Combining class-specific fragments for object classification

Erez Sali and Shimon Ullman
Department of Computer Science and Applied Mathematics.
The Weizmann Institute of Science.
Rehovot, Israel 76100
email: erez,shimon@wisdom.weizmann.ac.il

Abstract

We describe an approach to object classification based on the conjunction of multiple class-specific object fragments detected in the image. The method represents members of a given class (such as a face, or a car) using combinations of common sub-structures, termed fragments. These fragments are partial 2-D patterns extracted from examples views of objects belonging to the class in question. An object view is covered by multiple, overlapping fragments of several types, and at multiple levels of complexity. We describe the detection of the individual fragments, and the combination of the fragments to detect complete objects. The combination of fragments to form a consistent overall arrangement is based in this scheme on a number of simple mechanisms: the use of overlapping fragments, a spatial `voting' scheme, and by imposing some constraints on the tolerated location of the fragments within the overall object view.

We present experimental results of the application of the method to the detection of face and car views in cluttered scenes and to partially occluded objects. We present evidence that by combining fragments from different objects the method can deal successfully with intra-object variability within a class. The method is more economical and more resistant to occlusion and deformations than methods relying on global object views.

Introduction

In this paper we study the challenging task of detecting different objects from a given class (such as a face or a car) in an image. In addition to the unknown location and illumination of the object, the method must deal with intra-class variability between objects from the same class. The detection process must therefore cover a range of possible shapes, missing parts and additional clutter.

To deal with the problem of shape variability and detect novel shapes of a given class, Turk & Pentland (1991) used the principal components of registered face views. Views of novel objects can be approximated by the superposition of several basis functions, or 'eigenfaces'. Poggio & Sung (1994) used a distribution-based modeling scheme for detecting faces in cluttered scenes. They represented a face view as a gray level vector with 283 components, and trained a multi-layered preceptron network to classify such data as face/non-face vectors. The generalization to novel shapes within the class is obtained in these schemes by the inherent generalization capacity of the neural network mechanism. Rowley, Baluja & Kanade (1995) and Lin, Kung & Lin (1996) performed face detection by training a neural network to distinguish between face and non-face intensity windows of 20x20 pixels. They applied the network classifier to enhanced 20x20 windows that were taken from the input image at several resolutions.

A number of earlier schemes suggested the use of 3-D parts such as generalized cylinders (Binford 1971, Marr and Nishihara 1978, Marr 1982), Geons (Beiderman 1985) or superquadratics (Bajcsy and Solina 1987, Pentland 1987) for the detection of objects from a given class. They suggested the extraction of 3-D shape and the construction of object-centered description of the object for recognition under new viewing positions. Our approach is different in that it uses 2-D fragments rather then 3-D parts, and it does not use fixed and distinct parts (such as generalized cylinders), but a large set of overlapping fragments. Figure 2 shows examples of car- and face-fragments of this type, used in our scheme.

Other past methods suggested to describe object classes using parameterized description of 2-D parts. For example, several schemes (Brookes 1981, Grimson 1987, Yuille, Cohen and Hallinan 1989, Cootes et. al. 1992, Jain et. al. 1996, and Baker, Nayar and Murase 1997) suggested the detection of parts by deformable templates for object classification. A number of other methods suggested the representation of object classes in terms of their decomposition into skeletal parts and the relations between the parts. (Kupeev and Wolfson 1994, Zhu and Yuille 1994 and Siddiqi and Kimia 1996), or by partitioning the silhouette of objects from the class to parts (Latecki and Lakämper 1998).

The combination of simple local features was also suggested for object classification. Amit, Geman and Wilder (1997) used simple feature detectors and decision trees classifier for the classification of objects. Nelson and Selinger (1998) suggested the extraction of boundary segments from training object-views and associating them with the objects viewing parameters.

The use of class-specific fragments

Unlike other methods that use local 2-D features we do not employ universal shape features. Instead, we use object fragments that are specific to a class of objects, taken directly from example views of objects in the same class. The fragments we detect are divided to equivalence sets that contain views of the same region in the object under different transformations and viewing conditions. As discussed later, the use of fragment views achieves better generalization with a smaller number of example views.

