Computational theories of object recognition

Shimon Edelman
School of Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton BN1 9QH, UK
email: shimone@cogs.susx.ac.uk
tel: +44 1273 678659
fax: +44 1273 671320

June 1997. Revised September 1997

SUMMARY. This paper examines four current theoretical approaches to the representation and recognition of visual objects: structural descriptions, geometric constraints, multidimensional feature spaces, and shape-space approximation. The strengths and the weaknesses of the theories are considered, with a special focus on their approach to categorization — a computationally challenging task which is not widely addressed in computer vision (where the stress is rather on the generalization of recognition across changes of viewpoint).

Keywords: shape representation, categorization, structural descriptions, feature spaces, geometric constraints, shape space.

The study of visual object recognition has seen such rapid development lately that its comprehensive survey would not fit within the confines of a journal paper. In this short review (which, in places, is little more than an annotated bibliography), I concentrate on some aspects of what Marr [1] termed the *computational theory* of object representation. Recognition *algorithms* stemming from the different computational formulations of the problem of representation are also mentioned. Very little space is devoted to implementational issues, and none at all to the evaluation of various theories as models of human performance or as explanations of the functional neurobiology of object recognition in primates (for these, see [2, 3, 4, 5], and the forthcoming special issues of *Vision Research* and *Cognition*).

In cognitive science, debates concerning theories of object representation traditionally center on computational problems stemming from the effect of viewpoint on the appearance of objects [6, 7, 8, 9, 10, 11, 12]. The emergence of powerful formal methods for overcoming the effect of viewpoint (e.g., [13, 7, 14]), and the recent successes of surprisingly simple empirical approaches to recognition (e.g., [15, 16]) are likely to shift the focus of theoretical discussion to other topics. Indeed, my chief aim here is to bring to the foreground a class of computational problems that differ from those related to viewpoint dependence, yet confront any recognition system. These problems arise from the need to *categorize*, or make sense of, novel objects.

Perception, recognition, and categorization

The spectrum of problems arising in connection with object recognition is best understood in terms of two basic distinctions. The first of these has to do with the *perception* of the shape of an observed object on the one hand, and the *recognition* of objects on the basis of their shapes on the other hand. A classical observation of this distinction was made by Wittgenstein ([17], II, xi), who discussed at length the difference between seeing a shape and seeing it *as* something. Unlike "merely" perceiving a shape (a problem not addressed in the present review), recognizing it as something involves memory, that is, representations of shapes seen in the past. The form of these representations is constrained by the various factors such as orientation and illumination that affect the appearance of objects. Because of the effects of orientation, for instance, simply storing a particular snapshot of an object for future reference would not do: another view of the same object may turn out to be less similar to the stored view than to a view of a different object, leading to an erroneous recognition. As noted above, contemporary theoretical treatments of recognition concentrate precisely on this problem; state of the art algorithms [18] are capable of overcoming it, provided that several views per object are available, and that certain auxiliary problems such as correspondence[1] are solved.

These algorithms, however, do not necessarily observe another distinction: that between recognition and categorization. In most everyday recognition tasks, memory is involved in a peculiar manner: whereas recognition is (literally) remembering a thing one saw before, making sense of a novel object (as when a child realizes that a giraffe, glimpsed at a zoo for the first time, is a quadruped animal) is more like remembering a thing never seen before. Obviously, to be capable of categorization either the system's memory traces, or the process whereby these are compared to the stimulus, or both, must be structured appropriately; no algorithm designed merely to counter the effects of viewpoint would do.

One may note that the distinction between familiar and novel shapes need not be of any consequence at the "front end" of a visual system — the processing stage whose task is to form a description of the current stimulus. To wit, completely unfamiliar objects can be, in principle, described in terms of their constituent

[1] Words printed in sans-serif can be looked up in the Glossary, at the end of this article.

edges, surfaces and other spatial features, without recourse to memory. However, even a complete specification of the shape of an object does not qualify as its categorization: the latter has to do with other objects which the stimulus resembles, rather than with the details of the shape of the stimulus itself. In other words, faithful geometrical description is no substitute for recognition or categorization; computing such a description constitutes a separate problem, which may have little to do with the problem of representing objects in a form suitable for recognition or categorization.

Among the various computational approaches to representation proposed in the past, the one intuitively most suitable for categorization is the variant of structural decomposition usually attributed to Marr and Nishihara [19], and subsequently popularized as a psychological theory by Biederman [8]. Until very recently, it seemed that this structural theory was the only one offering a principled treatment of novel objects. For this reason, it is the first theory I discuss; the emerging alternatives are presented in later sections.

Structural decomposition

In any structural decomposition model, the shape of an object is described in terms of relatively few generic components, joined by spatial relationships that are chosen from an equally small fixed set (see Box 1). The standardization of the primitives (the components and their relationships) is crucial in that it allows representation of novel objects. Mathematically, comparison between objects then amounts to labeled graph matching (a difficult combinatorial problem [20]), and categorization — to the attribution of an object to an equivalence class of graphs corresponding to shapes that are structurally identical, yet may be geometrically distinct.

