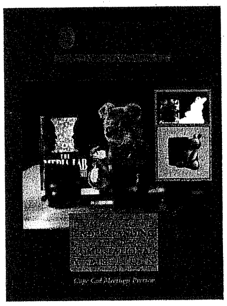
OPTICS NEWS

DEPARTMENTS



ABOUT THE COVER . . .

The cover illustration shows an inexpensive passive ranging device (a "range camera") that produces range estimates by measuring the point-by-point change in focus between two images that are identical except for their depth of field, a difference caused by using different aperture settings on the two cameras. The top inset shows a range image produced by this device, and the bottom inset shows a 3-D computer model produced from this range image. For more details, see the article, "Progress toward parallel vision machine" in this issue.

Integrating visual cues for object segmentation and recognition

By S. Edelman and T. Poggio

Looking around, we perceive the world as a collection of objects rather than as an amorphous aggregate of texture and color. The proficiency of our visual system in constructing a sensible interpretation of the surrounding scene tends to mask the enormous complexity of vision, considered as an information processing task. Stated concisely, this task is to interpret the image formed on the retinal mosaic in terms of physical objects and situations, arriving eventually at a description that in most cases can be put into words.

We recognize an object when it appears to us sufficiently similar to its representation in our memory. This common-sense notion of what it means to recognize something serves well to stress that there are two aspects to the recognition problem: representation and comparison. Representation has to do with the storage of visual information; comparison relates to its processing. The representation/processing duality is one of the foundations of Marr's approach to understanding vision.[1] According to Marr, solving a problem in computer vision involves combining an appropriate scheme for information representation with computational algorithms capable of deriving that representation, and, finally, finding computer architecture suitable for both tasks.

The problem of visual recognition

Theories of visual recognition may be roughly classified according to the kind of object representation they postulate. One possible dimension for such classification is the extent to which a representation is *pictorial*, that is, "looks" like the real-world object it stands for. (See Ref. 2 for a discussion of related issues.) So far, two approaches that may be placed at the extremes of the pictoriality scale have drawn most of the research efforts.

The first approach assumes that objects have certain invariant visual features that are commom to all their views and to all admissible variants within an object category, and differ between categories.[3] If n different properties are measured for each view of an object, the results may be considered as points in an n-dimensional real space R^n. The measurement can then be described by a mapping f: $R^2 \rightarrow R^n$. The image of the set of all possible views of an object and its admissible variations under the mapping f defines an invariant representation of that object. This representation could be useful for object recognition, provided that the subspaces of R^n that correspond to different objects are easily separable. A viewed object would then be recognized by determining the subspace of R^n within which its image under f falls. Note that the feature space approach is non-pictorial. It represents objects as lists of features that are disembodied in the sense that their posi-

Continued on page 10

S. Edelman and T. Poggio are with the Center for Biological Information Processing, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Mass.

Continued from page 8

tion (either in the image plane or in the real world) is not explicitly used in recognition.

The other extreme calls for maintaining a three-dimensional model, or at least a sufficiently complete set of possible views, for each object.[4-6] Typically, an object is recognized by a three-stage process. First, a set of key features is identified in the image. Many candidate sets are usually detected. Second, the pose of the object is computed from the relative positions of the features and the 3-D model is appropriately transformed and projected into the image plane, or an appropriate view is chosen. Finally, the degree of fit between the transformed model and the image is assessed. The second and the third stages must be repeated for every candidate feature set and for every possible model. Recognition occurs for those models that lead to a close enough fit.

A biologically plausible recognition scheme

The main factor limiting the theoretical and practical value of the feature-space approach is the difficulty of choosing features that are invariant with respect to object pose and at the same time allow fast and robust distinction among the possible objects. The pictorial approach is more successful in practice, but seems in several respects incomplete as a general theory of object recognition in the human visual system.

For one thing, it is clear that no such theory can be purely shape-based. It is hard to believe that we rely by and large on three-dimensional shape comparison to tell, say, a cat from a banana. On the contrary, it seems that people normally use every possible shortcut to arrive at a quick classification. Progressively more and more information may be used if the outcome is ambiguous, or if the task demands it.

