Object Classification Using a Fragment-Based Representation

Shimon Ullman and Erez Sali

The Weizmann Institute of Science
Rehvot 76100, Israel
shimon@wisdom.weizmann.ac.il

Abstract. The tasks of visual object recognition and classification are natural and effortless for biological visual systems, but exceedingly difficult to replicate in computer vision systems. This difficulty arises from the large variability in images of different objects within a class, and variability in viewing conditions. In this paper we describe a fragment-based method for object classification. In this approach objects within a class are represented in terms of common image fragments, that are used as building blocks for representing a large variety of different objects that belong to a common class, such as a face or a car. Optimal fragments are selected from a training set of images based on a criterion of maximizing the mutual information of the fragments and the class they represent. For the purpose of classification the fragments are also organized into types, where each type is a collection of alternative fragments, such as different hairline or eye regions for face classification. During classification, the algorithm detects fragments of the different types, and then combines the evidence for the detected fragments to reach a final decision. The algorithm verifies the proper arrangement of the fragments and the consistency of the viewing conditions primarily by the conjunction of overlapping fragments. The method is different from previous part-based methods in using class-specific overlapping object fragments of varying complexity, and in verifying the consistent arrangement of the fragments primarily by the conjunction of overlapping detected fragments. Experimental results on the detection of face and car views show that the fragment-based approach can generalize well to completely novel image views within a class while maintaining low mis-classification error rates. We briefly discuss relationships between the proposed method and properties of parts of the primate visual system involved in object perception.

1 Classification and the Generalization Problem

Object classification is a natural task for our visual system: we effortlessly classify a novel object as a person, dog, car, house, and the like, based on its appearance. Even a three-year old child can easily classify a large variety of images of many natural classes. In contrast, visual classification proved extremely difficult to reproduce in artificial computer vision system. It is therefore natural to study the mechanisms and processes used by biological visual system for object classification, and to examine the applicability of similar methods to computer vision system. Such studies may lead to the development of better artificial systems dealing with natural images, and can also shed light on what appears to be fundamental differences in the processes of

S.-W. Lee, H.H. Bülthoff, T. Poggio (Eds.): BMCV 2000, LNCS 1811, pp. 73-87, 2000.

visual information by current computer systems on the one hand, and by biological systems on the other.

It is interesting to note in this context the difference between general object classification and the specific identification of individual objects. Classification is concerned with the general description of an object as belonging to a natural class of similar objects, such as a face or a dog, whereas identification involves the recognition of a specific individual within a class, such as the face of a particular person, or the make of a particular car. For human vision, the general classification of an object as a car, for example, is usually easier than the identification of the specific make of the car (Rosch *et al.*1976). In contrast, current computer vision systems can deal more successfully with the task of recognition compared with classification. This may appear surprising, because specific identification requires finer distinctions between objects compared with general classification, and therefore the task appears to be more demanding.

The main difficulty faced by a recognition and classification system is the problem of variability, and the need to generalize across variations in the appearance of objects belonging to the same class. Different dog images, for example, can vary widely, because they can represent different kinds of dogs, and for each particular dog, the appearance will change with the imaging conditions, such as the viewing angle, distance, and illumination conditions, with the animal's posture, and so on. The visual system is therefore constantly faced with views that are different from all other views seen in the past, and it is required to generalize correctly from past experience and classify correctly the novel image. The variability is complex in nature: it is difficult to provide, for instance, a precise definition for all the allowed variations of dog images. The human visual system somehow learns the characteristics of the allowed variability from experience. This makes classification more difficult for artificial system than individual identification. In performing identification of a specific car, say, one can supply the system with a full and exact model of the object, and the expected variations can be described with precision. This is the basis for several approaches to identification, for example, methods that use image combinations (Ullman & Basri 1991) or interpolation (Poggio & Edelman 1990) to predict the appearance of a known object under given viewing conditions. In classification, the range of possible variations is wider, since now, in addition to variations in the viewing condition, one must also contend with variations in shape of different objects within the same class.