Recognition By Components

A typical structural theory, Biederman's [8] Recognition By Components (RBC), postulates a set of 30 or so primitive shapes (geons), claimed to be easily detected in images due to their nonaccidental properties. The latter are 3D features that are almost always (that is, barring an accident of viewpoint) preserved by the imaging (projection) process [21].

The use of nonaccidental features to infer the presence of geons, and the distributed computation of the graph structure of the input object are the cornerstones of the implementation of RBC described in [22]. This work demonstrated the ability of a carefully engineered multilayer neural network to derive structural representations from labeled line drawings. In many respects, however, it also served to highlight the shortcomings of RBC, three of which are discussed below (see Figure 1).

Computational problems

The need for metric information. Although RBC is, in principle, capable of representing novel shapes (via their structural decomposition), this ability comes at the expense of ignoring fine (metric, or quantitative, as opposed to structural, or qualitative) distinctions among shapes. This shortcoming was recognized and amended by Stankiewicz and Hummel [23], who augmented RBC by quantitative variables, encoding, for instance, the lengths of the various parts of an object, in addition to their qualitative characteristics such as convexity or cross-section shape.

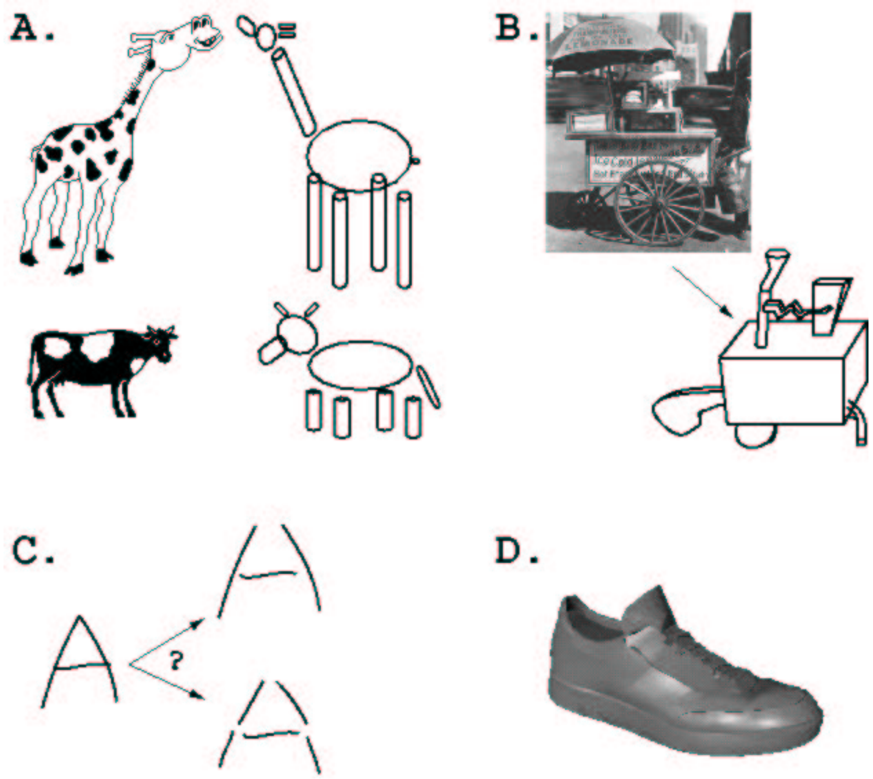

Figure 1: Computational problems with structural representations. A. Structural descriptions must be accompanied by metric information, to represent differences among commonly encountered categories. The inclusion of metric details reduces the ability of structural methods to deal with novel objects. B. A picture of a New York City street-corner hot dog cart, and a stylized object, which, as Biederman [8] suggests, may be described as such following a structural decomposition in the visual system. At present, there is no reliable method for mapping a gray-level image into a collection of (labeled) primitives (lines, corners, etc.) from which RBC's geons are constructed. Thus, although a carefully engineered system such as that described in [22] can form a structural description of the line drawing of a cart-like object, the goal of deriving such a description directly from an image remains elusive. C. Even in simpler tasks (e.g., in character recognition, where the figure is readily separable from the ground), the derivation of a structural description is problematic. The difficulty here stems from the possibility to assign multiple structural descriptions to the same image. D. In some tasks, coming up even with one structural description is problematic; how does one represent a shoe in terms of RBC's geons [7]?

Difficulties with the recovery of parts. A more severe problem faced by the structural approaches is the need for reliable detection of parts in images. One aspect of the detection problem — inferring 3D structure from a 2D projection — is essentially solved by the use of nonaccidental features. Nevertheless, the difficulty of finding in the image lines and junctions whose nonaccidental relationships are to be used to infer the presence of geons has so far precluded RBC from being applied to the recognition of objects in gray-level images. The model described in [22] worked from hand-labeled line drawings; attempts to apply Recognition By Components to images of real objects invariably involve highly simplified shapes consisting of 2-3 clearly distinguishable parts, and are likely to use range data instead of photometric images.

