VISION REANIMATED

Shimon Edelman[⋆]

ABSTRACT

Computer vision systems are, on most counts, poor performers, when compared to their biological counterparts. The reason for this may be that computer vision is handicapped by an unreasonable assumption regarding what it means to see, which became prevalent as the notions of intrinsic images and of representation by reconstruction took over the field in the late 1970's. Learning from biological vision may help us to overcome this handicap.

1. INTRODUCTION

Although the field of computer vision is only a few decades old, its current agenda can be traced back a long way: "What does it mean to see? The plain man's answer (and Aristotle's, too) would be, to know what is where by looking," (Marr, 1982, p.3). Attempts to specify, in unambiguous computational terms, just *how* should one go about finding out what is where have met so far with extreme difficulties, provoking in the process considerable controversy. Although building a general-purpose computer vision system proved to be much more difficult than making a computer play championship-level chess, there is no consensus as to why this should be so. In fact, some people believe that the same brute-force approach that led to successes in computer chess or in automatic theorem-proving will soon succeed also in vision.

I contend that throwing sheer computer power at the problem of vision is somewhat like banging against a sliding door: given enough force, it may succeed, but there is a cheaper way to get through. In the rest of this paper, I will try to substantiate this claim, first by reviewing where the banging has got us so far, then by pointing to what seems to me a handle that may help slide the door open — lessons from biological vision.

2. COMPUTER VISION

The predisposition to try every trick in the trade on the problem of visual recognition is not peculiar to the present time: a similar attitude has been perceived in the field by Marr in the beginning of 1970's. Whereas neither this attitude, nor the considerable increase in the available computer power that has occurred since then, has led to the development of a general-purpose object recognition system, the more successful systems all seem to use the same kinds of tricks — ones that mimic biological vision, and run counter to the hierarchical 3D reconstruction-oriented framework for computational vision, stated in (Marr, 1982).

[⋆] Unpublished manuscript, 1995. The author's present address: Department of Psychology, 232 Uris Hall, Cornell University, Ithaca, NY 14853-7601, USA, http://kybele.psych.cornell.edu/ edelman.

Slow progress in the face of the brute-force onslaught seems to indicate a need for a revision of the basic assumptions of the current paradigm. This paradigm has been established when Marr, Poggio, and others pointed out the lack of principles behind the "bag of tricks" approach that prevailed at that time (Marr, 1976; Marr and Poggio, 1977). Their reaction took the form of an attempt to organize the thoughts before sitting down to program. This resulted in a call to (1) define the desirable features of any possible solution, and (2) formulate the problem for which a solution is sought. The rest of this section discusses the two parts of Marr's research program that were derived from these two observations.

2.1. The spirit of '76

Two of the most desirable principles of any information processing system, stated in (Marr, 1976), are graceful degradation and least commitment. According to the first principle, a recognition system should be tolerant to inputs that are noisy, contain objects that are only partially visible, or objects that are similar, rather than identical, to the familiar ones. The second principle concerns the mode of operation of the system, rather than its input/output specification. Effectively, it states that good performance (and, in particular, graceful degradation) is not to be expected of a system that commits itself early to a subset of possible solutions by discarding information that cannot be recovered later.

2.2. Ungraceful degradation

The principles of graceful degradation and least commitment fell into oblivion almost as soon as they were formulated. By the end of 1970's, Marr was advocating a hierarchical approach to vision that started with edge detection and was supposed to culminate in a complete 3D reconstruction of the visual world.

Edge detection. This operation, which is the first stage in an overwhelming majority of contemporary computer vision systems, constitutes a typical example of blatant violation of the principle of least commitment. The decision whether or not to label a certain location in the image as an edge is a commitment *par excellence*: the presence (or the absence) of an abrupt change of intensity there constitutes a small proportion of the information present in the original distribution of intensities.

In some cases, the exact distribution of intensity in an image can be recovered from its edge map. For the zero-crossings of the Laplacian of Gaussian operator (Marr and Hildreth, 1980), the conditions allowing such recovery have been stated in Logan's theorem. Nevertheless, even if such information is present in the edge map, it is usually promptly discarded, when the individual edge elements are "linked" to form a higher-level representation of longer contours, or interpreted symbolically, as in Mirage, a detailed derivative of Marr's theory of edge detection (Watt and Morgan, 1985).

