Computational approaches to shape constancy

Shimon Edelman
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel
Internet: edelman@wisdom.weizmann.ac.il

Daphna Weinshall
Dept. of Computer Science
The Hebrew University
Jerusalem 91904, Israel
Internet: daphna@cs.huji.ac.il

1 Introduction

The appearance of a three-dimensional object (that is, the pattern formed by its projection onto the retina of an eye or onto the imaging plane of a camera) depends on the point of view of the observer. The collective human awareness of this dependence is attested to by the widespread use of expressions that involve the metaphor of *point of view*, in languages as different as English, Russian, and Hebrew. Nevertheless, as far as recognition is concerned, the matters of viewpoint seem to be of secondary importance: the human visual system exhibits an impressive ability to recognize a familiar object viewed from an unfamiliar perspective. This phenomenon has been termed shape constancy, by analogy with other perceptual constancies.[1]

Computational understanding of shape constancy can be gained both by attempting to build artificial vision systems for object recognition, and by modeling human performance in this task. Maintaining a constant interpretation of the three-dimensional world in the face of changing viewing conditions has long been a major goal of computer vision. The first part of this chapter classifies and reviews several approaches to 3D object recognition developed within this field. In the second part of the chapter, we list the central characteristics of shape constancy in human vision, and compare the virtues and the shortcomings of the computational approaches, this time considered as models of human performance. We conclude with a general discussion of the phenomenon of shape constancy within the framework of the computational study of perception.

[1] Shepard (1987) noted that perceptual generalization — assigning different stimuli to the same class — implies not the inability to distinguish between them, but rather a decision, usually subconscious, that these stimuli are really the same in some deeper sense. In the case of recognition, such decision is based on the identification of clearly and manifestly different images as "merely" different views of the same 3D object.

Discussion of any kind of constancy exhibited by a perceptual system only makes sense insofar as there is *change* in the proximal stimulus (say, an image of an object) with which the percept of a constant distal entity (the 3D structure of the object) is contrasted. Accordingly, the manifestations of shape constancy can be divided into those in which the proximal change is present and explicit, as in the kinetic depth effect, and those in which the change is only implicit, as in the recognition of a familiar object from a novel viewpoint. Computationally, this corresponds to the distinction between multiple-view and single-view approaches to constancy. The present survey concentrates on the latter, although relationships between recognition schemes and structure from motion algorithms will be pointed out where appropriate.

Real-world objects change their appearance not only due to geometrical factors such as their position relative to the observer, but also due to the photometric factors: the number, the location, the nature, and the intensity of the illumination sources. Photometry is, however, irrelevant in the case of "pure" shapes, which may be defined as collections of points placed in 3D space. Moreover, an important example of shape constancy involves the perception of the 3D structure of line drawings, in which case, again, photometric factors are excluded. Thus, in what follows we chose to discuss mainly the geometrical aspects of shape constancy.

1.1 Classification of computational approaches to shape constancy

Computational approaches to geometrical shape constancy can be classified according to their stance on two major issues: degree of reliance on reconstruction and availability of object models.

1.1.1 Shape reconstruction

Assuming that constancy involves the comparison of the stimulus to some immutable ideal (leading to a realization that the former is an instance of the latter), it is possible to distinguish between two stages in the computation leading to constancy: representation and matching. Different approaches to constancy call for a different distribution of the processing load between these two stages. On the one hand, there are approaches that depend on a reconstruction (leading to a quasi-3D representation) of the stimulus. The comparison between the recovered representation and the internal model in such a scheme may be quite simple (a kind of template matching), although the reconstruction itself is computationally challenging. On the other hand, there are approaches that employ rather simple representations (e.g., 2D image-based measurements or features), for which they compensate by having to rely on complicated matching decision procedures. Examples of these two classes of approaches will be discussed in sections 2 and 3.

1.1.2 The availability of a model

In the above discussion we assumed that constancy is achieved via a comparison between the input and a model. In fact, the involvement of a well-defined model is not a prerequisite for constancy (consider the drawing in Figure 1, devised by Biederman (1987) as an example of the ability of human observers to assign a correct 3D interpretation to a novel view of an arbitrary object). In the following two sections, we will show how by relaxing the assumption of the availability of a model one can affect the conceptual underpinnings and the computational properties of constancy algorithms, both of the reconstruction and of the feature variety.

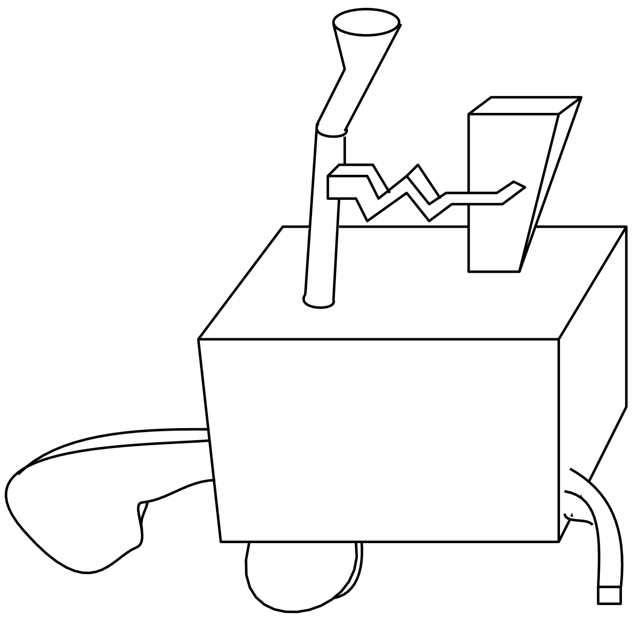

Figure 1: A do-it-yourself object (adapted with permission from Biederman, 1987). There is a strong consensus in the perception of the 3D structure of this object.

2 Constancy via reconstruction

Assuming that the goal is to reconstruct the shape of a 3D object from a 2D image or images, the issue still remains of whether and how to compute the pose of the object (its orientation with respect to the camera). Knowledge of the pose is usually not necessary for the recognition of the object. However, separating the pose from the shape may be difficult, and the ability to do that may be equivalent to the ability to perceive shape constancy. One may compare this ambiguity to the problem of color constancy, where the color of an object may be due to its surface pigmentation, or to the spectral composition of the illumination source. Similarly, the shape of an image of an object may be due to the object's geometrical structure, or to the particular viewpoint from which it is observed. The issue here is whether to reconstruct shape and pose together, or only shape. In the color analogy, the issue is whether to compute the color of the object and the color of the light source, or just the color of the object.