Instability of description in terms of parts. Even if the input to a structural decomposition system is given in the form of a collection of labeled lines, its interpretation in all but the most artificial examples is problematic, because of an inherent instability that affects all structural approaches. The instability stems from the possibility to decompose any shape in a number of ways, depending on the primitives that are assumed to exist. For example, a handprinted letter A can be decomposed into either three or into five approximately straight lines, depending on whether the sides of the A are represented as single long lines or concatenations of two shorter segments each. It should be realized that the same problem arises in any combinatorially structured domain. For example, the problem of basis pursuit in signal processing [24] consists of choosing a subset of basis functions, whose weighted sum best approximates a given signal. The difficulty here stems from the possibility of decomposing the signal in many different ways, depending on the choice of basis functions and on the optimization criterion. Likewise, in latent-variable analysis in statistics [25] one is faced with the problem of selecting (or rather, postulating) the set of variables, and of estimating their contributions to the observables.

It has been suggested that the instability problem may be alleviated by imposing a prior expectation (in the Bayesian sense) on each possible solution, at a number of levels of a structural hierarchy, as in pixel – edge element – curve [26, 27]. This *compositional* approach attempts to combat instability by regularization of the solution, and by using top-down expectations descending from the higher levels of the hierarchy to help disambiguate the interpretations at the lower levels.

Because structural interpretation guided by these principles has not been attempted for unconstrained object classes or for “raw” gray-level images, it is difficult to estimate the effectiveness of the compositional approach in overcoming the instability problem, or the sheer combinatorics of representing moderately complex objects as structural hierarchies (cf. [28, 29]). Experience with large-scale projects that adopted this approach has not been encouraging. A vintage example, the MIT Vision Machine program [30], which explored bottom-up structured solutions to low-level visual tasks such as edge detection, confronted computational difficulties and ran into a representational dead end. On the one hand, its recovery of shape from various low-level cues proved to be unreliable. On the other hand, nobody seemed to have any use for the recovered shape even when the low-level computations did work. The abandonment of this and similar projects in the early 1990’s suggests that the current attempts at reviving the structural interpretation methods [26, 27] are about to face severe difficulties, both computational and conceptual.

Geometric constraints

Whereas structural methods ignore much of the quantitative information inherent in the image locations of object features, geometric methods such as alignment [7] use this information to identify the object and to compute its pose with respect to the observer (Box 2). These methods rely on the following viewpoint consistency constraint [31]: the establishment of correspondence between localized features of the object and of the image constrains the relative placement of the object features, and, therefore, the object's geometry.

Varieties of alignment

Given a library of object models, each accompanied by a set of fiducial geometric features, and a corresponding set of features in the images, one can compute the hypothetical viewing position of each candidate object, and verify the hypothesis of its presence in the image by transforming the model (aligning it to the image) and evaluating the goodness of the resulting match. Ullman [7] proved that the locations of as few as three features in the image suffice, under certain conditions, for unique alignment (the robustness of the method can be improved by using more features, or anchor points, than strictly necessary). Soon afterwards, Ullman and Basri [14] realized that storing a few views per object (with full correspondence) obviates the need for maintaining 3D models of objects. This work prompted the development of a variety of algebraic methods for view-based recognition, all based on the observation that views (i.e., vectors of image coordinates of a set of fiducial points) of a rigid object reside in a low-dimensional (linear, if the projection is orthographic) subspace of views of all possible objects.

Computational problems

Need for feature correspondence. Because the establishment of correspondence is an absolute prerequisite for all the above methods, poor performance of the feature extraction and correspondence stage can completely disrupt subsequent recognition. Given the difficulty of detecting features (either points [13, 7] or regions [32]) reliably in a bottom-up fashion, it seems that alignment will remain practical only in the context of industrial object identification.

Lack of abstraction of category information. A more serious problem with alignment-like methods is their too literal treatment of object geometry. Alignment attempts to account for the observed location of every feature of the object; in comparison, categorization of novel objects requires abstraction of geometrical detail. This seems to call for a conceptual framework that is inherently statistical (cf. [33, 34]). Mere tolerance to certain variation in model parameters does not seem to suffice, as indicated by the relatively disappointing performance of a version of alignment that adopted this approach [35].

Lack of an explicit representation of object statistics. From a statistical standpoint, alignment is deficient because it is geared to treating two objects at a time, instead of capturing several dimensions of variation within an ensemble. A step towards combining alignment with statistics has been made by Basri [36], who developed an algorithm for representing object classes by their prototypes (defined as statistical averages of exemplars). However, because this method assigns the input to the closest known category (an approach known in pattern recognition as the nearest-neighbor decision), it is essentially limited to the processing of objects that resemble one of the prototypes much more closely than any of the others.

