COGNITION

Cognition 67 (1998) 21–44

Three-dimensional object recognition based on the combination of views

Shimon Ullman*

Weizmann Institute of Science, Department of Applied Mathematics and Computer Science, POB 26, 76100 Rehovot, Israel

Abstract

Visual object recognition is complicated by the fact that the same 3D object can give rise to a large variety of projected images that depend on the viewing conditions, such as viewing direction, distance, and illumination. This paper describes a computational approach that uses combinations of a small number of object views to deal with the effects of viewing direction. The first part of the paper is an overview of the approach based on previous work. It is then shown that, in agreement with psychophysical evidence, the view-combinations approach can use views of different class members rather than multiple views of a single object, to obtain class-based generalization. A number of extensions to the basic scheme are considered, including the use of non-linear combinations, using 3D versus 2D information, and the role of coarse classification on the way to precise identification. Finally, psychophysical and biological aspects of the view-combination approach are discussed. Compared with approaches that treat object recognition as a symbolic high-level activity, in the view-combination approach the emphasis is on processes that are simpler and pictorial in nature. © 1998 Elsevier Science B.V. All rights reserved

Keywords: Three-dimensional object recognition; View combinations; Classification

1. Recognition and the variability of object views

For biological visual systems, visual object recognition is a spontaneous, natural activity. In contrast, the recognition of common objects is still beyond the capabilities of current computer vision systems. In this paper I will examine certain aspects of the recognition problem and outline an approach to recognition based on the

* Tel.: +972 8 9343545; fax: +972 8 9342945; e-mail: shimon@wisdom.weizmann.ac.il

PII S0010-0277(98)00013-4

combination of object views. The discussion of the recognition problem will be limited in a number of ways. In particular, it will focus on shape-based recognition, and it will consider primarily object identification rather than classification. In general, objects can be recognized not only by their shape, but also based on other visual cues, such as color, texture, characteristic motion, their location relative to other objects in the scene, context information, and expectation. Here, I will focus on the recognition of isolated objects, using shape information alone.

Why is visual recognition difficult? It may appear that the problem could be approached by using a sufficiently large and efficient memory system. In performing recognition, we are trying to determine whether an image we currently see corresponds to an object we have seen in the past. It might be possible, therefore, to approach object recognition by storing a sufficient number of different views associated with each object, and then comparing the image of the currently viewed object with all the views stored in memory (Abu-Mostafa and Psaltis, 1987). Models of so-called associative memories have been proposed for implementing this ‘direct’ approach to recognition (Willshaw et al., 1969; Kohonen, 1978; Hopfield, 1982).

Although direct comparison to stored views can play a useful role, especially for the recognition of highly familiar objects, this direct approach by itself is insufficient for recognition in general. One reason is that the space of all possible views of all the objects to be recognized is likely to be prohibitively large. A second, and more fundamental reason, is the problem of generalization, that is, recognizing an object under novel viewing conditions. Object views are highly variable and depend on the viewing direction and distance, the effects of illumination direction, shadowing and highlights, partial occlusion by other objects, and possible changes and distortions in the object itself. As a result, the image to be recognized will often not be sufficiently similar to any image seen in the past.

1.1. An empirical comparison of intra- and inter-object variability

The effects of these sources of variation, in particular, viewing position and illumination direction, were evaluated quantitatively in the domain of face images (Adini et al., 1997). In the study, twenty-six different individuals were imaged from a number of viewing directions and under different illumination conditions. The goal was to compare images of different individuals, with images of the same individual but under different viewing conditions. In this manner, it becomes possible to examine whether the differences induced by mere changes in the viewing conditions are large or small compared with the differences between distinct individuals. The images used were of males, with no glasses, beards, etc., and with the hairline covered, taken under five viewing conditions. The first was from a frontal view, with left illumination (45°),and neutral expression. The other four differed from the first by changing either the illumination (45° right), the viewing direction (17° to the right), or the facial expression. The different images were taken by moving a robotic arm with a TV camera to different locations in space.

To compare the different images, one needs to define a measure of similarity. The study employed a number of commonly used measures for comparing images. The simplest measure used the average absolute difference between the image intensity levels of corresponding points. The face images were normalized in size, orientation, and position, before this measure was computed. A more flexible measure allowed local distortions between the two images: the image intensity value at a given point was compared not only to a single corresponding location in the second image, but to all the points within a given neighborhood, and the best-matching value within this neighborhood was selected. The computation of image differences also included compensation for changes in overall intensity level and linear intensity gradients, so that the difference measure became insensitive to these global parameters. Another type of difference measure used transformed versions of the images, obtained by applying to the images various filters, such as difference-of-gaussians, (DOG) filters (Marr and Hildreth, 1980), Gabor filters (Daugman, 1989), and using directional derivatives of the gray level images (Koenderink and Van Doorn, 1990). Filtering of this kind appears to take place in the early processing stages of the mammalian visual system, and it is also often used in artificial image processing systems. Finally, images were also compared by first obtaining an edge map from each image, and then comparing the resulting contour maps. The use of such edge maps is also a standard procedure in image processing, partly because they are less sensitive to illumination conditions than the original gray-level images. For each of the different measures, comparisons were made between the full face images, but also between partial face images, such as the upper or lower parts.

The main result that emerged from these comparisons is that the differences induced by changes in the viewing conditions are large compared with the differences between different individuals. An example is illustrated in Table 1 (after Moses, 1993), comparing the effects of changes in illumination and in viewing direction, with differences between individuals, for 11 different faces (F_1–F_{11}).

Table 1
Distances between face images

FC	IC	VP	Pairs of faces										
			F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11
F1	59	25		21	25	26	24	39	25	22	23	22	17
F2	56	23			16	18	27	25	16	14	16	15	21
F3	56	26				17	27	27	14	15	20	16	34
F4	58	27					26	26	18	18	20	20	26
F5	54	30						38	29	26	23	15	15
F6	50	25							26	28	25	24	40
F7	57	30								17	20	17	23
F8	51	23									18	17	22
F9	54	24										7	25
F10	53	21											23
F11	57	25											

Larger values indicate increased dissimilarity, for units see text. Column FC lists 11 different faces. IC, distances between images of the same face, but under different illumination; VP, distances between images of the same face taken from frontal and 17° side view; ‘pairs of faces’, distances between all 11 face pairs.

Each face is compared to an image of the same face under a different illumination (first column, marked ‘IC’, for illumination condition), to an image of the same face viewed from a different direction (17° change in direction, column marked ‘VP’, for viewing position), and then to ten other faces viewed under the original viewing conditions. For example, for the first face, F_1, a change in illumination induced a difference measure of 59 units between the two images, and for a change in viewing direction the difference measure was 25 units. (Units are the average difference between normalized gray level images, other measures were also used, as discussed below.) When this face image was compared with another face image, F_2, the difference measure was 21 units. This means that the changes induced by variations in the viewing conditions were larger than the difference between the two individuals. In fact, for eight out of the ten face images, the differences due to viewing conditions were the same or larger than the differences between distinct individuals.

One conclusion from these comparisons is that in this domain recognition based on direct view-comparisons will result in severely limited generalization to viewing direction and illumination. Suppose, for example, that we attempt to identify the faces of just the limited set of 26 individuals in the study by comparing an input image to a set of 26 images, one for each face, stored in memory, and then selecting the stored image that most closely resembles the input image. The results show that such a scheme will be highly inadequate: for the images in the face database, the wrong answer will often be selected. For the best performing comparison scheme (of a total of 107 tested) the error rates were above 20% for illumination changes and about 50% for changes in viewing direction. These results were also compared with the performance of the human visual system, using the same as well as additional test images (Moses et al., 1996). It was found that the variations tested in the study, as well as larger ones (up to 51° in viewing positions) were easily compensated for by human observers.