Intrinsic images. Just as edge detection is the common initial stage in computer vision systems, a 3D description of a scene is the common ultimate representation, which, in recognition tasks, is compared to a library of stored representations of the same kind.

The viewer-centered $2\frac{1}{2}$D and the object-centered 3D representations in (Marr, 1982) are closely related to the notion of an *intrinsic image* — a description of the true properties of the scene (geometry and layout of the object surfaces, and their color and texture), from which the influence of extraneous factors such as illumination and pose have been removed (Barrow and Tenenbaum, 1978; Barrow and Tenenbaum, 1981; Tenenbaum et al., 1981; Witkin, 1981).

The reconstructionist approach to visual representation, popularized by Marr (1982), runs counter to the principle of least commitment on all levels. At the lower levels, this is expressed in the choice of edge detection and symbolic primal sketch over continued use of intensity information. At the higher levels, problems arise from the insistence to describe visible surfaces as fully as possible, before moving on to classification or recognition. In fact, the very decision that the ultimate representation of the visual world should be its geometric replica, constitutes a meta-level commitment, which, once made, is very difficult to undo.

The insistence on reconstructing the world would not have been so detrimental to the progress in computer vision, were the task feasible. As things stand now, there are serious doubts regarding the feasibility of the reconstructionist approach. These are echoed in a recent retrospective by two of the originators of the concept of intrinsic images (Barrow and Tenenbaum, 1993):

> *Ten years after the publication of* [Barrow and Tenenbaum, 1981], *many issues remain open. Perhaps the most direct issue is, simply, can the recovery process be made to work either for line drawings or for real world images? [p.75]*
>
> *. . . It may also be the case that we have placed too much emphasis on analytical recovery models and exact recovery . . . [p.77]*

3. LEARNING FROM BIOLOGY

The most encouraging thing about computer vision at present is the certainty of our knowledge that general-purpose vision is possible. The proof of that is the existence of a large variety of successful living visual systems. This observation begs the question: *what are we doing wrong that the biological visual systems do right?*

The answer to this question is not to be found in Marr's 1982 book, which inspired so much of the subsequent research in vision: at the time of its writing, "virtually nothing [was] known about the physiological and anatomical arrangements that mediate the construction of three-dimensional visual descriptions of the world" (Marr, 1982, p.326).[2] Thus, despite Marr's background in neurobiology and his recognition of the importance of learning from biological information processing systems, the foundations of his approach reflect the paucity of the information available at that time regarding both the performance of and the mechanisms behind object representation and recognition in biological systems.

[2] The only illustration pertaining to biological pattern recognition in (Marr, 1982) has to do with grouping and refers to Attneave's work, dating back to the 1950's.

3.1. PERFORMANCE OF THE RECOGNITION SUBSYSTEM IN PRIMATE VISION

Behavioral studies conducted in the past decade reveal that the performance of the primate visual system in recognition, while undoubtedly impressive, is limited in a number of telling respects. I chose to concentrate here on three classes of findings that support this claim: qualitative nature of visual perception, non-invariant performance in visual recognition, and selective retention of visual information.

Qualitative perception. The human visual system appears to be more sensitive to qualitative rather than quantitative or metric properties of visible surfaces. For example, when asked to carry out judgments based on their perception of surface depth from shading, subjects behave as if they employ a qualitative “shortcut” algorithm, yielding the relative order of points in depth, rather that their absolute depth values (Todd and Reichel, 1989). This seems to be the rule and not an exception in the perception of surface shape: although the recovery of 3D metric properties of surfaces is possible, given time and motivation (Koenderink et al., 1993), such recovery is not necessary either for recognition (Edelman and Bülthoff, 1992), or for grasping (Goodale et al., 1991; Kamon et al., 1994). Interestingly, calls for a more qualitative treatment of the problem of recovery of 3D surface properties have been sounded in the computer vision community at about the same time the psychophysical study of qualitative vision intensified (Aloimonos, 1990).

