A Fragment-Based Approach to Object Representation and Classification

Shimon Ullman, Erez Sali, and Michel Vidal-Naquet

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

Abstract. The task of visual classification is the recognition of an object in the image as belonging to a general class of similar objects, such as a face, a car, a dog, and the like. This is a fundamental and natural task for biological visual systems, but it has proven difficult to perform visual classification by artificial computer vision systems. The main reason for this difficulty is the variability of shape within a class: different objects vary widely in appearance, and it is difficult to capture the essential shape features that characterize the members of one category and distinguish them from another, such as dogs from cats.
In this paper we describe an approach to classification using a fragment-based representation. 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. The 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. Experiments indicate that it is possible to trade off the complexity of fragments with the complexity of the combination and decision stage, and this tradeoff is discussed.
The method is different from previous part-based methods in using class-specific object fragments of varying complexity, the method of selecting fragments, and the organization into fragment types. Experimental results of detecting face and car views show that the fragment-based approach can generalize well to a variety of 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 Introduction

The general task of visual object recognition can be divided into two related, but somewhat different tasks – classification and identification. 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. Identification is a more specific level of

C. Arcelli et al. (Eds.): IWVF4, LNCS 2059, pp. 85–100, 2001.

recognition, that is, 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, classification is a natural task: 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. Furthermore, the general classification of an object as a member of a general class such as a car, for example, is usually easier than the identification of the specific make of the car [25]. In contrast, current computer vision systems can deal more successfully with the task of identification 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 systems 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 [29] or interpolation [22] 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 outline an approach to the representation of object classes, and to the task of visual classification, that we call 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. Following a brief review of related approaches, 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. Finally, we conclude with some comments about similarities between the proposed approach and aspects of the human visual system.

2 A Brief Review of Related Past Approaches

A variety of different approaches have been proposed in the past to deal with visual recognition, including the tasks of general classification and the identification of individual objects. We will review here briefly some of the main approaches, focusing in particular on methods that are applicable to classification, and that are related 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 [17], separating hyperplanes performed, for example, by Perceptron-type algorithms and their extensions [15], or, more optimally, by support vector machines [30]. 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 [21].

More directly related to our approach are methods that attempt to describe all object views belonging to the same class using a collection of some basic building blocks and their configuration. One well-known example is the Recognition By Components (RBC) method [3] and related schemes using generalized cylinders as building blocks [4, 12, 13]. 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 image features as the underlying building blocks. These building blocks were typically small simple image features, together with a description of their qualitative spatial relations. Examples of such features include local image patches [1, 18], corners, direct output of local receptive fields of the type found in primary visual cortex [7], wavelet functions [26], simple line or edge configurations [23], or small texture patches [32].

The eigenspace approach [28] can also be considered as belonging to this general approach. In this method, a collection of objects within a class, such as a set of faces, are constructed using combinations of a fixed set of building blocks. The training images are described as a set of grey-level vectors, and the principle components of the training images are extracted. The principal components are then 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 rather than local in nature. As we shall see in the next section, the building blocks selected by our method are intermediate in complexity: they are considerably more complex than simple local features used in previous approaches, but they still correspond to partial rather than global object views.

3 The Selection of Class-Based Fragments

The classification of objects using a fragment-based representation raises two main problems. The first is the selection of appropriate fragments to represent a given class, based on a set of image examples. The second is performing the actual classification based on the fragment representation. In this section we deal with the first of these problems, the selection of fragments that are well-suited for the classification task. Subsequent sections will then deal with the classification process.

Our method for the selection and use of basic building blocks for classification is different from previous approaches in several respects. First, unlike other methods that use local 2-D features, we do not employ universal shape features. That is, the set of basic building blocks used as shape primitives are not the same for all classes, as used, for instance, in the RBC approach. Instead, we use object fragments that are specific to a class of objects, taken directly from example views of objects in the same class. As a result, 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 then used as a set of common building blocks to represent, by different combinations of the fragments, different objects belonging to the class. Second, the fragments we use as building blocks are extracted using an optimization process that is driven directly by requirements of the classification task. This is in contrast with other scheme where the basic building elements are selected on the basis of other criteria, such as faithful reconstruction of the input image. Third, the fragments we detect are organized into equivalence sets that contain views of the same general region in the objects under different transformations and viewing conditions. As we will see later, this novel organization plays a useful role in performing the classification task.