Similar results were reached in a study by Liu et al. (1995). The study, described in Section 5 below, also concluded that generalization to novel views cannot be explained in terms of independent comparisons to stored views.

Taken together, these results indicate that the differences induced by changes in the viewing conditions are large compared with the differences between different individuals, and that direct image comparisons, even in combination with the pre-normalization used for size, orientation, position, and intensity level, are not sufficient for recognition. To obtain reliable recognition, some processes that can compensate for the effects of viewing conditions are required. The nature of these processes is a fundamental problem in the study of visual recognition. A review of the main approaches to this problem, including the use of invariances and the construction of object-centered structural description can be found in (Ullman, 1989, 1996). In this paper I will focus on an approach that compensates for the effects of viewing directions by comparing the novel image with certain combinations of previously stored views. The approach is somewhat similar to the direct comparison scheme in that it is based on the comparison of picture-like representations. However, by using previously stored views not independently but in certain combinations, significant generalization across viewing direction can be obtained.

The discussion will focus on theoretical aspects of the scheme. Psychophysical and physiological aspects of this approach are reviewed in Section 5.

2. The combination of object views

2.1. The view-combination property

In many recognition theories it is assumed that the visual system somehow stores and manipulates 3D object models (Biederman, 1985; Lowe, 1985; Ullman, 1989). When confronted with a novel 2D image of the object, the system deduces whether it is a possible view of one of the already stored 3D objects. In contrast, the method outlined in this section does not use explicit 3D models. Instead, it uses small collections of object views directly, without the need to explicitly recover and represent the 3D structure of objects.

In this approach, a 3D object is represented by the linear combination of 2D views of the object. If $M = M_1,\ldots,M_k$ is the set of views representing a given object, and P is the 2D image of an object to be recognized, then P is considered an instance of M if

$$P = \sum_{i=1}^{k} \alpha_i M_i$$

for some constants α_i.

The linear combination of views has the following meaning. Suppose that (x_i, y_i), (x_i', y_i'), (x_i'', y_i'') are the coordinates of corresponding points (i.e. points in the image that arise from the same point on the object) in three different views. Let X_1, X_2, X_3, be the vectors of x-coordinates of the points in the three views. Suppose that we are now confronted with a new image, and X' is the vector of the x-coordinates of the points in this new view. If X' arises from the same object represented by the original three views, then it will be possible to express X' as the linear combination of X_1, X_2, X_3. That is, $X' = a_1X_1 + a_2X_2 + a_3X_3$ for some constants a_1, a_2, a_3. Similarly, for the y-coordinates, $Y' = b_1Y_1 + b_2Y_2 + b_3Y_3$ for some constants b_1, b_2, b_3. In general, different coefficients will be required for the x and y components, and therefore the total number of coefficients is six. In more pictorial terms, we can imagine that each of the three points x_i, x_i', x_i'' has a mass associated with it. The mass at x_i, x_i', x_i'' is a_1, a_2, a_3, respectively (the same weights are used for all triplets). The linear combination of the points is now their center of mass. The linear combination property is expressed by the following mathematical statement: all possible views of a rigid object that can undergo rotation in space, translation, and scaling, are spanned by the linear combinations of three views of the object.

The proposition assumes orthographic projection and objects with sharp bounding contours. For objects with smooth bounding contours, the number of views required is five rather than three. Objects with smooth bounding contours, such as an egg or a football, require more views because the object's silhouette is not generated by fixed contours on the object. The bounding contours generating the silhouette move continuously on the object as the viewing position changes. Finally, it should be noted that, due to self occlusion, three views are insufficient for representing an object from all orientations. That is, a different set of views will be required to represent, e.g. the ‘front’ and the ‘back’ of the same object. For a proof of the proposition and further details of its implications see Ullman and Basri (1991).

Fig. 1 shows an example of using linear combinations of views to compensate for changes in viewing direction. Fig. 1a shows three different views of a car (a VW). The figure shows only those edges that were extracted in all three views; as a result, some of the edges are missing. This illustrates that reliable identification can be obtained on the basis of partial image data (as may happen due to noise and partial occlusion). Fig. 1b shows two new views of the VW car. These new images were not obtained from novel views of the car, but were generated by using linear combinations of the first three views. Fig. 1c shows two new views of the VW, obtained from new viewing positions. Fig. 1d superimposes these new views and the linear combinations obtained in Fig. 1c. It can be seen that the novel views are matched well by linear combinations of the three original views. For comparison, Fig. 1e shows the superposition of a different, but similar car (a Saab), with the best matching linear combinations of the VW images. As expected, the match is not as good. This illustrates that the linear combination method can be used to make fine distinctions between similar 3D objects in novel viewing directions. Although the two objects being compared have complex 3D shapes, and are quite similar, they were reliably discriminable by the view-combination method within the entire 60° rotation range. To represent the object from a wider range of viewing directions a number of different models of this type will be required. This notion is similar to the use of different object aspects suggested by Koenderink and Van Doorn (1979). It is worth noting that although the objects in this example are quite similar in shape, recognizing them reliably over a large range of viewing directions is a task mastered easily by human observers.

Fig. 2 shows another example of a view combination, applied to a gray-scale image. The two images in the top row are two input images of the same individual from different viewing directions. The bottom row shows two images produced by the view-combination method, depicting the same individual from different viewing directions. One is an intermediate view, between the two original viewing directions, the other is an extrapolation beyond the range spanned by the original views. The gray levels in the combined images were taken as the average of the corresponding points in the original views, for a detailed discussion of combining views containing gray-level information see Shashua (1992).

2.2. Using two views only

A novel object view was expressed in the scheme described above as the linear combination of three fixed views of the object. The three views are necessary if the transformations that the object is allowed to undergo are restricted to rigid transformations. In this case, the coefficients of the view combination are required to satisfy certain functional constraints that can be tested to verify whether the object trans-

