Matching with Shape Contexts

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

Abstract

We introduce a new shape descriptor, the shape context*, for measuring shape similarity and recovering point correspondences. The shape context describes the coarse arrangement of the shape with respect to a point inside or on the boundary of the shape. We use the shape context as a vector-valued attribute in a bipartite graph matching framework. Our proposed method makes use of a relatively small number of sample points selected from the set of detected edges; no special landmarks or keypoints are necessary. Tolerance and/or invariance to common image transformations are available within our framework. Using examples involving both silhouettes and edge images, we demonstrate how the solution to the graph matching problem provides us with correspondences and a dissimilarity score that can be used for object recognition and similarity-based retrieval.*

1. Introduction

This paper addresses two related problems: measuring shape similarity and recovering point correspondences. The shapes in our study include silhouettes as well as line drawings with internal markings. We introduce a new shape descriptor, the *shape context*, to describe the coarse arrangement of the shape with respect to a point inside or on the boundary of the shape. The shape context can be combined with a conventional appearance-based descriptor, such as local orientation. We then use the combined descriptor as a vector-valued attribute in a bipartite graph matching framework. Our proposed method makes use of a relatively small number of sample points selected from the set of detected edges. No special landmarks or keypoints are necessary. We demonstrate by means of example how the solution to the graph matching problem provides us with correspondences and a similarity score that can be used for object recognition and similarity-based querying. In this regard, our contribution can be seen as a module that could add functionality in a standard content-based retrieval system (e.g. [3, 17, 12, 1]).

The matching method we propose operates on 2D images. Though there are certain applications in image retrieval where the media is inherently 2D (e.g trademarks or fonts), ultimately our interest lies in matching 3D objects from multiple 2D views. Since it is generally impractical to have access to arbitrarily many views of an object, we need a metric for similarity that is reasonably tolerant to pose variations and occlusion.

The structure of this paper is as follows. We first discuss related work in Section 2. Next, we introduce the shape context in Section 3 and our method for matching in Section 4. We then summarize our experimental results in Section 5. Finally, we conclude in Section 6, where we outline how this approach might be integrated into a general image retrieval system using automatically segmented images.

2 Related Work

The related work on shape matching divides into three main categories, which we discuss next.

2.1 Silhouettes

A great deal of research on shape similarity has been done using the boundaries of silhouette images. Since silhouettes do not have holes or internal markings, the associated boundaries are conveniently represented by a single closed curve. Since this constraint is too restrictive for the aims of our present work, we will only briefly discuss a few works in this area.

Much of the early work on silhouette boundaries was done using Fourier descriptors, e.g. [32, 18]. A representative of more recent works is that of Gdalyahu and Weinshall [5], which is a dynamic programming-based approach that uses the edit distance between curves. Their algorithm is fast and invariant to several kinds of transformation including some articulation and occlusion.

A dual approach to the problem of matching silhouette boundaries can be found, for example, in the work of Kimia et al. [24]. Their approach makes use of the medial axis representation and achieves similar results to those in [5]. The use of the medial axis shows promise from the point of view of exploiting topology and reasoning about parts, but is inherently sensitive to occlusion and noise.

2.2 Image Matching

Image matching or appearance-based methods offer a complementary view to silhouette-based methods. Rather than focus on the shape of the occluding contour, these approaches make direct use of the brightness values within the visible portion of the object. A measure of similarity can be computed via simple correlation or, as is done in the so-called "eigenimage" approaches, by comparing projections onto a set of principal components. Much of the work in the latter area involves face recognition, e.g. [25, 26]. Murase and Nayar have applied these ideas to multiple 2D views of 3D objects [14], and this has been an area of continued activity.

Though the eigenimage framework will most certainly persist at least as a component of future image retrieval and object recognition systems, it is inherently too sensitive to small changes in illumination and pose to be practicable as a general solution.

2.3 2D Point Sets

The last category is that of 2D point set methods. We use this designation to refer to approaches for which the input representation is a set of extracted feature points (e.g. edge elements or corners) which are regarded as lying in the plane rather than along a parameterized curve.