Our approach initially detects multiple fragments of object views taken from the same general class (a face, car, etc.), divided into several fragment-types. In the second stage the algorithm verifies that fragments from all the equivalence sets were detected in the image, that the fragments are properly aligned and taken under the same viewing conditions. This is obtained by the combination of two simple methods: by detecting overlapping fragments that are composed of several basic fragments and thus "bind" the basic fragments, and by a "pointing" method that verifies that the fragments are properly aligned.

The fragment-based approach is robust to occlusions since it can rely on the detection of a sufficient subset of the fragments. It can also use high-resolution input and remain efficient because the input size of each detector is much smaller then the size of the entire object.

The use of fragments for the detection of objects poses an inherent problem: how to make sure that the fragments share the same viewing parameters and that they combine coherently into a complete object. This can be approached in two different ways: binding the fragments together by using overlapping fragments as will be demonstrated below, or by extracting the viewing parameters (illumination, rotation, relative position and etc.) and verifying directly that all the fragments share the same parameters. Minsky and Papert's (1969) Perception work gave interesting support to the power of using multiple overlapping fragments to enforce consistency. They proved that patterns in binary images can be recognized uniquely by detecting all possible triplets of points in the image. That is, the collection of all black triplets, without any explicit representation of their spatial relation, provides a unique `signature' of the object. This demonstrates how the recognition of large and complicated object views can be approached by using the detection of relatively small and simple fragments in the object. Mel and Fiser (1998) analyzed the use of subsets of all the possible detectors in text recognition as a simplified example for visual recognition using fragments detection. They analyzed the detection of words by detecting the presence of embedded n-grams of letters. They tested the rate of correct word recognition as a function of the number of n-grams, the number of letters in the detectors, the number of words and the clutter, and showed that reliable detection is obtained even for a small portion of the possible detectors set. Our scheme also uses a similar approach of enforcing the consistent arrangement of the individual fragments by the use of multiple, overlapping fragments.

Overview of the algorithm

Our detection algorithm consists of two main stages. In the first stage basic fragments are detected by comparing the image at each location with several sets of stored fragment views. Each set contains fragments of objects in a class, seen under various viewing conditions. The comparison is performed by combining the results of three comparison criteria: qualitative grey-level based representation, gradient and orientation measures. The second stage verifies that a sufficient subset of fragment-types were detected, and enforces the consistency of the fragments viewing parameters. In addition to the use of overlapping fragments, we check that the fragments are properly aligned by a simple test of their relative position. The consistency of the rest of the parameters such as rotation and illumination is ensured only by the detection of overlapping fragments.

The algorithm is applied to the image at several scales so that object views at different scales can be detected. Each level detects objects at scale differences of ±35%. The combination of several scales enables the detection of objects under considerable changes in their size.

In the following sections we describe the details of the algorithm. We begin by describing the similarity measure used for the fragments detection.

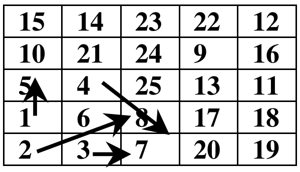
Similarity between image patches

We have evaluated several methods, both known and new, to measure similarity between gray level patches in the stored fragment views and patches in the input image. Many of the comparison methods we tested gave satisfactory results, but we found that a method that combined qualitative image based representation suggested by Bhat & Nayar (1997) with gradient and orientation measures gave the best results. The method measured the qualitative shape similarity using the ordinal order of the pixels in the regions, and measured the orientation difference using gradient amplitude and direction. For the qualitative shape comparison we computed the ordinal order of the pixels in the two regions, and used the normalized sum of displacements of the pixels with the same ordinal order as the measure for the regions' similarity. (See Fig. 1).

The similarity measure $D(F,H)$ between an image patch H and a fragment patch F is a weighted sum of their sum of ordinal displacements d_i, their absolute orientation difference $|\alpha_F - \alpha_H|$ and their absolute gradient difference $|G_F - G_H|$:

$$D(F,H) = k_1 \sum_i d_i + k_2|\alpha_F - \alpha_H| + k_3|G_F - G_H|$$

This measure appears to be successful because it is mainly sensitive to the local structure of the patch and less to absolute intensity values.