In this paper we propose an approach to classification that uses a fragment-based representation. In this approach, images of objects within a class are represented in terms of class-specific fragments. These fragments provide common building blocks that can be used, in different combinations, to represent a large variety of different images of objects within the class. In the next section we discuss the problem of selecting a set of fragments that are best suited for representing a class of related objects, given a set of example images. We then illustrate the use of these fragments to perform classification and deal with the variability in shape between different objects of the same class. We also discuss the problem of coping with variability in the viewing conditions, focusing on the problem of position invariance, with possible application to other aspects of object recognition. Finally, we conclude with some comments about similarities between the proposed approach and aspects of the human visual system.

2 The Selection of Class-Based Fragments

2.1 A Brief Review of Related Past Approaches

Before describing the selection of fragment from a collection of example images, we will review briefly past approaches to recognition, focusing on methods that bear relevance to the approach developed here.

A popular framework to classification is based on representing object views as points in a high-dimensional feature space, and then performing some partitioning of the space into regions corresponding to the different classes. Typically, a set of n different measurements are applied to the image, and the results constitute an n-dimensional vector representing the image A variety of different measures have been proposed, including using the raw image as a vector of grey-level values, using global measures such as the overall area of the object's image, different moments, Fourier coefficients describing the object's boundary, or the results of applying selected templates to the image. Partitioning of the space is then performed using different techniques. Some of the frequently used techniques include nearest-neighbor classification to class representatives using, for example, vector quantization techniques, nearest-neighbor to a manifold representing a collection of object or class views (Murase & Nayar 1995,), separating hyperplanes performed, for example, by Perceptron-type algorithms and their extensions (Minsky & Papert 1969), or, more optimally, by support vector machines (Vapnik 1995). The vector of measurements may also serve as an input to a neural network algorithm that is trained to produce different outputs for inputs belonging to different classes (Poggio & Sung 1995).

More directly related to our approach are methods that attempt to describe all object views belonging to the same class using a collection of fundamental building blocks. The eigenspace approach (Turk & Pentland 1990) belongs to this general approach. In this method, a collection of objects within a class, such as a set of faces, are viewed as a set of vectors constructed from the raw grey level values. A set of principal components is extracted from these images to describe the images economically with minimal residual error. The principal components are used as the building blocks for describing new images within the class, using linear combination of the basic images. For example, a set of `eigenfaces' is extracted and used to represent a large space of possible faces.

In this approach the building blocks are global in nature. Other approaches have used more localized building blocks that represent smaller parts of the objects in question. One well-known scheme is the Recognition By Components (RBC) method (Biederman 1985) and related schemes using generalized cylinders as building blocks (Binford 1971, Marr 1982, Marr & Nishihara 1978). The RBC scheme uses a small number of generic 3-D parts such as cubes, cones, and cylinders. Objects are described in terms of their main 3-D parts, and the qualitative spatial relations between parts. Other part-based schemes have used 2-D local features as the underlying building blocks. These building blocks were typically small simple image features, such as local image patches of 2-5 pixels together with their qualitative spatial relations, (Amit & Geman 1997, Nelson & Selinger 1998), corners the direct output of local receptive fields of the type found in primary visual cortex (Edelman 1993).

2.2 Fragments of Intermediate Complexity in Size and Resolution

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. That is, the shape fragments used to represent faces, for instance, would be different from shape fragments used to represent cars, or letters in the alphabet. These fragments are used as a set of common building blocks to represent, by different combinations of the fragments, different objects belonging to the class. The fragments we detect are divided into equivalence sets that contain views of the same general region in the objects under different transformations and viewing conditions. As discussed later, the use of fragment views achieves superior generalization capability with a smaller number of example views compared with more global methods.

The use of the combination of image fragments to deal with intra-class variability is based on the notion that images of different objects within a class have a particular structural similarity -- they can be expressed as combinations of common substructures. Roughly speaking, the idea is to approximate a new image of a face, say, by a combination of images of partial regions, such as eyes, hairline etc. of previously seen faces. In this section we describe briefly the process of selecting class-based fragments for representing a collection of images within a class. This procedure will be described in more detail elsewhere. In the following section, we describe the use of the fragment representation for performing classification tasks.

Examples of fragments for the class of human faces (roughly frontal) and the class of cars (sedans, roughly side views) are illustrated in Figure 1. The fragments used as a basis for the representation were selected by the principle of maximizing mutual information I(C,F) between a class C and a fragment F. This is a natural measure to employ, because it measures how much information is added about the class once we know whether the fragment F is present or absent in the image. In the ensemble of natural images in general, prior to the detection of any fragment, there is an a-priori probability p(C) for the appearance of an image of a given class C. The detection of a fragment F adds information and reduces the uncertainty (measured by the entropy) of the image. We select fragments that will increase the information regarding the presence of an image from the class C by as much as possible, or, equivalently, reduce the uncertainty by as much as possible. This depends on p(F|C), the probabilities of detecting the fragment F in images that come from the class C, and on p(F|NC) where NC is the complement of C.