These methods benefit from being abstracted away from the raw pixel brightnesses since the feature detection can be made invariant to common variations in illumination. Moreover, the lack of dependence on a 1D arclength parameterization makes these methods in principle more tolerant to linebreaks due to small occlusions or noise.

Huttenlocher et al. developed a method in this category based on the Hausdorff distance [8], yielding a significant improvement over binary correlation. The method can be extended readily to deal with partial matching and clutter. One drawback, however, is that it is highly dependent on having closely spaced poses in the training set [9], as was true for Murase and Nayar [14]. The entities considered in the Hausdorff distances are the points themselves – edge elements, in this case – with no associated local or global encoding of shape, the inclusion of which could allow one to encode invariance to changes in pose. Another drawback for our purposes is that the method does not return correspondences, though this can be seen as a benefit from the standpoint of speed. Methods based on distance transforms, such as [4], are similar in spirit and behavior.

The work of Sclaroff and Pentland [21] makes the shape description for points on an object more explicit through the physical analogy of a finite element spring-mass model. The corresponding eigenmodes were shown to be effective for recovering accurate correspondences and indexing into databases based on shape similarity. Though the method appears to be highly tolerant to deformation, the problem of occlusion is not addressed. Earlier examples of approaches in this vein can be found in [23, 22, 27]. An alternative approach along the same lines has been developed by Gold and Rangarajan using deterministic annealing [7, 6].

The method of deformable templates [31] lies somewhere between the 2D point set and appearance-based categories. A deformable template is a parameterized model, e.g. of a human eye, that searches for a good fit in the surface of the grayscale image. This is a very flexible approach in that invariance to certain kinds of transformations can be built into the measure of model similarity, but it suffers from the need for hand-designed templates and the sensitivity to initialization when searching via gradient descent.

Geometric hashing and indexing techniques, e.g. [2, 11] have also seen some success in the shape matching application. Since these methods can be viewed as speeding up algorithms that can be described in more straightforward but less efficient terms, we do not focus on the details of these systems here.

For a more detailed discussion of the variety of shape matching techniques, the reader is referred to [28].

3. Introducing the Shape Context

In order to compute shape correspondences and similarity, one must start by defining a shape descriptor. In analogy to the stereo matching problem, we would like to have descriptors that can be computed in one image and used to find corresponding points, if visible, in another image.

3.1 The Basic Idea

The shape context analysis begins by taking N samples from the edge elements on the shape. These points can be on internal or external contours. They need not, and typically will not, correspond to keypoints such as maxima of curvature or inflection points. We prefer that the samples be roughly uniform in 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 use of the silhouette-based methods mentioned in Section 2.1 due to its internal contour.

Now consider the vectors originating from a point in Figure 1(d) 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 vectors is a rich description, since as N gets large, the representation of the shape becomes exact.[1]

[1] This assumes the contours have bounded curvature; we are not considering fractals and such.