Another characteristic of the pictorial approach that is biologically implausible is its assignment of an equal status to all known objects when trying to interpret a scene: every known object is potentially present as its presence is tested for.* Again, we do not seem even to consider comparing a yellow-green gently curing elongated blob to our model of a cat to find out what it is.

These arguments suggest the following three-stage scheme as a model of the recognition process in human vision (see also Ref. 7):

- **Selection:** segmenting the image into regions that are likely to correspond to single objects. In the banana recognition example, the peculiar color would probably suffice to distinguish it from the background, although other cues, such as intensity and depth, might be necessary if there is an entire bunch of bananas out there.
- **Indexing:** defining a small set of candidate objects that are likely to be present in the image. In the case of a banana, the color can serve for that purpose, too, but note also how its outline narrows down the set candidates to exclude all complex and articulated objects.
- **Verification:** testing each of the candidates to find the best match to the image. At this stage, the system can afford to perform complicated tests, since the number of candidate objects is small. For example, if there are grounds to believe that a banana-colored boomerang may be present in the image, the system could perform three-dimensional shape matching of a boomerang model, assisted by depth information obtained from the image.

So far, computer vision systems have not been overly successful in segmenting single static images of natural scenes.

■ *Segmentation*

The utility of an early segmentation of a scene into meaningful entities lies in the great reduction of complexity of scene interpretation. Each of the detected objects can in turn be subjected to separate recognition, by comparing it with object models stored in memory. Without prior segmentation, every possible combination of image primitives such as lines and blobs can in principle constitute an object and must be checked out.

At times, achieving early segmentation can be quite difficult. It is still disputed, for example, whether the human visual system segments a black-and-white photograph of a scene before recognizing objects present in it, or whether the objects only appear to be distinct because we become aware of their separate identities through recognition. So far, computer vision systems have not been overly successful in segmenting single static images of natural scenes.[8,9] It is reasonable to assume, however, that biological visual systems work reliably in the real world mainly because they never depend on any single visual cue. The amount of segmentation-related information that can be extracted even from a single gray-scale image may be greater than it originally appears. [10,11]

Ideally, the outcome of the segmentation stage is several simple closed curves, the interiors of which correspond to single objects (as yet unidentified) and the exterior of which constitutes the background. Sharp changes of brightness (brightness edges) and discontinuities in depth,

*In addition, if no segmentation information is available prior to recognition, every possible combination of the key features must be processed by the pose recovery and fit assessment mechanism.

motion, texture, and color can all arise from physical boundaries between objects and are therefore good candidates for image segmentation boundaries. Segmented regions may be labeled with locally constant or slowly changing information such as color or texture. Thus, the representation on which the subsequent processing operates looks much like a cartoon. It is important to realize that segmentation does not need to be always correct to be useful. In fact, it is impossible to develop a perfect, low-level segmentation scheme.

■ *Indexing*

Although one cannot hope to achieve an ideal segmentation in real-world situations, partial success is sufficient if the indexing process is robust. Assuming that most objects in the real world are redundantly specified by their local features, a good indexing mechanism would use such features to overcome changes in viewpoint and illumination, occlusion, and noise.

What kind of feature is good for indexing? Reliably detected lines provided by the integration of several low-level cues in the process of segmentation may suffice in many cases. Simple viewpoint-invariant combinations of primitive elements, such as two lines forming a corner, parallel lines, and symmetry are also likely to be useful. [4,12,13] Ideally, only 2-D information should be used for indexing, although it may sometimes be augmented by qualitative 3-D cues such as relative depth.

The power of feature-based indexing would be increased if a shallow hierarchy is introduced into the concept of "feature," that is, if relatively stable and coherent parts of objects (such as the tail of a horse or the pointed ears of a cat) are considered features just like the characteristic shape and color of a banana.

■ *Verification*

We conjecture that hierarchical indexing by a small number (two or three) features that are spatially localized in 2-D suffices to achieve useful interpretations of most everyday scenes. In general, however, further verification by task-dependent routines[14] or precise shape matching, possibly involving 3-D information [4-7,15-17] is required.

