointly determine point correspondences and underlying image transformations, where typically some generic transformation class is assumed, e.g. affine or, more generally, thin plate splines. This formulation leads to a difficult non–convex optimization problem which is solved using deterministic annealing. [3].

As we will show, shape contexts will greatly simplify the matching part, leading to a very robust point registration technique. It is invariant to scale and translation and to a large extent robust to rotation and deformation. Extensions incorporating rotational invariance and local appearance features may be found in [2].

3 Shape Context

Shape context analysis begins by converting the edge elements of a shape into a set of N feature points. These points can be on internal or external contours. They need not, and typically will not, correspond to key-points such as maxima of curvature or inflection points. We prefer to sample the shape with roughly uniform spacing, though this is also not critical. An example using the shape in Figure 1(a) is shown in Figure 1(c). Note that this shape, despite being very simple, does not admit the

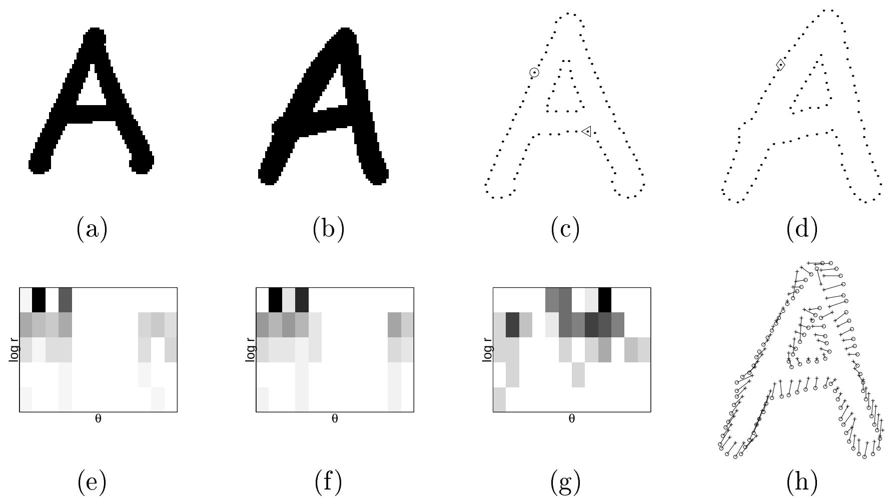

Figure 1: Shape context computation and matching. (a,b) Original shapes. (c,d) Sampled edge points. (e-g) Example shape contexts for reference samples marked by $\circ, \diamond, \triangleleft$ in (c,d). Each shape context is a log-polar histogram of the coordinates of the rest of the point set measured using the reference point as the origin. Here we have used 5 and 12 bins for $\log r$ and θ, respectively. (Dark=large value.) Note the visual similarity of the shape contexts for $\circ$ and $\diamond$, which were computed for relatively similar points on the two shapes. By contrast, the shape context for $\triangleleft$ is quite different. (g) Correspondences found using bipartite matching, with weights defined by the χ^2 distance between histograms.

use of silhouette-based methods due to its internal contour. Now consider the set of vectors originating from a point in Figure 1(c) to all other points in the shape. These vectors express the appearance of the entire shape relative to the reference point. Obviously, this set of $N-1$ vectors is a rich description, since as N gets large, the representation of the shape becomes exact.

The full set of vectors as a shape descriptor is inappropriate since shapes and their sampled representation may vary from one instance to another. In contrast, we identify the *distribution* over relative positions as a robust and compact, yet discriminative descriptor. For a point P on the shape, we compute a coarse histogram of the relative coordinates of the remaining $N-1$ points. This histogram is defined to be the *shape context* of P. The reference orientation for the coordinate system can be absolute or relative to a given axis. In this paper we will assume an absolute reference orientation, i.e. angles measured relative to the positive x-axis. The descriptor should be more sensitive to differences in nearby pixels. We thus propose to use a log-polar coordinate system. An example is shown in Fig. 1(e). Throughout this paper we have used 12 equally spaced angle bins and 5 equally spaced log-radius bins.