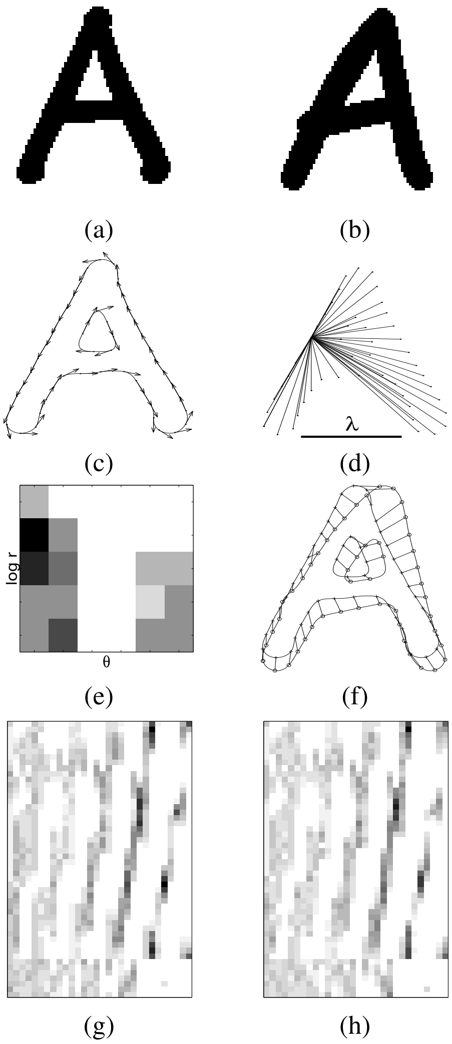

Figure 1. Shape context computation and graph matching. (a,b) Original image pair. (c) Edges and tangents of first letter with 50 sample points. (d) Vectors from a sample point (at left, middle) to all other points. The median distance λ for all N^2 point pairs is shown at bottom for reference. (e) $\log r, \theta$ histogram of vectors in (d), with 5 and 6 bins, respectively. (Dark=large value.) (f) Correspondences found using Hungarian method, with weights given by sum of two terms: histogram dissimilarity and tangent angle dissimilarity. (g,h) The "shape contexts" for the two letters, formed by flattening and concatenating the histograms for all points in each image; each shape context has 50 rows, one per sample point, and 30 columns, one for each histogram bin.

Since our goal is to match shapes that vary from one instance to another, to use the full set of vectors as a shape descriptor is inappropriate.

Our solution is to produce a compact descriptor for each sample point by computing a coarse histogram of the relative coordinates of the remaining points. The reference orientation for the coordinate system can be absolute or relative to a given axis; this choice depends on the problem setting. For now we will assume an absolute reference orientation, i.e. angles measured relative to the positive x-axis.

Since we would like the histogram to distinguish more finely among differences in nearby pixels, we propose to use a log-polar coordinate system. An example of a such a histogram computed for the set of vectors in Figure 1(c) is shown in Figure 1(e); dark pixels denote larger values. We call this histogram the *shape context*. In this example we have used 6 equally spaced angle bins and 5 equally spaced log-radius bins. The complete set of N shape contexts are shown, flattened and concatenated, in Figures 1(g) and (h) for the two sample images in (a) and (b).

3.2 Incorporating Local Appearance

While the shape context encodes a description of the density of boundary points at various distances and angles, it may be desirable to include an additional feature that accounts for the local appearance of the reference point itself. Such features might include local orientation, vectors or filter outputs, color histograms, and so on (see e.g. [19, 29, 10, 20]). In our experiments, we illustrate this concept using the simple feature of local orientation.

3.3 Incorporating Invariances

Depending on the application, it may be necessary to have invariance to certain image transformations. Note that in some cases, as when distinguishing 6 from 9, complete invariance (in this case, to rotation) impedes recognition performance. We now discuss how these invariances can be addressed by our system.

3.3.1 Translation

Invariance to translation is intrinsic to the shape context definition since everything is measured with respect to points on the object. If translation-variant descriptors are desired, an absolute reference frame with a fixed origin can be used.

3.3.2 Scale

A number of choices are available for achieving varying degrees of scale invariance. The method we use is to normalize all radial distances by the median distance λ between all N^2 point pairs in the shape, since the median is reasonably robust to outliers. An example of the median length is shown in Figure 1(d).

3.3.3 Rotation

Invariance to rotation can be obtained in several ways. The simplest method is to measure all angles for an object relative to its axis of least inertia. This axis will be arbitrary, however, for shapes that are not elongated. A more general solution is to measure all angles relative to the tangent angle at each sample point. In this frame, which we refer to as the *relative frame*, the shape context definition becomes totally rotation invariant. The drawback is that if one wishes to incorporate rotation-variant local appearance features (e.g. local orientation), their discriminative power is lost. In order to fix this problem, one might make use of a voting scheme such as pose clustering to guess a global reference frame. We intend to explore solutions to this problem in future work.