15	14	23	22	12
10	21	24	9	16
5	4	25	13	11
1	6	8	17	18
2	3	7	20	19

Fig. 1. *Displacement vectors for four pixels with the highest ordinal order. The displacement vectors are vectors connecting the locations of pixels with similar gray-level ordinal order in the two compared regions. The sum of the four displacements in this case is* $1+\sqrt{5}+1+\sqrt{5}$.

The detection of fragments

For the detection of fragment views in the images we compared the 5x5 gray level patches in each fragment view to the image using the above similarity measure. Only regions with sufficient variability were compared, since in flat regions the gradient, orientation and ordinal-order have little meaning. We allowed flexibility in the comparison of the fragment view to the image by matching each pixel in the fragment view to the best pixel in some neighborhood around its corresponding location. Most of the computations of the entire algorithm are performed at this stage. To speed up the application we reduced the search regions for fragments of each type as the search proceeded. We implemented the ordinal measure calculation on ASP's associative processor (ASP, 1998) and achieved an eight times improvement in speed. An associative processor is especially suitable for such computations since it can process in parallel thousands of pixels.

Merging the detection of the different fragment types

Following the detection of the individual fragments, the final stage of merging the results for the entire object detection is performed. To detect an object only if the fragments are organized properly and are consistent in their viewing conditions, This stage uses the detection of 'binding' fragments as well as the so-called 'pointing' method. It also verifies that fragments from a sufficient subset of fragment-types have been detected, although some occlusion is also allowed for. The 'binding fragments' are fragments with large overlap with other basic fragments such as an eye with a part of a nose, or a lower resolution view of a large part of the object.

In the 'pointing' method each detected fragment (with similarity value above a threshold) "points" to a common anchor region of a possible object. In face detection, for example, we used the tip of the nose as an anchor. A mouth fragment will therefore point up and a forehead fragment down. The total pointing magnitude of the different fragment types is summed for every location in the image. The pointing mechanism is used to integrate the information from all the fragments pointing to a particular location. Each fragment-type points to a particular location with an associated magnitude of $M = W_{Type} \cdot \underset{All\ Part\ ViewsS_i}{Max} (S_i - S_{TH})$ where W_{type} is the weighting factor of the fragment type, S_{TH} a threshold similarity value and S_i the similarity value of all the fragments of that type that point to that location.

Locations that are pointed to by most of the fragment types with high similarity values are candidate object locations. At the final stage we reject some of these locations according to the following rules. First, we reject locations where less then 3/4 of the fragment types were detected. We also compare the image fragments contributing to the candidate location to several low-resolution example views of objects from the class in question. This global filtering proved useful in further enforcing the consistent arrangement of the fragments.

After the restrictions are applied to the merged detection results, we mark locations where the merged results exceed a threshold as final detection locations.

Experimental results

We have tested our algorithm on face and car views. For faces we used a set of 1104 part views, taken from a set of 23 male face views under three illuminations and three horizontal rotations. The parts were grouped by 8 types – eye pair, nose, mouth, forehead, low-resolution view, mouth and chin, single eye and face outline. For cars, we used 153 parts of 6 types. Several examples of the fragments are shown in Figure 2. Figure 3A is the result of the individual fragment detectors that were then merged to yield full face detection in Figures 3B. The images in Figure 6 are additional examples. Note that although the system used fragments taken only from male views under a few illuminations and rotations, it detects successfully male and female face views under a much larger variability in the viewing conditions. Figure 4 demonstrates the detection of a partly occluded face view.

We have tested the rate of correct face detection vs. false detection by applying the algorithm to two images – a complex image of a cathedral (Fig. 7A) that does not contain faces and an image that contains multiple faces (Fig. 7B). We tested our fragment-based method and a full-face detection vs. the number of training examples. First we tuned the detection threshold for both methods so that it will not detect any face in the cathedral image while detecting as many faces as possible in the other image. Then we measured the number of faces that were detected by both methods vs. the number of example faces that were used. The results are shown in the graph 7C. Each of the face views was used both for extracting a template for the detection of a whole face view and for the extraction of eight fragments. The full face templates were of 32x22 pixels. This low-resolution representation is commonly used in other face detectors and yield better detection results then in using full resolution face views. The number of faces that were detected in Fig. 7B vs. the number example views that were used in both methods indicates that the use of multiple fragments performs better then the use of global views for the same amount of training data.

