[35] T. Poggio, M. Fahle, and S. Edelman. Synthesis of visual modules from examples: learning hyperacuity. A.I. Memo No. 1271, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

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

[37] J. R. Pomerantz and E. A. Pristach. Emergent features, attention, and perceptual glue in visual form perception. *Journal of Experimental Psychology: Human Perception and Performance*, 15:635–649, 1989.

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

[39] I. Rock, D. Wheeler, and L. Tudor. Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, 21:185–210, 1989.

[40] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[41] T. Sanger. Optimal unsupervised learning in feedforward neural networks. AI Lab TR 1086, MIT, 1989.

[42] R. N. Shepard and L. A. Cooper. *Mental images and their transformations*. MIT Press, Cambridge, MA, 1982.

[43] S. B. Steinman. Serial and parallel search in pattern vision? *Perception*, 16:389–398, 1987.

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

[45] D. W. Thompson and J. L. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In *Proceedings of IEEE Conference on Robotics and Automation*, pages 208–220, Raleigh, NC, 1987.

[46] A. Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977.

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

[48] S. Ullman and R. Basri. Recognition by linear combinations of models. A.I. Memo No. 1152, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[49] R. J. Watt and M. J. Morgan. A theory of primitive spatial code in human vision. *Vision Research*, 25:1661–1674, 1985.

[50] L. E. Wixson and D. H. Ballard. Real-time qualitative detection of multi-colored objects for object search. In *Proc. AAAI-90 Workshop on Qualitative Vision*, pages 46–50, San Mateo, CA, 1990. Morgan Kaufmann Publishers, Inc.

[51] R. S. Zemel and G. E. Hinton. Discovering viewpoint-invariant relationships that characterize objects. In D. Touretzky, editor, *Neural Information Processing Systems*, volume 3, San Mateo, CA, 1991. Morgan Kaufmann.

[18] M. W. Fahle. Parallel, semi-parallel, and serial processing of visual hyperacuity. In *Proc. SPIE Conf. on Electronic Imaging: science and technology*, Santa Clara, CA, February 1990. to appear.

[19] P. C. Gerhardstein and I. Biederman. 3D orientation invariance in visual object recognition. *Invest. Ophthalm. Vis. Science Suppl.*, 32:338, 1991.

[20] G. E. Hinton and L. M. Parsons. Scene-based and viewer-centered representations for comparing shapes. *Cognition*, 30:1–35, 1988.

[21] D. D. Hoffman and W. A. Richards. Parts of recognition. *Cognition*, 18:65–96, 1984.

[22] A. Hurlbert and T. Poggio. Synthesizing a color algorithm from examples. *Science*, 239:482–485, 1988.

[23] N. Intrator. Unsupervised adaptive classification of gray-scale images of 3D objects, 1990. Brown University TR, in preparation.

[24] P. Jolicoeur and S. M. Kosslyn. Coordinate systems in the long-term memory representation of three-dimensional shapes. *Cognitive Psychology*, 15:301–345, 1983.

[25] A. Karni and D. Sagi. Human texture discrimination learning — evidence for low-level neuronal plasticity in adults. *Perception*, 19:335, 1990.

[26] J. J. Koenderink and A. J. van Doorn. Receptive field families. *Biological Cybernetics*, 63:291–297, 1990.

[27] D. LaBerge. Perceptual learning and attention. In W. K. Estes, editor, *Handbook of learning and cognitive processes*, volume 4, pages 237–273. Erlbaum, Hillsdale, NJ, 1976.

[28] R. Linsker. Perceptual neural organization: some approaches based on network models and information theory. *Ann. Rev. Neurosci.*, 13:257–281, 1990.

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

[30] L. T. Maloney and M. S. Landy. A statistical framework for robust fusion of depth information. In *Proc. SPIE: Visual Communications and Image Processing*, pages 1154–1163, 1989. Part 2.

[31] D. Marr. Early processing of visual information. *Phil. Trans. R. Soc. Lond. B*, 275:483–524, 1976.

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

[33] T. Poggio. A theory of how the brain might work. Proc. Summer 1990 Cold Spring Harbor meeting on Quantitative Biology and the Brain.

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

[2] S. Amari. Feature spaces which admit and detect invariant signal transformations. In *Proc. 4th Intl. Conf. Pattern Recognition*, pages 452–456, Tokyo, 1978.

