PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Low-Level Aspects of Segmentation and Recognition [and Discussion]

Shimon Ullman, R. L. Gregory and J. Atkinson

Phil. Trans. R. Soc. Lond. B 1992 **337**, 371-379
doi: 10.1098/rstb.1992.0115

Low-level aspects of segmentation and recognition

SHIMON ULLMAN

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, U.S.A. and Department of Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel

SUMMARY

This paper discusses two problems related to three-dimensional object recognition. The first is segmentation and the selection of a candidate object in the image, the second is the recognition of a three-dimensional object from different viewing positions.

Regarding segmentation, it is shown how globally salient structures can be extracted from a contour image based on geometrical attributes, including smoothness and contour length. This computation is performed by a parallel network of locally connected neuron-like elements. With respect to the effect of viewing, it is shown how the problem can be overcome by using the linear combinations of a small number of two-dimensional object views.

In both problems the emphasis is on methods that are relatively low level in nature. Segmentation is performed using a bottom-up process, driven by the geometry of image contours. Recognition is performed without using explicit three-dimensional models, but by the direct manipulation of two-dimensional images.

1. LOW-LEVEL AND HIGH-LEVEL VISION

Although there is no sharp and universally accepted distinction between the domains of 'low-level' and 'high-level' vision, these terms are nevertheless useful and broadly used in analysing the processing of visual information. Low-level vision is usually associated with bottom-up and spatially uniform processing. High-level vision is associated with top-down processing that often employs object-specific knowledge, and is often regarded as more symbolic or reasoning-like in nature.

The process of object recognition certainly involves the use of stored knowledge about specific objects, and it is therefore often regarded as a quintessential example of high-level visual processing. Yet within the spectrum of visual recognition theories, some schemes are distinctly lower-level in nature than others. The use of template-matching, or of a large scale associative memory (Abu-Mostafa & Psaltis 1987; Hopfield 1982; Kohonen 1978; Willshaw *et al.* 1969) are examples of relatively simple and low-level schemes, that compare directly two-dimensional images with stored templates. Examples of recognition theories that can be considered higher level in nature are structural description schemes, such as the generalized cylinders scheme of Marr & Nishihara (1978) or Biederman's recognition-by-components (Biederman 1985). These methods produce a highly abstract description of the input shape in terms of its constituent parts and the spatial relations among them, and compare the resulting structural descriptions with similar object descriptions in long term store. Another example of a high-level recognition method is a scheme that attempts to recognize objects visually by reasoning about their possible function (Stark & Bowyer 1991).

In the context of segmentation, a similar distinction is often made between low-level and high-level components. Low-level components perform grouping and segmentation operations on the basis of image properties, such as proximity, similarity of color, motion, texture, etc. High-level processes in segmentation are those that use known shapes to perform segmentation and grouping (e.g. by looking for a particular object in the image, and using the identified object in the segmentation process).

In this paper I will discuss low-level aspects of image segmentation and object recognition. I will first discuss the problem of segmentation as a low-level autonomous process, and briefly describe a procedure that can extract fairly complex image structures, often disconnected, based on local geometrical properties of image contours. As for recognition, I will outline a scheme that recognizes three-dimensional objects without storing explicitly three-dimensional object models, but by using instead the linear combination of a small number of two-dimensional images of a given object. Clearly, higher level processes also play a role in segmentation and recognition, but, as we shall see, it appears that relatively low-level processes can be surprisingly powerful and solve significant aspects of the segmentation and recognition problems. In fact, in view of the performance of relatively simple biological visual system, such as the pigeon's for instance (Hernstein 1984), one might suspect that relatively

Phil. Trans. R. Soc. Lond. B (1992) **337,** 371–379
Printed in Great Britain

direct and low-level methods, relying on a significant memory store, might go a long way towards solving difficult visual recognition problems.

2. IMAGE SEGMENTATION

The process of segmentation means the partitioning of the image into different parts or structures (either by starting at the image level and breaking it down into smaller parts, or by starting with small image elements and grouping them into larger structures), and it also includes the problem of selecting one of the image structures for furhter analysis. This process plays an important role in recognition, because to recognize an object in a scene it is often necessary to select a portion of the image, and to apply to it additional processing stages, that will lead eventually to recognition. In natural environments that contain multiple objects, and where objects are often partially occluded, a considerable amount of processing may have to be completed before the image of an object or its parts can be interpreted. These and related problems have been studied extensively under different headings, such as 'segmentation', 'grouping', 'perceptual grouping' and 'figure-ground separation'.