Multidimensional feature spaces

The limitation imposed by the use of the nearest-neighbor decision procedure can be removed by reformulating the problems of recognition and categorization as clustering in multidimensional feature spaces ([37]; see Box 3). Importantly, this framework facilitates the representation of a novel object by its membership in a number of clusters (categories) simultaneously. The idea of using multiple reference classes may be illustrated by thinking of a giraffe: one may imagine that its ancient Roman name, *camelopardalis*, reflects the observation of its similarity to a camel (in its shape), and to a leopard (in its visual texture); cf. Figure 2.

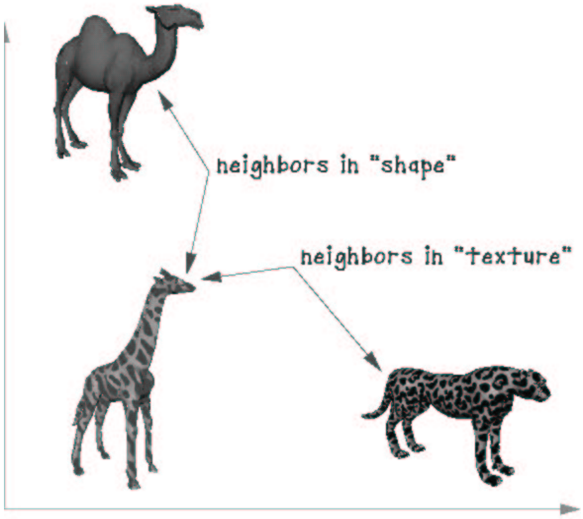

Figure 2: A novel object (*camelopardalis*), represented by two kinds of features: those having to do with shape (which make it a neighbor of camel), and those related to visual texture (according to which it is a neighbor of leopard).

Multidimensional histograms

A judicious choice of features in this classical pattern-recognition approach to vision abolishes the need for precise correspondence, and can lead to considerable invariance with respect to object transformations. This has been demonstrated by the success of systems that represent objects by histograms (computed over the entire input image) of multidimensional vectors of measurements related to color [38, 16] and to local distributions of intensity [16, 15].

Proponents of theories of representation based on encoding qualitative structure or metric details of objects may find it difficult to accept the possibility of representing shapes without explicitly specifying their geometry (e.g., by tallying the frequency of various measurements). It seems to me that this difficulty stems from a confusion between the phenomena of perception (seeing a shape) and of representation (seeing it *as* something; cf. Wittgenstein's point raised in the introduction). Whereas the structure delivered by a perceptual system should better be geometrically faithful to its object, there is no *a priori* computational reason to assume that this structure is to be retained in the memory trace laid down by the representational system into which perception feeds. In fact, the computational approaches discussed here and below postulate that representations are not geometrically or structurally analogous to percepts; whether or not the human visual system maintains structurally or geometrically faithful representations of objects is an empirical question that is not addressed here.

Computational problems

Combining diagnosticity with invariance. The main problem of feature-space methods is finding features that afford reliable discrimination among similar objects, along with invariance across object transformations (this corresponds to the issue of stability vs. sensitivity of features, mentioned by Marr [1]). Simple geometrical arguments [39, 37] can be used to show that these are conflicting requirements, which can be met jointly only as a result of a compromise, or following a special training [40]. That is, unless the features are both absolutely diagnostic and inherently invariant to the transformation in question, as in the case, say, of the bar codes used to label goods in stores.

Difficulty of learning from examples in multidimensional spaces. Because statistical representations such as those involving feature histograms must be learned, the issue of dimensionality assumes a central role in determining the viability of any given scheme. Learning from examples in a high-dimensional space is computationally problematic. The problem, known as the curse of dimensionality [41], lies in the exponential dependence of the required number of examples on the number of dimensions of the representation space. Dimensionality reduction thus becomes of primary importance [42]. The challenge, then, is to reduce dimensionality while preserving the ability of the representational system to deal with novel objects, without having to come up with novel features.

Approximation in feature spaces

An effective way to increase the likelihood that most possible differences between two objects would be captured by (that is, would have a non-vanishing projection onto) at least some of the dimensions of the feature space is, somewhat paradoxically, to *increase* the number of features (i.e., distinct measurements performed by the system on the image). The dimensionality of the resulting powerful yet unwieldy representation space must be reduced. The first stage in this process can be based on the observation that an object undergoing a transformation (e.g., rotation) is mapped into a low-dimensional manifold in the feature space [43, 44], provided that the measured features are smoothly related to the transformation parameters (Box 4).

Given a few points known to belong to the proper manifold (i.e., a few views of the object in question), it is possible to recognize any other view of the object by interpolation. Note that the nominal dimensionality of this representation is equal to the number of views used to interpolate its view space; the true dimensionality can be much lower, and is determined by the number of degrees of freedom of the object (equal to three for objects that are only allowed to rotate in space). Although this method seems to be irrelevant to the representation of novel objects (for which no example views are available in advance, by definition), it can in fact be extended to encompass both recognition and categorization.