2.1 Model-based approaches

Information about the 3D shape of the target object can be put to use in the recognition[2] process in a number of ways. We consider here two possibilities: methods that involve alignment between an internal 3D model and the input, and methods that operate directly on 2D images, without explicit reconstruction of a 3D model.

The approaches discussed in section 2.1.1 use knowledge of pose, whereas the approaches discussed in section 2.1.2 use only partial reconstruction, namely, reconstruction of the shape parameters only (as opposed to complete reconstruction, which seeks to compute the parameters of the shape and the pose).

2.1.1 Normalization via pose

Our first example of the normalization approach is the alignment method (Fischler and Bolles, 1981; Rives et al., 1981; Ullman, 1989). A generic alignment algorithm proceeds as follows. First, for each model available to the system, a viewing transformation is computed, subject to the requirement that the image and the appropriate view of the model match (i.e., become aligned) if the input view indeed belongs to the model. Second, the hypothesized transformation is carried out, and the degree of the resulting match is used to confirm or to refute the attribution of the input view to the model. The model that yields the best fit following alignment is declared the outcome of the recognition process.

As formulated above, the alignment approach appears to require an exhaustive search of the space of all possible transformations of each model, so that the optimal combination of model and transformation can be found. In fact, Ullman's alignment method actually obviates the need to search through all transformations, by specifying an algorithm for computing the aligning transformation given the correspondence between a small number of *anchor* image points and points in the model. Thus, the only search that remains to be done is over the possible models. Below we discuss a related approach, recognition by geometric hashing, which seeks to eliminate the need for this search as well, at the cost of using a large database.

A slightly different formulation of the basic idea behind the normalization approach — recovery and use of explicit information about the viewpoint — can be found in Lowe's (1987) notion of viewpoint consistency constraints. In Lowe's method, locations of a few key points in the image lead to a hypothesis regarding the viewing transformation, which is then used to predict iteratively the locations of additional object points in the image. An advantage of this method is the possibility to abort the processing of a given model in the presence of growing evidence against its suitability, provided by the incremental evaluation process.

The third approach of this variety that we mention is Lamdan and Wolfson's (1988) geometric hashing. Here, several of the object's features are chosen to serve as a linear basis in terms of which the locations of all other features are encoded. For any choice of the basis features it is possible to express all the other features of the object relative to the basis, i.e., in transformation-invariant coordinates. Accordingly, in the preprocessing step

[2] In the model-based case, the term "recognition" may be freely substituted for "constancy."

of the geometric hashing method, the coordinates of all the feature points of the object are precomputed for every possible choice of the basis feature set, and are used to construct a hash table that associates the pair (model, basis) with the set of coordinates.[3] Recognition then proceeds in several steps. First, features of interest are extracted. Second, a basis set is chosen, and the coordinates of all other features in this basis are computed. A counter associated with the hash table entry corresponding to the set of computed coordinates is then incremented, in a process that resembles "voting" for the appropriate (model, basis) pair. Finally, the pair with the most votes is chosen, the transformation between the model thus retrieved and the image is computed (as in the alignment scheme), and the transformed model is verified against the image.

An important advantage of geometric hashing over straightforward alignment is in its lower computation complexity (Lamdan and Wolfson, 1988). The reduction in complexity is achieved in a trade-off against potentially heavy memory usage dictated by the storage requirements of the hashing scheme. It should be noted, however, that a heavy reliance on memory, in conjunction with a relatively simple computation process (such as voting) may be considered an advantage rather than a handicap in a scheme that aims at modeling recognition in biological vision.

2.1.2 Normalization via formal invariance

The normalization methods discussed above either align the image with the model, a process which is equivalent to pose computation (where the transformation of the object with respect to the camera is computed), or use the pose of the object in a voting scheme. However, neither alignment nor geometric hashing (or generally any indexing technique) need to use the pose for recognition. To avoid pose dependency, it is necessary to use a representation of three dimensional shape which does not depend on its orientation with respect to the camera, or equivalently, which is invariant to the transformation of the camera (or the eye).

Koenderink and van Doorn (1991) described a hierarchy of representations determined by the particular selection of the group of three dimensional transformations of the camera, with respect to which invariance is sought:

Complete: The complete representation of the scene includes the $3D$ coordinates of each point, possibly as a depth map, and the pose of the camera relative to the object. This representation is typically sought in reconstruction algorithms in computer vision.

Euclidean invariant: A description of the three dimensional shape of objects which is invariant to the action of the group of similarity transformations in $3D$, which includes rotations, translations, and isotropic scaling.

Affine invariant: A description of the three dimensional shape of objects which is invariant to the action of the group of linear transformations in $3D$ and of translations. The group of affine transformations includes the group of similarity transformations.

[3] A hash table is a data structure that allows efficient access to information identified by the key associated with it. In the present case, the key is the set of coordinates of object features, and the information to be retrieved is the identity of the object and the basis features.

A *complete* representation describes uniquely the object and its orientation relative to the camera. It includes the largest possible amount of information concerning the scene, and therefore, not surprisingly, it is the most difficult one to compute. An *Euclidean* representation is unique for every general three dimensional object, but different scenes (where the same object is observed from different orientations) may have the same representation. This is a desirable ambiguity. An *affine* representation is not unique: objects which are related by a linear transformation have the same affine representation. For example, in two dimensions all squares and parallelograms have the same affine representation since they can be obtained from each other by a linear transformation (namely, by stretch and shear). This ambiguity is less desirable: the cost of using an affine representation is an increased number of false positive matches. On the other hand, affine three dimensional shape representations are easier to compute, because they can be computed from a sequence of 2D images by a linear algorithm (Shashua, 1991; Weinshall, 1993).

The normalization methods described in the previous section can be implemented without computing pose, using such invariant representations. Alignment of an image can be done with an invariant representation of the object, in which case there is no need to search the transformation space, although there is still a need to search the model space (Weinshall, 1993). Geometric hashing already relies on an affine invariant representation, and only uses the pose in the voting scheme. This involvement of the pose can be eliminated, and an object invariant quantity, such as the location of the center of gravity of the object, can be used instead during the voting (Mohan et al., 1993).