As far as recognition is concerned, the goal of the segmentation stage is to produce useful structures, but it is not required to delineate complete objects, nor is it required to segment the entire image into the union of disjoint objects. I will return to examine the requirements placed on the segmentation stage following the discussion of the recognition problem.

(*a*) *Bottom-up segmentation processes*

In some cases, the segmentation process in human vision appears to rely heavily on high-level processes, that employ knowledge about objects to identify certain image structures as corresponding to a particular object. In R. C. Jamses' well-known image of the Dalmation dog, for instance (Gregory 1970), it appears unlikely that the portion of the image containing the dog can be identified in a purely bottom-up manner, based on image properties alone. Segmentation and recognition in this case are strongly coupled, and both are aided by the knowledge that one is looking for a dog.

There is considerable evidence, however, that human vision also contains processes that perform grouping and segmentation prior to, and independent of, subsequent recognition processes. One type of evidence comes from brain lesions leading to recognition problems. For example, Luria (1980) has described cases where the recognition of isolated objects remained relatively intact, but it became severely impaired for non-isolated objects, when recognition also required the separation of an object from partially overlapping or touching distractors.

A psychophysical demonstration of the power of low-level segmentation processes is Bregman's (1984; reproduced also by Nakayama *et al.* (1989)) figure of several objects (instances of the letter 'B') partially occluded by an ink blot. With the occluder present, the image organizes itself effortlessly into distinct objects. When the occluder is removed, the scattered fragments of the letters fail to organize. Knowledge about the content of the scene is of limited help: it is as if the visual system makes a decision at a low level as to which parts of the scene belong together. A related demonstration is Nakayama's face-behind-the-fence demonstration (Nakayama *et al.* 1989). A face partially occluded by horizontal bars is easily recognizable when the unoccluded parts lie behind, but not in front of, the occluding bars. Knowledge about the face has again limited influence on the segmentation process. Additional support for segmentation processes operating in early vision comes form demonstrations of motion and stereo capture (Ramachandran 1986), where segmentation processes appear to precede and influence the organization of motion and binocular stero matching in the image. For example, the motion of the boundary of a square can influence the perceived motion inside, but not outside, the square. This points to an early separation between the inside of an object and the surrounding background.

From a physiological standpoint there is some evidence that segmentation processes may start as early as in visual area V2, that receives its input directly from the primary visual area V1, and is placed low in the stream of visual processing. In V2 (but not V1), some of the units are sensitive to subjective contours (von der Heydt *et al.* 1984), as well as to the coherent motion of a row of dots relative to the background (Paterhans & von der Heydt 1991), and it seems likely that the analysis of these cues is related to the perception of occlusion cues and to image segmentation.

(*b*) *Extracting globally salient structures*

This section outlines briefly a method for extracting from an image salient structures such as the ones shown in figure 1. When we look at an image such as figure 1*a*, it appears to us that our attention is somehow immediately drawn to the main object which we then recognize as a car. For most observers, the car is found immediately, without the need to scan the image systematically, and without first attempting to recognize some structures in other parts of the image.

Structures that attract our attention need not be recognizable objects. In figure 1*b–d*, for example, the round blobs are relatively easy to detect as the most salient, figure-like structures in these images. These are examples of segmentation and selection processes that appear to operate early and define certain parts of the image as figure-like based on geometric properties of the contours.

In examining the processes that make such structures salient in our perception, it is useful to draw a distinction between local and global (or structural) saliency. Our attention is sometimes drawn to an item in the image because this item differs in some local property from neighbouring elements; for example, a green dot in an image of red dots, or a vertical line segment surrounded by horizontal ones. This pheno-