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 and the like 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. We will focus here mainly on computational issues, possible biological implications will be

discussed elsewhere. In the following sections, we describe the use of the fragment representation for performing classification tasks.

3.1 Selecting Informative Fragments

Given a set of images that represent different objects from a given class, our scheme selects a set of fragments that are used as a basis for representing the different shapes within the class. 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 are selected using a criterion of maximizing the 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. 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

$$I(C,F) = H(C) - H(C/F) \quad (1)$$

where I is the mutual information, and H denotes entropy [6]. 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. Both the merit and the distinctiveness can be evaluated given the estimated value of p(C), p(F|C), p(F|NC). In summary,

fragments are selected on the basis of their merit, and then used on the basis of their distinctiveness.

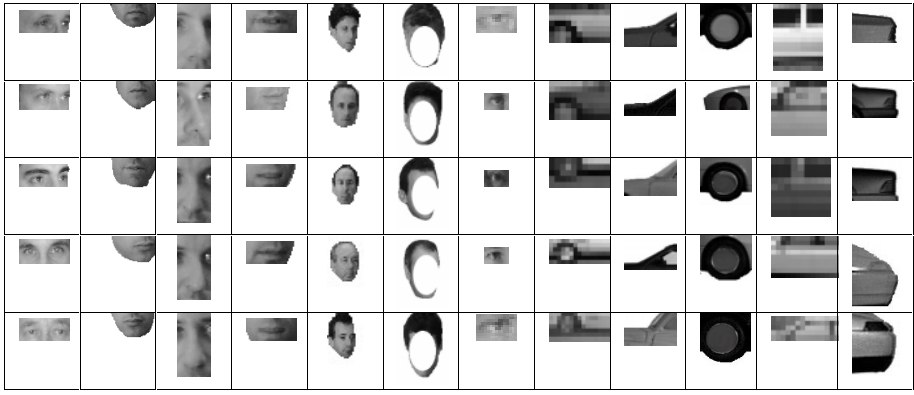

Fig. 1. Examples of face and car fragments

Our procedure for selecting the fragments with high mutual information is the following. Given a set of images, such as cars, we start by comparing the images in a pairwise manner. The reason is that a useful building block that appears in multiple car 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 cars, and also in a second database composed of many natural images that do not contain cars. 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 is 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 or resolution, and it includes full resolution fragments of intermediate size, and larger fragments of intermediate resolution.

4 Performing Classification

The set of fragments extracted from the training images during the learning stage are then used to classify new images. In this section we consider the problem of

performing object classification based on the fragment representation described above.

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: detecting fragments in the image, followed by a decision stage that combines the results of the individual fragment detectors. In the first stage, fragments are detected by comparing the image at each location with stored fragment views. The comparison is based on gray-level similarity, that is insensitive to small geometric distortions and gray level variations. As for the combination stage, we have compared two different approaches: a simple and a complex scheme. The simple scheme essentially tests that a sufficient number of basic fragments have been detected. The more complex scheme is based on a more complete probability distribution of the fragments, and takes into account dependencies between pairs of fragments. We found that, using the fragments extracted based on informativeness, the simple scheme is powerful and works almost as well as the more elaborate scheme. This finding is discussed later in this section.

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

5 Detecting Individual Fragments

The detection of the individual fragment is based on a direct gray-level comparison between stored fragments and the input image. To allow some distortions and scale changes of a fragment, the comparison is performed first on smaller parts of the fragments, that were taken in the implementation to be patches of size 5*5 pixels. We describe first the gray-level similarity measure used in the comparison, and then how the comparisons of the small patches are used to detect individual fragments.

We have evaluated several gray-level comparison 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 [2] with gradient and orientation measures gave the best results. The method measured the qualitative shape similarity using the ordinal ranking of the pixels in the

regions, and also measured the orientation difference using gradient amplitude and direction. For the qualitative shape comparison we computed the ordinal ranking of the pixels in the two regions, and used the normalized sum of displacements of the pixels with the same ordinal ranking as the measure for the regions similarity.

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

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

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

For the detection of fragments in the images we first compared local 5x5 gray level patches in each fragment 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 detect objects at different scale in the image, the algorithm is performed on the image at several scales. 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.

6 Combining the Fragments and Making a Decision

Following the detection of the individual fragments, we have a set of 'active' fragments, that is, fragments that have been detected in the image. We next need to combine the evidence from these fragments and reach a final decision as to whether the class of interest is present in the image.