Non-invariant recognition. A prime reason for the attempts to recover the 3D structure of objects prior to recognizing them is the belief that the availability of the 3D information will somehow contribute toward invariance with respect to viewpoint. While invariance is certainly a worthy goal, it is not at all clear that the human visual system perceives it as such: recent studies of object recognition revealed that the human recognition performance is far from invariant in many cases (Jolicoeur and Humphrey, 1998). In particular, recognition rate deteriorates with misorientation between familiar and novel views of the stimuli (Tarr and Pinker, 1989; Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Humphrey and Khan, 1992). Furthermore, the ability of the subjects to compensate for changes in the viewing conditions such as orientation and illumination depends on similarity between the stimuli that have to be discriminated (Edelman, 1995a), and on their familiarity (Moses et al., 1996). These psychophysical findings have been accompanied by simulation studies that explored the possible computational basis for the non-invariant performance; see (Bülthoff and Edelman, 1992; Edelman, 1995a; Lando and Edelman, 1995) for details.

Selective retention. The list of apparent shortcomings of the human visual system would be incomplete without mentioning the inability of subjects to retain detailed information about the perceived scene across saccades (Pollatsek et al., 1984; Blackmore et al., 1995; Rensink et al., 1995; Grimes, 1995). In a number of different psychophysical paradigms, researchers find that subjects tend to represent and remember the scene in qualitative terms, and are more often than not oblivious to large-scale changes (such as the disappearance of a chair from a room scene, or a radical change in the color of fabric in a swimsuit ad), if these are introduced during the blackout period accompanying a saccade.

Interestingly, the subjects in all the experiments mentioned in this section do not notice the limitations of their visual systems, and neither do the observers in uncontrolled, everyday situations. The visual world seems to us immutable, complete, perceived in full quantitative detail, despite the shortcuts and the guesses taken by our vision in judging the layout of the surrounding surfaces. Similarly, the lack of invariance in object recognition goes unnoticed in everyday life, although under laboratory conditions this effect is easily obtained with the kinds of objects that are likely to be encountered elsewhere. Finally, subjects in selective retention experiments are unaware of the fact that something changes between fixations, and not only of the visual features that undergo change.

Should this lack of awareness of the limitations of one's own visual system be interpreted as grave self-deception? The evolutionary success of primates suggests a more benign interpretation: we are unaware of the limitations of vision because it is limited in inessential ways. The contradiction between the objective deficiency and the subjective perfection of vision arises as a by-product of the inflated expectations stemming from the reconstructionist theories. One expects, no doubt, that a perfect internal representation of the 3D world support an equally perfect performance in 3D shape perception. For this very reason, performance that consistently falls short of the reconstructionist notion of perfection is merely a hint that it is time to reconsider the representation-as-replica theory.

3.2. MECHANISMS SUPPORTING RECOGNITION IN PRIMATE VISION

Anatomically, in all the visual areas, as in the entire neocortex in general, information is processed by the same few kinds of cells, arranged in the same laminar/columnar structure (Gilbert, 1988). The uniformity of the cortex is not limited to its anatomy: functional studies reveal a limited repertoire of computational mechanisms, of which tuned receptive fields (RFs) are probably the most ubiquitous one. In neurophysiology, the RF of a cell is defined as the part of the visual field in which a stimulus must appear to elicit a response from the cell (Kuffler and Nicholls, 1976). Together with the specification of the preferred stimulus of the cell, this constitutes a useful characterization of its input-related function.[3]

The characteristics of the receptive fields of cortical cells and their interconnection patterns constrain the kind of information processing that can be supported by the cortex. Some of the more prominent relevant observations are outlined briefly below.

Broad tuning. At all stages of visual processing, cells respond preferentially to some patterns compared to others (for example, in the primary visual cortex there are cells that are tuned to the orientation of intensity gradients and to spatial frequency). This tuning is usually broad: the relevant property of the stimulus may have to change considerably before the response is reduced to the baseline (Bishop et al., 1973). Another manifestation of this phenomenon is the high degree of spatial overlap (in retinotopic terms) between the RFs of neighboring cells — this corresponds to a broad tuning in retinal space.

[3] For a complete characterization of the cell's function, its lateral links and projective field should also be specified.

Graded response. As the stimulus moves (in retinal space or in the space of the relevant feature, such as orientation), the response of the cell changes, as a rule, gradually, rather than abruptly (Shapley and Victor, 1986). Functionally, a graded, broadly tuned RF can be considered a transducer, which smoothly maps changes in some feature space into changes along the dimensions represented in the output of the cell (a single dimension, if the response is taken to be the firing rate of the cell; possibly several dimensions, if the temporal properties of the firing are taken into account).