An attractive characteristic of the shape context is the invariance to common deformations. Invariance to translation is intrinsic to the shape context definition since everything is measured with respect to points on the object. To achieve scale invariance we normalize all radial distances by the median distance λ between all N^2 point pairs in the shape. Choosing the median provides robustness to outliers. Robustness to significant rotations can be achieved by iterating the steps of match-

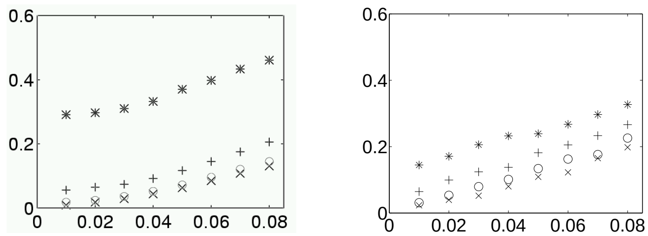

Figure 2: Randomized point set matching results. Left: Results from [3] Right: Results for shape context matching. The x axis shows σ and the y axis shows average parameter estimation error. $\times : p_d = 0.0, p_s = 0.0, \circ : p_d = 0.1, p_s = 0.1$, $+ : p_d = 0.3, p_s = 0.1, * : p_d = 0.5, p_s = 0.1$. Each data point represents the mean over 500 trial runs.

ing and point set alignment a few times, as shown in the evaluation below. As we will empirically demonstrate, shape contexts are robust to additions and deletions. In a companion paper [2] we extended the shape context descriptor to complete rotational invariance employing relative instead of absolute frames.

Matching Shape Contexts In determining shape correspondences, we aim to meet two criteria: (1) corresponding points should have very similar descriptors, and (2) the correspondences should be unique.

Consider a point i on the first shape and a point j on th second shape. We compare the shape contexts at i and j to come up with a cost $C_{i,j}$ for matching these two points. Let the K-bin (normalized) histogram at i be $g(k)$ and at j be $h(k)$. Then the cost $C_{i,j}$ is given by the χ^2 statistic

$$C_{i,j} = \frac{1}{2} \sum_{k=1}^{K} \frac{[g(k) - h(k)]^2}{g(k) + h(k)}$$

Given the set of costs $C_{i,j}$ between all pairs of points i on the first shape and j on the second shape we want to minimize the total cost of matching subject to the constraint that the matching be one-to-one. This is an instance of the square assignment (or weighted bipartite matching) problem, which can be solved in $O(N^3)$ time using the Hungarian method. In our experiments, we use the comparatively more efficient algorithm of [5]. The input to the assignment problem is a square cost matrix with entries $C_{i,j}$. The result is a permutation $\pi(i)$ such that the sum $\sum_i C_{i,\pi(i)}$ is a minimum. The result of applying this algorithm to the letter-A example is shown in Figure 1(h).

When the number of samples on two shapes is not equal, the cost matrix can be made square by adding "dummy" nodes to each point set with a constant matching cost of ϵ_d. The same technique may also be used even when the sample numbers are equal to allow for robust handling of outliers.

Once a correspondence between points is established, we can estimate the transformation between them. Assuming a noisy measurement model, one usually restricts the class of allowed transformations to obtain robust estimators. In this work, we restrict attention to affine transformations which consist of a translation followed by an arbitrary linear map. Since the correspondences are known the affine transformation is estimated using standard least squares methods. These two steps can be iterated to achieve additional precision. However, the initial estimate of correspondences is often sufficient to obtain an excellent estimate of the underlying affine

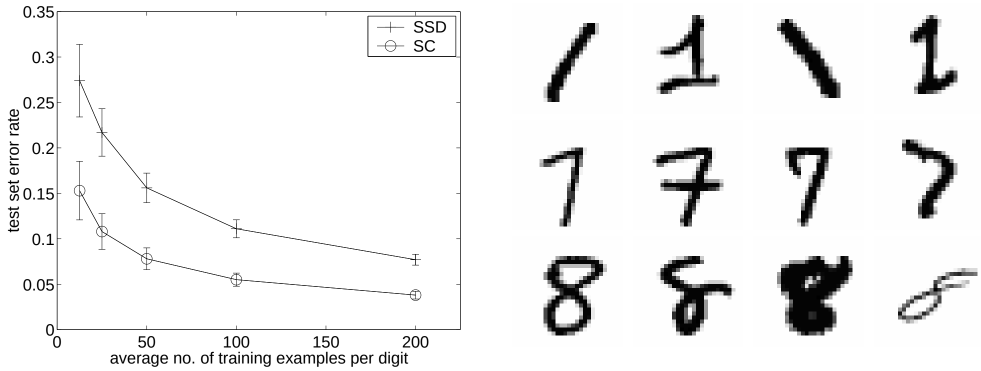

Figure 3: Handwritten digit recognition on the MNIST dataset. Left: Test set errors of a 1-NN classifier using SSD and Shape Contexts as distance measures. Right: some example ones, sevens, and eights, illustrating the high degree of intra-class variability.

transformation without any iteration, resulting in an extremely fast algorithm.