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

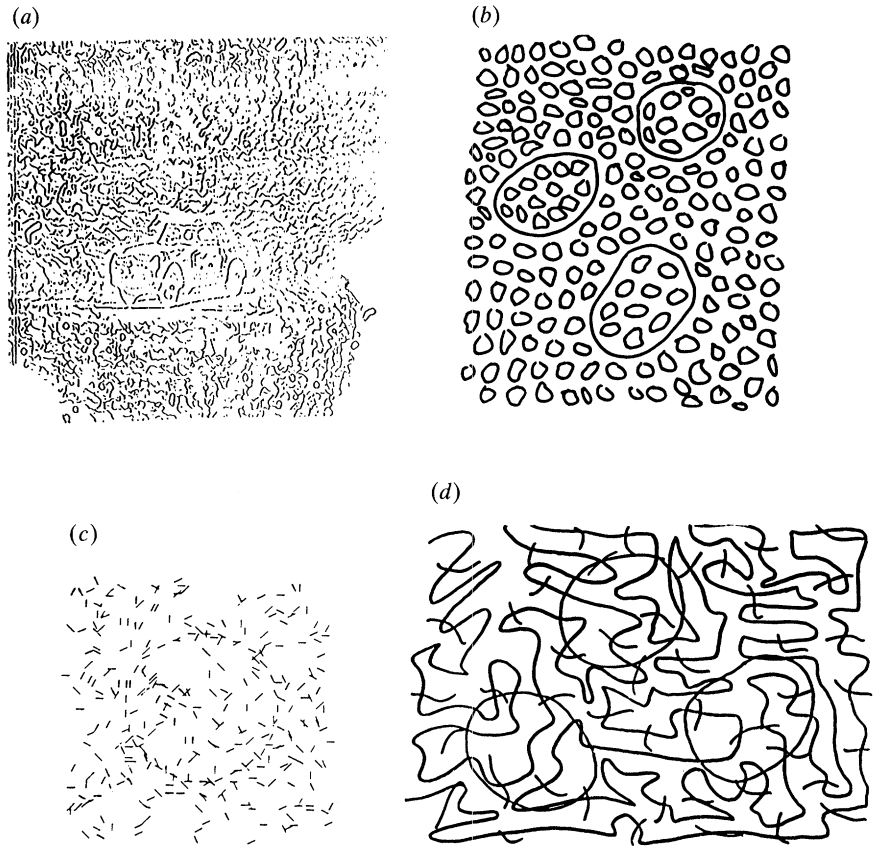

Figure 1. The perception of salient structures. The car in (*a*) attracts our attention, we can find it without scanning the image exhaustively. Figures (*b–d*) do not contain recognizable objects, but certain structures in the image are more salient and figure-like than others. The saliency in (*c*) and (*d*) is global rather than local.

menon of local saliency has been investigated in a number of psychological studies (e.g. Triesman & Gelade 1980; Julesz 1981). In other cases (such as figure 1*c*,*d*) the salient structure has no conspicuous local part, with a distinguishing local property such as colour, orientation, contrast, or cuvature. Although the elements comprising the structure are not individually salient, their arrangement makes the figure as a whole somehow globally conspicuous. In the more general case, the saliency of an image structure may be determined by the combination of both local and global aspects.

The section below describes a model that has been developed to extract certain classes of globally salient structures from images. This process is not intended to offer a full solution to the segmentation of such images, but to provide a useful stage in solving the problem.

(*c*) *Computing a saliency measure*

The model for extracting salient image structures proceeds by computing a measure of saliency at each point in the image. A successful model of this type will assign high saliency measure to image structures that are also salient in human perception, and should provide an efficient method for extracting from the image the conspicuous structures such as the car or the blobs in figure 1*b–d*.

For simplicity, the input image is assumed to be composed of contours. Such contours may be, for example, the lines and edges extracted from the image by line and edge detection processes. The saliency measure in the model increases with the contour's length, and decreases with its curvature or curvature variation; that is, the measure is designed to favor image contour that are long and smooth. In this section I will concentrate on a version that does not take curvature variation into account: the saliency measure increases with overall length, and decreases with total curvature. The use of length and curvature parameters was motivated by psychophysical observations regarding the effect of these parameters, and the exact form of the saliency measure was determined by computational considerations that are discussed in more detail below.

In defining the mathematical form of the saliency measure, it is convenient to consider first a single contour Γ in the image, and ignore all others. The contour is composed of a sequence of small line elements, that may be detected by local line (or edge) detecting units. Let p be a point on Γ, and $S_\Gamma(p)$ be the saliency measure at point p assuming that Γ is the only relevant curve. The saliency at p is then given by:

$$S_\Gamma(p) = \sum_i \omega_i \sigma_i.$$

In this expression σ_i is the local saliency of the ith edge element along the curve. For now, σ_i can be

thought as simply as having the value '1' for every edge element i, and '0' if the edge element is missing (i.e. there is a gap in the curve). More generally, the values of σ_i provide the link between local and global saliency. The idea is that the σ_is are determined by a local saliency measure; they increase, for example, for higher contrast, or when the ith edge element differs significantly from its neighbours in colour, or orientation, etc. In this manner the scheme can provide measure of the global saliency based on length and curvature, while at the same time taking into account the local saliency of the individual components.

In the expression above the overall saliency is obtained by a weighted sum of the local contributions σ_i along Γ. The weight ω_i of the ith element is:

$$e^{-c_i},$$

where c_i is the total curvature of the contour from p up to the ith element. (Mathematically, the total curvature is defined as $\int\kappa^2$, where κ is the curvature at a point.)