Parallel integration of cues for segmentation

Note that while the verification routines are situation-dependent and may involve serial attentional mechanisms, segmentation and indexing can in principle be carried out in parallel over large portions of the visual field.

In the last few years, the Vision Machine project has explored the idea that a major goal of the parallel integration stage of low-level vision modules is to compute a map of the discontinuities in the scene, somewhat similar to a cartoon or a line drawing. The description that we assumed is contour- and region-based: discontinuities in the physical properties of surfaces are sought together with properties of the enclosed regions. Later, recognition algorithms may make good use of the depth, color, and texture of a region, in addition to the outlines of the discontinuities in depth, color, and texture.

Poggio et al.[18,19] have argued that finding surface discontinuities in several early vision modules such as stereo, motion, texture, and color can be better achieved by integrating them with each other and with intensity edges. (In the following, edge detection will refer to the task of taking appropriate derivatives of the image data and possibly marking pixels that correspond to sharp changes of intensity.)

We have proposed[18,20] to expand the integration scheme to include the labeling of discontinuities according to their physical origin. To illustrate the importance of labeling via integration, consider an edge that is not detected by the stereo module. The existence of this edge in the color module will strongly suggest an albedo or specular discontinuity, whereas its absence in the color module will suggest a shadow or orientation discontinuity.

One may further argue that the flexibility and robustness of human vision in computing contour-like descriptions of a scene rely on the simultaneous use of several vision cues and their integration. It should be clear, however, that none of the vision cues is strictly necessary for a task such as recognition: humans manage to perform rather well on images devoid of color, motion texture, and depth cues. On the other hand, discontinuities of surface properties are convenient locations where the output of each of the vision modules may be coupled with the image data: surface discontinuities usually originate a sharp change in the image intensity, independent of the specific illumination, and dependent mainly upon shape and reflectance properties.

■ *The vision machine system*

In this section, we will review recent work on integrating visual modules detecting discontinuities[18,19,21], using the Vision Machine system.

. . . humans manage to perform rather well on images devoid of color, motion, texture, and depth cues.

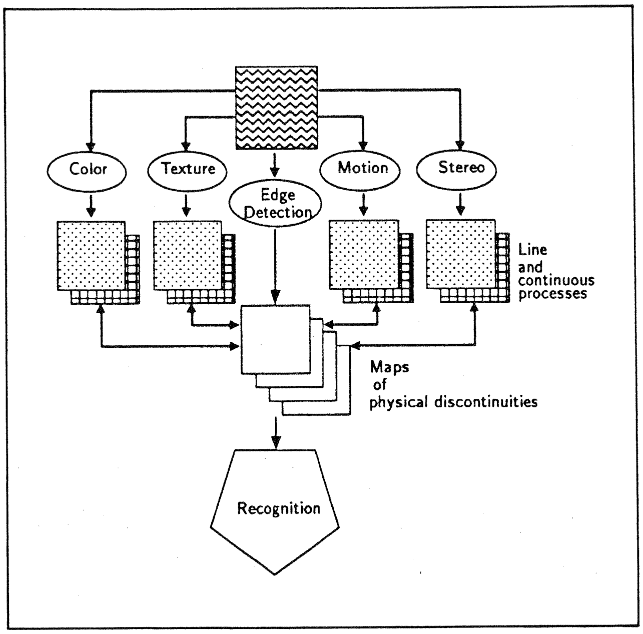

Color
Texture
Edge Detection
Motion
Stereo
Line and continuous processes
Maps of physical discontinuities
Recognition

FIGURE 1. Overall organization of the integration stage.

The overall organization of the system is shown in Fig. 1. The image(s) are processed through independent algorithms or modules corresponding to different visual cues, in parallel. Edges are extracted using Canny's edge detector. Stereo computes disparity from the left and right images. The motion module estimates an approximation to the optical flow from pairs of images in a time sequence. The texture module computes texture attributes, such as density and orientation of textons.[22] The color algorithm provides an estimate of the spectral albedo of the surfaces, independently of the *effective illumination*, that is, illumination gradients and shading effects, as suggested by Hurlbert and Poggio. [23]

Early vision modules—in this case, stereo, motion, color, and texture—operate on the image data. Their output is noisy, possibly sparse (for stereo and motion, for instance, depending on the specific version of the algorithm), sometimes not unique (some motion algorithms provide only one component of the optical flow), and does not explicitly represent discontinuities. Thus, the output of each module must be regularized to counteract noise, "fill in" the sparse data, and restore uniqueness.