In this section we will consider two alternative methods of combining the evidence from the individual fragments and reaching a decision. Previous approaches to visual recognition suggest there is a natural trade-off between the use of simple visual features that require a complex combination scheme, and the use of more complex features, but with a simpler combination scheme.

A number of recognition and classification schemes have used simple local image features such as short oriented lines, corners, Gabor or wavelet basis functions, or local texture patches. Such features are generic in nature, that is, common to all

visual classes. Consequently, the combination scheme must rely not only on the presence in the image of particular features, but also on their configurations, for example, their spatial relations, and pair-wise or higher statistical interdependencies between the features. A number of schemes using this approach [1, 14, 26, 33], therefore employ, in addition to the detection of the basic features, additional positional information, and probability distribution models of the features. In contrast, a classifier that uses more complex, class-specific visual features could employ a simpler combination scheme because the features themselves already provide good evidence about the presence of the class in question. In the next section, we first formulate the classification as a problem of reaching optimal decision using probabilistic evidence. We then compare experimentally the classification performance of a fragment-based classifier that uses a 'simple' combination method, and one that uses a more complex scheme.

6.1 Probability Distribution Models

We can consider in general terms the problem of reaching a decision about the presence of a class in the images based on some set of measurements denoted by $\boldsymbol{X}$. Under general conditions, the optimal decision is obtained by evaluating the likelihood ratio defined as:

$\frac{P(\boldsymbol{X} \mid C_1)}{P(\boldsymbol{X} \mid C_0)}$, where $P(\boldsymbol{X}|C_0)$, $P(\boldsymbol{X}|C_1)$ are the conditional probabilities of $\boldsymbol{X}$ within and outside the class. The elements of $\boldsymbol{X}$ express, in the fragment-based scheme, the fragments that have been detected in the image.

The direct use of this likelihood ratio in practice raises computational problems in learning and storing the probability functions involved. A common solution is to use restricted models using assumptions about the underlying probability distribution of the feature vector. In such models, the number of parameters used to encode the probability distribution is considerably smaller than for a complete look-up table representation, and these parameters can be estimated with a higher level of confidence.

A popular and useful method for generating a compact representation of a probability distribution is the use of Belief-Networks, or Bayesian-Networks. A Bayesian-Network is a directed graph where each node represents one of the variables used in the decision process. The directed edges correspond to dependency relationships between the variables, and the parameters are conditional probabilities between inter-connected nodes. A detailed description of Bayesian-Networks can be found in [19]. The popularity of Bayesian-Network methods is due in part to their flexibility and ability to represent probability distributions with dependencies between the variables. There are several methods that enable the construction of a network representation from training data, and algorithms that efficiently compute the probability of all the variables in the network given the observed values of some of the variables. In the following section, we use this formalism to compare two methods for combining the evidence from the fragments detected in the first stage.

One is a simple scheme sometimes called 'naïve Bayesian' method, and the second a more elaborate scheme using a Bayesian-Network type method.

6.2 Naïve-Bayes

The assumption underlying the naive-Bayes classifier is that the entries of the feature vector can be considered independent when the computing the likelihood ratio $\frac{P(\boldsymbol{X} \mid C_1)}{P(\boldsymbol{X} \mid C_0)}$. In this case, the class-conditional probabilities can be expressed by the product:

$$P(X_1,...X_N \mid C) = \prod_{i=1}^{N} P(X_i \mid C) \tag{3}$$

The values of the single probabilities are directly measured from the training data. In practice, this means that we first measure the probability of each fragment X_i within and outside the class. To reach a decision we simply multiply the relevant probabilities together. This method essentially assumes independence between the different fragments. (More precisely, it assumes conditional independence.) The actual computation in our classification scheme was performed using the fragment types, rather than the fragments themselves. This means that for each fragment type (such as a hairline or eye region), the best-matching fragment was selected. The combination then proceeds as above, by multiplying the probabilities of these fragments, one from each type.

6.3 Dependence-Tree Combination

The Dependence-tree model is a simple Bayesian-Network that describes a probability distribution which incorporates some relevant pairwise dependencies between variable, unlike the independence assumptions used in the naïve-Bayes scheme. The features are organized into a tree structure that represents statistical dependencies in a manner that allows an efficient computation of the probability of an input vector. The tree structure permits the use of some, but not all, of the dependencies between features. An optimal tree representation can be constructed from information regarding pairwise dependencies in the data [5]. The probability of an input vector is computed by multiplying together the probabilities of each node given the value of its parent. More formally:

$$P(X_1,...X_N \mid C) = P(X_1 \mid C) \times \prod_{i=2}^{N} P(X_i \mid X_{j(i)}, C) \tag{4}$$

where j(i) represents the parent of node i. X_1 is the root of the tree (which represent the class variable), that does not have a parent. The conditional probabilities are estimated during the learning phase directly from the training data.