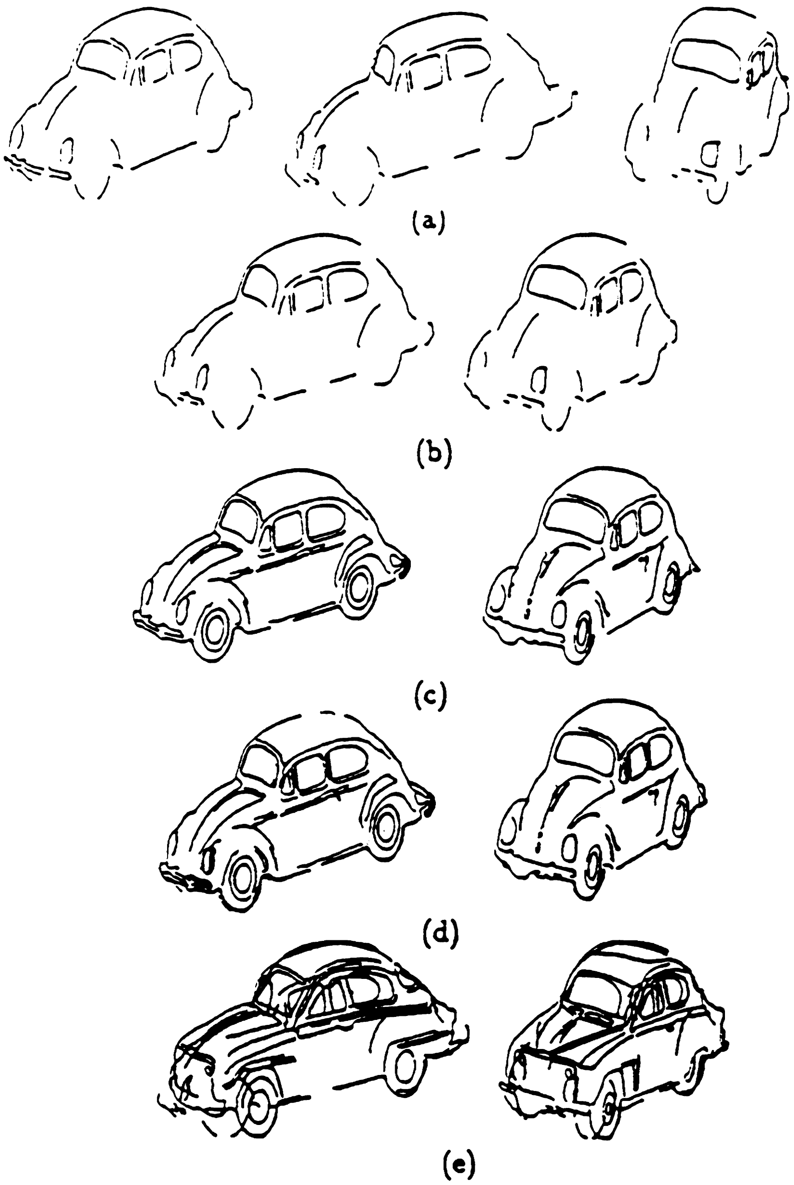

Fig. 1. Using view-combination to match a novel view. (a) Three model pictures of a VW car for ± 30° rotations around the vertical axis. Only a subset of the edges were used in the model. (b) Two linear combinations of the VW model. These are generated artificially by combinations of the first three views, rather than actual views. (c) Real novel images of the car. (d) Matching the linear combinations to the real images. The agreement is good within the entire range of ± 30°. (e) Matching the VW model to pictures of a similar car (Saab). (From Ullman and Basri, 1991; ©IEEE.)

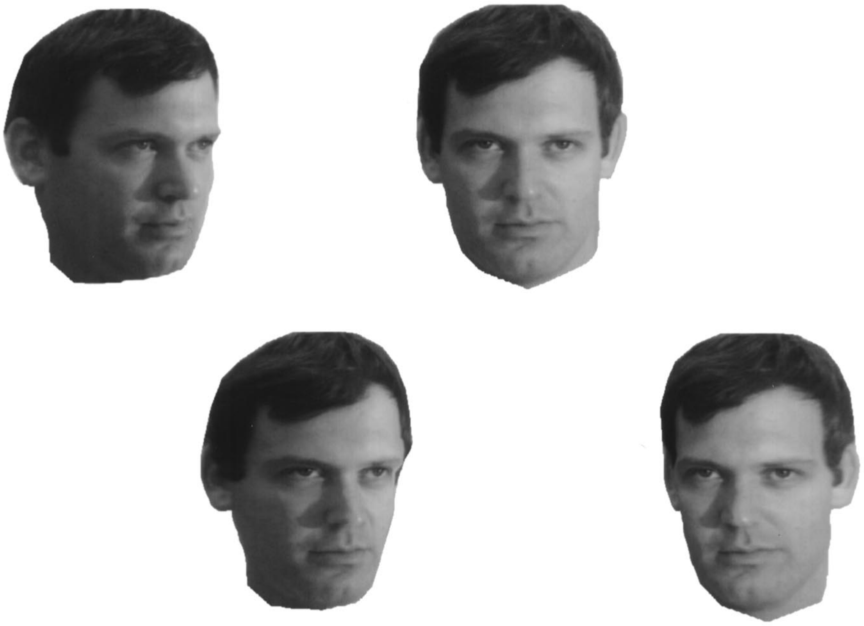

Fig. 2. A view-combination of gray level face images. On the top row are the two basic images. Bottom row: two view combinations, depicting the same person from new viewing directions. Left: an interpolated view. Right: an extrapolated view, beyond the range spanned by the original views.

formation was indeed rigid (Ullman and Basri, 1991). It is natural to consider also a more general case, in which objects are allowed to undergo more general distortions, for example, some stretch along one dimension. For general linear transformations of the object, it turns out that it becomes possible to use just two views of the object. (This observation was also made independently by T. Poggio.) Unlike the three-views formulation, this two-views formulation uses a mixture of *x* and *y* coordinates. This combination is somewhat less intuitive than the combination of three views, but the use of only two rather than three views is sometimes more convenient.

Mathematically, this combination has the following form. Let $\mathbf{x}_1$ be the vector of all the *x* coordinates of the points in the first view, $\mathbf{x}_2$ in the second, and $\hat{x}$ in the novel view; $\mathbf{y}_1$ is the vector of *y* coordinates in the first view. Then:

$$\hat{x} = \mathrm{a}_1 x_1 + \mathrm{a}_2 y_1 + \mathrm{a}_3 x_2 \tag{1}$$

For any novel view of the same object (subject to the usual self-occlusion limitations), the set of new coordinates *x* is the linear combination of corresponding coordinates in the first two views. The *y* coordinates can be expressed in a similar manner:

$$\hat{y} = \mathrm{b}_1 x_1 + \mathrm{b}_2 y_1 + \mathrm{b}_3 x_2 \tag{2}$$

In this version the basis vectors are the same for the *x* and *y* coordinates, and they are obtained from two rather than three views.

2.2.1. A single view of a symmetric object

In reducing the number of base images, it is possible under some conditions to take a further step, and perform image combination based on a single view. This extension was developed by Vetter et al. (1994) for the class of bilaterally symmetric objects. One intuitive way of looking at this special case of symmetric objects is by considering the two symmetric halves of the object as two distinct views of a single part. Under the assumptions discussed in the previous section (orthographic projections, affine object transformations), two views are sufficient to recognize the object in question.

Vetter et al. (1994) have shown that for human observers the recognition of symmetric 3D objects from a single training view is indeed better than for non-symmetric objects. For the symmetric objects, the experiments showed a broader generalization, that is, observers recognized the symmetric test objects correctly over a wider range of rotations.

2.3. Using view-combinations for recognition

A straightforward way of using the linear combination method in practice is to recover the coefficients of the combination, then use these coefficients to produce a new model image and compare it with the input image. One method of recovering the unknown coefficient is by using a small number of matching image and model features. For example, by using three corresponding features points in the image and the model, the coefficients can be recovered uniquely by solving linear equations (two simple systems of three unknowns each, for the x and y components). Mathematically, the procedure is the following. Let X be the matrix of the x-coordinates of the alignment points in the model. That is, x_{ij} is the x-coordinate of the jth point in the ith model picture. $\mathbf{p}x$ is the vector of x-coordinates of the alignment points in the image, and $\mathbf{a}$ is the vector of unknown combination parameters we wish to recover. The linear system to be solved is then simply $X\mathbf{a} = \mathbf{p}_x$. The combination parameters are given by $\mathbf{a} = X^{-1}\mathbf{p}x$ if an exact solution exists. We may use an overdetermined system (by using additional points), in which case $\mathbf{a} = X^{+}\mathbf{p}x$ (where X^{+} denotes the pseudo-inverse of X, (Albert, 1972)). A similar procedure is used to recover the coefficients in the Y direction. A convenient property of the scheme is that the matrix X^{+} used to derive the unknown coefficients does not depend on the image and can therefore be pre-computed and stored for a given model. Future recovery of the coefficients simply requires only a multiplication of $\mathbf{p}_{\mathrm{x}}$ by the stored matrix. Using the recovered coefficients, an internal combined image will be generated and compared with the viewed object.