Ensemble encoding. Because of the broad tuning and the graded response profile of individual RFs, the properties of the stimulus are best described by the population response, and not by the activity of any single RF in isolation. Population coding is a well-known concept in distributed information processing. In biological systems, in view of the properties of broad tuning and graded overlapping RFs, the population response confers the additional advantage of hyperacuity: the resolution (in retinal or feature space) supported by the ensemble response is likely to be far better than what can be derived from the responses of the individual RFs (Snippe and Koenderink, 1992; Weiss et al., 1993).

Selective invariance. Invariance to viewing conditions is a central prerequisite of any visual system. Paralleling the psychophysical findings, the degree to which the response of a cortical cell is invariant to extraneous variables affecting the appearance of a stimulus is usually quite limited. For example, a large majority of cells in the study of (Logothetis et al., 1994), which concentrated on the inferotemporal cortical area in monkeys, were found to be tuned to a contiguous subset of views of certain 3D objects; a very few cells responded invariantly to all views. Even for these cells, the invariance is stimulus-specific, because of their shape-space tuning (i.e., their selectivity for particular 3D shapes; cf. the interaction between invariance and familiarity mentioned in section 3.1.). Similar findings have been reported even for such basic properties as shift invariance (Ito et al., 1995).

Plasticity and learning. The stimulus-specific quasi-invariance of tuning properties of cortical cells is complemented by an ever-present capacity for modification of these properties in response to changing patterns of stimulation (Gilbert, 1994; Sagi and Tanne, 1994). A novel stimulus is not likely to be optimally processed; however, even short practice or mere exposure leads to a re-tuning of the system and results in improved performance. This mechanism may be operational also on a longer time scale: it has been hypothesized that the receptive fields at the initial stages of visual processing have evolved to match the statistics of the natural images (Field, 1994), so as to optimize various properties of the representations (e.g., the informativeness) evoked by natural stimuli.

3.3. AN INTERIM SUMMARY: THE WORLD AS ITS OWN REPRESENTATION

The above observations of the architecture and the function of the visual system raise serious doubts concerning the adequacy of the reconstructionist theory as a model of primate vision. As we have seen, psychophysical evidence suggests that subjects are not likely to be engaged in reconstructing the world internally; this is just as good, because, according to the neurobiological evidence, the visual system seems to be ill-suited for such an undertaking.

How, then, if not by virtue of reconstruction, does the human visual system come to represent the external world in all its apparent richness? One possibility has been discussed in (O'Regan, 1992):

> *I [...] suggest an alternative approach, in which the outside world is considered as a kind of external memory store, which can be accessed instantaneously by casting one's eyes (or one's attention) to some location. The feeling of the presence and extreme richness of the visual world is, under this view, a kind of illusion, created by the immediate availability of the information in this external store. [p.461]*

The upshot of this discussion is that the efforts of computer vision should be redirected from attempting to reconstruct the world internally to matching the unfaltering performance of human vision in everyday tasks such as shape discrimination and classification. It appears that human vision excels in these tasks despite its limited capacity for explicit recovery of 3D shape geometry. The next section discusses some possible solutions to shape-related problems, whose common feature is the reliance on approaches borrowed from biological vision.

4. PRACTICAL EXAMPLES

4.1. IMAGE REPRESENTATION

The information inherent in image intensities, if not spoiled by being forced into a symbolic straight-jacket, can directly support both matching of image regions, as required in binocular stereopsis, and matching of library images to the current input, as in spatial indexing.

Representation for stereo. Binocular stereo was the first problem on which Marr's notion of symbolic low-level representation (based on intensity edges) has been tried (Marr and Poggio, 1979). Although stereo systems following this approach have been subsequently improved and extended (Grimson, 1985), their performance is easily matched by simpler approaches, which do not rely on edge detection. Instead of edges, these approaches use raw pixel values (Cox et al., 1992) — an anathema to Marr — or vectors of responses of oriented filters (Jones and Malik, 1992). Note that both these approaches adhere to the principle of least commitment.

Representation for spatial indexing. In spatial indexing, the task of the system is to detect potential locations in the image that may contain an object of interest, taken from a library of familiar objects. Indexing reduces to simple search (which, furthermore, can be conducted in parallel over the image) if a set of feature values reliably associated with each object or object class can be determined. Surprisingly (or maybe not so, in view of the line of reasoning developed above), simple intensity and color-based features may suffice, as indicated by the results reported in (Wixson and Ballard, 1990; Swain and Ballard, 1991).