The saliency measure defined so far dependes on a particular curve Γ. The final saliency at p is given by:

$$S(p) = \max_\Gamma S_\Gamma(p).$$

The maximum is taken over all possible curves terminating at p. (This computes the contribution to p from one side of the curve, the contribution from the other side is determined in a similar way.) In practice, it is also convenient to use this definition limited to curves of length N:

$$S_N(p) = \max_{\Gamma_N} S_{\Gamma_N}(p)$$

where Γ_N stands for Γ restricted to length N. It is important to note that the optimum is sought over all possible curves, including fragmented ones. In the case of the fragmented circle in figure 1*c*, for example, the scheme will in effect consider all the possible curves running through any number of the individual line segments in the figure. This task may appear prohibitive: to determine the salient figure, one most consider all possible curves through all the elements in the image, and along each one integrate the curvature-based saliency measure described. As it turns out, the saliency measure described above can in fact be computed by a surprisingly simple, locally connected, network described below.

(*d*) *Computing global saliency by a simple local network*

To detect the globally salient structures in the image, the saliency measure $S(p)$ is computed at each image point p by a locally connected network of processing units. The processing elements can be thought of as local 'line detectors' that respond to the presence of lines or edges in the image. The entire image is covered by a grid of $n \times n$ points, where each point corresponds to a specific x,y location in the image. At each point p there are k 'orientation elements' coming into p from neighbouring points, and the same number of orientation elements leaving p to nearby points. Each orientation element p_i (the ith orientation element at point p) responds to an input image by signalling the presence of the corresponding line segment in the image. A lack of activity at p_i means that the corresponding line segment is not present in the image. The activity level of the element p_i, denoted by E_{pi}, will eventually correspond to the saliency of this line element. The initial activity is determined by the local saliency of the element, denoted by σ_i. This local saliency is determined by comparing the element in question with surrounding elements along a number of dimensions. For example, if the ith element has high contrast, or if it is very different from the surrounding elements in colour, orientation or direction of motion, then its local saliency will be high. To account for global rather than local saliency, the activity $E(p_i)$ is then modified by interactions with the neighbouring elements, so that eventually it also measures the length and the curvature of the contour passing through p_i.

The activity $E(p_i)$ is updated by the following simple local computation:

$$E_{p_i}^{(0)} = \sigma_i,$$

$$E_{p_i}^{(n+1)} = \sigma_i + \rho_i \max_{p_j} E_{p_i}^{(n)} f_{i,j},$$

where p_j is one of the k possible neighbours of p_i. In this formula, $f_{i,j}$ are 'coupling constants' between neighbouring line elements. The main property of these coupling constants is that they decrease with the angle between successive elements. A particular choice of the coupling constants above can ensure that the computed saliency will depend directly on the total curvature along the contour. The factors ρ_i are of secondary importance, they make the contributions of far-away locations smaller than that of nearby elements along the curve (for more detail, see Shashua & Ullman (1988)).

The saliency computation as defined above is simple and local: at each step in the iteration at a given element, the element simply adds the maximal contribution of its k neighbours to its original local saliency σ_i.

The interesting point about this computation is that by using this simple updating formula the quantity E_{pi} computes the desired measure $S(p)$, defined above, at every point p. It may appear surprising that such a simple local computation is sufficient for this task, since the saliency measure S at a point p in the image is in fact a rather elaborate measure. For each possible curve Γ passing through p, it must compute a measure S_Γ and select the best one (with highest S_Γ). The computation achieves all of this without explicitly tracing and then examining different curves. Although the number of possible curves of length N increases exponentially as the number N grows, the computation is only linear in the length N.

Figure 2 shows examples of the computation applied to three figures. The first is the car image in figure 1*a* (only a portion of the image is shown). The figure on the left is a 'saliency map' after 30 iterations of the computation. In this representation, the wider, lighter, contours are those with higher saliency $E(p)$.