A fragment F is highly representative of the class of faces if it is likely to be found in the class of faces, but not in images of non-faces. This can be measured by the likelihood ratio p(F|C) / p(F|NC). Fragments with a high likelihood ratio are highly distinctive for the presence of a face. However, highly distinctive features are not necessarily useful fragments for face representation. The reason is that a fragment can be highly distinctive, but very rare. For example, a template depicting an individual face is highly distinctive: its presence in the image means that a face is virtually certain to be present in the image. However, the probability of finding this particular fragment in an image and using it for making classification is low. On the other hand, a simple local feature, such as a single eyebrow, will appear in many more face images, but it will appear in non-face images as well. The most informative features are therefore fragments of intermediate size, as can be seen in Figures 2. In selecting and using optimal fragments for classification, we distinguish between what we call the `merit’ of a fragment and its `distinctiveness’. The merit is defined by the mutual information

$$I(C,F) = H(C) - H(C/F) \qquad \mathbf{(1)}$$

where I is the mutual information, and H denotes entropy (Cover & Thomas 1991). The merit measures the usefulness of a fragment F to represent a class C, and the fragments with maximal merit are selected as a basis for the class representation. The distinctiveness is defined by the likelihood ratio above, and it is used in reaching the final classification decision, as explained in more detail below. In summary, fragments are selected on the basis of their merit, and then used on the basis of their distinctiveness.

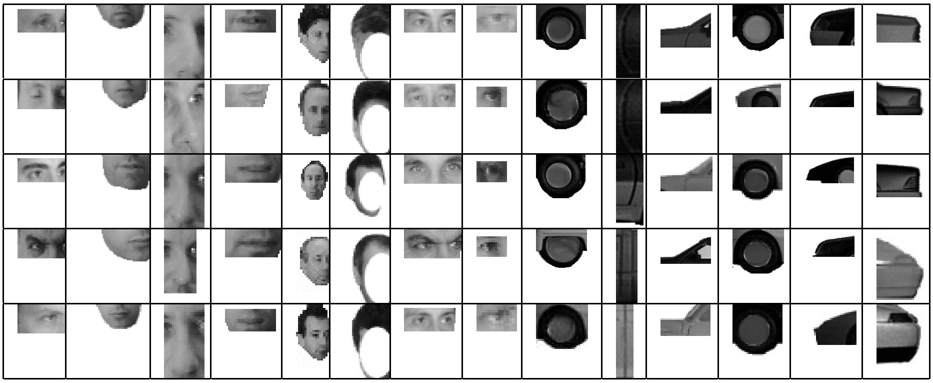


Fig. 1. Examples of face and car fragments

Our procedure for selecting the fragments with high mutual information is the following. Given a set of images, we start by comparing the images in a pairwise manner. The reason is that a useful building block that appears in multiple face images must appear, in particular, in two or more images, and therefore the pairwise comparison can be used as an initial filter for identifying image regions that are likely to serve as useful fragments. We then perform a search of the candidate fragments in the entire database of faces, and also in a second database composed of many natural images that do not contain faces. In this manner we obtain estimations for p(F|C) and p(F|NC) and, assuming a particular p(C), we can compute the fragment’s mutual information. For each fragment selected in this manner, we extend the search for optimal fragments by testing additional fragments centered at the same location, but at different sizes, to make sure that we have selected fragments of optimal size. The procedure is also repeated for searching an optimal resolution rather than size.

We will not describe this procedure in further detail, except to note that large fragments of reduced resolution are also highly informative. For example, a full-face fragment at high resolution in non-optimal because the probability of finding this exact high-resolution fragment in the image is low. However, at a reduced resolution, the merit of this fragment is increased up to an optimal value, at which it starts to decrease. In our representation we use fragments of intermediate complexity in either size of resolution, and it includes full resolution fragments of intermediate size, and larger fragments of intermediate resolution.