The constraints that can be exploited to achieve this goal are smoothness (depth, motion, texture) or piecewise constancy (color) on the output of each module. Ideally, one would like to impose smoothness or piecewide constancy everywhere but at discontinuities. This is a stage of *approximation* and *restoration* of data, performed by using a Markov Random Field model. Simultaneously, discontinuities are found in each cue. Prior knowledge of the behavior of discontinuities exploits, for instance, the fact that they are continuous lines, not isolated points. Detection of discontinuities is aided by the information provided by brightness edges. Thus, each cue—disparity, optical flow, texture, and color—is coupled to the edges in brightness.

The full scheme involves finding the various types of physical discontinuities in the surfaces—depth discontinuities (extremal edges and blades), orientation discontinuities, specular edges, albedo edges (or marks), shadow edges—and coupling them with each other and back to the discontinuities in the visual cues, as illustrated in Fig. 1. So far, we have implemented only the coupling of brightness edges to each of the cues provided by the early algorithm. As we will discuss later, the technique we use to approximate, to simultaneously detect discontinuities, and to couple the different processes is based on MRF models. The output of the system is a set of labeled discontinuities of the surfaces around the viewer. In our implemented version of the system, we find discontinuities in disparity, motion, texture, and color. These discontinuities, taken together, represent a "cartoon" of the original scene that can be used for recognition and navigation (along with interpolated depth, motion texture, and color regions).

■ *Integration and segmentation with MRFs*

How can this be done? We have chosen to use the machinery of Markov Random Fields (MRFs), initially suggested for image processing by Geman and Geman.[24] Consider the prototypical problem of approximating a surface given sparse and noisy data (depth data), on a regular 2-D lattice of sites. We first define the prior probability of the class of surfaces in which we are interested. The probability of a certain depth at any given site in the lattice depends only upon neighboring sites (the Markov property). Because of the Clifford-Hammersley theorem, the prior probability has the Gibbs form:

$$P(f) = \frac{1}{Z} e^{-\frac{U(f)}{T}} \tag{1}$$

where Z is a normalization constant, T is called temperature, and $U(f) \;=\; \sigma_i U_i(f)$ is an energy function that can be computed as the sum of local contributions from each lattice site i. The energy at each lattice site $U_i(f)$ is, itself, a sum of the potentials, $U_c(f)$, of each site's cliques. A clique is either a single lattice site or a set of lattice sites such that any two sites belonging to it are neighbors of one another.[25,26] As a simple example, when the surfaces are expected to be smooth (like a membrane), the prior energy can be given in terms of:

$$U_i(f) = \sum_j (f_i - f_j)^2 \tag{2}$$

where j is a neighboring site to i (that is, i and j belong to the same clique).

If a model of the observation process is available (that is, a model of the noise), then one can write the conditional probability $P(g|f)$ of the sparse observation g for any given surface f. Bayes's theorem then allows one to write the posterior distribution:

$$P(f\,|g) = \frac{1}{Z}\,e^{-\frac{U(f|g)}{T}} \qquad (3)$$

In the example of Eq. 2, we have (for Gaussian noise):

$$U_i(f\,|g) = \sum_j\,(f_i - f_j)^2 + \alpha\gamma_i\,(f_i - g_i)^2 \qquad (4)$$

where $\gamma_i = 1$ only where data are available, and otherwise $\gamma_i = 0$. More complicated cases can be handled in a similar manner.[25]

The maximum of the posterior distribution or other related estimates cannot be computed analytically, but sample distributions with the probability distribution of Eq. 3 can be obtained by means of Monte Carlo techniques such as the Metropolis algorithm.[27] These algorithms sample the space of possible surfaces according to the probability distribution $P(f|g)$ that is determined by the prior knowledge of the allowed class of surfaces, the model of noise, and the observed data.