2.1.3 Direct methods equivalent to reconstruction

In all the examples of normalization, the achievement of constancy hinges on the availability of an explicit 3D model of the object in question, which is used to relate different views together via viewpoint consistency constraints of some kind or another. A recent work by Ullman and Basri (1991) demonstrated that one need not represent the model of an object explicitly to determine whether an input image belongs to the object. Let each view of a 3D object be represented by a list of x, y coordinates of the orthographic projections of certain object features onto the image plane. Ullman and Basri showed that in this case the representation of any view of the object can be written down as a linear combination (weighted sum) of the representations of a small number of fixed views of the same object.

A practical recognition scheme based on this result can be obtained, for example, as follows (Ullman and Basri, 1991). First, the fixed views of an object are used to precompute a matrix specific for that object. That matrix, when multiplied by the vector representing a novel view, yields a vector whose norm (magnitude) can be thresholded to decide whether the novel view belongs to the given object. Note that in this case the shape of the object (i.e., its model) is represented implicitly in the elements of the matrix. Indeed, according to the structure from motion theorems (Ullman, 1979), the shape can be recovered from the set of given views used to construct the matrix. Therefore, the linear combination scheme belongs properly to the class of model-based reconstructive approaches.

2.2 Model-free approaches

In all the model-based approaches discussed in the preceding section, constancy is essentially a matter of decision as to which of a number of competing objects provides the best account for the structure of the input view. This notion of object constancy, while computationally powerful, seems to run counter to the intuition regarding what constancy is supposed to be about: perceiving the "right" shape of a not necessarily familiar object from an arbitrary viewpoint. It should be realized that it does not make sense to speak about *the* right shape if the object is indeed unfamiliar to the observer: in that case, the loss of information in the process of imaging ensures the existence of an infinity of solutions to the problem of recovery of the object shape from the image.[4] Because of the infinite number of solutions, the problem of model-free shape recovery from a single image is mathematically ill-posed (Poggio et al., 1985).

2.2.1 Assumptions regarding shape

Despite the ill-posed nature of model-free shape recovery, people do tend to agree as to the shape of the object depicted in Figure 1. Computational accounts of this feat of shape constancy are usually formulated in terms of the invocation of prior constraints on the solution. A classical example of such a prior is the recovery of the "correct" interpretation of an ellipse as the image of a tilted circle. Brady and Yuille (1984) showed that the circle is obtained in this case by requiring that the solution maximize the ratio between area and squared perimeter, a common simple shape descriptor in computer vision.

In the more complicated example involving polyhedral scenes, one may assume that the 2D angles between edges belonging to the same face in the shape to be recovered are all straight (as in many man-made objects) or, more generally, that the angles are such that their variance is minimal (Marill, 1989). Algorithms based on constraints of this kind, when applied to line drawings of polyhedral scenes, arrive at solutions that frequently are intuitively satisfying.

An interesting variation on the theme of prior constraints on the sought shape is the assumption of bilateral symmetry. This assumption, which seems to be true of many natural objects as well as artificial ones, allows the recovery of an object's shape from a single view (Poggio and Vetter, 1992). This scheme can be seen as a development of the linear combination approach, in which one view is given, and another view, necessary for constructing a model of the object, is obtained by an essentially geometrical procedure based on a plane symmetry constraint (Mitsumoto et al., 1992).

2.2.2 Assumptions regarding viewpoint

In the theory of ill-posed problems, the imposition of constraints intended to make the problem well-posed (e.g., have a unique solution) is called regularization. In the recovery of object shape, regularization amounts to constraining the possible shapes, so that the indeterminacy associated with the unknown viewpoint is reduced. An alternative approach

[4] Loss of geometric information; confounding between surface reflectance, surface orientation, illumination direction, and illumination intensity.

is to make assumptions regarding the viewpoint itself. Probably the least exacting assumption of this kind is that of generic or nonaccidental viewpoint (Binford, 1971; Witkin and Tenenbaum, 1986; Lowe, 1986).

To illustrate this concept, consider an image consisting of a single point. This could be the projection onto the image plane of a point-like object, or of a straight line seen end-on. Note that the second interpretation involves a (rather unlikely) condition on the viewpoint, while the first one does not. Thus, by a simple probabilistic argument, the first interpretation — that of a point-like object — is to be preferred (unless prior information is available which can bias the relative likelihood of the two interpretations; see Richards and Jepson, 1992). In other words, one is advised to favor the interpretations that do not presuppose an accident of viewpoint (such as seeing a line end-on, something that happens for a set of viewpoints whose measure relative to the set of all possible viewpoints is zero).

The notion of generic viewpoint can be defined more formally and more precisely (Weinshall et al., 1993). As the camera is rotated around an object, its image changes. From some viewpoints it changes slowly, while from other viewpoints the change is much more rapid. Depending on how the image is represented, it is possible to compute these changes. We may then refer to the viewpoints where the image changes the fastest as non-generic views. Viewpoints where the image changes the least may be called generic viewpoints. More generally we have here a continuous quantitative measure of viewpoint "genericity," or the likelihood of seeing the object as it appears from a given viewpoint.

A quantitative probabilistic framework for exploiting the notion of generic views has been proposed by Freeman (1993). This framework applies Bayesian tools to estimate the relative likelihood of various possible interpretations of the input scene. A number of previously developed Bayesian approaches to shape recovery and scene understanding expressed the likelihood of a solution — a description of the geometry of a scene — in terms of fidelity to the observed image and prior probability of that scene. Freeman's formulation adds a third term to the scene probability equation: the generic view term, which quantifies how accidental the observed view of the scene is. This is done by introducing generic variables, such as viewpoint, object pose, and illumination direction, and considering the image as a function of these variables. The main idea that leads to the scene probability equation is that large values of the derivatives of the image with respect to the generic variables correspond to unlikely scenes. Although some initial experiments with applying this insight to the recovery of shape from shading (Horn and Brooks, 1989) seem promising (Freeman, 1993), it is not clear whether this approach can be made general enough to help solve the model-free version of the shape constancy problem in its full generality, including the recovery of shape from line drawings.