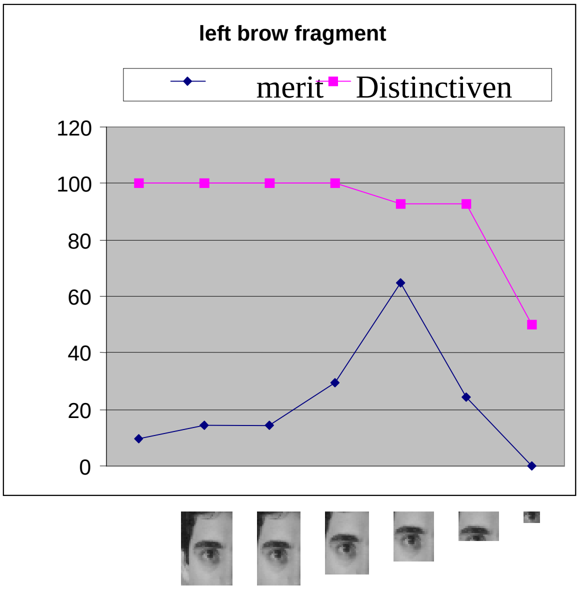


Fig. 2. Selecting optimal fragments by maximizing mutual information. The fragment's merit (based in mutual information) and distinctiveness (based on likelihood ratio) as a function of size. The fragment is optimal at an intermediate size.

3 Performing Classification

In performing classification, the task is to assign the image to one of a known set of classes (or decide that the image does not depict any known class). In the following discussion, we consider a single class, such as a face or a car, and the task is to decide whether or not the input image belongs to this class. This binary decision can also be extended to deal with multiple classes. We do not assume that the image contains a single object at a precisely known position, consequently the task includes a search over a region in the image. We can therefore view the classification task also as a detection task, that is, deciding whether the input image contains a face, and locating the position of the face if it is detected in the image.

The 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 gray-level based representation, gradient and orientation measures. The second stage combines the results of the individual fragment detectors. It verifies that a sufficient subset of fragment-types have been detected, and enforces the consistency of the fragments viewing parameters. The main tool for verifying the consistency is the use of multiple overlapping fragments. With respect to position in the image, we also incorporate a test for rough position that we found experimentally to be helpful. The consistency of the other viewing parameters such as rotation and illumination is ensured only by the detection of overlapping fragments. The algorithm is performed on 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 detection of the basic fragments.

3.1 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 within the subsequent classification algorithm, but we found that a method that combined qualitative image based representation suggested by Bhat and 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. 3).

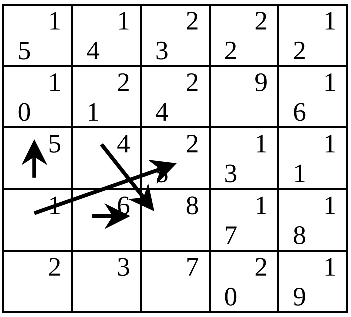
The similarity measure $D(F,H)$ between an image patch H and a fragment patch F is a weighted sum of their sum of ordinal displacements d_i, their absolute orientation difference $|\alpha_F - \alpha_H|$ and their absolute gradient difference $|G_F - G_H|$:

$$D(F,H) = k_1 \sum_i d_i + k_2 |\alpha_F - \alpha_H| + k_3 |G_F - G_H| \qquad \textbf{(2)}$$

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

3.2 The Detection of Fragments

For the detection of fragment views in the images we compared local 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-intensity 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 a speed factor of approximately 8.5. An associative processor is especially suitable for such computations since it can process in parallel thousands of pixels.



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. 3. Displacement vectors for four pixels with the highest ordinal order. The displacement vectors connect 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 $1+\sqrt{5}+1+\sqrt{5}$.

3.3 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 the anchor. A mouth fragment will therefore point up and a forehead fragment down. The overall contributions of the different fragment types are summed for every location in the image. This procedure is used to integrate the information from all the fragments that are arranged in roughly the expected positions around the anchor location. Each fragment-type contributes to the overall sum associated with a particular location with an associated magnitude of M computed by:

$$M = W_{Type} \cdot \underset{All\ Part\ Views S_i}{Max} (S_i - S_{TH}) \tag{3}$$

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 the location in question. The natural choice for the weighting factor is the distinctiveness defined above based on the likelihood ratio.

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 collection of 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.