In our implementation, a highly parallel computer generates a sequence of surfaces from which, for instance, the surface corresponding to the maximum of $P(f|g)$ can be found. This corresponds to finding the global minimum of $U(f|g)$ (simulated annealing is one of the possible techniques). Other criteria can be used: Marroquin has shown that the average surface f under the posterior distribution is often a better estimate that can be obtained more efficiently simply by finding the average value of f at each lattice site.[28]

One of the main attractions of MRF models is that the prior probability distribution can be made to embed more sophisticated assumptions about the world. Geman and Geman[24] introduced the idea of another process—the line process—located on the dual lattice and explicitlh representing the presence or absence of discontinuities that break the smoothness assumption (Eq. 2). It is, in fact, possible to extend the energy function of Eq. 4 to accommodate the interaction of more processes and of their discontinuities. In particular, we have extended the energy function to couple several of the early vision modules (depth, motion, texture, and color) to sharp changes of brightness in the image.[25]

This is a central point in our integration scheme: here we assume that changes of brightness guide the computation of discontinuities in the physical properties of the surface, thereby coupling surface depth, surface orientation, motion, texture, and color each to the image brightness data and to each other. The reason for the primary role of the gradient of brightness, as conjectured here, is that changes in surface properties usually produce large brightness gradients in the image.

As already mentioned, we have been using the MRF machinery with appropriate prior energies to integrate edge brightness data with stereo, motion, color, and texture information on the MIT Vision Machine System described earlier. Figure 2 shows some results. The union of the discontinuities in depth, motion, and texture for the scene gives a "cartoon" of the original scene. Notice that this "cartoon" represents discontinuities in the physical properties of 3-D surfaces that are well defined, whereas brightness "discontinuities" are not. Our integration algorithm achieves a preliminary classification of the brightness edges in the image, in terms of their physical origin.

A more complete classification may be achieved by implementing the full scheme of Fig. 1. The lattices at the top classify the different types of discontinuities in the scene: depth discontinuities, orientation discontinuities, albedo edges, specular edges, and shadow edges.[20] The set of such discontinuities in the various physical processes seems to represent a good set of data for later recognition. In some preliminary experiments, we have successfully used a parallel, model-based recognition system[29] on the discontinuities (stereo and motion) provided by our MRF scheme.

Our present implementation represents a subset of the possible interactions shown in Fig. 1, itself only a simplified version of the organization of the likely integration process. As described elsewhere,[21,25] the system will be im-

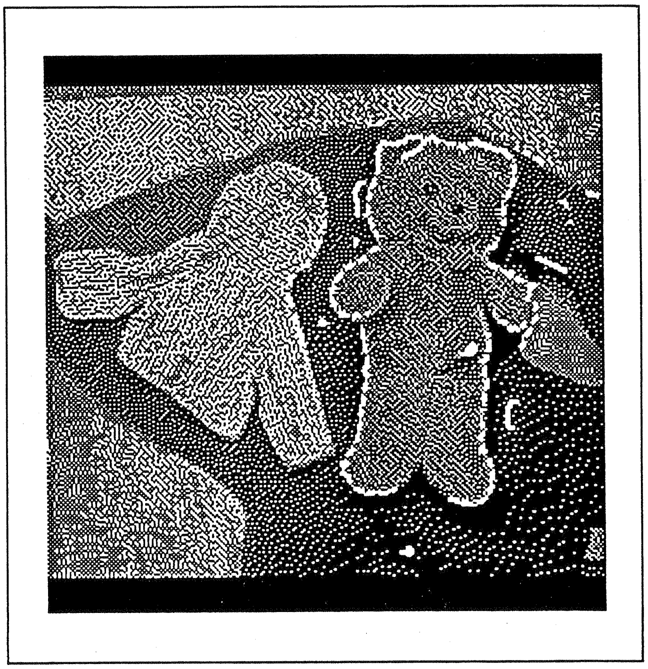


FIGURE 2. A cartoon-like representation of two objects segmented from the background using depth, motion and texture cues.