The graph in figure 8 shows the best similarity obtained between a novel fragment of a given type and a set of stored example fragments. It shows that for smaller fragments such as an eye or nose, a small set is sufficient for good approximation, while for large fragments or a full face a large set is required. This illustrates an advantage in the representation of objects as the conjunction of example fragments rather then a global representation

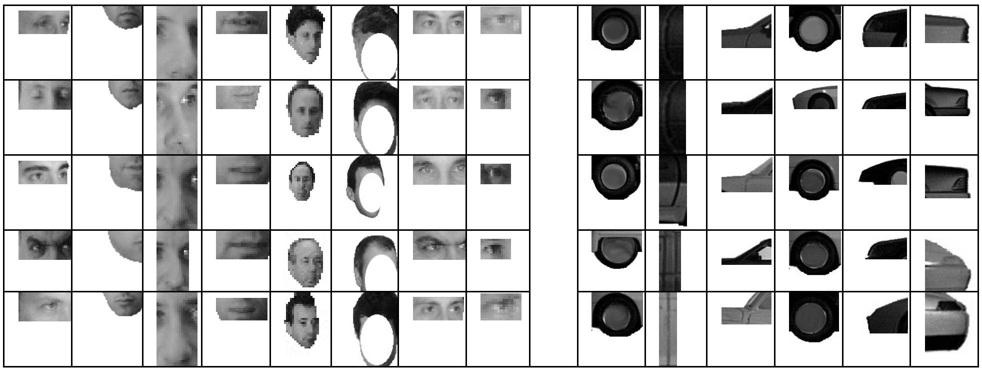


Fig. 2. *Example of fragment views and low resolution views. The fragment views are taken from car and face views of several objects under three illuminations and six horizontal rotations.*

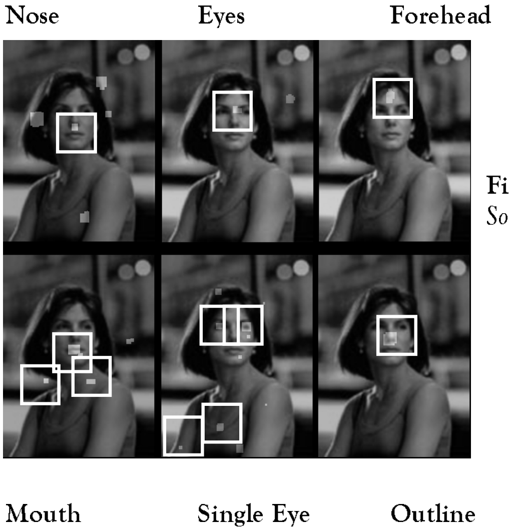


Fig. 3A. *Detection of fragments in an image. Some of the detection is marked in white square*