4 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 used by the algorithm are shown in Figure 1. Figure 4 is the result of the individual fragment detectors that were then merged to yield full-face detection in Figure 5. The images in Figure 6 are additional examples. Note that although the system used only male views in few illuminations and rotations, it detects male and female face views under various viewing conditions. Figure 7 demonstrates the detection of a partly occluded face view.

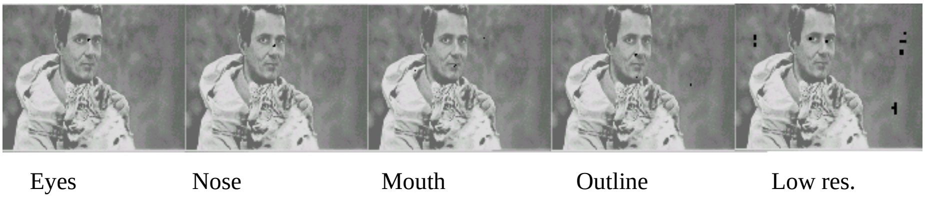


Fig. 4. Detection of individual fragments

We have tested the rate of face detection vs. false detection by applying the algorithm to two images – a complex image of a cathedral (Fig. 8A) that does not contain faces, and an image that contains multiple faces (Fig. 8B). We first fixed the system thresholds so that it will not detect any face in the cathedral image while detecting as many faces as possible in the other image. The number of faces that were detected vs. the number of fragments used in each of the eight fragment types is shown in (Fig. 8C). The fragments were taken from the same set of global views and had less overall "image area cover" thenthe global views. We also tried to detect faces by measuring similarity to low resolution (32x22) face views. The number of faces that were detected in Fig. 8B vs. the number of low-resolution example views that were used for the extraction of fragments and as low-resolution patterns is shown in (Fig. 8C). It indicates that the use of multiple fragments performs much better then the use of global views.