This scheme is simple and efficient for an artificial recognition system. However, for a biological system, it may not be straightforward to implement the required processes, such as the matching of corresponding features, solving for the coefficients, or using them for generating internal images. There are alternative ways in which the view-combination property can be used in the recognition process that I will mention here only briefly.

One possibility is to use an iterative method that starts with one of the stored images and successively refines it by combining it with additional images (Lipson, 1993). This method requires only approximate contour matches rather than precise pointwise matches between corresponding features.

An alternative approach is to use a 'correspondence-less' scheme that does not rely on matching image and model features. This can be done by performing a search in the space of possible coefficients. In this method, we first choose some initial values for the set of coefficients, and then apply a linear combination to the model using these values. We repeat this process using a different set of coefficients, and finally choose the coefficient values that produced the best match of the model to the image. The search can be guided by an optimization procedure, by measuring the residual discrepancy between the model and viewed image, and using minimization techniques to reduce this error. This procedure is similar to the approach taken by the deformable template method (Yuille and Hallinan, 1992). The advantage of the search approach is that it does not rely on the establishment of feature correspondence between the image and a stored model. The main disadvantage is that the search will typically require the generation and comparison of multiple internal patterns. In computational experiments with this scheme (Ullman and Zeira, 1997),a total of several hundred intermediate patterns were generated in the course of the recognition processes. For a biological system, performing in parallel multiple pattern comparisons may have an advantage over the explicit recovery of the transformation parameters. These considerations have discussed in more detail previously (Ullman, 1995, 1996).

2.4. Adding abstract descriptions

In the discussion so far we have treated object views in a simplified form. Object models consisted of image contours and similar features, without defining larger structures such as object parts, as used in the structural description approach to recognition (Biederman, 1985),or using abstract descriptions, as in the invariant properties approach (Ullman, 1989; Mundy and Zisserman, 1992). It is possible, however, to combine the main advantages of the part decomposition and invariant properties approaches with the view-combination approach. The resulting scheme is likely to be more suitable for recognizing objects that cannot be handled easily by the simpler method alone.

To illustrate how abstract descriptions might be used, suppose that one is trying to recognize a familiar person with characteristic curly hair. In matching the novel view with a previously stored model, the contours comprising the hair region are not expected to match in detail. However, the corresponding regions in the model and the image are expected to have similar textural properties. To compare the image and model at a more abstract level, one can imagine a region descriptor, or 'label', describing, for example, texture and color properties, being overlaid over the hair region. This description is abstract in the sense that it is less specific than the original image itself; many different images will map onto the single label ('curly' in this case). When the internal model is manipulated and compared with the viewed object, the detailed internal contours in the two will not be in close agreement, but they will both have the same label in corresponding locations. Abstractions of this type can describe properties of 2D contours, but may also be 3D in nature, such as convex or concave regions. It appears that the inclusion of multiple levels of abstraction is an important future direction, that will make view-based schemes more flexible and robust.

3. Class-based view combinations

In the discussion above we have seen how novel views of an object can be recognized by combinations of a number of representative views. Although the view-combination scheme allows a small number of stored views to deal with a large range of novel viewing directions, it still requires, for each individual object, a sufficient number of representative views. In contrast, it appears that the human visual system can obtain substantial generalization on the basis of a single view of a novel object. Single-view generalization can sometimes be based on the presence of a distinctive feature, such as a scar or birthmark on a face. It appears, however, that significant generalization from a single view can be obtained even in the absence of such distinctive features. For example, a study by Moses et al. (1996), examined the ability of human observers to generalize in face recognition to novel viewing directions and illumination conditions on the basis of a single example view. In this study, subjects were presented with a single image of each of a number (three or more) of individuals. They were later tested with additional images of the same individuals, but under novel viewing conditions. The results showed almost error-free recognition across wide changes in viewing direction (51°) and illumination conditions (e.g. left vs. right). Generalization was also tested for inverted face images. Recognition of inverted faces is known to be more difficult than of upright faces. In this study, however, the focus was not on the overall difficulty of the task, but on the ability to generalize from one view to another. It turned out that for inverted faces, even after training that made the training images easily recognizable and without error, generalization from a single view to novel conditions was limited.

Generalization from a single view was also examined in a study by Tarr and Gauthier (1998) using a set of artificial objects. Subjects were trained with multiple views of several similar objects. A novel object from the same general class was shown under a single viewing direction, but tested with different views. They found that the training objects, even a single similar object, could facilitate the recognition of the novel object under the new viewing directions.

At an intuitive level, the results are perhaps not surprising. For example, observers have seen in the past many different face images. Upon seeing a novel face image illuminated from the right, say, the visual system might be able to use its past experience to deal with the same face, only illuminated from the left. The question then arises as to how prior experience with different objects in the same general class might be used to facilitate the recognition of new individual members of the class.

Within the view-combination approach, such class-based generalization can be obtained by using views of different objects instead of different views of the same single object. Mathematically, the main process proceeds as follows (for a fuller description see Beymer and Poggio (1995) and Sali and Ullman (1998)).

Suppose that we have seen a number of objects, such as faces, where each of the objects is seen under two different viewing conditions, that we will call ‘frontal’ and ‘non-frontal’. Let $V_1,...,V_k$ be the frontal views and $U_1,...,U_k$ the corresponding non-frontal views. We now have a single frontal view V' of a novel object, and we wish to predict U', the appearance of the novel object in the non-frontal view, based on its single view as well as the views of the other objects. We start by approximating the novel frontal view as a combination of the frontal views of the other k objects:

$$V' = \Sigma \mathrm{a_i} V_\mathrm{i} + \Delta \tag{3}$$

$\mathrm{V_i}$ are the k known views, $\mathrm{a_i}$ are the coefficients of the combination that are chosen to obtain the closest possible approximation to the novel view. The quantity Δ is the residual error that can be significant, especially if the number of examples used is small. To predict the new appearance U' we express it as a combination of the non-frontal views U_i with the same coefficients $\mathrm{a_i}$ (Beymer and Poggio, 1995) and the same residual Δ (Sali and Ullman, 1998):

$$U' = \Sigma \mathrm{a_i} U_\mathrm{i} + \Delta \tag{4}$$

This process can be used to obtain class-based generalization to changes in viewing direction as well as illumination changes. An example is shown in Fig. 3, depicting the result of this process applied to a novel face image on the basis of known face images. In this example, generalization to a new viewing direction is obtained from a single image of the novel object on the basis of three other objects in the same class.

In most approaches to recognition, generalization is obtained based on information associated with a single object. For example, in the view-combination approach, different views of the same object are combined to deal with novel views of the object in question. Similarly, in the structural descriptions approach, an object description is constructed for a given object, independent of the descriptions of similar objects. The discussion above illustrates that generalization in recognition can be class-based rather than object-based. Computationally, class-based schemes provide means for dealing with novel objects by using past experience with similar objects. Psychophysically, it appears that such processes play an important role in human object recognition. It seems, therefore, that class-based recognition is an important direction for further study in both the theoretical and the empirical studies of visual object recognition.