[3] A. Beinglass and H. Wolfson. Articulated object recognition, or, how to generalize the Generalized Hough Transform. Institute of Computer Sciences TR 194/90, Tel Aviv University, 1990.

[4] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.

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

[6] I. Biederman. Aspects and extensions of a theory of human image understanding. In Z. Pylyshyn, editor, *Computational processes in human vision: an interdisciplinary perspective*, pages 370–428. Ablex, Norwood, NJ, 1988.

[7] I. Biederman and E. E. Cooper. Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20, 1991.

[8] M. Bravo and R. Blake. Preattentive vision and perceptual groups. *Perception*, 19:515–522, 1990.

[9] T. M. Breuel. Adaptive model base indexing. A.I. Memo No. 1008, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[10] H. H. Bülthoff and S. Edelman. Psychophysical support for a 2D interpolation theory of object recognition, 1990. submitted.

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

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

[13] S. Edelman and H. H. Bülthoff. Viewpoint-specific representations in 3D object recognition. A.I. Memo No. 1239, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[14] S. Edelman, H. H. Bülthoff, and E. Sklar. Task and object learning in visual recognition. CBIP Memo 63, MIT, April 1991.

[15] S. Edelman and T. Poggio. Bringing the Grandmother back into the picture: a memory-based view of object recognition. A.I. Memo No. 1181, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[16] S. Edelman and T. Poggio. Representations in high-level vision: reassessing the inverse optics paradigm. In *Proc. DARPA Image Understanding Workshop*, pages 944–949, San Mateo, CA, May 1989. Morgan Kaufman.

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

be based on recent studies of unsupervised feature extraction. This issue has been approached using tools of statistics (principal component analysis of image intensities [41], exploratory projection pursuit [23]) and information theory (maximizing mutual information between the input and the output of a layer of feature detectors [28], or between two separate groups of detectors [51]).

The proposed scheme relies on learning also for the synthesis of object-specific recognition modules (cf. [22,35,33].) Learning recognition from examples has been demonstrated in [34,17] for synthetic 3D objects, represented by image locations of simple features such as corners and terminators. Its extension to real-world objects may depend on the use of more appropriate features [15], a conclusion not unexpected in view of the preceding discussion.

4 Summary

The present paper proposes a unified approach to the understanding and modeling of human performance in object recognition. The highlights of the proposed approach are as follows:

1. *Versatility:* Recognition starts with the extraction of a large variety of image-based features.

2. *Plasticity:* Recognition procedure for a given object at a given category level is synthesized at need and is optimized with practice.

3. *Hierarchy:* One of the ways of optimizing recognition performance involves formation of compound features out of simpler ones, and subsequent reliance on such features.

4. *Invariance/diagnosticity trade-off:* Some of the features are well-localized within the 2D image-based reference frame. Exclusive reliance on such features under certain circumstances makes recognition viewpoint-dependent. Other features, such as color, are "diffuse", which makes them relatively viewpoint-invariant. Yet another kind, the nonaccidental features, are viewpoint-invariant for geometrical reasons. There is a trade-off between viewpoint invariance of a feature and its diagnosticity, or the degree of discrimination among object instances that it affords.

Parts of the feature-based model of recognition have been already implemented and found to replicate human performance characteristics at the subordinate level. These include learning of spatially defined features, and learning to recognize objects from image feature locations. Extension of these results to basic-level recognition and to the automatic selection of relevant features will make a unified account of recognition possible.

Acknowledgements

I am grateful to Heinrich Bülthoff and Shimon Ullman for their comments on a draft of this paper, and to Barton Rubenstein for several useful references.

References

[1] D. W. Aha, D. Kibler, and M. A. Albert. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

features are built on top of simpler ones. Prime candidates for lower-level features are simple functions of image intensity (e.g., edges, extrema, etc.; see [49]). As to the higher levels of the hierarchy, some of the possibilities for their formation are outlined below.

Subsumption by symbolic labeling

Marr's notion of the primal sketch [31], in which primitive place tokens are grouped according to rules that resemble the Gestalt laws of perceptual organization, is an early example of what may be considered a hierarchy of spatially localized features. In basic-level recognition by alignment groups of place tokens carrying relatively complex shape information can be more useful if they are assigned symbolic labels. In Ullman's example ([47], p.233; cf. [16]), in the representation of a rooster the label "wiggly" could subsume the complex shape of the crown in most cases (except, possibly, when a particular rooster is to be recognized, and the precise shape details matter).

Local elaboration

When fine details of local image structure are important for the task at hand (as in the case of individual exemplar recognition), such details can be provided by computing higher order derivatives of image intensity than the first derivative, which is strictly necessary for the detection of intensity edges. A specific computational hypothesis regarding such higher-order local features has been advanced by Koenderink and van Doorn [26], who suggested that the various receptive fields found in the primary visual area of mammalian brain can be described as derivative operators.

Spatial agglomeration

An alternative to local elaboration of image description through the computation of higher-order derivatives is joining together of spatially distinct features on a larger scale. A well-known example of a method that yields spatially distributed features is the Fourier transform. If one considers space instead of spatial frequency, compound features may be obtained by encoding conjunctions of the occurrences of simple features that satisfy certain constraints (e.g., constraints on type and location), or groups of place tokens [8]. Biederman's geons fall into this category of compound features. Another example of such features can be found in [3].

Hashing

Under certain conditions (in particular, for the purpose of initial fast indexing into a large database of object representations), one may wish to define features whose only required property is that they allow efficient object access. In this case, object recognition may be likened to texture discrimination, with the individual objects regarded as textons (which must be classified statistically when present in quantity, but need not be described in detail). Linear receptive fields [2], considered as implementations of hashing functions (cf. [9]), are a possible candidate for a biologically motivated computational approach to the extraction of indexing features. Diffuse features such as color also belong to this category [50].

3.7 The role of learning

It seems plausible to assume that the visual system acquires many of the possible features of recognition through experience (psychophysical evidence for perceptual learning at the feature level can be found in [25,18,43]). Computational theory showing how this may happen could

3.5 Computational issues

Recently developed computational theories of recognition can be divided into two broad classes, depending on the type of representation they use (volumetric or image-based).

3.5.1 Theories that rely on volumetric object-centered representations

Recognition by alignment [47] is an example of a computational theory that agrees with human performance in basic-level tasks in its prediction of near independence of recognition rate on viewpoint. According to this theory, the input image is compared with the projection of a stored model in a process resembling template matching, but only after the two are brought into register. The transformation necessary to achieve alignment is computed by matching a small number of "anchor" features in the image with the corresponding features in the model. The aligning transformation is computed separately for each of the models stored in the system. The outcome of the recognition process is the model that fits the input most closely after the two are aligned. Other similar schemes [29,45] choose the best model using viewpoint consistency constraints, which relate projected locations of the features of a model to its 3D structure, given a hypothesized viewpoint.

3.5.2 Theories that rely on image-based representations

Although the basic version of alignment involves 3D object models, this method can be reformulated so that it relies on multiple 2D view representations ([47], p.243), while retaining the stage in which 3D transformations of the input objects are explicitly compensated for. Several recently proposed approaches to recognition dispense both with the need for storing 3D models, and with explicit viewpoint normalization [48,34,17]. In particular, the regularization network scheme [34,36] represents objects by sets of 2D views (vectors of image feature locations), and regards generalization from familiar to novel views as a problem of nonlinear hypersurface interpolation in the space of all possible views.

This scheme was recently shown to replicate central characteristics of human performance in subordinate-level recognition, two of which were mentioned in section 3.2. Within the interpolation framework, canonical views are determined by the spatial distribution of the object's features in the stored representation [17], while the observed pattern of generalization to novel views[5] is obtained as a natural consequence of the reliance on view interpolation [13].

3.6 A taxonomy of compound features

The possibility of extending current interpolation models to include basic-level recognition appears to depend to a large extent on the availability of a wide variety of image-based features of different diagnosticity and complexity. Compound features, discussed in this section, may provide the necessary variety.

A basic observation regarding computational aspects of the compound feature issue is that any recurring image pattern whose identification, in turn, does not presuppose prior recognition of the entire object it is included in should be considered a potential feature. This suggests that the features of recognition may form a loose hierarchy, in which higher-order

[5] The increase in the error rate with misorientation relative to a familiar view depends on the arrangement of familiar views with respect to each other, and not just on the distance between the test view and the nearest familiar view. Specifically, interpolation among familiar views obtained by rotating the object in a fixed plane appears to be easier than extrapolation, which, in turn, is easier than recognition of views that lie outside the plane of rotation [10,13].

3.2.2 Limited generalization

As mentioned in section 3.1, when asked to give a relatively broad classification of an object seen from an odd viewpoint, people virtually never err. In comparison, when the task can only be solved through relatively precise shape matching, the error rate reaches chance level already at a misorientation of about 40° relative to a familiar attitude [38,13]. The detrimental effect of misorientation persists in the presence of depth cues such as binocular disparity, which reduces somewhat the mean error rate, but does not cancel the dependency of error rate on viewpoint [13].

3.3 Psychophysics of recognition: a unified view

In the preceding brief overview one may identify several points of similarity between performance characteristics at the two levels of recognition. First, objects possess canonical views, whether they are considered at the basic ([4], p.66) or at a subordinate [13,14] level. Second, the same processes may underlie the emergence of canonical views at both levels. For instance, an examination of the experimental data on the subordinate-level recognition of computer-generated novel objects [13] reveals that their canonical views can be frequently characterized as displaying one or more of the nonaccidental features mentioned by Biederman ([4], p.36). Third, at the basic level, viewpoint-invariant performance breaks down when key information (the nonaccidental features in [4], or the anchor features in [47]) cannot be reliably extracted from the image. These similarities support the unified feature-based account outlined in section 2.1. In particular, the dichotomy between viewpoint invariance of basic-level recognition and viewpoint sensitivity of subordinate-level recognition does not stand up to scrutiny.

3.4 Open problems

The major psychophysical issues raised by the proposed feature-based approach are the representation and relative importance of different feature types, and the limitations on performance stemming from the reliance on image-based features. Preliminary psychophysical evidence on the role of compound (geon-like) features in object representation is provided by the recent results on the priming of contour-deleted images ([7]; see section 3.1.3), which seems to be mediated by entities that are more complex than mere contour segments. In 2D form perception, similar evidence for compound features such as symmetry and closure may be found in [37].

The second as yet unexplored (but see [46]) issue is that of the relative importance of the various features and feature types for recognition. A possible starting point for the study of this question can be provided by the recent work on the integration of visual submodalities [11,30], and on the role of different depth cues in recognition [13].

The third major experimental issue concerning the features of recognition has to do with the central role assigned by this approach to a viewer-based frame of reference. Previous studies of the frames of reference in object representation involved subjective judgement [24,39] or pairwise shape comparison [20], rather than recognition. At present, there are indications that, at least at the subordinate level, the visual system has no direct access to shape attributes defined in an object-centered frame [38,13]. At the basic level, it appears that such access is facilitated by the use of appropriate (i.e., viewpoint-invariant) compound features, or geons [19], in a manner compatible with the predictions of the feature-based approach.

curvilinearity, symmetry, parallelism and cotermination of contours. It has been shown that human subjects behave as if they take into account nonaccidental properties in their inference of 3D relations among objects and their parts (see the review in [5]). In particular, perceived symmetry contributes to shape constancy and supports fast shape discrimination, vertex configurations play an important role in the interpretation of polyhedral scenes, and, on the other hand, specially arranged accidental cotermination and parallelism give rise to compelling illusions such as the Ames chair and the Penrose triangle.

3.1.3 Explicit representation of intermediate-level features

A class of experiments that provide the most direct evidence in favor of explicit representation of geon-like features in the human visual system involves priming of contour-deleted images [7]. In these experiments priming has been demonstrated between two versions of the same image, from which different contour parts were deleted, so that the two new images had no contour in common, but could, in principle, support a common geon representation (built from the partial contours). In comparison, when the deletion of contours disrupted potential extraction of geons, no priming was observed. This finding can be interpreted in terms of the existence of intermediate representation (such as the compound 2D features proposed in section 2.1) situated between the levels of raw contours and of entire objects.

3.2 Psychophysics of subordinate-level recognition

Unlike basic-level categorization, which is, under normal circumstances,[3] largely viewpoint-invariant, object recognition at the subordinate levels depends on the observer's point of view in several ways. Two major classes of phenomena relevant to the viewpoint dependency of subordinate-level recognition are canonical views and limited generalization.

3.2.1 Canonical views

Common objects are easier to recognize from certain viewpoints, called canonical, than from other, random viewpoints. The canonical views phenomenon, originally demonstrated for basic-level naming of a few everyday objects [32], is found also in subordinate-level recognition of synthetic 3D stimuli such as bent tubes [13] or randomly perturbed smooth spheroids [14]. Importantly, experience with synthetic objects shows that the existence of canonical views is not predicated on preferential prior exposure of the subject to these views.

While uniform initial exposure does not preclude the formation of canonical views, repeated presentation of the same stimulus eventually increases the uniformity of response time over different views of the stimulus. Thus, practice affects the response time aspect of the canonical views phenomenon: after only a few trials, the differences in response time between canonical and random views diminish significantly, even in the absence of any feedback to the subject [13].

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 [44,13]. This dependency of response time on misorientation disappears with practice, along with the general weakening of canonical views effects.[4]

[3] Except when the geon structure of the object is not readily apparent, e.g., due to foreshortening; see [5], p.140ff.

[4] For details regarding this phenomenon and its relationship to classical mental rotation [42], see [17].

compensation from the rest of the process of recognition. According to this approach, features serving as the substrate of recognition are first extracted from the image. The location of some of these features in the image is well-defined, and carries information about the viewpoint. Such features are then used to compensate for the variability of the object's appearance due to potentially unfamiliar viewpoint. This step is especially important in subordinate-level recognition, where shape details matter the most. The compensation process is implicit[1] and relies on interpolation among several stored views [17,34]. At the basic level, where the need for precise shape matching is less severe, relatively invariant features can be used.[2]

3 Discussion

To assess the viability of the proposed feature-based approach, and to identify major problems that remain open, I shall now discuss several experimental and computational aspects of basic and subordinate-level recognition. The discussion starts with the psychophysics, proceeds to mention recent theories of recognition, and concludes with a characterization of some possible types of compound features, which play a central role in the proposed model.

3.1 Psychophysics of basic-level recognition

According to Biederman [5], at the basic level objects can be recognized rapidly, when viewed from most novel orientations, under moderate levels of visual noise, when partially occluded, and when they are new exemplars of a category. Many of experiments studying basic-level recognition in the recent years have been inspired by a theory, called recognition by components (RBC), which was advanced by Biederman as an account of the above phenomena.

3.1.1 Recognition by components

Within the RBC framework, the first stage of recognition is the extraction of luminance, texture and color edges from the image, which effectively provides a line drawing description of the object. Next, these are grouped into (projections of) volumetric components, using constraints provided by *nonaccidental* properties [29] of the arrangement of edges in the image. At the same time, components constituting an object are separated from each other (using heuristics such as the transversality principle [21]) and are classified in terms of a small set of volumetric primitives or *geons*. Finally, a description including the list of geons and their relative disposition in space is matched to memory. Two central aspects of the RBC theory are briefly discussed below.

3.1.2 Nonaccidental properties

Certain properties of edges in a 2D image can be relatively safely interpreted in terms of similar properties of the corresponding 3D edges in the world, due to the low probability of their emergence by accident of a viewpoint. Such nonaccidental properties include collinearity,

[1] As opposed to the explicit normalization in the alignment scheme (see section 3.5.1 [47]), which also separates viewpoint compensation from object identity determination, and relies on image locations of key features to do that.

[2] Main candidates for these are the 3D shape primitives or geons suggested by Biederman ([5]; see section 3.1.1). These can be regarded as complex 2D features: since the information sufficient for geon detection is present in the image ([6] p.395), there is no need to postulate the existence of volumetric object-centered representations over and above the 2D image-based ones.