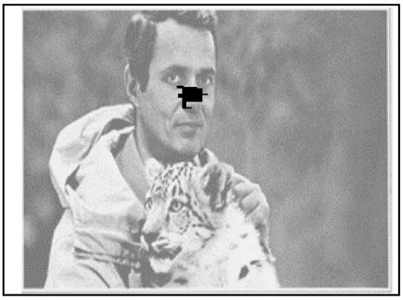


Fig.5. Final face detection

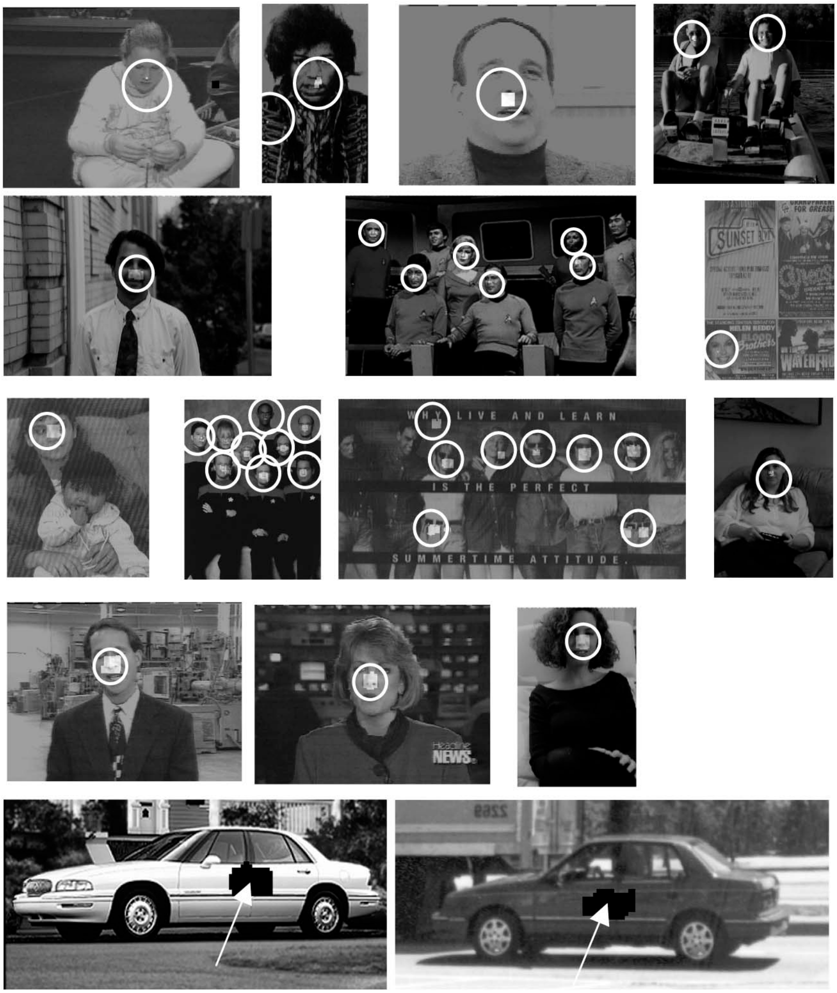


Fig. 6. Examples of face and car detection (Most of the images are taken from the CMU face detector gallery.) Note the detection under different rotations, illuminations and in cluttered scenes.

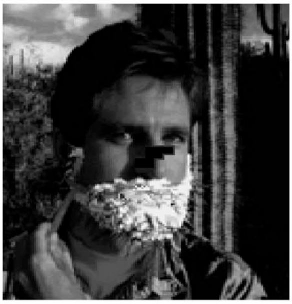
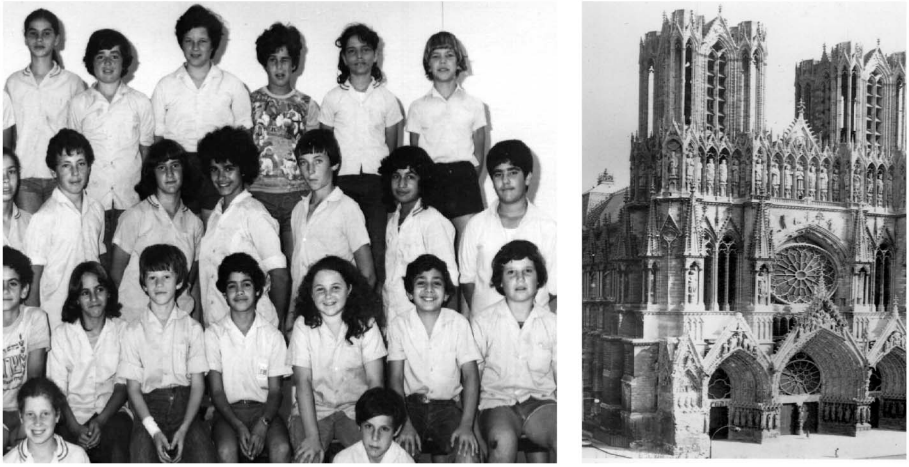


Fig. 7. Detection of partly occluded face.



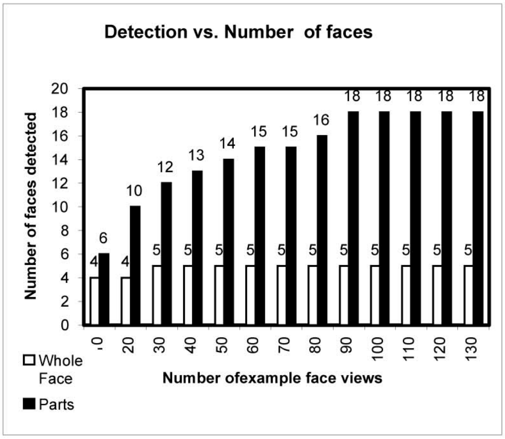


Fig. 8. Sensitivity vs. false detection rate. Image (A) contains 20 faces, (B) contains no faces. System parameters were tuned so that no false detection will occur in image (B), while the maximal number of faces will be detected in (A).

The number of faces detected in (A) vs. the number of training fragments is shown in (C) for the use of fragments (dark bars) and whole views (white bars).

The results of applying the method to these and other images indicate that the fragment-based representation generalizes well to novel objects within the class of interest. Using face fragments obtained from 26 individuals it was possible to classify correctly diverse images of males and females, in both real images and drawings, that are very different from the faces in the original training set. This was achieved while maintaining low false alarm rates on images that did not contain faces. Using a modest number of informative fragments, in different combinations, appears to have an inherent capability to deal with shape variability within the class. The fragment-based scheme was also capable of obtaining significant position invariance, without using explicit representation of the spatial relationships between fragments. The insensitivity to position as well as to other viewing parameters was obtained primarily by the use of a redundant set of overlapping fragments, including fragments of intermediate size and higher resolution, and fragments of larger size and lower resolution.