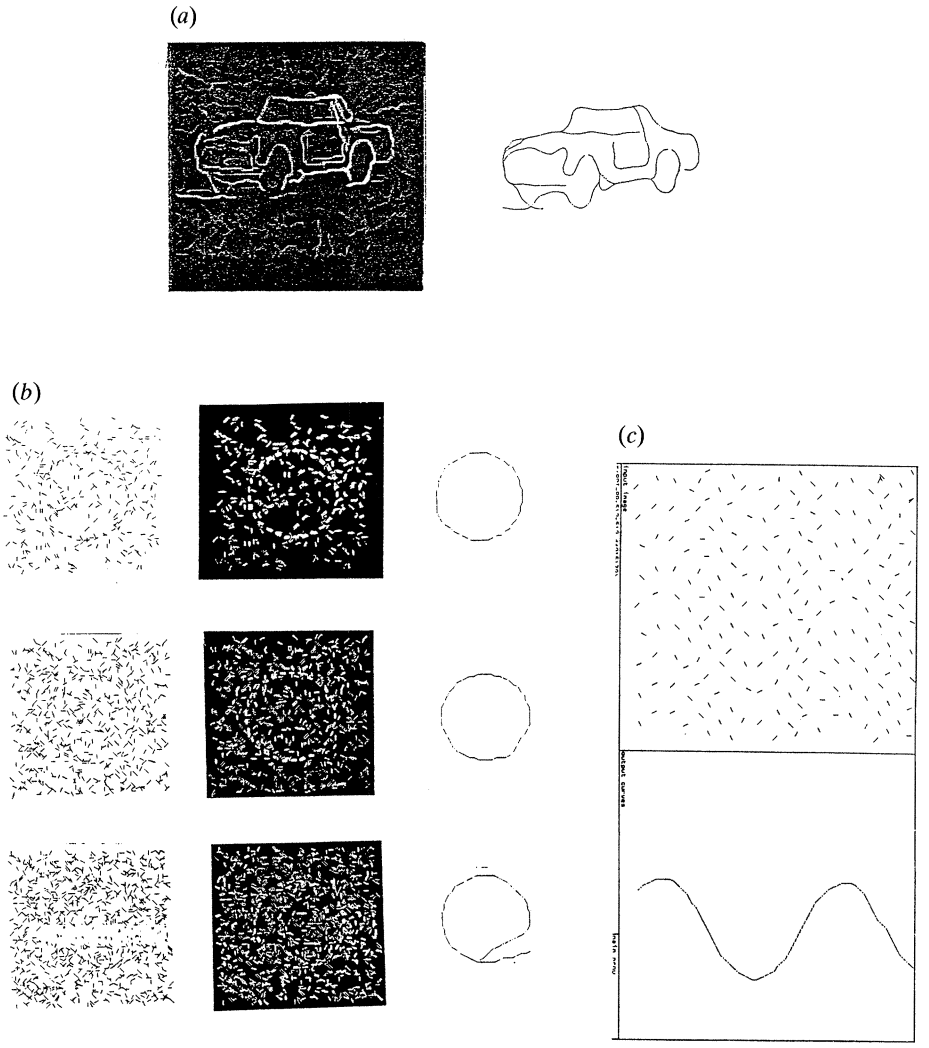

Figure 2. Results of the saliency computation. (*a*) A portion of the car image in 2*a*. Left: the saliency map after 30 iterations. Right: the five most active (salient) contours after 30 iterations. (*b*) A fragmented circle embedded in an increasing amount of noise. Left: original images. Centre: the saliency map after 10 iterations. Right: the most salient contour following 10 iterations. (*c*) A sinusoidal curve in background noise. Top: the input figure. Bottom: the most salient contour in the input image.

It can be seen that the activity in the background is reduced compared with the activity in the figure. The figure on the right shows the five most salient contours by the end of the 30 iterations.

Although the process successfully selects the main object in this image, it should be noted that, in general, the saliency computation described above is not intended to model the entire process of selecting out a candidate object from the image. A more plausible view is that such a selection process is obtained in two stages. The first stage, which is applied uniformly and in parallel across the entire image, selects and 'highlights' a small number of contours. A candidate object can then be selected out by processing these preferred contours further. This second stage can be more serial, and applied preferentially to the contours selected in the first stage, rather than to all the contours in the image.

Figure 2*b* shows a fragmented circle embedded in an increasing amounts of noise. The left column illustrates the input figures. It can be noted that in the first two images the circle is immediately discernible by our perceptual system despite the gaps and the high noise level. The second column shows the saliency map after 10 iterations, and the right column shows the most active (salient) contour by the end of the 10 iterations. The performance of the scheme appears to be comparable to human perception. It is also worth noting that the gaps in the original figure are filled-in in the course of the computation.

Figure 2*c* was devised by J. Beck from the University of Oregon. Beck has noted that the figure is a challenging one, but still perceivable by human observers. It is also interesting because it is not a simple closed compact figure. Schemes that are sensitive specifically to blob-like structures will not be able to extract such long curved structures. The scheme described above has some preference for closed figures, but, like human vision, can detect any smooth extended structure.

In summary of this part, regarding segmentation and selection:

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

1. There are good reasons to believe, based on psychophysical and computational considerations, that processes involved in segmenting the image and selecting structures for further processing are important in the early stages of visual information processing. Low-level visual areas, such as V2, may be involved in this type of processing.
2. Segmentation and selection involve a number of different processes, that depend on both local (e.g. contrast, colour, and motion) and global (e.g. length and overall curvature) properties.
3. The global saliency of a contour in the image increases when the contour is long and smooth.
4. The extraction of smooth long contours can be obtained by a simple network of locally interacting line elements.