The main idea behind the approach of Ballard et al. is to represent objects by the distribution of their colors, and to seek invariance with respect to factors unrelated to object identity through histogramming the color distributions over the extent of the object. A match is declared for a certain region of the input image if its histogram matches that of the object.

This approach has been recently extended to deal with features other that color distributions, such as oriented intensity energy, contrast, etc. (Mel, 1996). Given $12 - 36$ training views of 100 objects, Mel's histogram-based system learned to recognize test views of objects that could vary in position, orientation in the image plane and scale; for nonrigid objects, recognition was also tested under deformations. The system's performance on 600 novel object views was 97% (chance-level performance would be 1%), and was comparable for the subset of 15 nonrigid objects. The generalization behavior and classification errors of the system reported by Mel indicate that it may have learned several natural shape categories that were not explicitly encoded in the dimensions of the feature space. For example, the first few candidate matches for an image of a book were typically other books, followed by objects similar to books in general appearance, such as cereal boxes.

4.2. OBJECT PROCESSING

Although histogram-based indexing approaches are useful for signaling potential membership in a class of objects, they must be complemented by a more rigorous comparison to the internally represented shapes, if the human capacity for shape categorization and for object recognition is to be matched. As we shall see next, both categorization and recognition can be approached using the same biologically inspired building blocks — receptive fields tuned to a variety of patterns, ranging from simple (oriented intensity energy) to complex (entire objects).

Face recognition. As an example of a simpler problem in object processing, consider face recognition. As with general objects, computer vision systems first attempted to deal with face recognition by representing faces symbolically. A typical system would look for the eyes and the mouth in an edge map, then would quantify and process their spatial relationships. The unreliability of this route to representation has been already discussed above; in face processing too, better results were achieved when researchers switched to raw intensity-based representations derived from RF responses (von der Malsburg, 1995).

Information available in the high-dimensional RF-based feature spaces can be cast into a form better suited for classification if the dimensionality of the feature space is reduced. It turns out that an efficient dimensionality reduction scheme for a given set of stimuli can be learned from examples, if several of the examples are stored and used as reference patterns, to which new incoming stimuli are compared (Edelman et al., 1992). Under this scheme, the new inputs are represented by the vector of their similarities to the reference patterns, as explained below.

Object recognition and categorization. Consider a number of classifiers acting in parallel, each tuned to a particular optimal stimulus, with the response falling off gradually and monotonically with dissimilarity between the actual and the optimal stimulus. In such a system, the category to which the stimulus belongs is signaled by the identity of the classifiers that respond above threshold, while its exact characterization is encoded in the distribution of responses of the active classifiers (Edelman et al., 1996). The system, thus, effectively reduces the dimensionality of the stimulus from that of the RF-based feature space to a low value, equal to the number of active classifiers.

A convenient framework for describing this approach to representation in terms used in computer vision is provided by the Pandemonium — one of the first explicit proposals for an object recognition scheme based on feature detectors (Selfridge, 1959; Lindsay and Norman, 1977). The original Pandemonium consisted of a three-level hierarchy: feature demons (responsible for the detection of lines, corners, etc.), cognitive demons (signaling the presence of entire objects) and a master demon (responsible for the recognition decision). To conform to the principles of least commitment and graceful degradation, this scheme has to be modified on all three of its levels. Some of the required modifications are the introduction of cooperating feature demons with graded-profile overlapping RFs (Edelman, 1996), making the cognitive demons probabilistic (Barlow, 1990), and replacing the winner-take-all decision demon by a multidimensional mechanism that would take into consideration the relative response levels of the cognitive demons, and not merely signify the strongest-responding one (Edelman, 1995b). The resulting scheme bears a significant resemblance to some of the current characterizations of higher-level visual function of the primate brain (Edelman, 1996). For example, the graded-profile shape-space RFs, proposed to serve as feature detectors, may be compared to the shape-specific RFs reported in (Sakai et al., 1994; Logothetis and Pauls, 1995), and to the representation of shapes in cortical columns, described elsewhere (Fujita et al., 1992; Tanaka, 1992).