4. Extensions to the basic scheme

The image combination scheme outlined above is restricted in a number of ways, and it will be of interest to extend it in several directions. One attractive extension that will not be considered here is the use of object parts and partial views. The scheme discussed so far used views of the entire objects. It may be advantageous,

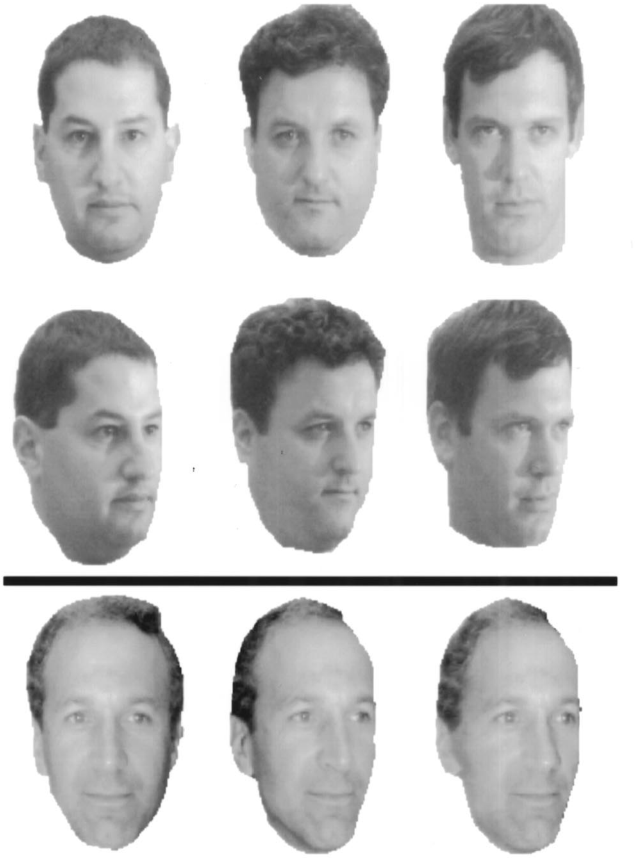

Fig. 3. Class-based generalization from a single view. Three face examples are shown above the horizontal line, in both frontal and side views. Below the line: a novel face, frontal view (left), a non-frontal view generated from the three examples and the frontal view (right). The view in the center is an actual side view, for comparison.

however, to use in a similar manner views covering only parts of the objects. Another important issue is the problem of model selection. The discussion so far has assumed that a candidate model has been selected, and the scheme evaluates the agreement between this selected model and the viewed object. If many object models are stored in memory, do we have to examine all of them in this manner, or can we somehow focus on a smaller number of potential models? Other issues include non-orthographic image views, non-linear image combinations, and dealing effectively with occlusion. This section contains a brief discussion of recent progress and future directions in these areas.

4.1. Perspective projections

The scheme as presented assumes rigid transformations and an orthographic projection. Under these conditions, all the views of a given object are embedded in a low-dimensional linear subspace of a much larger space. What happens if the projection is perspective rather than orthographic, or if the transformations are not entirely rigid?

The effect of perspectivity appears to be quite limited. Ullman and Basri (1991) applied the linear combination scheme to objects with a ratio of distance-to-camera to object-size down to 4:1, with only minor effects on the results. Instead of using the mathematically convenient approximation of orthographic projection it is also possible to perform image combination directly using perspective views. The view-combination scheme provides a method for predicting a new view of an object based on a model constructed from two or more corresponding views. Shashua (1995) developed a similar method for perspective views. The direct use of perspective views increases the accuracy of the reconstruction, but a larger set of corresponding features (seven or more) is required.

4.2. Non-linear image combinations

As for non-rigid transformations and other possible distortions, an interesting general extension to consider is where the set of views is no longer a linear subspace, but still occupies a low-dimensional manifold within a much higher-dimensional space. This manifold resembles locally a linear subspace, but it is no longer ‘globally straight’. By analogy, one can visualize the simple linear combinations case in terms of a 3D space, in which all the orthographic views of a rigid object are restricted to some 2D plane. In the more general case, the plane will bend, to become a curved 2D surface within the 3D space.

This issue of dealing effectively with lower-dimensional subspaces appears to be a general case of interest for recognition as well as for other learning tasks. The general reason can be explained by the following consideration. For recognition to be feasible, the set of views $\{V\}$ corresponding to a given object cannot be arbitrary, but must obey some constraints, that may be expressed in general in the form $F(V_{\mathrm{i}}) = 0$. Under general conditions, these restrictions will define locally a manifold embedded in the larger space (as implied by the implicit function theorem). Algorithms that can learn to classify efficiently sets that form low-dimensional manifolds embedded in high-dimensional spaces will therefore be of general value.

An elegant approach that can use general, non-linear combination of images, is the radial basis functions (RBF) method, developed by Poggio and his collaborators (Poggio and Edelman, 1990; Poggio and Girosi, 1990). This method uses non-linear interpolation between 2D images for the purpose of recognizing 3D objects, as well as for other tasks that involve learning from examples.

4.3. Occlusion

The problem of occlusion is an important issue in any theory of visual recognition. The recognition process must be able to deal with incomplete data, and to distinguish between two sources of image-to-model mismatch: an incomplete match that results from occlusion, and mismatches that indicate the use of an inappropriate model.

In the view-combination scheme, combined views consist of object points that are visible in the generating views. Problems may arise, therefore, either because of self-occlusions (when object points are occluded from view by the object itself), or because of occlusion by other objects.

The problem of self-occlusion is handled in the view-combination approach by representing an object not by a single model, but by a number of models covering its different 'aspects' (Koenderink and Van Doorn, 1979). To cover the object completely, the main mathematical requirements are that each object point will be visible from at least two views, and, roughly, that each view will have at least four points in common (and with known correspondence) with two or more views. The overall number of required views is not fixed, but depends on the object. In practice, it was found that to distinguish between similar car models, for example, typically ten views or fewer were sufficient.

As for occlusion by other objects, the method is somewhat less sensitive to this problem than other methods such as the use of invariants or structural descriptions. In the simplest version of using view combinations, a small number of corresponding features are used to recover the coefficients of the combination. Occlusion will not present a major difficulty to this method, provided that the visible part of the object is sufficient for obtaining correspondence and thus for recovering the required coefficients. After performing the view combination, a good match will be obtained between the transformed model and the visible part of the object. The search method mentioned above is the least affected by occlusion. When the correct parameters are reached, a good match will be obtained between the model and the unoccluded part of the object.

To deal effectively with occlusion, the matching function used by the scheme must be able to distinguish between two different cases of a partial match between the image and the model. Occlusion results in a close match, restricted to a part of the object. The use of an inappropriate model can lead to some moderate agreement over a large portion of the object. A close agreement over a sufficient part of the object should therefore provide a stronger indication for the presence of an object than an inaccurate match spread over a larger region.

4.4. Multiple models and the role of classification

Objects can be recognized at different levels of specificity. Sometimes they are assigned to a general class, such as a 'house', 'dog', 'face' – classes that contain a variety of objects, of many different shapes. Objects can also be identified as unique individuals, such as someone's house, or a particular friend's face.

The scheme considered so far was aimed primarily at object identification. The image combination scheme can distinguish well between individual objects, such as two cars that have closely similar shapes. It will have a harder time, however, classifying a new car, whose shape is not yet represented in the system.

Classification is an important problem in its own right. It is clearly useful to be able to classify a novel object as a car, or a person, and so on, even if we have not seen the particular individual before. The ability to recognize objects at different levels of generality is therefore an important aspect of recognition.

From the point of view of individual identification, classification can also serve a useful role, particularly in dealing with large collections of stored objects. Classification can be used on the way to more specific identification in a number of ways. First, classification can reduce the number of candidate object models. The view-combination scheme discussed above assumed that a candidate object model has been selected, and the task of the recognition system is then to compare the internal model with a novel view of the object. When the number of object models is large, it becomes desirable to reduce the number of candidate models and allow the system to focus on the more likely object models. Classification can be useful in this process, by directing subsequent processing to a restricted class of models. If the image can be classified, for example, as representing a face, without identifying the individual face, then subsequent processing stages can be restricted to face models, ignoring models of other 3D objects.