Real-world objects never present their entire surface to an observer at the same time.

proved in an incremental fashion, including pathways not shown in Fig. 1, such as feedbacks from the results of integration into the matching stage of the stereo and motion algorithms.

The highly parallel algorithms we have outlined map quite naturally onto an architecture such as the Connection Machine, which consists of 64 K simple 1-bit processors with local and global connection capabilities. The same algorithms also map onto VLSI architectures of fully analog elements (we have successfully experimented with a version of Eqs. 5 and 6, in which l is a continuous variable), mixed analog and digital components, and purely digital processors (similar to a much simplified and specialized Connection Machine).

The next step: parallel indexing by localized features

Real-world objects never present their entire surface to an observer at the same time. Thus, from any given vantage point, only a part of an object's feature set is visible. By rotating the object in 3-D, the observer can make some of the visible features become occluded and others appear. Equivalently, the same results may be achieved by changing the vantage point of the observer with respect to the object. Vantage point is fully specified by two parameters, corresponding to the latitude and the longitude of the eye positioned on an imaginary sphere centered at the object. The remaining degree of freedom in this system corresponds to a rotation of the eye around the viewing direction. Such rotation does not affect the apparent shape of the object and is of no concern to us here.

The viewing sphere may be naturally divided into regions or aspects over each of which the set of visible features is constant.[30] Boundaries of these regions are defined by visual events: occlusion or appearance of features. A list of all aspects of an object, along with feature visibility information for each aspect, constitutes a compact feature-based description of the object. Let the indexing be performed by counting matching features for each model and carrying out a "winner take all" (WTA) operation. The result will then be relatively insensitive to noise and partial occlusion, because of the locality of the features and because of the dampening action of the WTA step.

The aspect-specific feature-based representation has been shown to work for recognizing polyhedra.[31] Can this approach be extended to real-world vision without losing its inherent insensitivity to noise and occlusion? One possibility involves using features that are localized in the two dimensions of the image plane. In this scheme, called CLF for conjunction of localized features, a possible presence of an object is signaled if most of its characteristic features are detected at their proper positions with respect to one another. The basic tenets of the biologically motivated CLF scheme are as follows:

- The visual system uses converging data from feature detection units to capture and encode significant events in terms of conjunction of their characteristic features.[32]
- Representation of objects that is invariant with respect to certain transformations (such as the change of viewpoint) may be achieved by explicitly tying together representations of specific appearances of the objects, through an automatic learning process.[33]
- The number of units needed for such explicit representation may be significantly reduced through preprocessing.

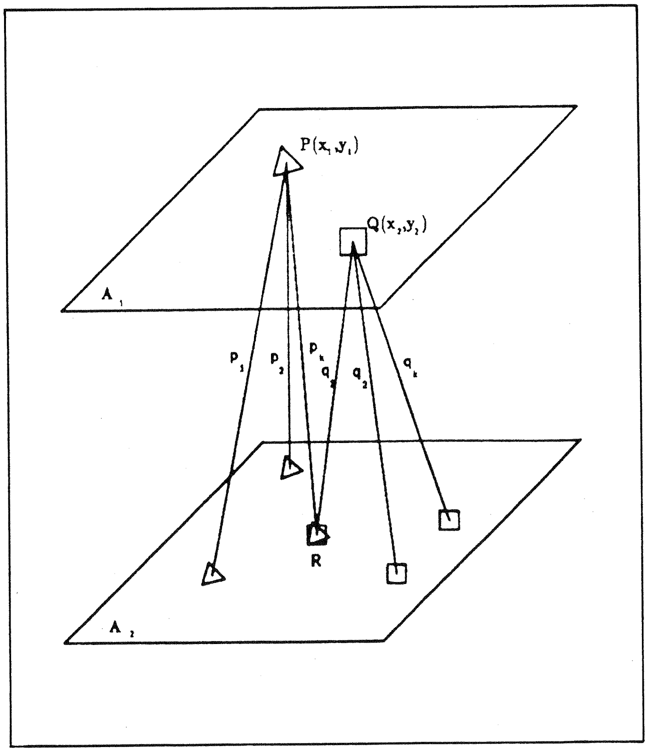


FIGURE 3. The unit R in A_2 represents the conjunction $P(x_1,y_1)$ & $Q(x_2,y_2)$. It is likely to arise by chance, if the number k of projections emanating from each of the units in A_1 is right. An optimal value for k can be computed, based on a trade-off between low cross-talk among representations and high probability of any given pattern in A_1 having a representation in A_2.