3. RECOGNITION BY THE COMBINATION OF TWO-DIMENSIONAL IMAGES

Three-dimensional object recognition is a challenging task that at present cannot be performed efficiently by artificial systems. What makes object recognition so difficult, is mainly the variability of object views. There are four main sources for this variability, which are (i) viewing position (different views give rise to different two-dimensional projections); (ii) photometric effects (illumination and shadows); (iii) effects of occlusion and different settings of the object in the scene; and (iv) shape changes (of articulated or flexible objects). In this section I will consider one of these aspects, the problem of viewing position, and outline a relatively low-level method that approaches the problem by interpolating between two-dimensional images.

The problem of compensating for changes in the image induced by different orientations of objects in three-dimensional space is a difficult one, and a number of schemes have been proposed over the years as possible models for how the human visual system may overcome this problem. The large number of recognition schemes that have been proposed in the past can be divided into three broad classes, which are: (i) Invariant properties methods; (ii) object decomposition methods; and (iii) alignment methods.

For a review of these methods, see Ullman (1989). In the following sections I will present a scheme that belongs to the family of alignment methods. It differs from previous ones in that it uses collections of two-dimensional images instead of storing three-dimensional object models. This scheme appears to be simpler and more direct than alternative ones, and might correspond more closely to some of the processes used by biological visual systems in recognizing three-dimensional objects.

The combination of two-dimensional views

In most recognition theories it is assumed that the visual system somehow stores and manipulates three-dimensional object models. When confronted with a novel two-dimensional image of the object, the system deduces whether it is a possible view of one of the already stored three dimensional objects.

The method outlined in this section does not use explicit three-dimensional models. Instead, it uses directly small collections of two-dimensional images. The approach has several possible advantages. First, there is no need to explicitly recover and represent the three-dimensional structure of objects. Second, it handles all the rigid three-dimensional transformations, but it is not restricted to such transformations. Third, the processes involved are often simpler than in previous schemes. Fourth, it becomes easier to acquire new object models.

In this approach, a three-dimensional object is represented by the linear combination of two-dimensional images of the object. If $M = M_1, \ldots, M_k$ is the set of pictures representing a given object, and P is the two-dimensional image of an object to be recognized, then P is considered an instance of M if

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

for some constants α_i.

What I mean by a linear combination of views is the following. 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. Note that, in general, different coefficients are required for the x and y components. 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 proposition: 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, are more complex 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,

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

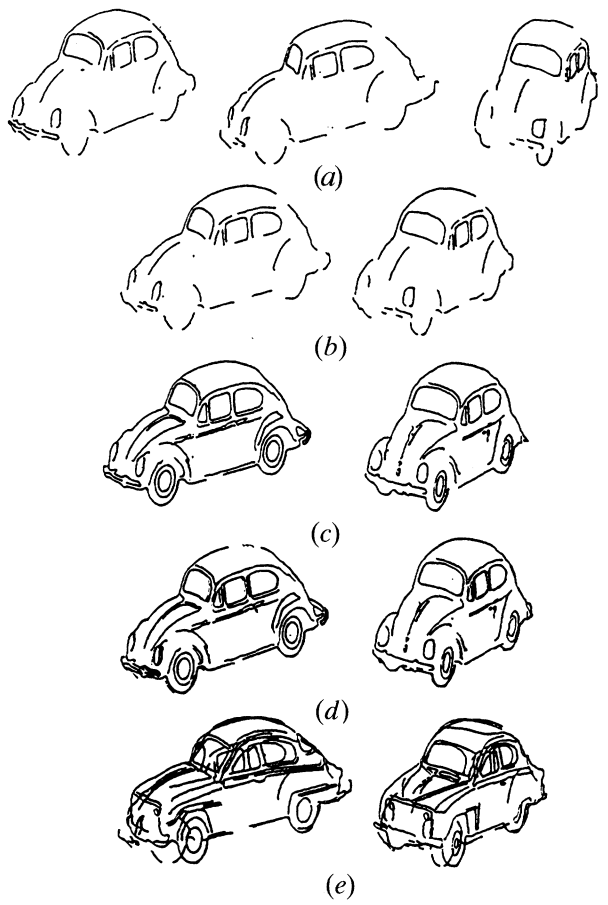

Figure 3. The linear combination of two-dimensional views. (*a*) Three views of a car (VW). Owing to technical reasons, only some of the edges are illustrated. (*b*) Two new views of the VW. These new views were obtained by linear combinations of the views in (*a*). (*c*) Two novel views of the VW. (*d*) Superposition of the images in (*c*) and the linear combinations in (*b*). The new views are matched well by linear combinations of the original views. (*e*). The best matching linear combination to a similar, but different, car (a Saab). The match is less precise, illustrating that the linear combination method can make fine distinctions between similar objects seen from novel viewing positions. Adapted from (Ullman & Basri 1991).

for example, the 'front' and the 'back' of the same object.