2.2.3 Nonaccidental properties

The idea of nonaccidental shape properties (that is, image features from which 3D properties of objects can be safely inferred, barring the unlikely accident of a viewpoint), which originated in the work of Binford and was developed by Lowe and others (Binford, 1971; Lowe, 1986), serves as a foundation to Biederman's theory of object recognition by components (Biederman, 1987). Some of the nonaccidental properties of contour segments considered by this theory are straightness, sense of curvature, parallelism, and co-termination (for example, two segments with endpoints near each other are to be interpreted as touching in 3D).

From a list of such properties, Biederman composes a small set of generic shape primitives (which he calls geons), which (1) can be readily detected in single images of objects, including line drawings, and (2) can be combined to form a faithful description of 3D objects at the basic level of categorization. According to this theory, the drawing in Figure 1 is perceived as a composition of simple volumetric parts, which, along with their specified spatial relationships, give rise to the perception of 3D shape which does not depend on the viewpoint.

3 Constancy without reconstruction

Recent developments in computer vision indicate that it may be possible to come up with a viable alternative to the reconstructionist approach to constancy. In this section, we survey the major computational approaches to shape constancy that do not require 3D reconstruction of the shape. One way to avoid the need to reconstruct the viewed shape in 3D before recognizing it is by characterizing the shape in terms of a set of measurements or features that are provably invariant with respect to viewpoint changes (or any other factor that may affect the appearance of the shape in the observed image). As in the discussion of the reconstructionist theories, we classify the various approaches into model-based vs. model-free ones.

3.1 Model-based invariants

In section 2.1.2 we described representations which are invariant to changes of the viewing position. Any object can be represented in such a way. As the representation is invariant to the viewing position, it relates to all the images of the object in an inherently similar way. We can therefore take an object and one of its invariant representations, and define an operator that acts on the image and performs verification: it returns one value (say 0) if the image is an instance of the invariant model, and another value otherwise. A verification operator of this sort is called a model-based invariant operator. Model-based invariant functions, for objects composed of n fiducial points, were described in (Weinshall, 1993). The question remains as to which features should be used in the description of the object, a question which is discussed below.

The main drawback of this approach is that it is not very different from the reconstruction approach. Although no shape is reconstructed, not much saving is obtained either, since the image still needs to be compared sequentially to the different models in the database via the model-based invariant operators.

3.2 The model-free case

Ideally, a recognition operator is desirable that would operate on an image and return a certain fixed value for all the views of the same object, and different values for views of other objects. Such a model-free operator would be much more powerful, and more useful, than the model-based invariant function. For example, such an operator could be applied to an image to produce an index into a database, thereby identifying the object (or set of objects) present in the image.

Such operators can be defined only if there exist measurable quantities in the image that are invariant to changes in the orientation of the camera with respect to the object. Unfortunately, it has been shown that such invariants do not exist for general objects (Burns et al., 1990; Moses and Ullman, 1992). This fact makes intuitive sense, if one realizes that achieving shape constancy with unknown generic objects must necessarily involve classification (as opposed to the case when the objects have known models, which, as we pointed out above, may be properly termed recognition). In classification, the image is assigned to a family of objects. If model-free invariants exist for this family,[5] then identification may be attempted. In many situations, however, identification is unnecessary or meaningless (cf. Figure 1), and a general and qualitative description of the object's shape will suffice.

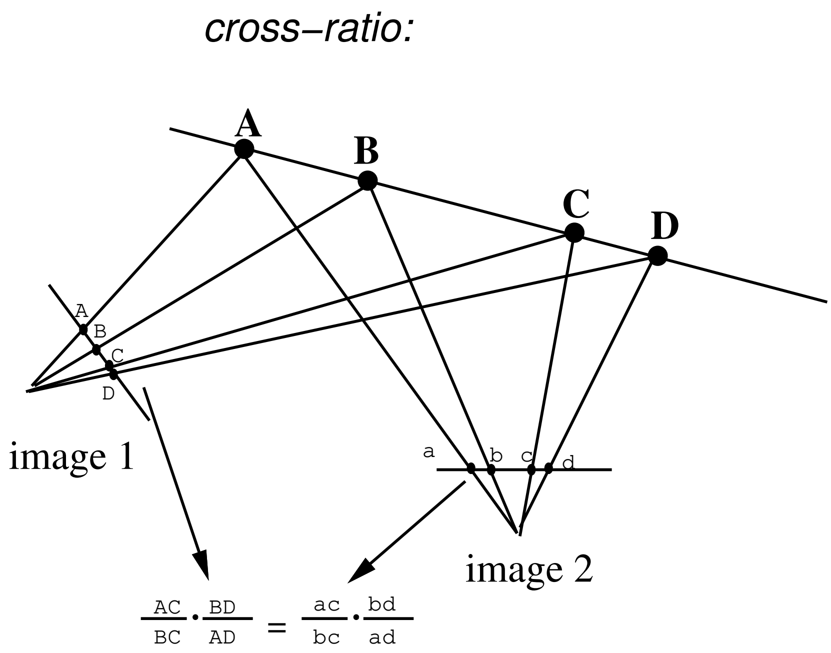

Figure 2: The cross ration remains constant for any image of the four coplanar space points A, B, C, D.

What would it mean for model-free invariant operators to exist? Despite the loss of information in projecting a 3D object onto a 2D image, there would have to be still enough

[5]As they do for special kinds of objects, such as planar (flat) (Forsyth et al., 1990), or symmetrical ones (Moses and Ullman, 1992).

information left to identify **any** possible object. The following argument shows that the requirement of identifiability of any possible object is too strong. A model-free invariant function must return the same value for any two objects composed of the same number of features, say, all objects with 10 features. Let us choose two objects comprising 10 points each. Either these two objects have one image (projection) in common, or there exists a sequence of less than 10 objects linking one of them to the other, such that each two consecutive objects on the list have an image in common. A model-free invariant function would have to return the same value for all the objects on the list, and therefore it would not be able to discriminate between the original two objects.