5. SUMMARY

The consistent difficulties encountered by the reconstructionist program in computer vision, as well as the recent advances in understanding biological vision and in its computational modeling, suggest that a new interpretation should be sought for Aristotle's notion of what it means to see. Such an interpretation is likely to be closer to Gibson's (1966) idea of a representational system "resonating" to the world, and to Shepard's representation by "second-order isomorphism" (Shepard, 1968; Shepard and Chipman, 1970) than to Marr's (1982) proposal of representation by reconstruction. Although the implementation of the lessons from biological vision is only beginning, its initial results give reason to believe that the new approach may succeed where the old one faltered. The main guideline behind this approach is to let the visual world bear the burden of its own representation. Artificial visual systems should be capable of matching the performance of their biological counterparts, if they imitate nature in settling for stimulus-specific invariance (coupled with attention and active exploration; cf. Bajcsy, 1988; Ballard, 1991), in learning from examples (Poggio, 1990), and, generally, using the world as an external memory.

ACKNOWLEDGMENTS

This article is based on a talk given at the 1995 Rosenön Workshop on Computer Vision, organized by Jan-Olof Eklundh, and supported by the EC-IS Cooperation Program.

Thanks to Heinrich H. Bülthoff for the hospitality of Max Planck Institute for Biological Cybernetics in Tübingen, where this talk was prepared, and Dana Ballard, Freddy Bruckstein, John Mayhew, Hezy Yeshurun, and other participants of the Rosenön Workshop for their comments. SE is an incumbent of the Sir Charles Clore Career Development Chair at the Weizmann Institute.

REFERENCES

Aloimonos, J. Y. (1990). Purposive and qualitative vision. In *Proc. AAAI-90 Workshop on Qualitative Vision*, pages 1–5, San Mateo, CA. Morgan Kaufmann.

Bajcsy, R. (1988). Active perception. *Proc. IEEE*, 76(8):996–1005. special issue on Computer Vision.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48:57–86.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30:1561–1571.

Barrow, H. G. and Tenenbaum, J. M. (1978). Recovering intrinsic scene characteristics from images. In Hanson, A. R. and Riseman, E. M., editors, *Computer Vision Systems*, pages 3–26. Academic Press, New York, NY.

Barrow, H. G. and Tenenbaum, J. M. (1981). Interpreting line drawings as three-dimensional surfaces. *Artificial Intelligence*, 17:75–116.

Barrow, H. G. and Tenenbaum, J. M. (1993). Retrospective on "Interpreting line drawings as three-dimensional surfaces". *Artificial Intelligence*, 59:71–80.

Bishop, P. O., Coombs, J. S., and Henry, G. H. (1973). Receptive fields of simple cells in the cat striate cortex. *J. Physiol. (London)*, 231:31–60.

Blackmore, S. J., Brelstaff, G., Nelson, K., and Troscianko, T. (1995). Is the richness of our visual world an illusion? Transsaccadic memory for complex scenes. *Perception*, 24:1075–1081.

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

Cox, I. J., Hingorani, S., Maggs, B. M., and Rao, S. B. (1992). Stereo without disparity gradient smoothing: a Bayesian sensor fusion solution. In *British Machine Vision Conf.*, pages 337–346, Berlin. Springer-Verlag.

Edelman, S. (1995a). Class similarity and viewpoint invariance in the recognition of 3D objects. *Biological Cybernetics*, 72:207–220.

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

Edelman, S. (1996). Receptive fields for vision: from hyperacuity to object recognition. unpublished manuscript.

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

Edelman, S., Cutzu, F., and Duvdevani-Bar, S. (1996). Similarity to reference shapes as a basis for shape representation. In Cottrell, G. W., editor, *Proceedings of 18th Annual Conf. of the Cognitive Science Society*, pages 260–265, San Diego, CA.

Edelman, S., Reisfeld, D., and Yeshurun, Y. (1992). Learning to recognize faces from examples. In Sandini, G., editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 787–791. Springer Verlag.

Field, D. J. (1994). What is the goal of sensory coding? *Neural Computation*, 6:559–601.

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin, Boston, MA.

Gilbert, C. D. (1988). Neuronal and synaptic organization in the cortex. In Rakic, P. and Singer, W., editors, *Neurobiology of Neocortex*, pages 219–240. Wiley, New York, NY.

Gilbert, C. D. (1994). Neuronal dynamics and perceptual learning. *Current Biology*, 4:627–629.

Goodale, M. A., Milner, A. D., Jakobson, L. S., and Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, 349:154–156.

Grimes, J. (1995). On the failure to detect changes in scenes across saccades. In Akins, K., editor, *Perception*, volume 5 of *Vancouver Studies in Cognitive Science*, chapter 4. Oxford University Press, New York.