6.4 The Trade-Off between Feature Complexity and Combination Complexity

We have implemented and compared the two schemes outlined above. This allows us to compare a simple combination scheme that is based primarily on the presence or absence of fragments in the image, and a more elaborate scheme that uses a more refined model of the probability distribution of the fragments. Figure 3 shows the performance of a fragment-based classifier trained to detect side views of cars in low-resolution images, using both combination schemes.

The results are presented in the form of Receiver Operating Characteristic (ROC) curves [9]. Each point in an ROC curve represents a specific pair of false-alarms and hit-rate of the classifier, for a given likelihood ratio threshold. The efficiency of a classifier can be evaluated by the 'height' of its ROC curve: for a given false-alarm rate, the better classifier will be the one with higher hit probability. The overall performance of a classifier can be measured by the area under the performance curve in the ROC graph.

The curves for both methods are almost identical, showing that including pairwise dependencies in the combination scheme, rather than using the information of each feature individually, has a marginal effect on the classifier performance. This suggests that most of the useful information for classification is encoded in the image fragments themselves, rather than their inter-dependencies. This property of the classifier depends on the features used for classification. When simple generic features are used, the dependencies between features at different locations play an important role in the classification process. However, when more complex features are used, such as the ones selected by our information criterion, then a simpler combination scheme will suffice.

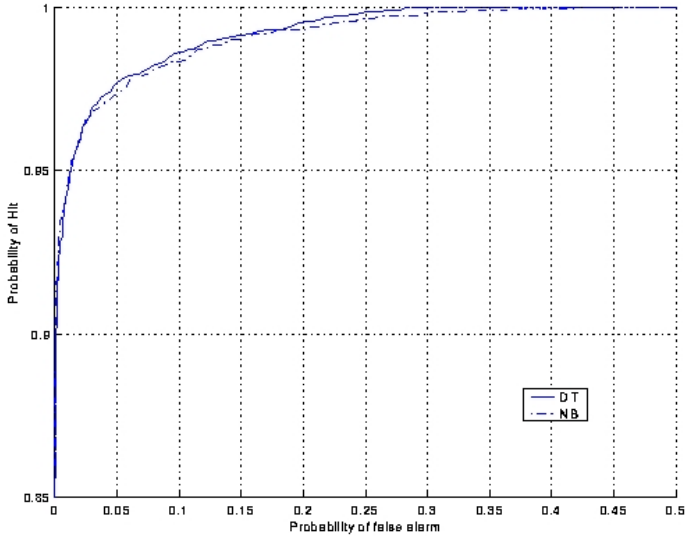

Fig. 2. Receiver Operating Characteristic curves for both classifiers. NB: Naïve-Bayes classifier. DT: Dependence-Tree classifier.

7 Experimental Results

We have tested our algorithm on face and car views. A number of similar experiments were performed, using somewhat different databases and different details of the fragment extraction and combination procedures. In an example experiment, we tested face detection, using 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 into eight fragment 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. Figure 5 shows several 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.

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 a small set of examples 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.

8 Some Analogies with the Human Visual System

In visual areas of the primate cortex neurons respond optimally to increasingly complex features in the input. Simple and complex cells in the primary visual area (V1) responds best to a line or edge at a particular orientation and location in the visual field [10]. 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 [31], some V4 units respond to spiral, polar and other local shapes [8], TE units respond to moderately complex features that may resemble e.g. a lip or an eyebrow [27], and anterior IT units often respond to complete or partial object views [11, 24]. 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 [11, 20, 24, 27].

Fig. 3. Examples of face and car detection. The images are from the Weizmann image database, from the CMU face detector gallery, and the last two from www.motorcities.com.

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 were found that respond to specific full or partial views of such objects [11]. In animals trained with fractal-like images, units were subsequently found that respond to one or more of the images in the training set [16].

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 lower level features are simple generic features. The preferred stimuli of the more complex units are dependent upon the family of training stimuli, and appear to be class-dependent rather than, for example, a small set of universal building blocks. 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.