Empirical Robustness Evaluation In order to study the robustness of shape contexts for recovering correspondences, we performed the random point set matching experiment described in [3, Sect. 5.2]. This experiment consists of repeatedly generating a random point set and matching it to a distorted version of itself. The model point set is made by choosing 50 points uniformly at random in a unit square. The parameter values for the distorting transformation are drawn independently and uniformly at random from the following intervals: $-0.5 < t_x, t_y < 0.5$ (translation), $-27^\circ < \theta < 27^\circ$ (rotation), and $0.5 \leq e^a \leq 2$ (scale). Points in the transformed set are deleted and spurious points added according to the fractions $p_d \in \{0, 0.1, 0.3, 0.5\}$ and $p_s \in \{0, 0.1\}$, respectively. Jitter is introduced by adding independent Gaussian noise with $\sigma = \{0.01, 0.02, \ldots, 0.08\}$ to each coordinate before transformation. The measure of performance is based on the average error between the actual and the estimated transformation parameters. To obtain our parameter estimates, we iterated the steps of matching and least-squares alignment recovery four times. We added dummy nodes with $\epsilon_d = 0.15$ to make the total number of nodes in each point set 60. A comparison of the two sets of results is shown in Fig. 2.

4 Results

A straightforward strategy for recognition is to use a 1-NN classifier with shape context dissimilarity as the distance measure. The overall algorithm has 3 steps: (1) estimate affine transforms between a prototype and a query shape, (2) apply the affine transform and recompute the shape contexts for the transformed point set, and (3) score the match by summing up the shape context distances between each point on a shape to its most similar point on the other shape[1].

Case study 1: Digit recognition The first experiment is concerned with the MNIST dataset of handwritten digits, which consists of 60,000 training and 10,000

[1] We actually obtain two scores, one projecting reference shape onto query shape and one vice versa. The final score is obtained by taking the maximum.

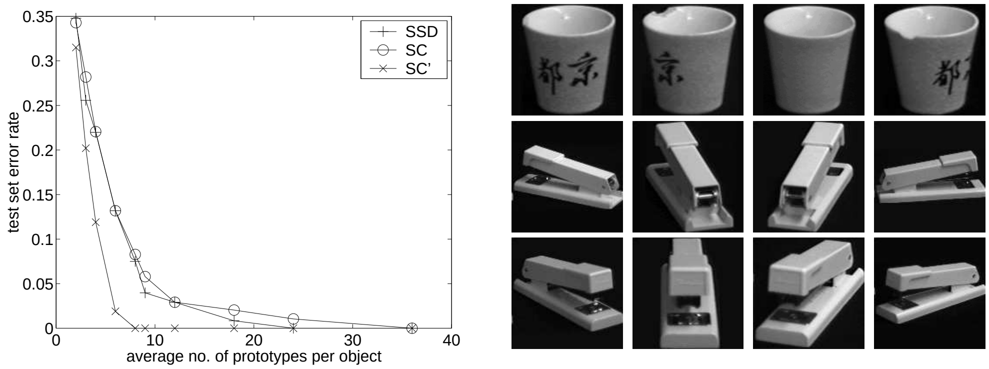

Figure 4: 3D object recognition. Left: comparison of test set error for SSD, Shape Contexts (SC), and Shape Contexts with K-medoid prototypes (SC') vs. number of prototype views. For SSD and SC, we varied the number of prototypes uniformly for all objects. For SC', the number of prototypes per object has been chosen adaptively, see text. Right: K-medoid prototype views for two different examples, using an average of 6 prototypes per object.