Grimson, W. E. L. (1985). Computational experiments with a feature-based stereo algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7:17–34.

Humphrey, G. K. and Khan, S. C. (1992). Recognizing novel views of three-dimensional objects. *Can. J. Psychol.*, 46:170–190.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.*, 73:218–226.

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

Jones, D. G. and Malik, J. (1992). A computational framework for determining stereo correspondence from a set of linear spatial filters. In Sandini, G., editor, *Proc. ECCV-92*, pages 395–410, Berlin. Springer.

Kamon, Y., Flash, T., and Edelman, S. (1994). Learning to grasp using visual information. CS-TR 94-04, Weizmann Institute of Science. also in Proc. Intl. Conf. on Robotics and Automation, Minneapolis, April 1996.

Koenderink, J. J., van Doorn, A. J., and Kappers, A. M. L. (1993). Depth and viewing conditions: pictures versus real scenes. *Perception*, 22 (suppl.):98. Proc. ECVP'93.

Kuffler, S. W. and Nicholls, J. G. (1976). *From neuron to brain*. Sinauer, Sunderland, MA.

Lando, M. and Edelman, S. (1995). Generalization from a single view in face recognition. CS-TR 95-02, Weizmann Institute of Science.

Lindsay, P. H. and Norman, D. A. (1977). *Human information processing: an introduction to psychology*. Academic Press, New York.

Logothetis, N. and Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex*, 3:270–288.

Logothetis, N. K., Pauls, J., Poggio, T., and Bülthoff, H. H. (1994). View dependent object recognition by monkeys. *Current Biology*, 4:404–41.

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

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

Marr, D. and Hildreth, E. (1980). Theory of edge detection. *Proc. R. Soc. Lond. B*, 207:187–217.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, 15:470–488.

Marr, D. and Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204:301–328.

Mel, B. (1996). SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. Technical report, University of South California, Los Angeles, CA.

Moses, Y., Ullman, S., and Edelman, S. (1996). Generalization to novel images in upright and inverted faces. *Perception*, 25:443–462.

O'Regan, J. K. (1992). Solving the real mysteries of visual perception: The world as an outside memory. *Canadian J. of Psychology*, 46:461–488.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910.

Pollatsek, A., Rayner, K., and Collins, W. E. (1984). Integrating pictorial information across eye movements. *J. Exp. Psychol.: General*, 113:426–442.

Rensink, R., O'Regan, K., and Clark, J. J. (1995). Image flicker is as good as saccades in making large scene changes invisible. *Perception*, 24 (suppl.):26–27.

Sagi, D. and Tanne, D. (1994). Perceptual learning: learning to see. *Current opinion in neurobiology*, 4:195–199.

Sakai, K., Naya, Y., and Miyashita, Y. (1994). Neuronal tuning and associative mechanisms in form representation. *Learning and Memory*, 1:83–105.

Selfridge, O. G. (1959). Pandemonium: a paradigm for learning. In *The mechanisation of thought processes*. H.M.S.O., London.

Shapley, R. and Victor, J. (1986). Hyperacuity in cat retinal ganglion cells. *Science*, 231:999–1002.

Shepard, R. N. (1968). Cognitive psychology: A review of the book by U. Neisser. *Amer. J. Psychol.*, 81:285–289.

Shepard, R. N. and Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1:1–17.

Snippe, H. P. and Koenderink, J. J. (1992). Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, 66:543–551.

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

Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology*, 2:502–505.

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

Tenenbaum, J. M., Fischler, M. A., and Barrow, H. G. (1981). Scene modeling: a structural basis for image description. In Rosenfeld, A., editor, *Image Modeling*, pages 371–389. Academic Press, New York.

Todd, J. T. and Reichel, F. D. (1989). Ordinal structure in the visual perception and cognition of smoothly curved surfaces. *Psychological Review*, 96:643–657.

von der Malsburg, C. (1995). Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5:520–526.

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

Weiss, Y., Edelman, S., and Fahle, M. (1993). Models of perceptual learning in vernier hyperacuity. *Neural Computation*, 5:695–718.

Witkin, A. P. (1981). Recovering surface shape and orientation from texture. *Artificial Intelligence*, 17:17–45.

Wixson, L. E. and Ballard, D. H. (1990). 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. Morgan Kaufmann.