For example, foveation largely obviates the need for translation-invariant representation, while certain properties of the retinocortical mapping may partially counter the influence of size variation.[34] Figure-ground separation (e.g., by the method described in the previous section) is also important.

■ Under real-world, real-time conditions, the information necessary for associating different views of an object would be provided by the pattern of the optic flow due to the observer's movement with respect to the object.[35] Furthermore, the analysis of this flow[36-38] could endow the representation with metric structure, that is, representations of similar aspects of objects would be made to reside close to each other (closeness here is defined in terms of tight association rather than physical proximity).

CLF is easily implemented in parallel, by assigning processors to feature detectors and representation units and providing for the necessary connections (see Fig. 3). We believe that the values of connection strength needed to encode the conjunction mapping and its metric structure can be computed automatically. Encouragingly, conjunction Boolean formulae are shown to be effectively learnable from examples.[39] Learnability of CLF, as well as its computational aspects, are currently under investigation.

The human visual system is capable of forming split-second interpretations of complex scenes that may include any of thousands of possible objects, arranged in largely unpredictable configurations. To achieve comparable performance, machine vision systems may have to integrate several visual cues. Once computed, the integrated representation constitutes a natural solution to the old problem of object segmentation. It may also serve as a basis for development of simple, parallel schemes for indexing or object classification.

Acknowledgments

This paper describes research done within the Center for Biological Information Processing in the MIT Department of Brain and Cognitive Sciences and at the MIT Artificial Intelligence Laboratory. The Center's research is sponsored by grants from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Alfred P. Sloan Foundation; and by the National Science Foundation. Support for the A.I. Laboratory's artificial intelligence reseach is provided by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76-85—C—0010, and in part by ONR contract N00014-85—K—0124. Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at the Massachusetts Institute of Technology, Whitaker College, and by the Artificial Intelligence Center of Hughes Aircraft Corp. (S1-801534-2). Shimon Edelman is supported by a Chaim Weizmann Postdoctoral Fellowship and by a grant from the National Science Foundation.