3.3.4 Occlusion

Occlusion presents a comparatively more difficult problem than the preceding cases since it involves a loss of information. Much work is still required in the area of object recognition to fully address this problem; we do not claim to have solved it. However, certain characteristics of our approach help to minimize its impact. The use of log-distances places more dependence on nearby pixels, so that pixels far from the occluded edge still have a chance of finding a good correspondence in the model image. Though the correspondences will be poor for the occluded points, we can exploit the good correspondences either to iterate and repeat the match only using the hypothesized unoccluded regions, or simply use a partial cost to achieve some robustness to outliers. We intend to explore this in future work.

4. 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. We will start with the first criterion.

The matching cost is comprised of two terms: one for shape context and the other for local appearance. The shape context term C_S is given by the χ^2 distance between the two histograms. If we denote the two K-bin (normalized) histograms by $g(k)$ and $h(k)$, the χ^2 distance is given by

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

Its value ranges from 0 to 1.

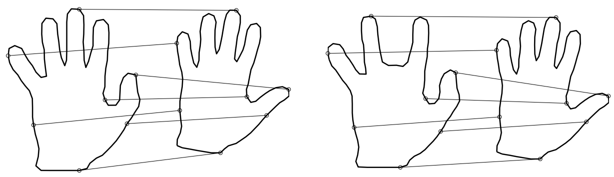

Figure 2. Correspondences between selected points on a pair of hands, with and without "digital surgery" (after [21]). The number of samples used was 100.

The local appearance term C_A, which in our case represents a measure of tangent angle dissimilarity, is defined by

$$C_A = \frac{1}{2}\left\| \begin{pmatrix} \cos(\theta_1) \\ \sin(\theta_1) \end{pmatrix} - \begin{pmatrix} \cos(\theta_2) \\ \sin(\theta_2) \end{pmatrix} \right\|$$

In other words, it is half the length of the chord in the unit circle between the unit vectors with angles θ_1 and θ_2. Its value ranges from 0 to 1.

The combined matching cost is computed via a weighted sum: $C = (1-\beta)C_S + \beta C_A$. We assume that the same number of samples (N) are taken from each shape.

Given the set of costs $C_{i,j}$ between all pairs of points, we can proceed to address the uniqueness criterion. Our objective is 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, and it is solved in $O(N^3)$ time by the Hungarian method [16].

The input to the Hungarian method is a square cost matrix with entries $C_{i,j}$ representing the cost of matching node i in the first image to node j in the second. 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 method to the letter-A example is shown in Figure 1(f), where we have used $\beta = 0.3$. Another example with the same settings (except that $N = 100$) is shown in Figure 2 for a few selected points on a pair of hands.

5. Experiments

In the following experiments, the dissimilarity score is defined as the total cost $\sum_i C_{i,\pi(i)}$ and the bin definitions are as follows: 12 equally spaced angle bins, from 0° to 360°, and 5 log-spaced radius bins from 0.125λ to 2λ. Radius values inside 0.125λ or outside 2λ are assigned to the first or last radius bin, respectively. In each experiment the number of samples is $N = 100$.

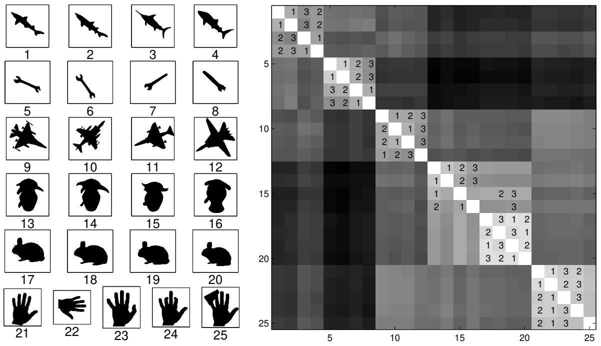

Figure 3. Left: Kimia dataset. Right: Dissimilarity matrix. The digits in each row indicate the three nearest neighbors.