If we know that the image cannot be an image of just anything, by virtue of certain conditions on the class of permitted objects, then restricted model-free invariant functions may exist. Projective invariant functions of planar objects have been discussed in the classical literature on projective geometry. One well-known example of such an invariant is the cross-ratio. For simplicity, we shall describe the cross-ratio of four collinear points (but this can be generalized to five coplanar points). Let A, B, C, D denote four such points (see Figure 2), and let $\bar{PQ}$ denote the length of the segment between points P and Q. Then the following cross ratio of the four points remains the same for all their images:

$$\frac{\bar{AB}}{\bar{BC}} \cdot \frac{\bar{BD}}{\bar{AD}}$$

4 Computational models of shape constancy in human vision

4.1 Background

To what extent can the computational schemes for shape constancy discussed in the preceding sections serve as models of visual object perception and recognition in humans? The first thing to note about shape constancy in human vision is that it is far from perfect. To provide a background for discussion, we list here the major manifestations of the imperfections of shape constancy (see also the chapter by Jolicoeur and Humphrey in this collection).[6]

- *Canonical views.* Perceiving the shape of an object irrespective of the viewing conditions such as its orientation in space and its distance from the observer frequently incurs a certain information-processing cost, over and above what it takes to recognize the same object in its most familiar appearance (see Figure 3). This additional processing cost is reflected in longer response times and in higher error rates evoked by randomly chosen views of the object, as compared to certain so-called canonical views (Palmer, Rosch and Chase, 1981).

[6] Failures of object constancy are particularly revealing as test cases for various theories of recognition, because a model based on any such theory, if it can be made to fail at all, is likely to do so in a peculiar way that can be compared to human performance in an appropriately designed experiment. Moreover, the human performance can also be manipulated so as to reveal as much as possible about the processes involved in recognizing objects under changes in viewpoint.

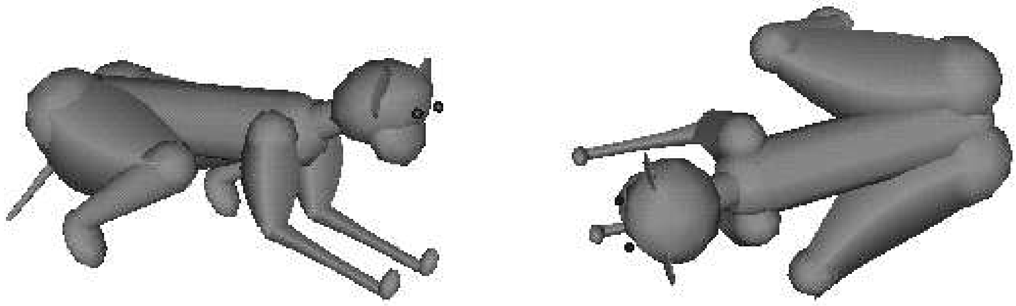

Figure 3: Canonical views: certain views of 3D objects are consistently easier to recognize or process in a variety of visual tasks. For example, a "three quarters" view of an animal-like shape (left) is bound to yield lower response time and error rate and to receive higher subjective "goodness" score than other views of the same object, such as the view on the right. Such differences may exist even among views that are seen equally often.

- *Mental rotation* Transition from a canonical to a non-canonical view of an object does not merely increase the expected recognition time: response latency depends on the viewpoint in an orderly fashion, growing monotonically with misorientation relative to the nearest canonical view (Tarr and Pinker, 1989; Edelman and Bülthoff, 1992). This dependency of response time on misorientation resembles the celebrated finding by Shepard and Metzler (1971) of a class of phenomena that became known as mental rotation.

- *Limited generalization* When asked to give a relatively broad classification of an object seen from an odd viewpoint (that is, when the task requires basic-level categorization), people virtually never err, except when the object appears severely foreshortened; see (Biederman, 1987). In comparison, when the task can only be solved through relatively precise shape matching, the error rate reaches chance level already at misorientation of about 40° relative to a familiar attitude (Rock and DiVita, 1987; Edelman and Bülthoff, 1992).

Figure 4: Generalization to novel views: people have a great difficulty in recognizing a novel view of an object, if the contrast set consists of similar objects. In this example, the leftmost and the central images are two views of the same wire-like object, but the rightmost image is not.

4.2 Computational models of recognition in human vision

We now turn to discuss three recognition schemes which were explicitly advanced by their proponents as models of human performance. The particular choice of models, while necessarily incomplete, has been dictated by the extent to which they explicitly address the relevant computational issues, and by their success to account for human performance in different shape constancy tasks.

4.2.1 Normalization to multiple stored views

The dependence of naming time for novel objects on their orientation with respect to the observer has been interpreted in the past as evidence for the involvement of mental rotation in recognition (Tarr and Pinker, 1989). According to one version of the alignment theory (see section 2.1.1), the human visual system tries to attain a match between the input and a model by a normalizing transformation that would align the input with the nearest stored view of the model. Presumably, the normalization process consumes time, and this is reflected in the dependence of the response time on the misorientation between the input and the nearest stored view. At some stage, the system may decide that it is worthwhile to adjoin the current view to the set of stored views, in which case subsequent responses to this view would become faster.

Can alignment constitute a viable model of shape constancy in human vision? The answer to this question would have been positive, were the constancy perfect. Indeed, the technique of alignment is so powerful that it is difficult to see how it could be made to account for recognition response time data without giving up some of its central characteristics. Once a rotation angle has been computed on the basis of the hypothesize match, the normalization process can, in principle, be carried out in a single step (the complexity of this step, which amounts to matrix multiplication, may depend on the complexity of the object, but certainly need not depend on the rotation angle). Furthermore, storing more and more views confers no computational advantage on the system (on the contrary, it taxes the system's memory requirements). Thus, alignment predicts neither the dependence of response time on misorientation, nor the need to store multiple views of the object beyond the small number that is strictly necessary for representing it 3D structure (as mentioned in section 2.1.1). Finally, the alignment theory cannot be easily brought to bear on the error rate data, because it does not constrain possible sources of errors, either in the normalization stage, or during matching.

It may be possible to reconcile the alignment theory with the experimental data on recognition, if one is prepared to make assumptions regarding the mechanism of the normalization. Suppose that the hardware limitations of the system are such that only rotations by a fixed discrete amount are supported, and, furthermore, that these rotations can only be carried out with respect to a few special axes.[7] Larger rotations will then take longer, because of required iteration of the elementary rotation step. Moreover, rotation time will depend on the orientation of the axis (as it indeed does in the human data; see S. Shepard and D. Metzler, 1988). We note that such stepwise alignment in effect gives up the main computational advantage of the original theory, namely, the freedom from the need to search exhaustively the space of all possible aligning transformations. Still, a recent development of the idea of alignment (Ullman, 1991) suggests that an appropriate distributed architecture and a combined bottom-up and top-down control structure may make alignment a useful model of the kind of shape constancy one finds in human vision.