5 Some Analogies with the Human Visual System

In visual areas of the primate cortex neurons respond optimally to increasingly complex features in the input. A simple cell in the primary visual area (V1) responds best to a line or edge at a particular orientation and location in the visual field (Hubel & Wiesel 1968). In higher-order visual areas of the cortex, units were found to respond to increasingly complex local patterns. For example, V2 units respond to collinear arrangements of features (Von der Heydt 1984), some V4 units respond to spiral, polar and other local shapes (Gallant *et. al.* 1993), TE units respond to moderately complex features that may resemble e.g. a lip or an eyebrow (Tanaka 1993), and anterior IT units often respond to complete or partial object views (Perorate *et. al.* 1982, Rolls 1984, Logothetis *et. al.* 1994). Together with the increase in the complexity of their preferred stimuli, units in higher order visual areas also show increased invariance to viewing parameters, such as position in the visual field, rotation in the image plane, rotation in space, and some changes in illumination (Logothetis *et al*. 1994, Perret, Rolls & Caan 1982, Rolls 1983, Tanaka 1993). The preferred stimuli of IT units are highly dependent upon the visual experience of the animal. In monkeys trained to recognize different wire objects, units are found that respond to specific full or partial views of such objects (Logothetis *et al*. 1994). In animals trained with fractal-like images, units are subsequently found that respond to on or more of the images in the training set (Miyashita & Chang 1988).

These findings are consistent with the view that the visual system uses object representations based on class related fragments of intermediate complexity, constructed hierarchically. The preferred stimuli of simple and intermediate complexity neurons in the visual system are specific 2-D patterns. Some binocular information can also influence the response, but this additional information, which adds 3-D information associated with a fragment under particular viewing conditions, can be incorporated in the fragment based representation. The preferred stimuli are dependent upon the family of training stimuli, and in this sense appear to be class-dependent rather than, for example, a small set of universal building blocks used by other classification schemes. Invariance to viewing parameters such as position in the visual field or spatial orientation appears gradually, possibly by the convergence of more elementary and less invariant fragments onto higher order units. From this theory we can anticipate the existence of two types of intermediate complexity units that have not been reported so far. First, for the purpose of classification, we expect to find units that respond to different types of partial views. As an example, a unit of this kind may respond to different shapes of hairline, but not to a mouth or nose regions. Second, because the invariance of complex shapes to different viewing parameters is inherited from the invariance of the more elementary fragment, we expect to find intermediate complexity units, responding to partial object views at a number of different spatial orientations and perhaps different illumination conditions.

In our implementation we used the notion of a `pointing' mechanism to limit the relative spatial displacements allowed for the different fragments within a more global shape. The particular implementation we used is not intended to be directly related to biological mechanisms. However, similar results can be obtained by processes that are more biological in nature. For example, attentional mechanisms, that limit the processing to restricted regions of visual space, can play a similar role in limiting the combination of fragments and ensure that all the fragments are detected within a common region. Together with the use of overlapping fragments, this can prevent the "illusory conjunction" of fragments that are not properly arranged.

6 Summary

Object classification is a challenging task because individual objects within the same class can have large variations is shape that are difficult to define precisely and to compensate for during the classification process. In our approach, objects within a class are represented by class-specific fragments. These are image fragments that appear in similar shape in multiple individual objects within the class, and that are highly informative about the class in questions. The fragments we obtain are typically of intermediate complexity, either in size or resolution, and they form a redundant, overlapping set of `building blocks' for representing individual objects within the class.

The experimental results support the view that the fragment-based approach can generalize well to novel object images, and it can detect and classify objects of a given class despite large variability in shape. It can also deal with changes in position, pose, and illumination, but these aspects will have to be extended and tested further in future work. The results support the notion that the intra-class variability of views can be expressed, in large part, in terms of different combinations of shared common sub-structures.

The detection algorithm initially detects fragment views of different fragment types, and then combines the results for the detection of the entire object. The consistency in the fragments viewing parameters is ensured 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 smaller fragments with coarser larger ones and this increases the efficiency of the scheme, as well as its ability to impose the proper arrangement of the components.

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 rather than universal fragments also has limitations, since it implies 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.