test digits, see [6] for a description and results. However, since we are mainly interested in understanding shape, and thus in generalizing from few examples, we present here results for small training sets chosen at random from the full set. The test set error plotted in Fig. 3 has been evaluated over 1000 randomly chosen test digits using a 1–NN classifier. Two different similarity measures, shape context (SC) and sum of squared differences (SSD) are used to provide a direct comparison. A significant improvement can be seen when shape contexts are used to provide the distance measure, resulting in an error rate as low as 3.8% compared to 7.7% for SSD for 2000 training images.[2]

Case study 2: 3D object recognition The second experiment involves 20 common household objects selected from the COIL-100 database [7]. Each object was photographed on a turntable with rotation increments of $5°$ for a total of 72 views per object. Each image is gray-scale and 128×128. We prepared our training sets by selecting a number of equally spaced views for each object. The remaining views were then used for testing. Fig. 4(a) shows the performance of shape context matching (SC) compared to SSD using 1-NN. The shape context tests were performed with the same settings as in the digit experiment using 100 points randomly sampled from the Canny edges of each image. SSD is known to perform very well on this database due to the lack of variation in lighting [4]. Our method, being dependent on features abstracted away from the raw image brightnesses, does not share this sensitivity. Naturally one could benefit from combining appearance based features with shape contexts, but in the present work we focus exclusively on shape.

Beyond recognition, shape context allows for the definition of a generic shape similarity measure. In [2] we exploited this property in the context of image retrieval. Here we demonstrate a clustering application which allows us to select a set of prototypical images for a given class, an application known as *editing*. We rely on a grouping technique for pairwise data known as K-medoids. K-medoids can be seen as a variant of K–means that restricts prototype positions to data points,

[2] For Euclidean k-NN an error rate of 5.0% using 60,000 training images is reported [6].

but it readily generalizes to arbitrary similarity data. Concretely, first a matrix of pairwise similarities between all possible prototypes is computed and stored. For a given number of K prototypes the K-medoid algorithm then iterates two steps: (i) For a given assignment of points to (abstract) clusters a prototype is selected by minimizing the average distance of the prototype to all elements in the cluster, and (ii) given the set of prototypes, points are then reassigned to clusters according to the nearest prototype. Though heuristic at a first glance this scheme can be made rigorous by deriving a joint cost function for both steps.

In the recognition context this technique can also be used to optimally allocate resources, i.e. more prototypes are allocated to difficult shapes. In this case we run separate clustering algorithms for each category. We employ a splitting strategy, however, we choose the cluster to split based on the associated overall misclassification error, thus coupling the different editing processes. Two examples of the prototypes selected using this method in the COIL experiment are shown in Fig. 4(b). The curve marked SC' in Fig. 4(a) shows the improved classification performance using this prototype selection strategy instead of equally-spaced views.

5 Conclusion

We have presented a new approach to computing shape similarity and correspondences based on the shape context descriptor. Shape context is simple and easy to apply, yet provides an extraordinarily rich descriptor for point sets greatly improving point set registration, shape matching and shape recognition. In our experiments we have demonstrated invariance to several common image transformations, including significant 3D rotations of real-world objects.

Acknowledgments This research is supported by (ARO) DAAH04-96-1-0341, the Digital Library Grant IRI-9411334, an NSF graduate Fellowship for S.B and the German Research Foundation by Emmy Noether grant PU-165/1.

References

[1] Y. Amit, D. Geman, and K. Wilder. Joint induction of shape features and tree classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(11):1300–1305, November 1997.

[2] S. Belongie and J. Malik. Matching with shape context. In *IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL-2000, to appear)*, 2000.

[3] S. Gold et al. New algorithms for 2D and 3D point matching: pose estimation and correspondence. *Pattern Recognition*, 31(8), 1998.

[4] D.P. Huttenlocher, R.H. Lilien, and C.F. Olson. View-based recognition using an eigenspace approximation to the Hausdorff measure. *PAMI*, 21(9):951–955, Sept. 1999.

[5] R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38:325–340, 1987.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

[7] H. Murase and S.K. Nayar. Visual learning and recognition of 3-d objects from appearance. *Int. Journal of Computer Vision*, 14(1):5–24, Jan. 1995.

[8] M. Turk and A.P. Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–96, 1991.

[9] R. C. Veltkamp and M. Hagedoorn. State of the art in shape matching. Technical Report UU-CS-1999-27, Utrecht, 1999.