Another way to reconcile the predictions of alignment with the human data is to give up the assumption that the recognition mechanism has access to precise 3D models of objects. A likely alternative is for the alignment process to use 3D models which are themselves computed from observed images of objects, and which are updated when additional images become available. One should, therefore, expect imperfect performance, which improves with the availability of additional images of the target. Moreover, one should expect less errors when the model is compared with images that were used to build the model (as opposed to novel images), as it indeed happens in the human data. This point is illustrated is (Weinshall and Tomasi, 1993).

4.2.2 Recognition By Components

Hummel and Biederman (1992) developed a detailed model of the phenomenon of constancy as it is illustrated in Figure 5. Their model is based on Biederman's theory of Recognition By Components (RBC). This detailed implementation of RBC attempts to address most of the computational issues involved in the perception of constant shape in simple cartoon-like line drawings of 3D objects. The processing stages that are explicitly modeled are:

1. Grouping of contour fragments according to the long-established principles of perceptual organization (Lowe, 1986; Witkin and Tenenbaum, 1986);

2. Grouping of entire segments into geons (generic volumetric primitives postulated by RBC), based on the nonaccidental properties (see section 2.2.3);

[7] A suggestion for a simple connectionist implementation of a rotation module satisfying similar requirements can be found in (Churchland, 1987).

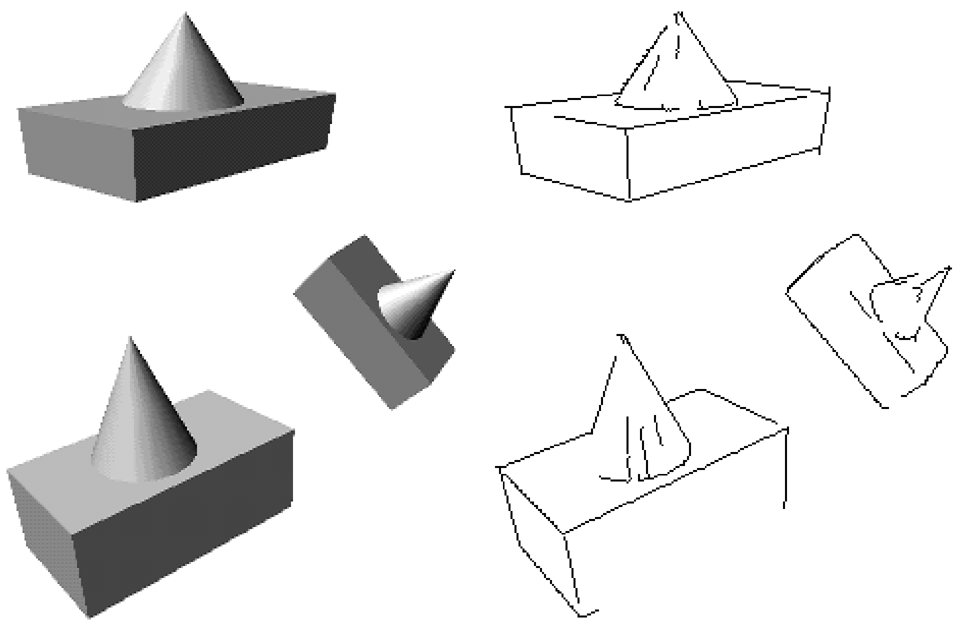

Figure 5: *Left:* The Recognition By Components model is designed to detect the constancy of objects such as this one across different views, despite their being unfamiliar (adapted with permission from Hummel and Biederman, 1992). The implemented model deals not with shaded images but with line drawings. *Right:* reliable and consistent recovery of a representation resembling a line drawing from a gray-level image is by itself a most challenging computational problem (as noted in Hummel and Biederman, 1992, p.514). The edge map shown here was obtained with an edge detection program that is part of a standard image-processing software package, using its default parameter settings.

3. Determination of the spatial relationships among geons and the formation of an appropriate shape representation.

The various intermediate representations employed by the model are distributed. For example, the presence of a contour in the input image is coded by a distribution of activities of units whose oriented receptive fields approximately match the location and the orientation of the contour. At a higher stage, units responsible for certain contour configurations such as the two ends of a partially occluded line cooperate to produce the representation of the entire line. Finally, at the top stage, simultaneous activation of appropriate geon and spatial relation units signals the "perceived" shape.

When tested on a set of ten objects, the model exhibited shape constancy by responding in an invariant manner to shapes undergoing translation, scaling, mirror reflection, and rotation (Hummel and Biederman, 1992). More importantly, the model represents one of the first attempts to address issues of object constancy computationally, on a level that is more concrete than the qualitative arguments normally employed by cognitive psychologists in the discussion of recognition.

4.2.3 Multiple-view interpolation

The distributed RBC model described in the preceding section relies, on the one hand, on graded-response mechanisms (e.g., receptive fields), which are perfectly credible both biologically and computationally. On the other hand, it postulates detectors sensitive to various configurations of line and curve segments, which are difficult to extract reliably from natural images (Hildreth, 1987; see Figure 5). The multiple-view interpolation model, described next, is more consistent, in that it uses receptive field-like mechanisms that operate directly on gray-level images. Unlike the RBC model, it concentrates on discrimination among similar objects.

Consider a receptive field-like structure connected to an array of image elements, which is capable of rapidly adjusting the weight assigned to each input element according to an activity-dependent rule (Edelman and Weinshall, 1991). It appears that such a structure can be made into a "detector" selective for a particular view of a 3D object. Because of the graded nature of the weight distribution, the receptive field for a given view will also exhibit a certain degree of generalization to similar views. Furthermore, providing a mechanism that incorporates several latent detectors of this kind with a sufficient number of training views would allow one to control the degree of generalization (Edelman, 1993). This control allows one to replicate the pattern of limited generalization to novel views exhibited by human subjects under conditions of high target-foil similarity (Bülthoff and Edelman, 1992). Moreover, if the representations of the stored views are linked to each other by an associative learning rule (cf. Sakai and Miyashita, 1992), the resulting structure will also replicate central features of the time course of recognition in humans, namely, canonical views and mental rotation.[8]

The computational rationale behind the multiple-view approach is derived from treating the problem of generalizing recognition across views as an instance of a standard problem in approximation theory – learning a function from examples (Poggio and Girosi, 1990). Clearly, the problem of shape constancy would be solved if it were possible to map any view of an object into a standard (constant) view, that is, to determine the function from any view to the standard one. Given a number of known views of an object as examples, it is only possible to learn such a function if its choice is properly constrained. Specifically, if it is assumed that the target function is smooth (changing the input view slightly will not change its classification), then regularization theory dictates a solution which is an interpolation of the example views. Moreover, an interpretation of this solution in terms of receptive fields yields exactly the mechanism outlined in the previous paragraph (Poggio and Edelman, 1990).