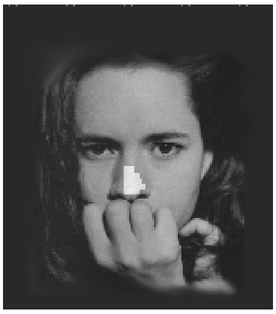


Fig. 3B. *The final face detection.*

Fig. 4. *The detection of occluded face views*

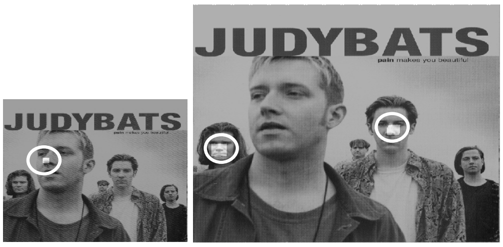


Fig. 5. *The detection of faces in several scales*

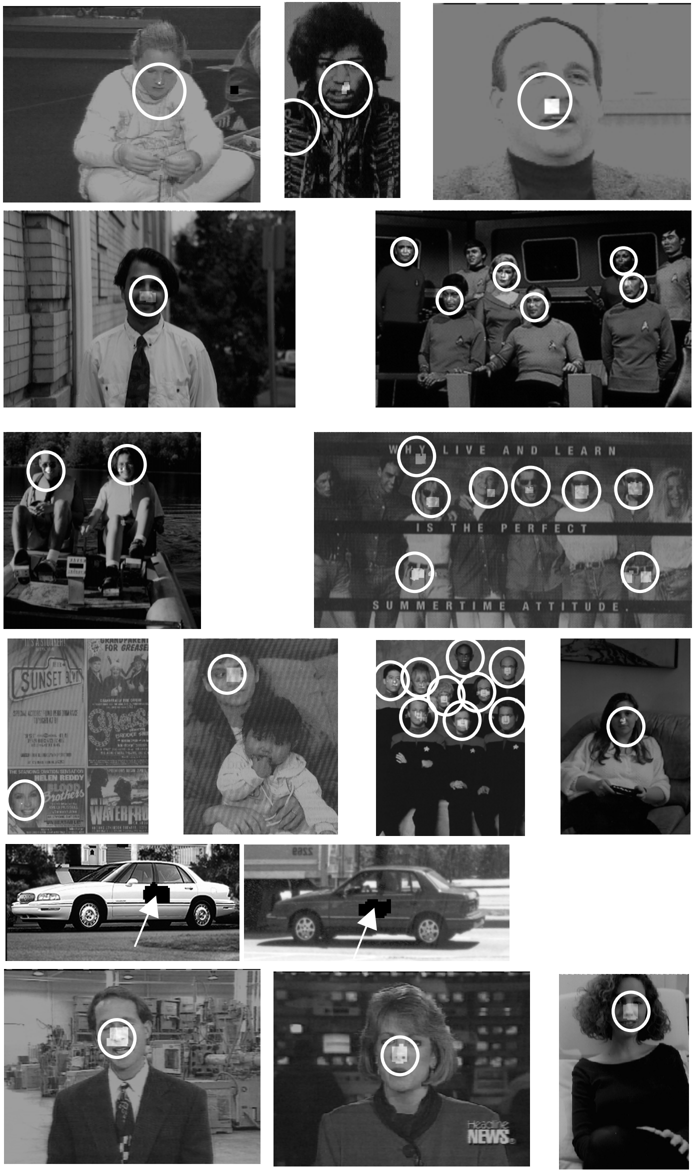
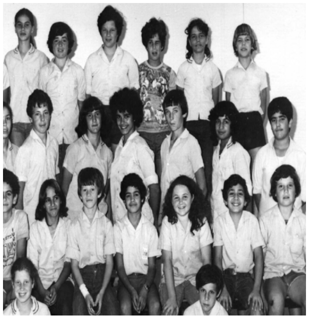
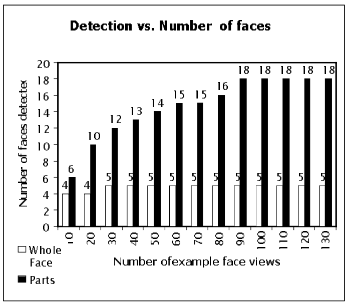


Fig. 6. *Examples of face and car detection (The framed images are taken from the CMU face detector gallery.) Note the detection under different rotations, illuminations and in cluttered scenes.*



(A)

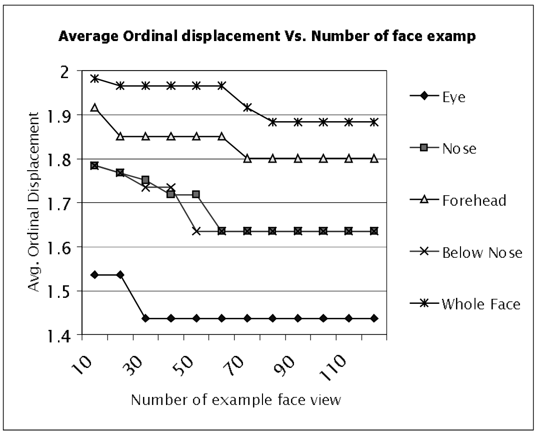
(B)



(C)

Fig. 7. *Detection rate of fragment-based and full face detection, while maintaining zero false detection.*
Image (A) containing 20 faces, (B) contain no faces. System threshold was tuned so that no false detection will occur in image (B), while the maximal number of faces will be detected in (A).
The number of faces detected in (A) vs. the number of training fragments is shown in the following graph (C) as a function of the number of example faces. Each face was used either as a global face template or for the extraction of fragments. The faces were of 23 people under 6 viewing conditions.