Figure 3 shows an example of using linear combinations of views to compensate for changes in viewing direction. Figure 3*a* 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 can be used to illustrate the point that reliable identification can be obtained on the basis of partial image data (as may happen due to noise and partial occlusion). Figure 3*b* shows two new views of the VW car. These new imges were not obtained from novel views of the car: they were generated instead by using linear combinations of the first three views. Figure 3*c* shows two new views of the VW, obtained from new viewing positions. Figure 3*d* superimposes these new views and the linear combinations obtained in figure 3*c*. It can be seen that the novel views are matched well by linear combinations of the three original views. For comparison, figure 3*e* 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 three-dimensional objects in novel viewing directions. (For a proof of the proposition and further details, see Ullman & Basri (1991).)

The 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 systems of three unknowns each, for the x and y components). There are alternative methods that will not be reviewed here.

It is worth noting in this regard that there is a link between the recognition method used, and the requirements placed upon the segmentation process discussed in the previous section. In the linear combination and related methods, for instance, the segmentation can be quite limited. It is sufficient to identify a small number of image features as corresponding to a single object, and there is no need to fully delineate the object's boundary. The internal model can then be aligned and matched with the image, and also be used to complete the segmentation process.

A related scheme for combining two-dimensional views is the Radial Basis Functions (RBF) method and its extensions developed by Poggio & Girosi (1989). The basic idea of the scheme, which will not be reviewed here, is to use again two-dimensional views, and then treat recognition as an approximation problem in the space of possible views. Given a number of two-dimensional views representing a single three-dimensional object, an approximation method is used to interpolate smoothly between the known views. The method used for interpolation is the so-called RBF method. This method interpolates a function between known data points by using a linear superposition of basis functions, centred on the known data points.

The schemes described above leave many problems unanswered. For example: how correspondence between image and model features may be established, the treatment of non-rigid transformations, general object classification as opposed to precise identification, and extending the scheme to deal with a large 'library' of internal models.

These problems 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 three-dimensional space: the percep-

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

tion of depth and three-dimensional 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 gradients. Similarly, visual object recognition is probably better viewed not as a single module, but as a collection of interacting processes. There is suggestive evidence, e.g. from the impressive ability of simpler animals to perform efficiently visual recognition, that these processes include a powerful component that is low-level and pictorial rather than abstract and symbolic in nature, and the combination or interpolation of two-dimensional views may be a candidate for such a component.

The following points summarize the main conclusions regarding the compensation for viewing direction.

1. In contrast with methods that use three-dimensional object models, the method outlined above uses directly small collections of two-dimensional images.
2. Applied to human vision, the scheme suggests that a particular view of a given three-dimensional object will be represented in the visual system by the combined activity of units, where each unit is tuned in a broad manner to a particular two-dimensional view. This seems to be in general agreement with physiological findings regarding face-selective cells in the primate visual cortex (Perret *et al.* 1985). These cells usually respond best to a particular two-dimensional view of a face, but the response is broadly tuned and usually covers similar faces as well as the same face from a range of viewing directions. A biological implementation of the scheme will also require a mechanism for generating combinations of existing views, and it will be of interest to explore in the future possible models for this operation.
3. The use of combinations of two-dimensional views in recognition appears more direct and straightforward than schemes that store and manipulate three-dimensional models, and may be more biologically plausible. The scheme is similar in certain respects to the direct use of an associative memory for two-dimensional patterns, but with a crucial difference. Given an input pattern, the system is not required to have an exact replica of the pattern already stored in memory. Instead, it is trying to establish whether the input pattern can be matched by combinations of small sets of stored patterns. As it turns out, such combinations are sufficient to compensate for the variations induced by changes in viewing direction.

I thank my students and collaborators in this work, Amnon Shashua and Ronen Basri. Support for this work came in part from NSF grant IRI-8900267.

This paper is based in part on material from the Proceedings of the Conference on Attention and Performance, Ann Arbor, Michigan, 1990, and the chapter 'Three dimensional object recognition' in the proceedings of the *Cold Spring Harb. Symp. quant. Biol.*, vol. LV (*The brain*), 1990 and of the Dahlem conference 1991. It also summarizes some of the material in Ullman & Basri (1991).