The multiple-view interpolation model combines psychophysical support with a solid computational grounding in recent developments in function approximation and in learning theory. However, this model will not be complete until it replicates the entire spectrum of human performance characteristics in recognition, from the strong viewpoint effects in the case of discrimination among highly similar objects to the complete constancy in the case of simple easily distinguished shapes. One possibility of extending the multiple-view

[8] Which is then seen to be merely a reflection of a step-by-step spread of activation throughout the web of view detectors exciting each other; see (Edelman and Weinshall, 1991; Edelman, 1991b).

model in this manner is outlined below.

4.2.4 Features of recognition

Does the visual system rely on distinct sets of representations and processes for subordinate and basic levels of recognition? A recent proposal (Edelman, 1991a) outlined a possible unified approach to the understanding both of shape constancy and of its failures. On the one hand, like the multiple-view model, the proposed approach requires neither 3D reconstruction of the stimulus, nor the maintenance of a library of 3D models of objects. On the other hand, the proposed model is related also to the RBC theory, in that its representation of object views is allowed to include nonaccidental features similar to those used to detect the presence of geons in the image. The choice of features and their complexity presumably vary among objects: recognition may start with the extraction of a large variety of image-based features, followed by an application of a recognition procedure appropriate for the given object set and for the required category level. Such procedures may be synthesized at need and optimized with practice (for example, compound features may be formed out of simpler ones, and subsequently relied upon).

This approach predicts a tradeoff between viewpoint invariance of a feature and its diagnosticity, or the degree of discrimination among object instances that it affords. Consider, for instance, a domain of objects composed of geon-like parts. To recognize such objects, a visual system can use local features such as image-plane positions of object corners or edges, as well as extended features such as patterns of shading over object surfaces. The two types of features will, in general, lead to different performance. When the pose of the object relative to the viewer changes, the projected locations of the corners will shift. Unless this shift is compensated for (e.g., by pose recovery and model alignment), recognition of unfamiliar views of the object will be poor. In comparison, the shape of a shaded patch can in principle be extracted regardless of the pose of the object to which it belongs (as long as the patch is visible). At the same time, if the pose is fixed, projected corners, edges, or other localized features offer better discrimination among similar shapes than shading (cf. Bülthoff and Mallot, 1988). The above considerations suggest that different patterns of performance emerge in response to the different levels of detail that must be addressed in subordinate and basic-level recognition. If this is true, then one would expect the extent of viewpoint invariance in subjects' performance to be affected by a manipulation of the relevant level of detail, determined by inter-object similarity. This effect was indeed found in (Edelman, 1992).

5 Conclusions

5.1 Constancy, representation and reconstruction

The central role of the choice of representation in computational approaches to problems such as recognition and constancy is by now generally accepted, largely due to the intellectual legacy of Marr (1982). Unfortunately, a lack of distinction between the concept of representation and the admittedly related, but rather more restricted, concept of reconstruction is nearly as general.

This lack of distinction characterizes well the approach taken by the mainstream of computer vision research in the past fifteen years. The importance of choosing the right representation for a given computational problem is widely acknowledged, and has been most forcefully argued for by Marr. However, many vision researchers, including Marr, essentially ignored the possibility that the best representation of the visual world may not be the same as the visual world reconstructed in full three-dimensional detail.

In the philosophy of mind, the equation of representation with reconstruction is in the spirit of a tradition dating back to Aristotle, Hume, and Berkeley (see (Cummins, 1989)). Computationally, a representation that is, in a sense, a geometrical replica of the thing being represented must be considered adequate for any visual task. From a pragmatic perspective, however, representation by reconstruction appears to be a poor choice, for several reasons. First, experience in computer vision shows that such a representation is notoriously difficult to recover from real unconstrained data. Second, even though all the information necessary for solving a given visual task is present in the geometrically reconstructed scene, it is not necessarily encoded there in the optimal form.[9] Moreover, a reconstructed scene would require a homunculus to make sense of it (Pylyshyn, 1973): of all possible approaches to scene interpretation, the one that starts with reconstruction is the most roundabout, because reconstruction per se contributes nothing towards interpretation. Third, as we mentioned in section 4, psychophysical evidence indicates that the human performance in recognition is in many cases inconsistent with the reconstructionist hypothesis.

5.2 Feature detectors and shape constancy

The roots of the alternative approach to representation taken by the multiple-views models can be traced back to Locke's *Essay concerning human understanding*. In the part that addressed the problem of semantics of mental representations, Locke suggested that an idea represents a thing in the world if it is naturally and predictably evoked by that thing, and not necessarily, as the Aristotelians would have it, if the idea resembles the thing in any sense. According to this suggestion, and using a standard terminology of vision research, the activity of a sufficiently reliable *feature detector* constitutes a representation of the presence of an appropriate feature out there in the visual world.[10] Representations of entire objects can then be built out of the available primitives following simple principles of association, if the repertoire of features is sufficiently rich (in particular, if, at some stage in the system, features correspond to entire views of objects). Shape constancy, therefore, may be a combined effect of knowing what the relevant objects look like from different vantage points, of internalized knowledge concerning what happens to a view of an object that undergoes a certain transformation in space (Shepard, 1983), and of judicious selection

[9] Marr, in his book, makes a very similar argument a mere fourteen pages before he discusses the work of Warrington (Warrington and Taylor, 1973) which swayed him in favor of the reconstructive approach.

[10] Cf. Putnam's dictum on the meaning of a representation being in the world, not in the head (Putnam, 1988).

of just the right things about the appearance of a 3D object that better be ignored.[11]

References

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychol. Review*, 94:115–147.

Binford, T. O. (1971). Visual perception by computer. In *IEEE Conference on Systems and Control*, Miami Beach, FL.

Brady, M. and Yuille, A. (1984). An extremum principle for shape from contour. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:288–301.