Classification can also allow the recognition system to use class-specific information in the recognition process. Different classes of objects can undergo some characteristic transformations, for example, faces can be transformed by facial expressions, that are specific to this class of objects. Following classification, information regarding the relevant set of transformations can be used for the recognition of a specific individual within the class. Finally, as we have seen in the discussion of class-based view combinations, classification can help to generalize from limited object-specific information.

Classification is therefore a useful intermediate stage within a recognition system on the way to more specific identification. It restricts the set of candidate models, allows the use of class-specific information, and makes it possible to perform broader generalization by supplementing object-specific information with class-based information. It appears that the problems of general classification – how it is performed, and how it is related to more specific identification – will be important issues in the study of visual recognition in the future. It also remains to be seen whether classification can be performed by view-based methods, or whether more abstract approaches, such as the use of structural descriptions, are required for this task.

5. Psychophysical and physiological aspects

In this section I review briefly psychophysical and physiological findings that are relevant to view-based recognition. The psychological and biological study of object recognition is not an easy task, because recognition is likely to involve a range of different and interacting processes. In an empirical setting, one can bias the recognition process to use different routes, by using different recognition tasks. For example, if in a recognition test one of the objects has a unique distinctive feature, this feature will often be used to distinguish it from other objects (Eley, 1982; Murray et al., 1993). Similarly, the distinction between a small number of highly different objects will produce different results compared with the recognition of a larger number of generally similar objects. In other situations, recognition may involve reasoning about the object's function rather than the direct use of visual cues (Warrington and Taylor, 1978).

The findings listed in this section are not intended, therefore, to argue that view-based mechanisms are used exclusively in visual recognition, but that they play an important, perhaps a major part, particularly in the fast identification of individual objects. It should also be noted that, given the current state of knowledge, it would be premature to consider in detail specific mechanisms. The focus should therefore be on the main underlying principles, for example, that the system stores a number of different views of an object, and that these views are used collectively to compensate for the variability across views.

5.1. Psychophysical evidence

A large body of psychophysical evidence has been accumulated regarding the processes of visual object recognition. I will list here mainly recent findings that are directly related to view-based recognition.

5.1.1. New views are more difficult than trained ones

A number of different studies have examined the dependence of recognition performance on the viewing direction. In such studies an object is usually presented at a single orientation, and recognition is subsequently tested with the object presented at novel 3D orientations. In many studies of this type it was found that recognition performance decreased with the departure from the original, trained orientation: the error rates typically increased, as well as the response time for correct recognition (Jolicoeur, 1985, 1990; Rock and Di Vita, 1987; Corballis, 1988; Tarr and Pinker, 1989, 1991; Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992). In structural description as well as in variance-based schemes, the precise view has no particular importance, since it is replaced during the processing by a view-invariant description. As long as the new view gives rise to the same structural description, or has the same invariances, no significant effect of viewing direction is expected. The findings are more consistent, therefore, with theories, such as the image combination scheme, that use multiple 2D views directly in the recognition process. For example, the RBF method, as well as some implementations of the linear combination scheme, will exhibit this superiority of the training views.

It should be noted, however, that not all the studies in this area show the 3D orientation effect. For example, Biederman and Gerhardstein (1993) found complete invariance of their test objects to viewing position. The objects in this study were designed to have a clear decomposition into a small number of simple parts, and this design may have contributed to the difference between this and other studies. In any case, the exact nature of the viewing position dependency is still a matter of some controversy, and further studies will be required to clarify the issue.

5.1.2. The difficulty persists when strong 3D cues are available

The dependence on viewing direction is present even when the object is seen in both the training and subsequent testing under conditions that facilitate the recovery of 3D shape, using stereo, shading, and motion (Edelman and Bülthoff, 1992). Overall recognition performance is somewhat improved under these conditions, but the decrease in performance with departure from the trained conditions is not significantly affected. A study by Sinha (1995) manipulated systematically 2D and 3D similarities and compared their effects on generalization in recognition. A training object was viewed in this study under good 3D viewing conditions. Subsequent test objects were either similar to the training object in their 2D view, but with different 3D structure, or similar in 3D shape but with a different 2D view. The results indicated that generalization was determined primarily by similarity of views, rather than of object-centered structure. These findings provide evidence that the generalization process to novel conditions does not benefit significantly from the availability of rich 3D information. In the view-combination approach, generalization depends primarily on the availability of additional views. At the same time, as noted in the theoretical discussion, additional 3D information can be used by a view-based approach to improve the recognition of the trained views themselves.

5.1.3. Generalization improves with additional views

Recognition improves after training with additional object views (Tarr and Pinker, 1989; Poggio and Edelman, 1990). A similar finding was also observed in monkeys trained for object recognition (Logothetis et al., 1994). This is again expected in the view-based approach, where generalization depends primarily on the availability of a sufficient number of representative views. The fact that for bilaterally symmetric objects generalization requires fewer views, and that good generalization can be obtained on the basis of a single view (Vetter et al., 1994), is also consistent with this point of view.

5.1.4. Better generalization to same-axis rotation

In a study by Bülthoff and Edelman (1992) subjects were also presented with multiple views of the same object; however, the views were all obtained by rotations of the object about a fixed axis in the image plane, such as the horizontal or the vertical axis. The generalization to novel views proved to be better for views obtained by further rotations about the same axis, compared with rotations about the orthogonal axis. This effect appears unexpected, except for the image combination approach. In this approach, combination of images obtained from rotations about, say, the vertical axis, produce new images that are also constrained to rotations about the same axis.

5.1.5. Recognition is better than an ‘ideal 2D observer’

The evidence listed above argues in favor of the direct use of object views in recognition, but does not address the issue of using the views independently or in some sort of view-combination scheme. A study by Liu et al. (1995) addressed this issue by comparing recognition performance of human observers with what they termed a ‘2D ideal observer’. The ideal observer compares a novel object view with all the previously seen views of the same object. The comparison is performed with each of the stored views separately, as opposed to the use of view combinations. They found that humans performed better than the ideal observer in generalizing to new views, demonstrating that independent comparisons to stored views are insufficient to account for human recognition performance.

5.2. Physiological aspects

The primate visual cortex contains multiple areas, but not all of them appear to be directly related to shape-based object recognition. A division has been suggested between two main processing streams, a ventral stream leading from V1 to the inferotemporal cortex (IT), where shape processing, leading to object recognition, seems to take place (Ungerleider and Mishkin, 1982), and a more dorsal stream, going to the parietal cortex, that may be related to object-directed action (Goodale and Humphrey, 1998).

The notion that IT cortex is involved with object recognition is supported by brain lesion studies, functional MRI (fMRI) studies, and by single cell recordings. Damage to IT cortex can cause deficits in object recognition. More restricted lesions usually affect mainly the precise identification of individual objects, and more extensive lesions can also affect the ability to perform broader classification (Damasio et al., 1990). Studies using fMRI techniques also indicate that these areas are involved in shape processing and object recognition (Tootel et al., 1996).