A number of natural extensions to the basic classification scheme outlined above will be considered in the future. One is the construction of fragments in a hierarchical manner. In the algorithm described above objects views are represented in terms of intermediate complexity fragments. These fragments in turn can be constructed from simpler fragments. Starting from simple local fragments, at the level of complexity of features found in primary visual cortex, more complex fragments can be constructed hierarchically. It will be of interest to consider the effects of having different number of levels. In the primate visual system, the hierarchy includes about five levels from V1 to anterior IT, and this may be a reasonable guideline to consider. As in our scheme, at each level it will be useful to extract the most informative fragments possible at that level, and to use a redundant, overlapping set of fragments. Fragments that can be considered equivalent, will be grouped together (as in the fragment types used above) to generate increased invariance to shape variations and to changes in viewing conditions. In such a hierarchical scheme, it is likely that the similarity scheme used above for the purpose of detecting fragments in the image will be modified: the similarity between an image region and a stored fragments could be based instead on the more basic sub-fragments they have in common.

A second area of further development has to do with invariance to viewing conditions, including rotations in space and changes in illumination. The current algorithm already exhibits some invariance to these parameters, because each fragment type includes fragments at somewhat different orientations and illuminations. However, additional fragments covering a wider range of changes will be required to reach invariance comparable to human perception. It is not entirely clear whether the mechanism of using multiple overlapping fragments will be sufficient to impose the correct overall configuration when the number of alternative fragments will be substantially larger, and it is conceivable that additional mechanisms, possible including top-down verification, will be required.

Finally, the scheme should be extended to deal effectively with multiple classes. Each class added to the system will be represented by a set of class-specific fragments that are optimal for the class in question. Some of them may be similar to already existing fragments for other classes, but others will be new. The overall scheme will remain similar, however, it remains to be seen how to best organize the system to deal as efficiently as possible with a large number of known classes.

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. **

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

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

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) (1981) 285-348.

Cootes T.F., Taylor C.J., Cooper D.H., Graham J., Active shape models -- their training and applications. *Computer Vision and Image Understanding*, 61 (1995) 38-59.

Cover, T.M. & Thomas, J.A. *Elements of Information Theory*. Wiley Series in Telecommunication, New York, 1991.

Edelman, S. Representing 3D objects by sets of activities of receptive fields 70, 37-45. *Biological cybernitics,* 70, (1993) 37-45.

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

Grimson, E.W.L., & Lozano-Perez, T. Localizing overlapping parts by searching the interpretation tree. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 9 (1987) 469-482.

Hubel, D. H., Wiesel, T. N. "Receptive fields and functional architecture of monkey striate cortex", *Journal of physiology,* 195 (1968) 215-243.

Logothetis N. K., Pauls J., Bülthoff H. H., Poggio T., "View-dependent object recognition in monkeys", *Current biology,* 4 (1994) 401-414.

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 (1978) 269-294.

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

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

Miyashita, Y. & Chang, H.S. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature, 331, (1988) 68-70.

Murase, H. & Nayar, S.K. Visual learning and recognition of 3-D objects from appearance. *International J. of Com. Vision*, 14 (1995) 5-24.

Nelson C. R., and Selinger A., "A Cubist approach to object recognition", ICCV-98 (1998) 614-621.

Perret D. I., Rolls E. T. Caan W., "Visual neurons responsive to faces in the monkey temporal cortex", *Experimental brain research,* 47 (1982) 329-342.

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

Poggio, T. \& Edelman, S. A network that learns to recognize three-dimensional objects. Nature, 343 (1990) 263-266.

Rolls E. T., "Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces", *Human neurobiology,* 3 (1984) 209-222.

Rosch, E. Mervis, C.B., Gray, W.D., Johnson, S.M. & Boyes-Braem, P. Basic objects in natural carogories. *Cognitive Psychology,* 8 (1976) 382-439.

Tanaka, K., "Neural mechanisms of object recognition", *Science*, Vol. 262 (1993) 685-688.

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

Ullman, S. \& Basri, R. Recognition by linear combination of models. *IEEE PAMI*, 13(10) (1991) 992-1006.

Vapnik, V. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

von der Heydt R., Peterhans E., Baumgartner G., "Illusory contours and cortical neuron responses", *Science,* 224 (1984) 1260-1262.