Representation by similarities to prototypes

The main requirement for such an extension is that the mechanism used for interpolating the view-space manifolds of familiar objects (“prototypes”) respond to unfamiliar objects as well [45]. Under this scheme, novel objects are represented by their similarities to the prototypical or reference shapes (cf. Figure 3), which, in turn, are represented by stored chosen views. The nominal dimensionality of the resulting shape space is determined by the number of reference shapes. An implementation of this scheme described in [46] contains 10 reference-shape detection modules, each of which computes the similarity of its preferred shape to the input. The vector of similarities is then subjected to further processing (e.g., compared to stored vectors to determine category membership). This system was tested on tens of gray-level shaded images of each of 50 novel objects (man-made and natural), achieving satisfactory results both in recognition and in categorization.

The most prominent feature of this approach (and the source of a potentially serious shortcoming, as discussed below) is its focus on similarities among shapes. Under certain conditions, its representation of similarities is formally veridical — a property that holds a great philosophical appeal (for some of the mathematics behind this, see [47, 48]). However, the geometry of individual shapes is not made explicit; the present method shares this characteristic with all other approaches based on abstract feature spaces.

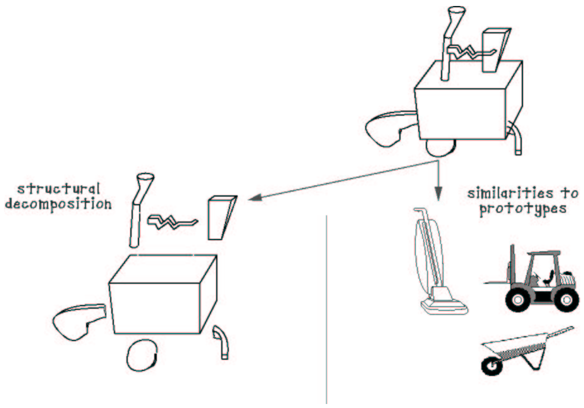

Figure 3: Representing Biederman's [8] "hot dog cart" in terms of spatial relations among abstract or generic parts (left), and by similarities to a number of concrete entire objects (right).

Computational problems

Potential proliferation of prototypes. Representation by similarities to prototypes has been only implemented so far on a rather limited scale. It is not clear, therefore, how well will this approach scale up with the number of possible input objects. Although it is easy to program a system to acquire new prototypes at need, the rate of such acquisition must decrease to zero, and the number of prototypes must asymptote at some small fraction of the total number of objects, if the system is to be viable. Another aspect of this problem is related to the instability of interpretation that occurs in the structural methods: if too many prototypes are found nearly equally similar to the input, it may be difficult to choose the best subset while attempting to lower the dimensionality of the representation.

Lack of representation of structure. A further challenge for this approach is posed by the need to make explicit the dimensions of similarity. Returning to the example of Figure 2, it seems reasonable to require of a representational system not only to note the resemblance of a giraffe to the camel and to the leopard, but also to realize that the former has to do with certain dimensions of the giraffe's shape and the latter – to its color and visual texture. The structural underpinnings of similarity need also be represented explicitly.

Consider two objects: a sphere attached to the top of a cube, and a cube on top of a sphere. By all accounts, both these objects are equally similar to a sphere; one expects that their structural difference be represented as well. It has been claimed that this can be done by associating similarity with certain locations in the image, and by maintaining pointers to the relevant locations along with the similarity values [49]; this approach is yet to be tested in practice.

Conclusions

This paper presented a critical review of four theoretical approaches to object representation: structural descriptions, geometrical models, high-dimensional feature spaces, and a low-dimensional representation based on similarity of the object to several prototypes. When judged by the criteria of computational plausibility and functional adequateness for the purposes of recognition and categorization, all the approaches were found to be deficient, although the nature and the severity of the problems that were identified varied from one theory to another.

On purely computational grounds, the best choice available to a designer of a visual system is probably a library of geometric models of objects accompanied by an alignment mechanism — but only if the range of tasks is restricted to the identification of one of the stored models at a time. If the system is to carry out categorization in addition to identification, a structural approach may be resorted to. Unfortunately, the extraction of fiducial features and the establishment of feature correspondence required for alignment are not always sufficiently reliable, while the structural decomposition and the matching algorithms needed for recognition by components tend to suffer from instability and from combinatorial problems. These difficulties, which are a matter of computational principle rather than implementational detail, limit the appeal of these two theories (especially as far as categorization is concerned). This, in turn, casts a certain doubt (which may or may not prove to be well-founded by an empirical investigation) on their validity as models of object representation in biological systems.

The other two theories discussed above take a completely different route to representation, by adopting the feature space paradigm. The early work in pattern recognition abstracted away the issues surrounding the choice of features and their behavior under object transformations. In contrast, contemporary approaches rely on feature spaces derived from studies of low-level vision, and use mathematical concepts and techniques that only recently became known outside their narrow fields of study (e.g., shape space approximation, dimensionality reduction). These foundations may be partly responsible for the impressive performance of some recently implemented feature-based systems in tasks that have confounded computer vision researchers for decades.