Fig. 8. *This graph shows the best similarity of a fragment of each type as a function of the number of example views. For smaller fragments a smaller number of views is sufficient to approximate novel fragments but for larger fragments or full-face views, a larger set of example views is required for a good approximation. This measurement was performed for several fragments of each type and the values presented are average of values obtained in each measurement.*



Applying position limitations on the fragments

To deal with classes of objects with high shape variability, we found it useful to introduce additional constraints on the location of the individual fragments contributing to the over all shape. In classes with high shape variability (for example, flexible and articulated objects), the relative location of the different regions in the object is highly variable, and this limits the use of large fragments with large overlapping areas with other fragments. To deal with this we constrained the location of the individual fragments. These restriction were qualitative, and did not require a strict position of the fragments. For example, in a side view an airplane (Fig. 10), the nose and cockpit fragments were limited to be in some region in the front of the figure, the tail was constrained to a region in the back, and so on .In classes with less variable structure such as faces or passenger cars these positional limitation proved unnecessary, but when we applied our method to objects such as birds, planes, butterflies and the like, the use of some additional limitations proved to be essential. To test these constrains we collected a set of 110 line drawings of objects from nine different classes. Some of them are presented in figure 9. We also collected, on average, 27 fragments for each class, divided into six to 12 fragment types. The drawings were divided to training and testing groups. The fragments were collected from the training drawings only. The testing of the remaining images yielded 91.7% correct classification rate.

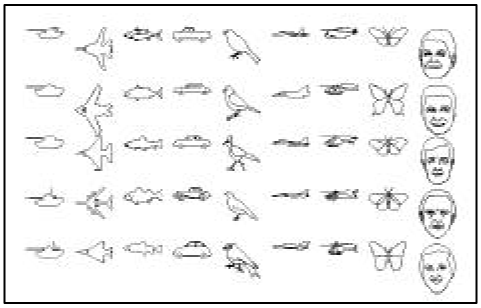


Fig 9. *Line drawing objects*

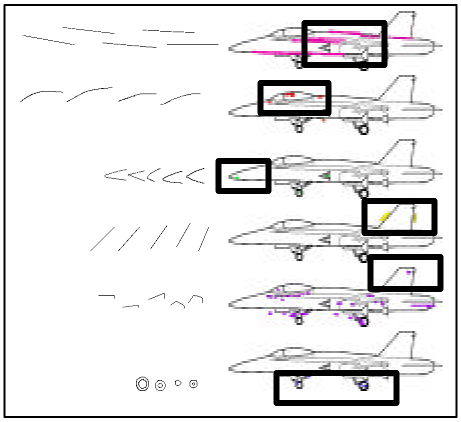


Fig 10. *Detection of plane-fragments in regions. The fragments on the left should be detected in the regions marked by black rectangles. The actual detection is marked by gray spots.*

Summary

The fragment-based method can detect objects of a given class despite large variations in pose, illumination, and object shape. The method is different from previous methods in that it uses multiple 2-D overlapping fragments as fundamental structural building block, that are object-specific and taken from a common class. The detection algorithm initially detects fragment views in the image, and then combines the fragments for the detection of the entire object. The consistency in the fragments viewing parameters is ensured both by the use of overlapping fragments that "bind" the parts together and by testing directly that the fragments are approximately aligned. An important advantage of this method is that it can compensate well for shape variations by matching a novel shape within a class with a new combination of stored fragments. The formation of a new combination also results in a system that uses less training examples compared with the use of global shapes. The scheme uses high-resolution small fragments combined with larger but coarser ones and this allows the implementation to be efficient.