Single-cell recordings in IT showed that the stimuli required to drive these cells are often complex compared with those of lower-level visual areas. A particular population of such cells, mainly in the STS region of IT, responds selectively to face images (Perret et al., 1982, 1985; Rolls, 1984; Gross, 1992; Young and Yamane, 1992); other cells in a nearby region have been reported to respond to hand images. Some of the face-selective cells respond best to complete face images, others prefer face parts, such as the eye or mouth regions. In experiments performed by Logothetis et al. (1995), where monkeys viewed novel wireframe and ameboid objects for an extended training period, cells in IT developed specific responses to these novel objects. Similar findings were reported by Miyashita (1988) after training with fractal-like patterns. In addition to cells responding to complex and meaningful stimuli, other cells in IT, especially the posterior region, have been reported to respond to more elementary shapes and shapes that do not correspond to familiar objects (Tanaka, 1996).

In considering the response of IT units to specific objects, it is worth noting that shape-selective cells in IT often respond in a graded fashion, and will respond not only to a single shape, but to similar shapes as well. Some units also respond to a number of different shapes (Rolls et al., 1996). The shape of a specific object, such as a face, may still be represented by such cells in the population response of a number of cells. This may be a general principle of encoding of objects’ shapes, that are represented by the combined activity of cells broadly tuned to different shapes, as well as units tuned to different parts of the entire shape.

5.2.1. Cells in IT are usually view-selective

Many IT cells show considerable selectivity in their responses when the stimulus changes in size, position, color, and sometimes orientation in the image plane. In terms of 3D viewing angle, most cells prefer a particular view of the objects. Some cells show a selective response to an individual face, or an artificial wireframe object, over a considerable range of viewing directions. At the same time, the response is usually not entirely object-centered. In face-selective cells, for instance, the selectivity does not cover the full range from frontal to profile view. The response is typically optimal for a particular orientation and decreases gradually for other orientations, but the decrease is sharper for some units and more gradual for others.

A possible interpretation of the increased tolerance to viewing direction is that a broadly-tuned cell receives converging input from a number of view-specific cells (Logothetis et al., 1995). This convergence of views belonging to the same object may be related to the view-combination approach; it may be a part of a biological implementation combining different views of the same object, leading to a broader generalization to novel views.

5.2.2. The response is determined by 2D similarity of views

The pattern of responses in IT to complex shapes appears to be consistent with the general notion of multiple pictorial representations. An object appears to be represented in IT by multiple units, tuned to different views of the object. The response of a face-selective unit, for example, to a given stimulus, appears to be governed by the overall similarity between the stimulus and the unit's preferred 2D pattern (Young and Yamane, 1992). Similarly, the units studied by Tanaka (1996) and his collaborators appear to be governed by 2D similarity of the test view to the view preferred by the unit. In these experiments, units were first tested using a large collection of different 3D objects. When a unit responded well to a particular object, additional attempts where made to characterize more precisely the effective stimulus for the unit. Units were typically driven by particular 2D patterns, rather than, for example, some preferred 3D shapes, regardless of their orientation in space.

5.2.3. Lesion evidence for the primacy of stored views

There is some evidence from animal lesion studies consistent with the notion that the system uses as a basic mechanism a direct comparison to stored views, augmented by mechanisms that are responsible for aspects of generalization to novel views.

For example, damage to area V4 and posterior IT (Weiskrantz, 1990; Schiller and Lee, 1991; Schiller, 1995) appears to affect especially the ability to compensate for transformations such as size, orientation, or illumination changes. The animal's ability to recognize the original views usually remains intact after the lesion. In the experiments carried out by Weiskrantz, monkeys were trained to recognize a set of test objects, under particular orientation and illumination conditions. They were then tested with the original objects as well as objects similar to the test objects except for changes in scale, orientation in the image plane, and illumination. Lesions to area AIT, the anterior part of IT, caused a general deterioration in recognition capacity. Lesions to the more posterior part of IT and some prestriate areas had a more specific effect on the ability to recognize the transformed views. Recognition of the original views were usually unaffected. Such results appear to support the notion that the most basic form of recognition relies on the direct comparison of stored views, as in the view-based approach, together with additional mechanisms that allow the undamaged system to also go beyond the stored views, and generalize to new ones.

6. Conclusions

A major problem in visual recognition comes from the fact that images of the same object are highly variable. To deal with this variability, one general approach has been to move away from the pictorial level and generate instead more abstract and view-independent representations. The view-based approach outlined here relies on the fact that the variability in the set of views belonging to a single object is still governed by regularities that can be captured at the pictorial level.

In the scheme presented above, an object is represented for the purpose of recognition by a number of its views, rather than, for instance, a single 3D object-centered representation. The views comprising the representation of a single object are not merely a collection of independent 2D object views. In the direct approach to recognition, objects are also represented by multiple views, but recognition is based simply on the best-matching individual view. In contrast, in the multiple views approach a number of object views are used collectively in the recognition process. The multiple views used to represent the object also include a known correspondence between individual views. As was discussed above, a set of corresponding 2D views provide a powerful and useful representation for the purpose of recognition. Without using explicit 3D information, this representation contains detailed information about the object’s structure, and this information is stored in a convenient form for the purpose of the recognition process. The object views used in this approach are not limited to simple images of the object. The use of abstract descriptions allows the scheme to incorporate in addition more abstract pictorial representations.

The approach described above leaves several problems unanswered. Some of the major ones listed in the paper include the use of abstractions, class-based recognition, and the problem of classification. These and other problems will require considerably more research, both empirical and computational. In considering these problems, it should be emphasized that recognition is probably more than a single process; there may be many and quite different processes used by the visual system to classify and identify visual stimuli. Object recognition may be analogous in this respect to the perception of 3D space: the perception of depth and 3D shape is not a single module, but is mediated by a number of interacting processes that utilize various sources of information, such as binocular disparity, motion parallax, surface shading, contour shape, and texture variations. Similarly, visual object recognition is probably better viewed not as a single module, but as a collection of interacting processes.

Computational studies outlined in this paper show that significant aspects of object recognition can be approached by using combinations of a small number of object views. In examining the human recognition system it is worth keeping in mind, however, the distinction between the mathematical formulation of particular algorithms and the more general properties of the approach. The image combination computation described above illustrates the approach; however, variations and extensions of the scheme are possible. The general suggestion, based on the computational studies, is that recognition by multiple pictorial representations and their combinations may constitute a major component of 3D object recognition. According to this view, the brain will store for each object a small number of pictorial descriptions, and the recognition process will involve the manipulation and some combination of these views.

Acknowledgements

This work was supported by the Brain and Computation grant of the Israel Science Foundation and by the Israel Ministry of Science under the Integrated Visual Platform Project. I thank R. Basri, Y. Moses, A. Zeira and E. Sali for their contributions. The face images are from the Weizmann Institute Face Base. Thanks to IEEE PAMI for the reproduction of Fig. 1.

References

Abu-Mostafa, Y.S., Psaltis, D., 1987. Optical neural computing. Scientific American 256, 66–73.

Adini, Y., Moses, Y., Ullman, S., 1997. Face recognition: the problem of compensating for illumination changes. IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7), 721–732.

Albert, A., 1972. Regression and the Moore–Penrose Pseudoinverse. Academic Press, New York.

Beymer, D., Poggio, T., 1995. Face recognition from one example view. Proceedings of the International Conference on Computer Vision ICCV – 1995, 500–507.

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

Biederman, I., Gerhardstein, P.C., 1993. Recognition of depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. Journal of Experimental Psychology: Human Perception and Performance 19, 1162–1182.

Bülthoff, H.H., Edelman, S., 1992. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proceedings of the National Academy of Science USA 89, 60–64.

Corballis, M.C., 1988. Recognition of disoriented shapes. Psychological Review 95, 115–123.

Daugman, J.G., 1989. Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. IEEE Transactions on Biomedical Engineering 36 (1), 107–114.