REFERENCES

Abu-Mostafa, Y.S. & Psaltis, D. 1987 Optical neural computing. *Scient. Am.* **256**, 66–73.

Biederman, I. 1985 Human image understanding: Recent research and a theory. *Comput. Vis. Graph. Image Process.* **32**, 29–73.

Bregman, A.S. 1984 Auditory scene analysis. *Proceedings of the 7th International Conference on Pattern Recognition*, Montreal, 168–175.

Gregory, R.L. 1970 *The intelligent eye*. London: Weidenfield & Nicholson.

Hernstein, R.J. 1984 Objects, categories, and discriminative stimuli. In *Animal Cognition* (ed. H. L. Roitblat, T. G. Bever & H. S. Terrace), pp. 233–261. Hillsdale, New Jersey: Lawrence Erlbaum Assoc.

Hopfield, J.J. 1982 Neural networks and physical systems with emergent collective computational abilities. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2554–2558.

Julesz, B. 1981 Textons, the elements of texture perception, and their interactions. *Nature, Lond.* **290**, 91–97.

Kohonen, T. 1978 *Associative memories: a system theoretic approach*. Berlin: Springer-Verlag.

Luria, A. 1980 *Higher cortical functions in man*. New York: Basic Books.

Marr, D. and Nishihara, H.K. 1978 Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond.* B **200**, 269–291.

Nakayama, K., Shimojo, S. & Silverman, G.H. 1989 Stereoscopic depth: its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception* **18**, 55–68.

Perret, D.I., Smith, P.A.J., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, A.D. & Jeeves, M.A. 1985 Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proc. R. Soc. Lond.* B **223**, 293–317.

Peterhans, E. & von der Heydt, R. 1991 Elements of form perception in monkey prestriate cortex. In *Representations of vision* (ed. A. Gorea), pp. 111–124. Cambridge University Press.

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

Ramachandran, V.S. 1986 Capture of stereopsis and apparent motion by illusory contours. *Percept. Psychophys.* **39** (5), 361–373.

Shashua, A. & Ullman, S. 1988 Structural saliency. *Proceedings of the International Conference on Computer Vision*. Tampa, Florida, 482–488.

Stark, L. & Bowyer, K. 1991 Achieving generalized object recognition through reasoning about association of function to structure. *IEEE Trans. PAMI*, **13** (10), 1097–1104.

Treisman, A. & Gelade, G. 1980 A feature integration theory of attention. *Cog. Psychol.* **12**, 97–136.

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

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

von der Heydt, R., Peterhans, E. & Baumgartner, G. 1984 Illusory contours and cortical neuron responses. *Science, Wash.* **224**, 1260–1262.

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

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Discussion

R. L. Gregory (*Department of Psychology, University of Bristol, U.K.*). Do supposed constraints – such as Marr's cylinders – make cylinder-shaped objects easier or quicker to see? If not, why not?

S. Ullman. The theory I have outlined based on the interpolation between two-dimensional views does not suggest that cylinder-shaped objects of the type described by Marr would be in general easier to recognize. In Marr's theory, recognition is based on object decomposition into particular type of parts (generalized cylinders). Objects constructed from such 'canonical' parts may be easier to recognize in this scheme.

The reason is that in this theory, unlike Marr's, identification is not based on object decomposition into 'canonical' parts, such as cylinders, and therefore there is no particular advantage to objects constructed out of these parts over others.

J. Atkinson (*Visual Development Unit, University of Cambridge, U.K.*). Although it is believed that, in general, classification is easier and faster (in terms of reaction times) than identification, neuropsychologists have identified patients who can name particular vegetables and fruit (identification) but cannot categorize them into fruit/vegetables and this has now been suggested for normal adults (i.e. fruit/vegetable classification slower than identification of particular fruit or vegetables). Has Professor Ullman applied his bottom-up algorithm to this particular classification (fruit/vegetables) and worked out its success for identification (controlling of course for context and frequency)? I wonder if another 'visual' example might be cars versus trucks as opposed to particular types of cars and trucks.

S. Ullman. The effects of brain lesions on naming, identification, and classification can be quite complex, and include both visual and linguistic components. Generally, classification appears more resistant to damage than the identification of unique entities (e.g. Damasio & Damasio 1989).

Classification and its relation to identification appear to me an important and challenging problem and I see this as a major area of future research in computational vision. However, the model I have outlined does not apply to this problem directly, but to the identification of specific rigid objects from different viewing position.

Reference

Damasio, H. & Damasio, A.R. 1989 *Lesion analysis in neuropsychology*. New York: Oxford University Press.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES