[32] J. Stewman and K. Bowyer. Creating the perspective projection aspect graph of polyhedral objects. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 494–500, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[33] 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.

[34] W. B. Thompson and J. K. Kearney. Inexact vision. In *Workshop on motion, representation and analysis*, pages 15–22, 1986.

[35] S. Ullman. Visual routines. *Cognition*, 18:97–159, 1984.

[36] S. Ullman. An approach to object recognition: Aligning pictorial descriptions. A.I. Memo No. 931, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Dec. 1986.

[37] A. Verri and T. Poggio. Against quantitative optical flow. In *Proceedings of the 1st International Conference on Computer Vision*, pages 171–180, London, England, June 1987. IEEE, Washington, DC.

[38] D. Weinshall. Application of qualitative depth and shape from stereo. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 144–148, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[14] K. Ikeuchi and T. Kanade. Applying sensor models to automatic generation of object recognition programs. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 228–237, Tarpon Springs, FL, 1988.

[15] D. W. Jacobs. The use of grouping in visual object recognition. TR 1023, MIT, January 1988.

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

[17] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21:105–117, March 1988.

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

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

[20] M. Minsky. A framework for representing knowledge. In P. H. Winston, editor, *The psychology of computer vision.* McGraw-Hill, New York, 1975.

[21] R. C. Nelson and J. Aloimonos. Using flow field divergence for obstacle avoidance: towards qualitative vision. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 188–196, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[22] S. M. Omohundro. Efficient algorithms with neural network behavior. UIUCDCS R-87-1331, Univ. of Illinois at Urbana-Champaign, April 1987.

[23] T. Poggio, E. B. Gamble, and J. J. Little. Parallel integration of vision modules. *Science*, 242:436–440, 1988.

[24] T. Poggio and the staff. MIT progress in understanding images. In *Proceedings Image Understanding Workshop*, Cambridge, MA,, Apr. 1988. Morgan Kaufmann, San Mateo, CA.

[25] J. R. Quinlan and R. L. Rivest. Inferring decision trees using the minimum description length principle, 1987. manuscript.

[26] W. Richards. How to play twenty questions with nature and win. A.I. Memo No. 660, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Dec. 1982.

[27] A. Rosenfeld. Recognizing unexpected objects: a proposed approach. *Int. J. of Pattern Recognition and Artificial Intelligence*, 1:71–84, 1987.

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

[29] A. Shashua and S. Ullman. Structural saliency: the detection of globally salient structures using a locally connected network. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 321–327, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[30] C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, 29:1213–1228, 1986.

[31] L. Stark, D. Eggert, and K. Bowyer. Aspect graphs and nonlinear optimization in 3-d object recognition. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 501–507, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

3.2 Summary

We have outlined an approach to the representation of objects and situations for visual recognition that combines detailed, integrated feature maps computed in a bottom-up fashion by low-level processes with qualitative, heterogeneous, open-ended high-level representations. Whether this approach will be more successful than the current paradigm that considers vision by and large as inverse optics is an empirical issue that remains to be settled.

References

[1] A. Abbott and N. Ahuja. Surface reconstruction by dynamic integration of focus, camera vergence and stereo. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 532–545, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[2] R. Basri and S. Ullman. The alignment of objects with smooth surfaces. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 482–488, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[3] E. B. Baum, J. Moody, and F. Wilczek. Internal representations for associative memory. *Biological Cybernetics*, 59:217–228, 1988.

[4] R. C. Bolles and R. A. Cain. Recognizing and locating partially visible objects: the local feature focus method. *International Journal of Robotics Research*, 1:57–82, 1982.

[5] R. Brooks. Intelligence without representation. In *Proc. of Foundations of AI Workshop*, MIT, June 1987.

[6] J. J. Clark and N. Ferrier. Modal control of an attentive visual system. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 514–523, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

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

[8] S. Edelman and T. Poggio. Integrating visual cues for object segmentation and recognition. *Optic News*, 15:8–15, May 1989.

[9] S. Edelman and S. Ullman. Reading cursive script by computer. In *Proc. 42nd SPSE Conference*, pages 179–182, Boston, MA, May 1989.

[10] S. E. Fahlman. *NETL: a system for representing and using real-world knowledge.* MIT Press, Cambridge, MA, 1979.

[11] Z. Gigus, J. Canny, and R. Seidel. Efficiently computing and representing aspect graphs of polyhedral objects. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 30–39, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[12] D. Heeger and G. Hager. Egomotion and the stabilized world. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 435–440, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[13] D. P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proceedings of the 1st International Conference on Computer Vision*, pages 102–111, London, England, June 1987. IEEE, Washington, DC.

In the domain of visual motion study, Thompson and Kearney [34] have argued recently in favor of "inexact" or qualitative vision. Their definition of a qualitative representation is as follows:

> If an attribute can take only a small number of values we say that a representation is qualitative. Qualitative representations define a set of equivalence classes on the quantitative scale. We will assume qualitative values correspond to disjoint, continuous intervals on the quantitative scale.

In other words, qualitative vision is equated with low-resolution vision, leaving the problem of (approximate) inverse optics in its place. Starting from an observation made by Thompson and Kearney, namely, that a representation should contain only as much information as is necessary for subsequent processing, we have attempted to revise the inverse optics paradigm. However, rather than abolishing the notion of inverse optics (along with the notion of representing the visual world in the usual sense; see [5]), we propose to confine it to the low-level processing responsible for the formation of integrated base representations.

3.1 Research directions

Exploration of the approach to visual knowledge representation that we have sketched above could proceed in four main directions. Development of robust algorithms for low-level vision, whose purpose is to compute the base representation, or the collection of feature maps, is perhaps the most important open problem. Some recent advances in this direction include the work on perceptual grouping ([18], [15] bottom-up detection of salient structures [29], integration of diverse visual cues, and detection and labeling of discontinuities in the resulting map ([24], [23]).

The second class of problems, which may be described as middle vision, is centered around further investigation of visual routines [35] that could be used to link low-level features. In particular, a suitable control structure has to be developed to guide the application of the routines in accordance with the host system's goals and subject to environmental pressure and constraints. This issue is related to that of active vision [21]: strategy employed by a system engaged in active exploration of its environment is likely to differ from that of a passive observer.

The third domain of exploration, traditionally referred to as high-level vision, includes improving recently developed recognition methods ([18], [13], [33]) and testing their limitations. In particular, approaches based on exact geometrical representation (such as the various alignment methods) may be subsumed by a general, qualitative scheme, which views recognition as a three-stage process [8]. In the first stage of this process, the system selects from its input a region of interest that would be subjected to further processing. Such a region may be defined by motion boundaries, structural saliency [29] and other information directly computable from the base representation. Next, an indexing stage could be introduced into the recognition process to narrow down the set of object models that would participate in a subsequent matching, or verification by alignment. Finally, instead of being precisely aligned with the image, an object model could be roughly rotated so that the image and the model are seen from the same general aspect. This coarse alignment would be followed by a qualitative comparison of the visible features of the kind discussed above. Alternatively, new verification methods could be devised that rely on processes other than the transformation of object models. All of the above suggestions amount to a reassessment of the issue of object representation, without which, we believe, little progress is to be expected in extending state of the art recognition techniques to real-world images.

Finally, work on knowledge acquisition and representation must be brought to bear on the issue of visual representation, if the latter is myto be considered in the context of general intelligent behavior. With the outline of section 2.2 in view, some topics that appear particularly relevant are semantic nets [10], similarity-based reasoning [30], efficient associative memory ([22], [3]) and decision trees [25].

- **Grounding in robust low-level representations.** The representation of objects and scenes should be based on maps of visual space that integrate different cues such as intensity, texture, stereo, and motion ([23], cf. the notion of base representations [35]).
- **Sparseness.** Objects should characterized by a small number of tags or features that come from a very large repertoire. A convenient analogy is description by short sentences in a natural language (cf. [26]).
- **Similar treatment objects and of static and dynamic scenes.** By further analogy, situations may be described using the same simple language. Dynamic situations (happenings and actions) may be distinguished by tags that have verb-like connotations.
- **Localized features.** The features should be two-dimensional (that is, detectable in 2D images, without depth recovery) and spatially localized (cf. [4]). The localization could be coarse and may employ relations computed by universal visual routines [35].
- **Shallow hierarchy.** Although nontrivial objects, themselves composed of several features, could in turn serve as features in a more complicated representation, the hierarchy should be shallow, e.g. *cat* to *ears* to low-level place tokens marking intensity edges, texture etc.
- **Heterogeneity and flexibility.** The scheme should not prescribe the precise format of or the amount of information present in the representations of various objects. Rather, it should be demand-driven and augment existing representations when the need arises for finer distinctions among some of the objects.
- **Robustness through representing similarity.** A system confronted with a novel object should come up with a plausible first guess (the "nearest neighbor" of the unknown object). This could be facilitated by maintaining an explicit representation of similarity relations among the feature sets that define objects.
- **Incremental learnability.** The scheme should allow the system to bootstrap itself by learning representations for novel objects, given only low-level routines (the apparatus for computing the base representations) and real-world input. Furthermore, learning should not require an oracle capable of specifying the desired representation for a given input. Self-organizing representations satisfying this constraint have been demonstrated for some low-level visual features ([17], [28]).

3 Discussion

The trend of reassessing the commitment to inverse optics can be discerned in the proceedings of a recent conference on computer vision. Some examples of this are mentioned below.

- The stress on obtaining qualitative rather than quantitative solutions to problems in vision (qualitative depth from stereo [38], qualitative description of motion for obstacle avoidance [21], see also [37]);
- Development of object representations that are 2D (or $2\frac{1}{2}$D) and viewer-centered, rather than 3D and object-centered (aspect graphs [11], [14], [32], [31]; a novel curvature-based representation [2]);
- Increasing interest in active vision (using egomotion information [12], attentive gaze control [6]);
- Efforts to integrate multiple sensors/cues (a general strategy for integration [14], integration of focus, vergence and stereo [1]).

2 A possible alternative

The main shortcoming of the invariant features and the structural decomposition approaches appears to be brittleness. It is difficult to find a set of features each of which is both simple enough to be computed reliably by a low-level process and sufficiently invariant over all possible views of an object. Thus, pure feature-based methods tend to fail when presented with scenes in which many similar objects may appear in widely varying poses. On the other hand, structural decomposition has difficulties with achieving stable descriptions of many real-world objects (such as a shoe or a rabbit, cf. [36]). In addition, this approach by definition must recover the three-dimensional structure of the image, to compare it with structurally specified object models, that is, it involves inverse optics.

Alignment of pictorial descriptions and similar schemes ([18], [13], [33] [16]) use 3D or $2\frac{1}{2}$D models but do not require inverse optics. Instead, alignment calls for computing the (2D) appearance of each model from the viewpoint constrained by key feature location in the image and for comparing that with the image. These methods appear to work under a variety of conditions. However, when applied to articulated or flexible objects, alignment had to be combined with some kind of structural decomposition. For example, it has been proposed to compute separate aligning transformations for different regions of articulated objects [36], and an alignment-based system for handwriting recognition had to decompose letters into smaller parts (strokes) to compensate for the variability of letter structure [9].

We interpret these compromises as arguments in favor of an approach to the representation problem that is more qualitative in that it is restricted neither to vectors of feature values, nor to syntactic constructs in a formal language, nor to pictorial descriptions. This approach combines elements of the three methods distinguished by Ullman and may be termed linguistic because it employs what looks like natural language descriptions of objects. It may also be called open-ended, because it does not prescribe a standard formula for representing different things, but rather leaves that to be determined by the interaction of the system with its environment.

2.1 Open-ended representations

Is there a good way to describe an object (say, a cat) to a person who never saw one, but who possesses much the same vocabulary as you do? There might be, if it is possible to summarize the differences between a cat and other, familiar, objects (say, dogs) that are similar to it in a few words. An example could be, "a cat looks like a dachshund, except that it has sort of pointed ears, its legs are straighter and longer," etc. This approach presupposes prior knowledge of somewhat complicated concepts such as an ear or a leg, but does not normally require exact decomposition of the object into parts corresponding to these concepts (there is no need to define the precise boundary between the head and the ear to see that (i) there is something like an ear on top of the head and (ii) it is pointed). It uses feature values (*pointed* ears) to distinguish between similar objects, but defines the features locally (pointed *ears*) rather than attempting to find out the range of the ratio of area to perimeter that is characteristic of an image of a cat. In some cases, this approach must resort to pictorial descriptions ("pointed like this: $\bigwedge$" — see Ullman [36], p.18, for a discussion of the need of detailed shape description in recognition). These descriptions, however, would also be local and might often be aimed at capturing ill-localized properties that are difficult to describe verbally, such as texture.

2.2 Synthesis

We propose to integrate the three approaches to representation (feature spaces, and structural and pictorial descriptions) by adopting the following as a basic set of requirements from a representation scheme:

we draw from this realization is that detailed geometrical representation of objects and scenes may not be necessary for recognition.

We illustrate this point by the simplified example in Figure 1. An object fragment such as the one that appears there may remind our system of a cat. In fact, if it does not remind it of much more than that, then it probably *is* a usable representation of a cat. Several such representations, each useful under different circumstances, may be kept by the system.

In real-world scenes, as opposed to Figure 1, abundant visual information is available. Most of this information is irrelevant for creating a "first-order" interpretation of the scene. Even if the system can select for further processing only those parts of the image that seem a priori most important, too much information may be still there that is both difficult to extract and unnecessary for recognition. For example, the knowledge (supplied by the stereo module) which of the two $\bigwedge$-like features detected in the image is farther from the cameras does not seem to increase or diminish the probability that the two $\bigwedge$s are in fact a cat's ears. Qualitative information, such as an assertion that the two $\bigwedge$s are several centimeters rather than several meters apart in depth, could be easier to compute and no less useful for recognition.

Difficulties with computing exact representations, and arguments similar to the one outlined above, have led in the past, on one hand, to feature-based approaches to recognition in which objects were represented by lists of values of several (often, many) perceptual variables [7], and, on the other hand, to top-down strategies that invoked general knowledge about the scene to assist object recognition [20]. However, Rosenfeld [27] pointed out that a sophisticated visual system should be able to recall by itself from memory the frame within which an image is to be interpreted (or to construct such a frame if a novel coincidence of an object persistently appearing in a certain context is noticed). Furthermore, recognition based on the invariant feature approach appeared to be infeasible, even when top-down constraints were invoked. Partly because of that, the paradigm of computer vision that emerged during the last decade calls for achieving representations through inverse optics. Its central tenet is that objects are best recognized only after their visible surfaces have been described geometrically, e.g. as sets of points in some (preferably, object-centered) coordinate system [19]. In the sequel, we argue that vision does not necessarily require *exact* representations. Accordingly, we see the goal of inverse optics in the derivation of the physical properties of 3D surfaces at a qualitative, not very precise level, as suggested by human vision.

A recent review of object recognition methodologies by Ullman [36] distinguishes three main classes of theories: (1) invariant properties methods, (2) parts decomposition methods, and (3) alignment methods.

> Theories in the first class assume that that certain simple properties remain invariant under the transformations that an object is allowed to make [such as the transformation arising from viewing a 3D object from different points in space]. This approach leads to the notion of invariances, feature spaces, clustering, and separation techniques. The second class relies on the decomposition of objects into parts. This leads into the notions of symbolic structural descriptions, feature hierarchies and syntactic pattern recognition. By and large, the first of these general approaches was the dominant one in the earlier days of pattern recognition and the second approach has become more popular in recent years.

Ullman argued that both these approaches are insufficient for the general problem of shape-based visual recognition. Instead, he proposed a method, called alignment, that involves more or less exact pictorial representations of object models. In Ullman's approach, a small number of corresponding key features in the image and in a model are used to compute the transformation that would bring the two into alignment. After normalization by alignment, the two shapes are compared to decide whether the image is actually an instance of the model.

Representations in high-level vision: reassessing the inverse optics paradigm

Shimon Edelman and Tomaso Poggio

1 Inverse optics

What performance is it reasonable to require from an artificial vision system before it can be properly regarded as seeing? Paraphrasing Marr, one would expect such a system to know what is there by looking ([19], p.3). This appears to imply that (1) the system must have a way to represent what is there in some fashion and (2) this representation must be effectively computable (in fact, it should better be *efficiently* computable, given the system's resources).

Marr stressed that the choice of representation is important because it can greatly influence the complexity of any subsequent processing of the information that is being represented. Thus, if intensity edges are used by some processing stage in the system, they should be made explicit in the representation computed by an earlier stage. In higher-level visual tasks such as scene understanding the choice of an output representation becomes less obvious if not downright arbitrary. For example, it appears to depend on the end to which the scene representation is the means: an "analog", map-like representation of a room would be useful for navigation but superfluous for merely finding out whether a specified object is present in the room (cf. [34]).

In object recognition (if it is not part of a navigation strategy), the computed representation must serve two related purposes. The first is identification: the representation must contain enough information to allow the system to choose from a set of known objects one that is most similar to the attended object, or to declare with a certain degree of confidence that the latter has not been previously encountered. The second purpose is that of association: the representation must allow access to additional information associated with the recognized object. This could be visual (e.g. what does the opposite side of the object look like), or general (e.g. its weight).

The idea of association is to rely on past experience to gain access to more information than is actually present in the visual input. Realizing that efficient association is, after all, the main objective of vision helps to define the requirements of an intelligent system from its vision module. The main conclusion

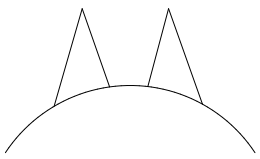

Figure 1: Two pointed things, one beside the other, on top of something rounded... Could be a cat's ears?