REFERENCES

1. D. Marr. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman and Company, San Francisco, 1982.
2. D. Perkins. Pictures and the real thing. In P. Kolers, M. Wrolstad, and H. Bouma, editors, Processing of visible language 2, 259–278. Plenum Press, New York, 1980.
3. R. Duda and P. Hart. Pattern classification and scene analysis. Wiley, New York, 1973.
4. D.G. Lowe. Perceptual Organization and Visual Recognition. Kluwer, Boston, 1986.
5. D. Thompson and J. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In Proceedings of IEEE Conference on Robotics and Automation, 208–220, Raleigh, N.C., 1987.
6. D. Huttenlocher and S. Ullman. Object recognition using alignment. Proceedings of the 1st International Conference on Computer Vision, 102–111, London, England, June 1987. IEEE, Washington, D.C.
7. S. Ulmann. An approach to object recognition: Aligning pictorial descriptions. A.I. Memo No. 931, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Dec. 1986.
8. T. Binford. Survey of model-based image analysis systems. International Journal of Robotics Research, 1:18–64, 1982.
9. R. Chin and C. Dyer. Model-based recognition in robot vision. ACM Comp. Surv., **18**:67–108, 1986.
10. A. Pentland. Shape information from shading: a theory about human perception. Proceedings of the 2nd International Conference on Computer Vision, 404–413, Tarpon Springs, Fla., 1988. IEEE, Washington, D.C.
11. A. Sha'ashua and S. Ullman. Structural saliency: the detection of globally salient structures using a locally connected network. Proceedings of the 2nd International Conference on Computer Vision, 321–327, Tarpon Springs, Fla., 1988. IEEE, Washington, D.C.
12. T. Kanade and J. Kender. Mapping image properties into image constraints: skewed symmetry, affine-transformable patterns and the shape from texture paradigm. In J. Beck, B. Hope, and A. Rosenfeld, eds., Human and machine vision, 237–258, New York, 1983. Academic Press.
13. A.P. Witkin and J.M. Tenenbaum. On perceptual organization. From Pixels to Predicates, 149–169. Ablex, Norwood, N.J., 1986.
14. S. Ullman. Visual routines. Cognition, **18**, 1984.
15. R.C. Bolles, P. Horaud, and M. Hannah. 3DPO: A three-dimensional part orientation system. Proceedings IJCAI, 1116–1120, 1983.
16. N. Ayache and O.D. Faugeras. Hyper: a new approach for the recognition and positioning of two-dimensional objects. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(1):44–45, 1986.
17. L.W. Tucker, C.R. Feynman, and D.M. Fritzsche. Object recognition using the Connection Machine. Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, 871–878, 1988.
18. T. Poggio. Integrating vision modules with coupled MRFs. Working Paper No. 285, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1985.
19. T. Poggio, E.B. Gamble, and J.J. Little. Parallel integration of vision modules. Science, **242**:436–440, 1988.
20. E. Gamble, D. Weinshall, D. Geiger, and T. Poggio. Labeling edges and the integration of low-level visual modules, 1989. in preparation.
21. T. Poggio. J. Little, E. Gamble. W. Gillett, D. Geiger, D. Weinshall, M. Villalba, N. Larson, T. Cass, H. Bulthoff, M. Drumheller, P. Oppenheimer, W. Yang, and A. Hurlbert. The MIT Vision Machine. Proceedings Image Understanding Workshop, Cambridge, Mass., April 1988, Kaufmann, San Mateo, Calif.
22. H.L. Voorhees. Finding texture boundaries in images. Master's thesis, Massachusetts Institute of Technology, 1987.

23. A. Hurlbert and T. Poggio. Synthetizing a color algorithm from examples. Science, **239**:482–485, 1988.
24. S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6:721–741, 1984.
25. E.B. Gamble and T. Poggio. Visual integration and detection of discontinuities: The key role of intensity edges. A.I. Memo No. 970, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Oct. 1987.
26. J.L. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. In L. Baumann, ed., Proceedings Image Understanding Workshop, 293–309, McLean, Va., Aug. 1985. Scientific Applications International Corp.
27. N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. J. Phys. Chemistry, **21**:1087, 1953.
28. J.L. Marroquin. Probabilistic Solution of Inverse Problems. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Mass., 1985.
29. T.A. Cass. Robust parallel computation of 2D model-based recognition. Proceedings Image Understanding Workshop, Boston, Mass., April Kaufman, San Mateo, Calif., 1988.
30. J. Koenderink and A. van Doorn. The internal representation of solid shape with respect to vision. Biological Cybernetics, **32**:211–217, 1979.
31. T. Breuel, 1989 (in preparation).
32. H. Barlow. Cerebral cortex as model builder. In D. Rose and V. Dobson, eds., Models of the Visual Cortex, 37–46. Wiley, New York, 1985.
33. W. Pitts and W. McCulloch. How we know universals: the perception of auditory and visual forms. Embodiments of Mind, 46–66. MIT Press, Cambridge, Mass., 1965.
34. E. Schwartz. Anatomical and physiological correlates of visual computation from striate to infero-temporal cortex. IEEE Trans. on Sys. Man Cybern, SMC-14:257–271, 1984.
35. J. Gibson. The Ecological Approach to Visual Perception. Houghton Mifflin, Boston, Mass., 1979.
36. J. Koenderink and A. van Doorn. Local structure of movement parallax of the plane. Journal of the Optical Society of America, **66**: 717–723, 1976.
37. J. Koenderink and A. van Doorn. Optic flow. Vision Research, **26**: 161–180, 1986.
38. H. Loguet-Higgins and K. Prazdny. The interpretation of a moving retinal image. Proceedings of the Royal Society of London B, **208**: 385–397, 1980.
39. L. Valiant. A theory of the learnable. Communications of the ACM, **27**:1134–1142, 1984.