5.1 Silhouettes I

The first silhouette database [24] consists of 25 objects divided into 6 classes. They are shown in Figure 3. In this experiment we used the relative frame (see Section 3.3.3) and $\beta = 0$.

The results are summarized in Figure 3, which shows the matching cost found using the Hungarian method for all 25^2 shape pairs. In [24] and [5], the authors summarize their performance on this dataset in terms of the number of 1st, 2nd, and 3rd nearest neighbors that fall into the correct category. Our performance measured in this way is 25/25, 24/25, 23/25. The results reported in [24] and [5] are 23/25, 21/25, 20/25 and 25/25, 21/25, 19/25, respectively.

5.2 Silhouettes II

Our next experiment makes use of a set of 56 silhouettes scanned from the index of the Audubon North American Mammals field guide [30]. Limited space does not permit us to show the complete set. Since these shapes do not divide into strict classes, our experiments are designed to show the ability of our system to sort shapes by similarity. A subset of the results, using absolute frame and $\beta = 0.3$, are shown in Figure 4.

5.3 Columbia Object Database

The COIL-100 database [15] consists of 100 color photographs of objects on a turntable with rotations in azimuth of $0°$ through $360°$ with $5°$ increments. For our tests, we converted the images to grayscale and selected three views per object $(-15°, 0°, 15°)$ for a total of 300 images. We extracted the Canny edges using the same threshold settings for the whole set. We then compared the 300 images

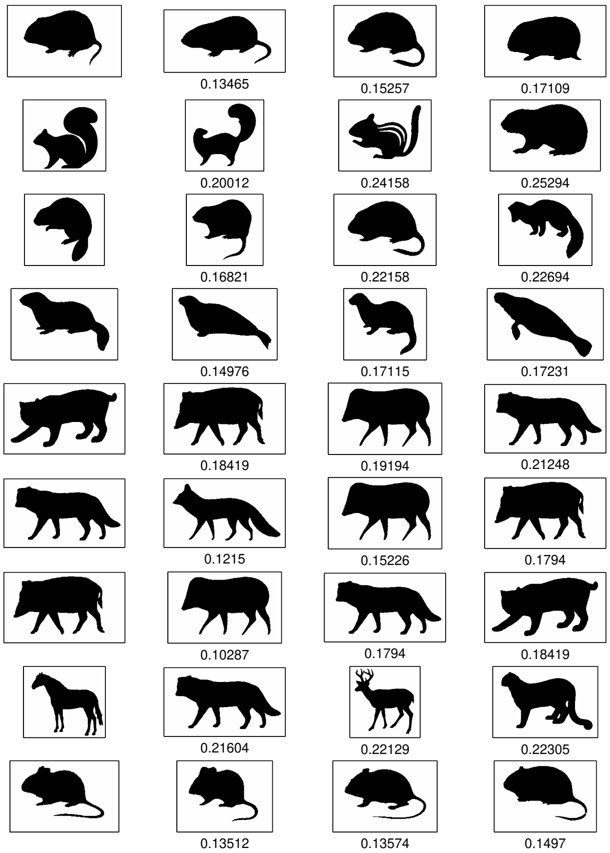

Figure 4. Mammal silhouettes. Column 1: query images. Columns 2-4: matching images, sorted by similarity.

exhaustively. For these tests, we used $\beta = 0.3$, absolute frame, and a fixed value of $\lambda = 50$.

To measure performance, we counted the number of times the closest match was a rotated view of the same object. The result is $280/300$. Of the 20 misses, 11 are due simply to the absence of color as a feature (e.g. red/green peppers, orange/green Rolaids bottles, yellow/green packs of gum). Of the remaining 9, the majority of the closest matches occur within similar object categories. The three closest matches for 10 examples are shown in Figure 5.

6. Conclusion

We have presented a new approach to computing shape similarity and correspondences in a graph-matching framework. In our experiments we have demonstrated invariance to several common image transformations, including significant 3D rotations of real-world objects. The problem of occlusion remains to be explored.