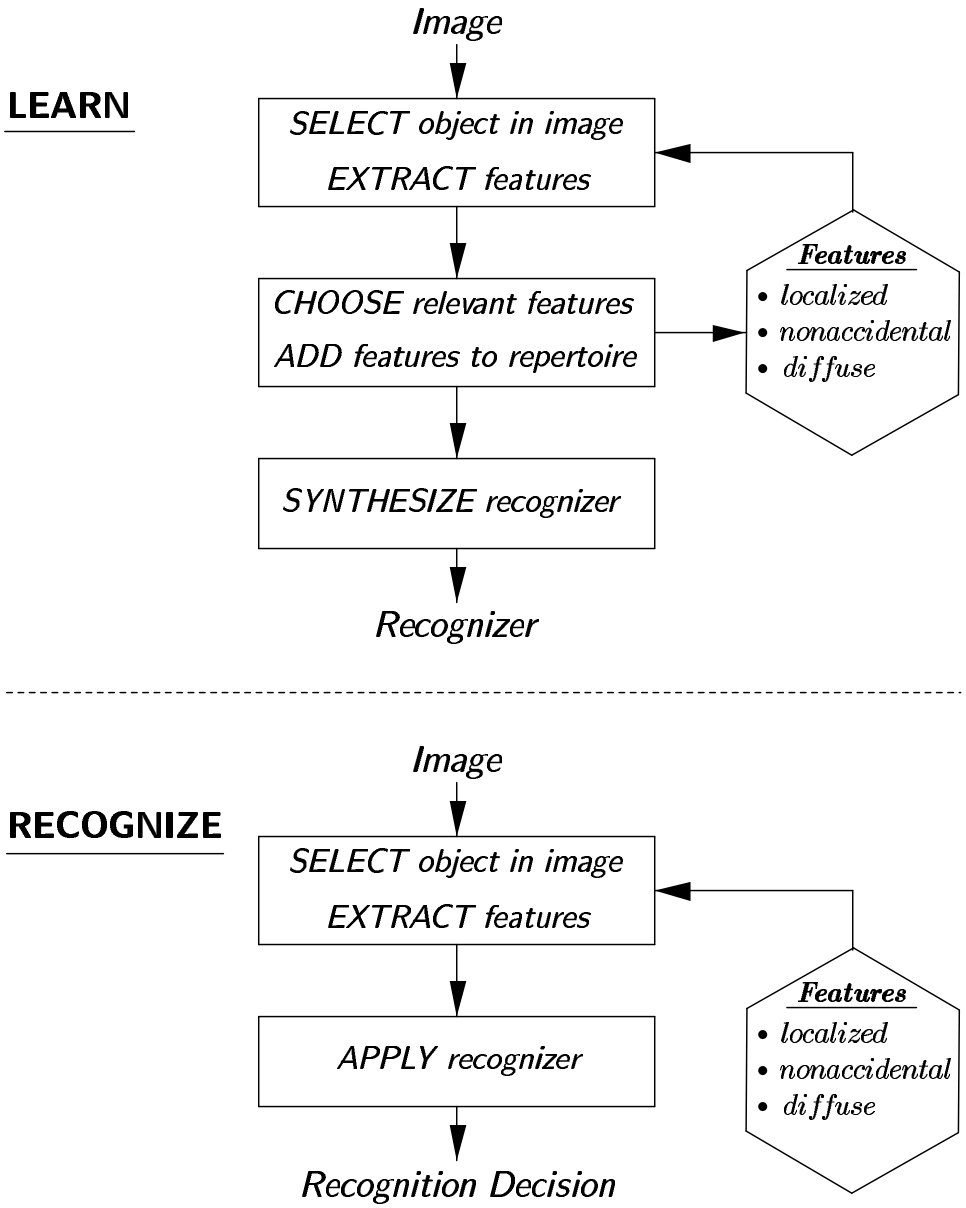

Figure 1: A generic recognition scheme that relies on a variety of potentially useful features of different types. Simple features such as corners convey information mainly through their locations in the image, which, however, are bound to vary with viewpoint. This variation can be largely compensated for by view interpolation [34,17]. Other, nonaccidental features (see [5]), are relatively invariant with respect to viewpoint. Spatial relationships among such features may carry information sufficient for basic-level categorization. Diffuse features such as distinctive coloration can also be useful in certain cases [50]. According to the proposed scheme, object-specific recognition modules are learned at need from examples [34,17]. New features that may be discovered during learning (see section 3.7) are added to the system's repertoire.

Instead, information sufficient for recognition can be found in the 2D image locations of object *features*. The choice of features and their complexity may vary between objects. For example, a pineapple can be recognized by its characteristic pattern of spiny scales. The main feature in this case is textural and is distributed over the object's surface. In comparison, the relevant features of a peanut are both its texture and a characteristic outline (in a line drawing, a round peanut can be confused with a golf ball). Finally, a road vehicle can be recognized as such by the presence of wheels (each of which may be considered a complex feature), but for the drawing of a vehicle to be classified, e.g., as a car, simple additional features such as contour elements and corners must be appropriately situated in the image (presumably, in the vicinity of the locations of corresponding features in the image of a prototypical car). The rest of this section describes a generic recognition scheme based on the idea of a hierarchy of image features.