Acknowledgement. This research was supported in part by the Israel Ministry of Science under the Scene Teleportation Research Project. We acknowledge the use of material reported in the Proceedings of the Korean BMCV meeting, Seoul, 1999.

References

1. Amit, Y.,Geman, D., Wilder, K.: Joint Induction of Shape Features and Tree Classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, 11 (1997) 1300-1306
2. Bhat, D., Nayar, K. S.: Ordinal Measures for Image Correspondence. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, 4 (1998) 415-423
3. Biederman, I.: Human image understanding: recent research and theory. Computer Vision, Graphics and Image Processing 32 (1985) 29-73
4. Binford, T. O.: Visual perception by computer. IEEE conf. on systems and control, Vol. 94, 2 (1971) 115-147

5. Chow, C.K., Liu, C.N.: Approximating Discrete Probability Distributions with Dependence Trees. IEEE Transactions on Information Theory, Vol. 14, 3 (1968) 462-467
6. Cover, T.M. & Thomas, J.A.: Elements of Information Theory. Wiley Series in Telecommunication, New York (1991)
7. Edelman, S.: Representing 3D objects by sets of activities of receptive fields. Biological cybernitics 70 (1993) 37-45
8. Gallant, J.L., Braun, J., Van Essen, D.C.: Selectivity for polar, hyperbolic, and cartesian gratings in macaque visual cortex. Science, 259 (1993) 100-103
9. Green, D. M., Swets, J. A.: Signal Detection Theory and Psychophysics. Wiley, Chichester New York Brisbane Toronto (1966). Reprinted by Krieger, Huntingdon, New York (1974)
10. Hubel, D. H., Wiesel, T. N.: Receptive fields and functional architecture of monkey striate cortex. Journal of physiology 195 (1968) 215-243
11. Logothetis, N. K., Pauls J., Bülthoff H. H., Poggio T.: View-dependent object recognition in monkeys. Current biology, 4 (1994) 401-414
12. Marr, D.: Vision. W.H. Freeman, San Francisco (1982)
13. 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
14. 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
15. Minsky, M., Papert, S.: Perceptrons. The MIT Press, Cambridge, Massachusetts (1969)
16. Miyashita, Y.,Chang, H.S.: Neuronal correlate of pictorial short-term memory in the primate temporal cortex. Nature, 331, (1988) 68-70
17. Murase, H.,Nayar, S.K.: Visual learning and recognition of 3-D objects from appearance. International J. of Com. Vision, 14 (1995) 5-24
18. Nelson, C. R., Selinger A.: A Cubist approach to object recognition. International Conference on Computer Vision '98 (1998) 614-621
19. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman Publishers, San Mateo, California (1988)
20. 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
21. Poggio, T., Sung, K.: Finding human faces with a gaussian mixture distribution-base face model. Computer analysis of image and patterns (1995) 432-439
22. Poggio, T., Edelman, S.: A network that learns to recognize three-dimensional objects. Nature, 343 (1990) 263-266
23. Riesenhuber, M.,Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience, Vol. 2, 11 (1999) 1019-1025
24. 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
25. Rosch, E. Mervis, C.B., Gray, W.D., Johnson, S.M., Boyes-Braem, P.: Basic objects in natural categories. Cognitive Psychology, 8 (1976) 382-439
26. Schneiderman, H., Kanade. T.: Probabilistic modeling of local appearance and spatial relationships for object recognition. Proc. IEEE Comp. Soc. Conference on Computer Vision and Pattern Recognition, CVPR98, (1998) 45-51
27. Tanaka, K.: Neural Mechanisms of Object Recognition. Science, Vol. 262 (1993) 685-688.

28. Turk M., Pentland A.: "Eigenfaces for recognition", Cognitive Neuroscience, 3 (1990) 71-86
29. Ullman, S., Basri, R.: Recognition by Linear Combination of Models. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 13, 10 (1991) 992-1006
30. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
31. von der Heydt, R., Peterhans, E., Baumgartner G.: Illusory Contours and Cortical Neuron Responses. Science, 224 (1984) 1260-1262
32. Weber, M, Welling M. & Perona, P.: Towards Automatic Discovery of Object Categories. Proc. IEEE Comp. Soc. Conference on Computer Vision and Pattern Recognition, CVPR2000, 2, (2000) 101-108
33. Wiskott, L., Fellous J. M., Krüger N., von der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching. Intelligent Biometric Techniques in Fingerprint and Face Recognition, 11 (1999) 355-396