Damasio, A.R., Damasio, H., Tranel, D., 1990. Impairment of visual recognition as clues to the processing of memory. In: Edelman, G.M., Gall, W.E., Cowan, W.M. (Eds.), Signal and Sense: Local and Global Order in Perceptual Maps. John Wiley, New York.

Edelman, S., Bülthoff, H.H., 1992. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. Vision Research 32, 2385–2400.

Eley, M.G., 1982. Identifying rotated letter-like symbols. Memory and Cognition 10, 25–32.

Goodale, M.A., Humphrey, G.K., 1998. The objects of action and perception. Cognition 67, 181–207.

Gross, C.G., 1992. Representation of visual stimuli in inferiortemporal cortex. Philosophical Transactions of the Royal Society, London B 335, 3–10.

Hopfield, J.J., 1982. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Science USA 79, 2554–2558.

Jolicoeur, P., 1985. The time to name disoriented natural objects. Memory and Cognition 13, 289–303.

Jolicoeur, P., 1990. Orientation congruency effects on the identification of disoriented shapes. Journal of Experimental Psychology: Human Perception and Performance 16, 351–364.

Koenderink, J.J., Van Doorn, A.J., 1979. The internal representation of solid shape with respect to vision. Biological Cybernetics 32, 211–216.

Koenderink, J.J., Van Doorn, A.J., 1990. Receptive field families. Biological Cybernetics 63, 291–297.

Kohonen, T., 1978. Associative Memories: a System Theoretic Approach. Springer, Berlin.

Lipson, P., 1993. Model Guided Correspondence. MSc Thesis, Computer Science, Massachusetts Institute of Technology.

Liu, Z., Knill, D.C., Kersten, D., 1995. Object classification for human and ideal observers. Vision Research 35 (4), 549–568.

Logothetis, N.K., Pauls, J., Bülthoff, H.H., Poggio, T., 1994. View-dependent object recognition in monkeys. Current Biology 4, 401–414.

Logothetis, N.K., Pauls, J., Bülthoff, H.H., Poggio, T., 1995. Shape representation in the inferior temporal cortex of monkeys. Current Biology 5, 552–563.

Lowe, D.G., 1985. Perceptual Organization and Visual Recognition. Kluwer Academic Publishing, Boston, MA.

Marr, D., Hildreth, E.C., 1980. Theory of edge detection. Proceedings of the Royal Society, London B 207, 187–217.

Miyashita, Y., 1988. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335, 817–820.

Moses, Y., 1993. Face Recognition: Generalization to Novel Images. PhD Thesis, Applied Mathematics and Computer Science, Weizmann Institute of Science, Israel.

Moses, Y., Ullman, S., Edelman, S., 1996. Generalization to novel images in upright and inverted faces. Perception 25, 443–461.

Mundy, J.L., Zisserman, A. (Eds.), 1992. Geometric Invariance in Computer Vision. MIT Press, Cambridge, MA.

Murray, J.E., Jolicoeur, P., McMullen, P.A., Ingleton, M., 1993. Orientation-invariant transfer of training in the identification of rotated natural object. Memory and Cognition 21, 604–610.

Perret, D.I., Rolls, E.T., Caan, W., 1982. Visual neurons responsive to faces in the monkey temporal cortex. Experimental Brain Research 47, 329–342.

Perret, D.I., Smith, P.A.J., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, A.D., Reeves, M.A., 1985. Visual cells in the temporal cortex sensitive to face view and gaze direction. Proceedings of the Royal Society B 223, 293–317.

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

Poggio, T., Girosi, F., 1990. Regularization algorithms for learning that are equivalent to multilayer networks. Science 247, 978–982.

Rock, I., Di Vita, J., 1987. A case of viewer-centered object perception. Cognitive Psychology 19, 280–293.

Rolls, E.T., 1984. Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. Human Neurobiology 3, 209–222.

Rolls, E.T., Booth, M.C.A., Treves, A., 1996. View-invariant representations of objects in the inferior temporal cortex. Society of Neurosci Abstracts 22, 1937.

Sali, E., Ullman, S., 1998. Recognizing novel 3-D objects under new illumination and viewing position using a small number of examples. Proceedings of the International Conference on Computer Vision ICCV – 1998, pp. 153–161.

Schiller, P.H., Lee, K., 1991. The role of the primate extrastriate area V4 in vision. Science 251, 1251–1253.

Schiller, P.H., 1995. Effect of lesion in visual cortical V4 on the recognition of transformed objects. Science 376, 342–344.

Shashua, A., 1992. Geometry and Phometry in 3-D Visual Recognition. PhD Thesis, Department of EECS, Massachusetts Institute of Technology.

Shashua, A., 1995. Algebraic function for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 17 (8), 779–789.

Sinha, P., 1995. Perceiving and Recognizing Three-dimensional Forms. PhD Thesis, Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Tanaka, K., 1996. Inferotemporal cortex and object vision. Annual Review of Neuroscience 19, 109–139.

Tarr, M.J., Gauthier, I., 1998. Do viewpoint-dependent mechanisms generalize across members of a class? Cognition 67, 71–109.

Tarr, M.J., Pinker, S., 1989. Mental rotation and orientation dependence in shape recognition. Cognitive Psychology 21, 233–282.

Tarr, M.J., Pinker, S., 1991. Orientation-dependent mechanisms in shape recognition: further issues. Psychological Science 2, 207–209.

Tootel, R.B.H., Dale, A.M., Sereno, M.I., Malach, R., 1996. New images from the human visual cortex. Trends in Neurosciences 19 (11), 481–489.

Ullman, S., 1989. Aligning pictorial descriptions: an approach to object recognition. Cognition 32 (3), 193–254.

Ullman, S., 1995. Sequence seeking and counter streams: a model for bi-directional information flow in the visual cortex. Cerebral Cortex 5 (1), 1–11.

Ullman, S., 1996. High-level Vision: Object Recognition and Visual Cognition. MIT Press, Cambridge, MA.

Ullman, S., Basri, R., 1991. Recognition by linear combinations of models. IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (10), 992–1006.

Ullman, S., Zeira, A., 1997. Object recognition using stochastic optimization. In: Pelillo, M., Hancock, E.R. (Eds.), Energy Minimization Methods in Computer Vision and Pattern Recognition, Lecture Notes in Computer Science. Springer, Berlin, pp. 329–344.

Ungerleider, L.G., Mishkin, M., 1982. Two cortical visual systems. In: Ingle, D.J., Goodale, M.A., Mansfield, R.J.W. (Eds.), Analysis of Visual Behavior. MIT Press, Cambridge, MA, pp. 549–586.

Vetter, T., Poggio, T., Bülthoff, H., 1994. The importance of symmetry and virtual views in three dimensional object recognition. Current Biology 4, 18–23.

Warrington, E.K., Taylor, A.M., 1978. Two categorical stages of object recognition. Perception 7, 152–164.

Weiskrantz, L., 1990. Visual prototypes, memory, and the inferotemporal lobe. In: Iwai, E., Mishkin, M. (Eds.), Vision, Memory and the Temporal Lobe. Elsevier, New York, pp. 13–28.

Willshaw, D.J., Buneman, O.P., Longuet-Higgins, H.C., 1969. Non-holographic associative memory. Nature 222, 960–962.

Young, M.P., Yamane, S., 1992. Sparse population coding of faces ininferotemporal cortex. Science 256, 1327–1331.

Yuille, A., Hallinan, P., 1992. Deformable templates. In: Blake, A., Yuille, A. (Eds.), Active Vision. MIT Press, Cambridge, MA.