2.1 The proposed scheme

The proposed scheme, outlined in Figure 1, learns to recognize new objects by constructing and storing their representations in terms of the current feature set. The first stage of processing both in learning and in subsequent recognition is feature extraction. A small sample of features extracted at this stage includes contour elements and events (corners, inflections, terminations), and color and texture labels, as well as *compound* features — recurring combinations of simple features proven useful by previous experience (as an example here, consider the elliptical blob or "wheel" feature in vehicle recognition).

Only a few of the potentially many features detected in the first stage are involved in the recognition of any given object. The relevance of a feature for the recognition of an object is determined by its diagnosticity. Thus, in basic-level recognition, a good feature is one that distinguishes the object's category from other basic categories, while on a subordinate level a distinction is sought between similar members of the same category [27].

Once the relevant features of an object are identified, the system synthesizes a recognizer for the object, essentially by storing the pattern of activities of the chosen feature detectors ([17]; for a background, see, e.g., [1]). Generalization to other views of the same object or to other members of the same basic category is achieved by interpolating among patterns corresponding to several different stored instances [34]. This interpolation can be carried out, e.g., by regularization networks [36]. The advantages of using this method include a natural distributed implementation by a network of "receptive fields", and the possibility of integrating a variety of feature types within the same interpolation mechanism [15].

2.2 Relationship to existing schemes

The proposed framework is only loosely related to the "classical" feature-based approach to recognition (see, e.g., [12]), which purports to represent objects by vectors of feature values in such a manner that different versions (say, views) of the same object are clustered together in the multidimensional feature space, while different objects always produce distinct clusters. Theories of recognition based on abstract feature spaces were gradually displaced in psychology by structuralist ones (e.g., [5]) when the difficulty of finding abstract features that combined viewpoint invariance with diagnosticity became clear (see the discussion in [47]).

The recognition scheme outlined in Figure 1 circumvents this difficulty (*i*) by maintaining a large pool of diverse features, in which invariance and diagnosticity are traded off for one another rather than combined, and (*ii*) by separating the mechanism responsible for viewpoint

Features of recognition

Shimon Edelman

Department of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel
edelman@wisdom.weizmann.ac.il

Abstract

The notion of features of recognition, of varying complexities and possessing different degrees of spatial localization, offers a unified approach to the understanding of object recognition in human vision. According to this approach, the visual system synthesizes a feature-based recognition module for a given object when it is first encountered (cf. [22,35]; the feasibility of such synthesis in 3D object recognition was demonstrated in [34,17]). The details of the synthesized module, such as the types of features it relies upon and the computations applied to the features, depend on the kind of the object and on the required level of recognition. Such flexibility may explain both the efficiency and speed of recognition in human vision (recognition of each object may involve only a few chosen features, combined in an optimal fashion), and its versatility (which may be due to an extensive repertoire of available features).

1 Introduction

The outcome of the process of visual recognition is routinely influenced by the definition of the task and by the expertise of the recognizer. Thus, for example, a feathered object perched on a tree branch can be perceived as a bird, in which case the act of recognition may be better described as classification and is said to happen on a basic category level [40]. The same object, under different circumstances, can be seen as a robin, a jay, or an instance of some other category subordinate to that of bird.

Experimental study of human object recognition at these two levels of categorization reveals differences, which, at the first glance, appear difficult to reconcile, and which have prompted the development of diverging theoretical accounts of recognition. On one hand, recently demonstrated viewpoint dependency of subordinate-level recognition is consistent with theories that hold that the visual system represents three-dimensional objects by storing several of their two-dimensional views ([38,44,13,17]; see also [32]). Developments in computational vision (specifically, new approaches to model-based recognition [48,34,15]) support such a possibility. On the other hand, psychophysical findings on basic-level recognition seem to point towards the representation of object prototypes by collections of volumetric primitives or components [5].

Does the visual system rely on distinct sets of representations and processes for subordinate and basic levels of recognition? In the rest of this paper I outline a unified account of human performance at both levels, based on the notion of features of recognition.

2 Features of recognition

The central tenet of the proposed account is that recognition normally requires neither 3D reconstruction of the stimulus, nor the maintenance of a library of 3D models of objects [16].