The most intriguing outstanding question in this context is whether the feature-based representations that hold promise for identification and categorization can also be made to support tasks that require an explicit manipulation of object structure or refer to the object's parts. A potentially fruitful approach here may be to label features by their approximate origin in the image [49]. For example, two instances of an "eye" in conjunction with a "nose" and a "mouth" (all properly positioned in the image) qualify as a representation of a face [50]. There are similarities here to the part-based structural approach, but there are also important differences. First, the features can be concrete shapes, as in the representation by similarities to prototypes. This allows for efficient recognition of structures, circumventing the problematic need for composition from generic primitives. Second, the features may span several levels of a hierarchy of parts and wholes, facilitating concise representations optimized for individual tasks. Third, the required structural relationships hold in the 2D image, not in the 3D object-centered space, and can be determined by an easily implementable attention-like mechanism. It is interesting to note that these modifications to feature-based representations correspond closely to Grenander's [51] distinction between pattern recognition (a process whereby patterns as wholes are attributed to various classes) and pattern theory — a computational framework that aims to account for the structure of each pattern and for the processes that may have generated it, and not merely classify it [52]. Further developments in this direction may lead to a more comprehensive yet computationally plausible computational theory of recognition and categorization.

Acknowledgments

I thank Nathan Intrator and Peter Dayan for useful discussions, and Sharon Duvdevani-Bar, Kalanit Grill-Spector and two anonymous reviewers for comments on earlier versions of the manuscript.

References

[1] D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.

[2] P. Jolicoeur and G. K. Humphrey. Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In V. Walsh and J. Kulikowski, editors, *Perceptual constancies*, chapter 10, pages 69–123. Cambridge University Press, Cambridge, UK, 1998.

[3] N. K. Logothetis and D. L. Sheinberg. Visual object recognition. *Annual Review of Neuroscience*, 19:577–621, 1996.

[4] K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139, 1996.

[5] E. T. Rolls. Visual processing in the temporal lobe for invariant object recognition. In V. Torre and T. Conti, editors, *Neurobiology*, pages 325–353. Plenum Press, New York, 1996.

[6] S. Pinker, editor. *Visual Cognition*. MIT Press, Cambridge, MA, 1985. special issue of *Cognition*.

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

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

[9] M. J. Tarr and H. H. Bülthoff. Is human object recognition better described by geon-structural-descriptions or by multiple-views? *Journal of Experimental Psychology: Human Perception and Performance*, 21:1494–1505, 1995.

[10] I. Biederman and P. C. Gerhardstein. Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff. *Journal of Experimental Psychology: Human Perception and Performance*, 21:1506–1514, 1995.

[11] H. H. Bülthoff, S. Edelman, and M. J. Tarr. How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5:247–260, 1995.

[12] S. Ullman. Sequence-seeking and counter-streams: a model for information flow in the cortex. *Cerebral Cortex*, 5:1–11, 1995.

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

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

[15] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In B. Buxton and R. Cipolla, editors, *Proc. ECCV'96*, volume 1 of *Lecture Notes in Computer Science*, pages 610–619, Berlin, 1996. Springer.

[16] B. Mel. SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9:777–804, 1997.

[17] L. Wittgenstein. *Philosophical Investigations*. Blackwell, London, 1973.

[18] S. Ullman. *High level vision*. MIT Press, Cambridge, MA, 1996.

[19] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B*, 200:269–294, 1978.

[20] M. R. Garey and David S. Johnson. *Computers and intractability: a guide to the theory of NP-completeness*. W. H. Freeman, San Francisco, CA, 1979.

[21] David G. Lowe and Thomas O. Binford. The Recovery of Three-Dimensional Structure from Image Curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(3):320–326, 1985.

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

[23] B. Stankiewicz and J. Hummel. MetriCat: a representation for basic and subordinate-level classification. In G. W. Cottrell, editor, *Proceedings of 18th Annual Conf. of the Cognitive Science Society*, pages 254–259, San Diego, CA, July 1996.

[24] S. Chen and D. L. Donoho. Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44, Pacific Grove, CA, 1994. IEEE Comput. Soc. Press.

[25] K. Jöreskog and H. Wold. *Systems under indirect observation: causality, structure, prediction*. North-Holland, Amsterdam, 1982.

[26] S. Geman. Minimum Description Length priors for object recognition. In *Challenging the frontiers of knowledge using statistical science (Proc. JSM'96)*, 1996.

[27] E. Bienenstock, S. Geman, and D. Potter. Compositionality, MDL priors, and object recognition. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Neural Information Processing Systems*, volume 9. MIT Press, 1997.

[28] R. A. Brooks. Model-based three-dimensional interpretations of two-dimensional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:140–149, 1983.

[29] J. H. Connell. Learning shape descriptions: generating and generalizing models of visual objects. A.I. TR No. 853, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1985.

[30] J. J. Little, T. Poggio, and E. B. Gamble Jr. Seeing in parallel: The vision machine. *International Journal of Supercomputing Applications*, 2:13–28, 1988.