Bülthoff, H. H. and Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64.

Bülthoff, H. H. and Mallot, H. A. (1988). Interaction of depth modules: stereo and shading. *Journal of the Optical Society of America*, 5:1749–1758.

Burns, J., Weiss, R., and Riseman, E. (1990). View variation of point-set and line segment features. In *Proceedings Image Understanding Workshop*, pages 650–659.

Churchland, P. S. (1987). *Neurophilosophy.* MIT Press, Cambridge, MA.

Cummins, R. (1989). *Meaning and mental representation.* MIT Press, Cambridge, MA.

Dennett, D. C. (1991). *Consciousness explained.* Little, Brown & Company, Boston, MA.

Edelman, S. (1991a). Features of recognition. CS-TR 91-10, Weizmann Institute of Science.

Edelman, S. (1991b). A network model of object recognition in human vision. In Wechsler, H., editor, *Neural networks for perception*, volume 1, pages 25–40. Academic Press, New York.

Edelman, S. (1992). Class similarity and viewpoint invariance in the recognition of 3D objects. CS-TR 92-17, Weizmann Institute of Science.

Edelman, S. (1993). On learning to recognize 3D objects from examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:833–837.

Edelman, S. and Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400.

[11]The choice of whether to ignore a troublesome aspect of the stimulus or to try to make sense of it has been given an illuminating treatment in the section on "Filling in versus finding out" in (Dennett, 1991, p.344). The problem of filling in the visual input across the blind spot or a scotoma (which, instead, may be merely labeled as being "the same" as the surrounding area), discussed by Dennett, corresponds in our case to the problem of reconstructing the 3D shape of the object, an alternative to which has been outlined in section 4.2.3.

Edelman, S. and Weinshall, D. (1991). A self-organizing multiple-view representation of 3D objects. *Biological Cybernetics*, 64:209–219.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395.

Forsyth, D., Mundy, J. L., Zisserman, A., and Brown, C. M. (1990). Invariance - a new framework for vision. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 598–605, Osaka. IEEE, Washington, DC.

Freeman, W. T. (1993). Exploiting the generic view assumption to estimate scene parameters. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 347–356, Washington, DC. IEEE.

Hildreth, E. C. (1987). Edge detection. In Shapiro, S., editor, *Encyclopedia of artificial intelligence*, pages 257–267. John Wiley, New-York, NY.

Horn, B. K. P. and Brooks, M. (1989). *Seeing shape from shading.* MIT Press, Cambridge, Mass.

Hummel, J. E. and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, 99:480–517.

Koenderink, J. J. and van Doorn, A. J. (1991). Affine structure from motion. *Journal of the Optical Society of America*, 8(2):377–385.

Lamdan, Y. and Wolfson, H. (1988). Geometric hashing: a general and efficient recognition scheme. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 238–251, Tarpon Springs, FL. IEEE, Washington, DC.

Lowe, D. G. (1986). *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston, MA.

Lowe, D. G. (1987). Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31:355–395.

Marill, T. (1989). Recognizing 3D objects without the use of models. A.I. Memo No. 1157, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

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

Mitsumoto, H., Tamura, S., Okazaki, K., Kajimi, N., and Fukui, Y. (1992). 3D reconstruction using mirror images based on a plane symmetry recovering method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:941–946.

Mohan, R., Weinshall, D., and Sarukkai, R. R. (1993). 3D object recognition by indexing structural invariants from multiple views. In *Proceedings of the 4th International Conference on Computer Vision*, pages 264–268, Berlin, Germany. IEEE, Washington, DC.

Moses, Y. and Ullman, S. (1992). Limitations of non model-based recognition schemes. In Sandini, G., editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 820–828. Springer Verlag.

Palmer, S. E., Rosch, E., and Chase, P. (1981). Canonical perspective and the perception of objects. In Long, J. and Baddeley, A., editors, *Attention and Performance IX*, pages 135–151. Erlbaum, Hillsdale, NJ.

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

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

Poggio, T., Torre, V., and Koch, C. (1985). Computational vision and regularization theory. *Nature*, 317:314–319.

Poggio, T. and Vetter, T. (1992). Recognition and structure from one 2d model view: observations on prototypes, object classes, and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Putnam, H. (1988). *Representation and reality.* MIT Press, Cambridge, MA.

Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin*, 80:1–24.

Richards, W. and Jepson, A. (1992). What makes a good feature? A.I. Memo No. 1356, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Rives, P., Bouthemy, B., Prasada, B., and Dubois, E. (1981). Recovering the orientation and the position of a rigid body in space from a single view. Technical report, INRS-Telecommunications, Quebec, Canada.

Rock, I. and DiVita, J. (1987). A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293.

Sakai, K. and Miyashita, Y. (1992). Neural origanization for the long-term memory of paired associates. *Nature*, 354:152–155.

Shashua, A. (1991). Correspondence and affine shape from two orthographic view: motion and recognition. A.I. Memo No. 1327, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Shepard, R. N. (1983). Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, 91:417–447.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Shepard, R. N. and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171:701–703.

Shepard, S. and Metzler, D. (1988). Mental rotation: effects of dimensionality of objects and type of task. *J. Exp. Psychol.: Human Perception and Performance*, 14:3–11.

Tarr, M. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282.

Ullman, S. (1979). *The interpretation of visual motion.* MIT Press, Cambridge, MA.

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

Ullman, S. (1991). Sequence-seeking and counter-streams: a model for information flow in the cortex. A.I. Memo No. 1311, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005.

Warrington, E. K. and Taylor, A. M. (1973). The contribution of the right parietal lobe to object recognition. *Cortex*, 9:152–164.

Weinshall, D. (1993). Model-based invariants for 3D vision. *International Journal of Computer Vision*, 10(1):27–42.

Weinshall, D. and Tomasi, C. (1993). Linear and incremental acquisition of invariant shape models from image sequences. In *Proceedings of the 4th International Conference on Computer Vision*, pages 675–682, Berlin, Germany. IEEE, Washington, DC.

Weinshall, D., Werman, M., and Tishby, N. (1993). Stability and likelihood of views of three dimensional objects. submitted for publication.

Witkin, A. P. and Tenenbaum, J. M. (1986). On perceptual organization. In *From Pixels to Predicates*, pages 149–169. Ablex Publishing Corporation, Norwood, New Jersey.