The method is relatively simple because it does not require the estimation of the viewing parameters and does not require the explicit representation and matching of spatial relations. The use of class specific fragments also has a limitation since it means that dealing with a new class of objects will require extending the set of stored object fragments. This raises interesting learning issues, currently under study, concerning the automatic extraction of useful fragments from a set of novel object views.

We also employed a new similarity measure to measure the similarity between a fragment views and regions in the input image. This method measures the change in the qualitative structure of the object view and the change in gradient and orientation. This comparison measure is highly invariant to appearance changes and more resistant to noise then the other methods we have tested.

Acknowledgements

The authors thank T. Kanade, H. A. Rowley and S. Baluja at CMU for providing face images.

This work was supported by the Israel Ministry of Science under the integrated visual platform project.

References

ASP associative processor, http://www.asp.co.il

Amit Y., Geman D., Wilder K., "Joint Induction of Shape Features and Tree Classifiers", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 11, November 1997

Baker S., Nayar K. S., Murase H., "Parametric feature detection", Tech. Rep. CUCS-028-95, school of computer science, Columbia University, 1995

Bajcsy R., Solina F., "Three dimensional object representation revisited", Proc. Of 1-st ICCV London 231-240, 1987

Bhat D., Nayar K. S., "Ordinal measures for image correspondence", *IEEE Trans. on PAMI* Vol. 20 No. 4, 415-423 1998

Biederman I., "Human image understanding: recent research and theory", *Computer Vision, Graphics and Image Processing,* 32:29-73, 1985

Binford T. O. "Visual perception by computer", IEEE conf. on systems and control 1971.

Brooks R., "Symbolic reasoning among 3-D models and 2-D images", Artificial intelligence (17):285-348 1981

Cootes T.F., Taylor C.J., Cooper D.H., Graham J., "Training models of shape from set of examples", Proceeding of the British machine vision conference 1992.

Grimson W. E. L., Recognition of Object Families Using Parametrized Models, Proc. First International Conference on Computer Vision, p. 93-101, 1987.

Jain A., Zhong Y., Lakshmanan S., "Object matching using deformable templates", IEEE on PAMI 18(3) 267-278 1996

Kupeev K. and Wolfson H., "On shape similarity", Proc. Int. Conf. On pattern Recognition 227-237 1994.

Latecki L. J. and Lakämper R., "Convexity Rule for Shape Decomposition Based on Discrete Contour Evolution", Computer Vision and Image Understanding 73, February 1999.

Lin S., Kung S. and Lin L., "Face recognition/detection by probabilistic decision based neural network", submitted to *IEEE trans. on neural networks special issue on artificial networks and pattern recognition.*

Marr D., *Vision,* W.H. Freeman, San Francisco CA, 1982

Marr D., Nishihara H. K. "Representation and recognition of the spatial organization of three dimensional structure" *Proceedings of the Royal Society of London B,* 200:269-294, 1978

Mel W. B., SEEMORE: "Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition", *Neural computation* 9, 777-804 1997

Mel W. B., Fiser J., "Seeing with spatially-invariant receptive fields: When the 'binding problem' isn't", Submitted.

Minsky M. and Papert S., *Perceptrons*, The MIT Press, Cambridge Massachusetts, 1969

Nelson C. R., and Selinger A., "A Cubist approach to object recognition", ICCV98 pp.614-621 1998.

Pentland A., "On recognition by parts", Proc. Of the first ICCV 612-620 1987

Poggio T. and Sung K., "Finding human faces with a gaussian mixture distribution-base face model", *computer analysis of image and patterns*, 432-439, 1995

Rowley H., Baluja S. and Kanade T., "Human face detection in visual scenes", Tech. Rep. CMU-CS-95-158R, school of computer science, Carengie Mellon University, 1995

Sidiqqi K, Kimia B., "A shock grammar for recognition", IEEE conf. On CVPR 1996

Turk M. and Pentland A., "Eigenfaces for recognition", *Cognitive Neuroscience*, 3:71-86, 1990

Yuille A., Cohen D., Hallinan P., "Feature extraction from faces using deformable templates" CVPR 104-109, 1989

Zhu S., Yuille A., "FORMS: A flexible object recognition and modeling system", Harvard TR-94-1 1994