[31] D. G. Lowe. *Perceptual organization and visual recognition*. Kluwer Academic Publishers, Boston, MA, 1986.

[32] R. Basri and D. W. Jacobs. Recognition using region correspondences. *Intl. J. Computer Vision*, 25:141–162, 1996.

[33] D. G. Kendall. Shape manifolds, Procrustean metrics and complex projective spaces. *Bull. Lond. Math. Soc.*, 16:81–121, 1984.

[34] F. L. Bookstein. Biometrics, biomathematics and the morphometric synthesis. *Bulletin of Mathematical Biology*, 58:313–365, 1996.

[35] Y. Shapira and S. Ullman. A pictorial approach to object classification. In *Proceedings IJCAI*, pages 1257–1263, 1991.

[36] R. Basri. Recognition by prototypes. *Intl. J. Computer Vision*, 19(147-168), 1996.

[37] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.

[38] M. J. Swain and D. H. Ballard. Color indexing. *Intl. J. Computer Vision*, 7:11–32, 1991.

[39] D. M. Green and J. A. Swets. *Signal detection theory and psychophysics*. Wiley, New York, 1966.

[40] N. Intrator and S. Edelman. Learning low dimensional representations of visual objects with extensive use of prior knowledge. *Network*, 8:259–281, 1997.

[41] R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.

[42] S. Edelman and N. Intrator. Learning as extraction of low-dimensional representations. In D. Medin, R. Goldstone, and P. Schyns, editors, *Mechanisms of Perceptual Learning*, pages 353–380. Academic Press, 1997.

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

[44] D. W. Jacobs. The space requirements of indexing under perspective projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:330–333, 1996.

[45] S. Edelman. Representation, Similarity, and the Chorus of Prototypes. *Minds and Machines*, 5:45–68, 1995.

[46] S. Edelman and S. Duvdevani-Bar. A model of visual recognition and categorization. *Phil. Trans. R. Soc. Lond. (B)*, 352(1358):1191–1202, 1997.

[47] S. Edelman. Representation is representation of similarity. *Behavioral and Brain Sciences*, 21:449–498, 1998.

[48] S. Edelman and S. Duvdevani-Bar. Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, 9:701–720, 1997.

[49] S. Edelman. Biological constraints and the representation of structure in vision and language. *Psycoloquy*, 5(57), September 1994. http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?5.57.

[50] M. Riesenhuber and P. Dayan. Neural models for the part-whole hierarchies. In M. Jordan, editor, *Advances in Neural Information Processing*, volume 9, pages 17–23. MIT Press, 1997.

[51] U. Grenander. *General pattern theory*. Oxford University Press, Oxford, UK, 1993.

[52] D. Mumford. Neuronal architectures for pattern-theoretic problems. In C. Koch and J. L. Davis, editors, *Large-scale neuronal theories of the brain*, chapter 7, pages 125–152. MIT Press, Cambridge, MA, 1994.

[53] W. E. L. Grimson and T. Lozano-Pérez. Localizing overlapping parts by searching the interpretation tree. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:469–482, 1987.

[54] A. Shashua. Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:779–789, 1995.

[55] M. Kirby and L. Sirovich. Application of the Karhunen-Loève procedure for characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.

[56] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3:71–86, 1991.

[57] H. Murase and S. Nayar. Visual learning and recognition of 3D objects from appearance. *Intl. J. Computer Vision*, 14:5–24, 1995.

[58] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of statistics*, 10:1040–1053, 1982.

[59] W. Richards, A. Jepson, and J. Feldman. Priors, preferences and categorical percepts. In D. Knill and W. Richards, editors, *Perception as Bayesian Inference*, pages 93–122. Cambridge University Press, 1996.

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

[61] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4:87–120, 1989.

1. STRUCTURAL DECOMPOSITION

Summary: an object is represented as a collection of parts (chosen from a small alphabet common to all objects), along with their spatial relationships.

Main mathematical tools used in representation and recognition: combinatorial optimization theory.

Examples: recognition by components [22], minimum description length interpretation [27].

Strengths: (1) conceptual parsimony – a handful of primitives can allow a very large number of object classes to be concisely represented; (2) invariance to changes in viewing conditions, as long as the parts and their relationships can be identified; (3) good support for categorization, stemming from the possibility to represent novel objects in terms of the same primitives as the familiar ones.

Weaknesses: (1) no reliable method for the extraction of structural primitives from raw images exists at present; (2) choosing the right structural interpretation among the many possible in a given situation is likely to be problematic.

2. GEOMETRIC CONSTRAINTS

Summary: an object is represented by the relative coordinates of a small set of its features.

Main mathematical tools: algebra and tensor calculus.

Examples: interpretation trees [53], varieties of alignment [13, 7], linear combination of views [14], trilinear tensor [54].

Strengths: (1) amenability to rigorous mathematical (algebraic) treatment; (2) invariance to changes in viewing conditions, as long as corresponding features in the input and in the stored representation can be identified; (3) demonstrable effectiveness in practical situations involving industrial objects.