A major focus in our ongoing work is to incorporate

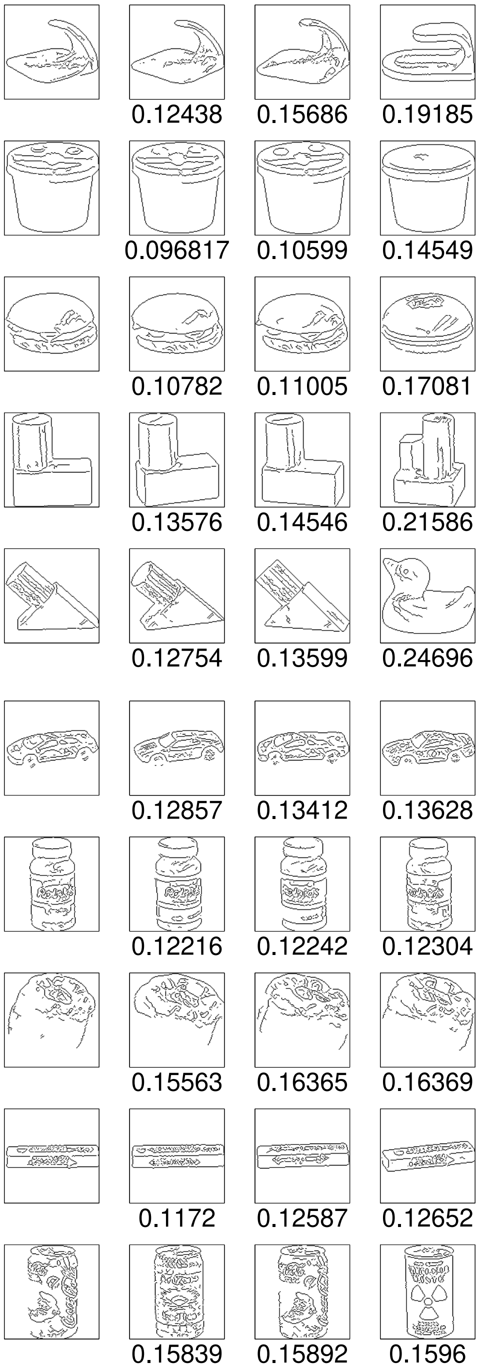

Figure 5. Three closest matches for ten example images in the COIL-100 database. Query images appear in the first column; matched images are shown to the right, with matching cost shown below. Rows 1-5: closest match correct. Rows 6-10: closest match incorrect. We observed that the majority of the matches labelled "incorrect" occurred between objects in similar categories (e.g. convertible/sedan car) or of different colors (e.g. red/green pepper).

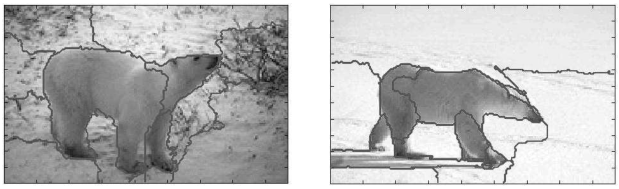

Figure 6. Future directions. Shown here is a pair of images segmented using the method of Malik et al [13]. By combinatorially hypothesizing merged regions, and with the help of color and texture cues, we aim to combine our shape matching technique with an automatic segmentation system to improve object recognition and retrieval performance.

the methods discussed here, which assume that the image has been segmented, into a system with automatically segmented real-world images. For illustrative purposes, Figure 6 shows a pair of images segmented using the algorithm from [13]. In order to apply a shape matching technique to this problem, one must first hypothesize merged regions from each image (or only one, if the model image is labeled) to compensate for oversegmentation, which may be due to error or to valid internal boundaries in multi-part objects. Since the number of segments is quite small relative to the original number of pixels, one can reasonably apply combinatorial methods, influenced of course by color or texture cues, contiguity, and so on. We intend to explore these directions in our ongoing work.