Weaknesses: (1) feature detection is unreliable for natural objects; (2) there is at present no clear extension from identification of individual shapes to categorization of shape classes; (3) feature correspondence must be established prior to recognition.

3. MULTIDIMENSIONAL FEATURE SPACES

Summary: an object is represented by a vector of feature values; the features can be geometric, photometric, or any others.

Main mathematical tools: multivariate statistics.

Examples: eigenfaces and related methods [55, 56, 57], histograms of local measurements [16, 15].

Strengths: (1) the features tend to be very easy to detect; (2) the statistical approach provides a common framework for recognition and categorization.

Weaknesses: (1) structure is represented implicitly rather than explicitly; (2) decision spaces tend to be high-dimensional, with the associated computational difficulties.

4. APPROXIMATION IN FEATURE SPACES

Summary: an object is represented as a low-dimensional “surface” (manifold) defined by prescribed points in a feature space.

Main mathematical tools: approximation theory, morphometrics (theory of shape spaces).

Examples: view space approximation [43]; representation by similarities to prototypes [45, 48].

Strengths: (1) the features tend to be easy to detect; (2) the statistical approach provides a common framework for recognition and categorization; (3) approximation by smooth functions alleviates the problems stemming from dimensionality.

Weaknesses: (1) structure is represented implicitly.

OUTSTANDING QUESTIONS

1. How can the alignment approach be made to represent qualitative structure and not only quantitative (metric) details?

2. How can the structural approach work be made to work in practice, on real images?

3. How can the feature space approaches be made to represent structure explicitly?

GLOSSARY

Basis pursuit: in signal processing, choosing the optimal "alphabet" of features for the description of a data set (e.g., basis functions in function approximation), and, simultaneously, estimating the optimal contribution (e.g., weight) of each feature to the observed data.

Correspondence: a mapping that assigns the same label to different projections of the same feature (e.g., a point on an object's surface), as seen in two images taken from different vantage points.

Curse of dimensionality: the exponential dependence of the number of examples required for learning a task on the number of dimensions of the representation space [41]. Suppose that filling a region in a 1-dimensional feature space with representative examples requires 10 data points; a comparable coverage of a 3-dimensional feature space would then require 1000 examples [58].

Fiducial geometric features: features that are associated with fixed locations on the object's surface and therefore can be trusted to convey information about its geometry and orientation. A surface marking or a corner formed by two surfaces meeting at an angle are good geometric features; a smooth bend in a surface is not.

Instability of structural interpretation: the possibility of assigning multiple structural interpretations to a given image, typically exacerbated by a sensitivity of the interpretation process to fine details of the data (which are most prone to corruption by noise).

Labeled graph matching: the establishment of a mapping between the vertices and the edges of two graphs in such a manner that the labels carried by the elements of one graph match those of the other one.

Latent varible analysis: in statistics, explaining the variation observed in a data set in terms of a few postulated "hidden" variables, which give rise to the data through the process of observation or measurement. The aim of the analysis is to specify both the variables and the observation process, to obtain an optimal fit to the data.

Manifold: intuitively, a smooth curve or surface embedded in a higher-dimensional space. Performing, say, a thousand simultaneous measurements on the image of an object effectively maps it to a point in a 1000-dimensional abstract space. If the object is then made to rotate around a fixed axis, the curve ascribed by that point as the values of the various measurements change with rotation is a 1-dimensional manifold embedded in the 1000 dimensions.

GLOSSARY (cont.)

Nonaccidental properties: 2D image features that can be used to make inferences about 3D object structure because of the low likelihood of the former to arise by chance [21]. A representative example of such a feature is a pair of parallel lines; because a chance image alignment of two segments that are in fact not parallel in 3D is unlikely, two parallel lines in the image are a good indicator of the presence of a 3D geon such as a cylinder “out there” in the scene. For a Bayesian treatment of related issues, see [59].

Regularization: a common mathematical technique applied to problems that are formally ill-posed. By extending the definition borrowed from the theory of differential equations, a problem is considered ill-posed if its solution does not depend continuously on the data, or if more than one solution exists, as in the case of structural interpretation. Regularization attempts to reduce the solution space, by imposing additional constraints, over and above those contained in the data. References to the mathematical literature on regularization and a discussion of its relevance to low-level visual tasks can be found in [60].

Shape space: an abstraction, introduced by D. G. Kendall [33, 61] to allow a rigorous statistical treatment of problems having to do with variation of shape in a sample of objects. Objects are treated as points in a metric space, where similar shapes correspond to nearby locations. Categorization thus becomes a matter of determining the location of the stimulus relative to other points in the shape space.

Viewpoint consistency constraint: the mathematical relationship between the projected (2D) co-ordinates of fiducial features, imposed by their arrangement in the 3D space. For a rigid object, the projected location of some initially detected features constrains the possible range of its orientation and predicts the location of other features. If this prediction is verified (as in the second stage of recognition by alignment [7]), the initial hypothesis of the presence of the object in the image is confirmed.