Acknowledgements

This research is supported by (ARO) DAAH04-96-1-0341, the Digital Library Grant IRI-9411334 and an NSF graduate Fellowship for S.B. We would like to thank Alyosha Efros, Thomas Leung, and Jan Puzicha for helpful discussions and Niclas Borlin of Umeå University for his Matlab code for the assignment problem.

References

[1] S. Belongie et al. Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In *Proc. 6th Int. Conf. Computer Vision*, Bombay, India, Jan. 1998.

[2] A. Califano and R. Mohan. Multidimensional indexing for recognizing visual shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(4):373–392, 1994.

[3] M. Flickner et al. Query by image and video content: The qbic system. *IEEE Trans. Computers*, 28(9):23–32, Sept. 1995.

[4] D. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In *Proc. 7th Int. Conf. Computer Vision*, pages 87–93, 1999.

[5] Y. Gdalyahu and D. Weinshall. Flexible syntactic matching of curves and its application to automatic hierarchical classification of silhouettes. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(12):1312–1328, 1999.

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

[7] S. Gold and A. Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(4), 1996.

[8] D. Huttenlocher, G. Klanderman, and W. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(9):850–863, Sept. 1993.

[9] D. Huttenlocher, R. Lilien, and C. Olson. View-based recognition using an eigenspace approximation to the hausdorff measure. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 21(9):951–955, Sept. 1999.

[10] D. Jones and J. Malik. Computational framework for determining stereo correspondence from a set of linear spatial filters. *Image and Vision Computing*, 10:699–708, 1992.

[11] Y. Lamdan, J. Schwartz, and H. Wolfson. Affine invariant model-based object recognition. *IEEE Trans. Robotics and Automation*, 6:578–589, 1990.

[12] W. Ma and B. Manjunath. Netra: A toolbox for navigating large image databases. In *Proc. Int. Conf. Image Processing*, 1997.

[13] J. Malik, S. Belongie, J. Shi, and T. Leung. Textons, contours and regions: Cue integration in image segmentation. In *Proc. 7th Int. Conf. Computer Vision*, pages 918–925, 1999.

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

[15] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (coil-100). Technical Report CUCS-006-96, Columbia Univ., 1996.

[16] C. Papadimitriou and K. Stieglitz. *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, 1982.

[17] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *Int. Journal of Computer Vision*, 18(3):233–254, June 1996.

[18] E. Persoon and K. Fu. Shape discrimination using Fourier descriptors. *IEEE Trans. Systems, Man and Cybernetics*, 7(3):170–179, Mar. 1977.

[19] B. Schiele and A. Pentland. Probabilistic object recognition and localization. In *Proc. 7th Int. Conf. Computer Vision*, pages 177–182, 1999.

[20] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(5):530–535, May 1997.

[21] S. Sclaroff and A. Pentland. Modal matching for correspondence and recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 17(6):545–561, June 1995.

[22] G. Scott and H. Longuet-Higgins. An algorithm for associating the features of two images. *Proc. of the Royal Soc. London*, B-244:21–26, 1991.

[23] L. S. Shapiro and J. M. Brady. Feature-based correspondence: an eigenvector approach. *Image and Vision Computing*, 10(5):283–288, June 1992.

[24] D. Sharvit, J. Chan, H. Tek, and B. Kimia. Symmetry-based indexing of image databases. *J. Visual Communication and Image Representation*, 1998.

[25] L. Sirovich and M. Kirby. Low dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A*, 4(3):519–524, 1987.

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

[27] S. Umeyama. An eigen decomposition approach to weighted graph matching problems. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(5):695–703, Sept. 1988.

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

[29] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *Proc. 6th Europ. Conf. Comput. Vision*, June 2000. Dublin, Ireland.

[30] J. Whitaker, Jr. *Natl. Audubon Soc. Field Guide to North American Mammals*. Knopf, 1997.

[31] A. Yuille. Deformable templates for face recognition. *J. Cognitive Neuroscience*, 3(1):59–71, 1991.

[32] C. Zahn and R. Roskies. Fourier descriptors for plane closed curves. *IEEE Trans. Computers*, 21(3):269–281, March 1972.