MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1158
C.B.I.P Memo No. 43

October 1989

Computational vision: a critical review

Shimon Edelman   Daphna Weinshall

Abstract

We review the progress made in computational vision, as represented by Marr's approach, in the last fifteen years. First, we briefly outline computational theories developed for low, middle and high-level vision. We then discuss in more detail solutions proposed to three representative problems in vision, each dealing with a different level of visual processing. Finally, we discuss modifications to the currently established computational paradigm that appear to be dictated by the recent developments in vision.

This report describes research done at the Massachusetts Institute of Technology within the Artificial Intelligence Laboratory and the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences and Whitaker College. The Center's research is sponsored by a grant from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Alfred P. Sloan Foundation; and by the National Science Foundation. The Artificial Intelligence Laboratory's research is sponsored 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. SE and DW are supported by Chaim Weizmann Postdoctoral Fellowships from the Weizmann Institute of Science.

1 Vision as an information-processing task

Contemporary cognitive science construes mental processes as computations defined over representations [126]. Under this view, visual perception, in particular, is defined as the process whereby light patterns impinging on the retina are transformed into internal representations. The problem of interpreting a retinal input is illustrated in figure 1, showing a photograph of a natural scene and an array of the intensity values corresponding to a part of it. These intensities are the raw data of vision, as they arrive in the camera or at the retina. To appreciate the difficulty of visual recognition, one may try to identify the object (present in the scene) described by this array.

Computational vision sets out to understand theoretical aspects of visual perception, to explicate possible strategies of solving the emerging problems, and to explore their biological and artificial implementations. This paradigm emphasizes the need for an interdisciplinary approach: machine vision can serve as a testbed for psychological and neural models of visual function, while biological vision can supply useful hints in choosing feasible approaches to problems in visual processing by providing examples of workable solutions.

The problem of understanding visual perception in information processing terms has been formulated most cogently by Marr and Poggio ([92], [96], [93]), who argued that vision (and, indeed, any information-processing task) may be studied at three distinct levels. At the top level they placed the computational theory of vision, in which the problem is characterized as a mapping from one kind of information to another, and the abstract properties of this mapping are analyzed. At the middle level, there is the choice of algorithm for the mapping, determined, among other factors, by the nature of representation at its input and output. At the lowest level, there are the implementational details, such as the physical realization of the representation and the algorithm.

In the original formulation, the three levels were only loosely coupled. Nevertheless, it appears that any commitment to a specific computational formulation of vision would inevitably affect the algorithm level. In particular, as we shall argue, Marr's definition of the computational purpose of vision have influenced the algorithm-level theories of an entire school of vision research. Marr held that the representation of the shape of an object is quite different from the representation of its use and purpose, and that vision alone can deliver an internal description of the shape of an object, even when the object is not recognized in the sense of understanding its use and purpose ([93], p.35). Together with the assumption that a good way to represent an object is to build a three-dimensional model of it, this view resulted in many current theories of visual recognition relying on the detection of 3D primitives [14], or requiring a library of 3D object models ([87], [156]). In biological visual systems, in comparison, the question of representation appears to be far from settled. As the retinal image is transformed with each successive processing stage, the theories concerning the nature of representation diverge more and more widely, reaching no consensus as to what is the ultimate visual representation of objects, scenes and events ([140], [130]).

Solutions to many visual tasks appear to be inherently ambiguous. For example, assuming, as Marr seems to have done, that the purpose of vision is to reconstruct the spatial layout of the outside world, one is confronted with the difficult problem of ambiguity of inferring the third dimension from retinal projection. This ambiguity is a necessary consequence of the imaging process, during which depth information is lost. Marr proposed to compensate for this loss by constraining the solution to the reconstruction problem to conform to a priori assumptions, dictated by our knowledge of the physical world. The research program based

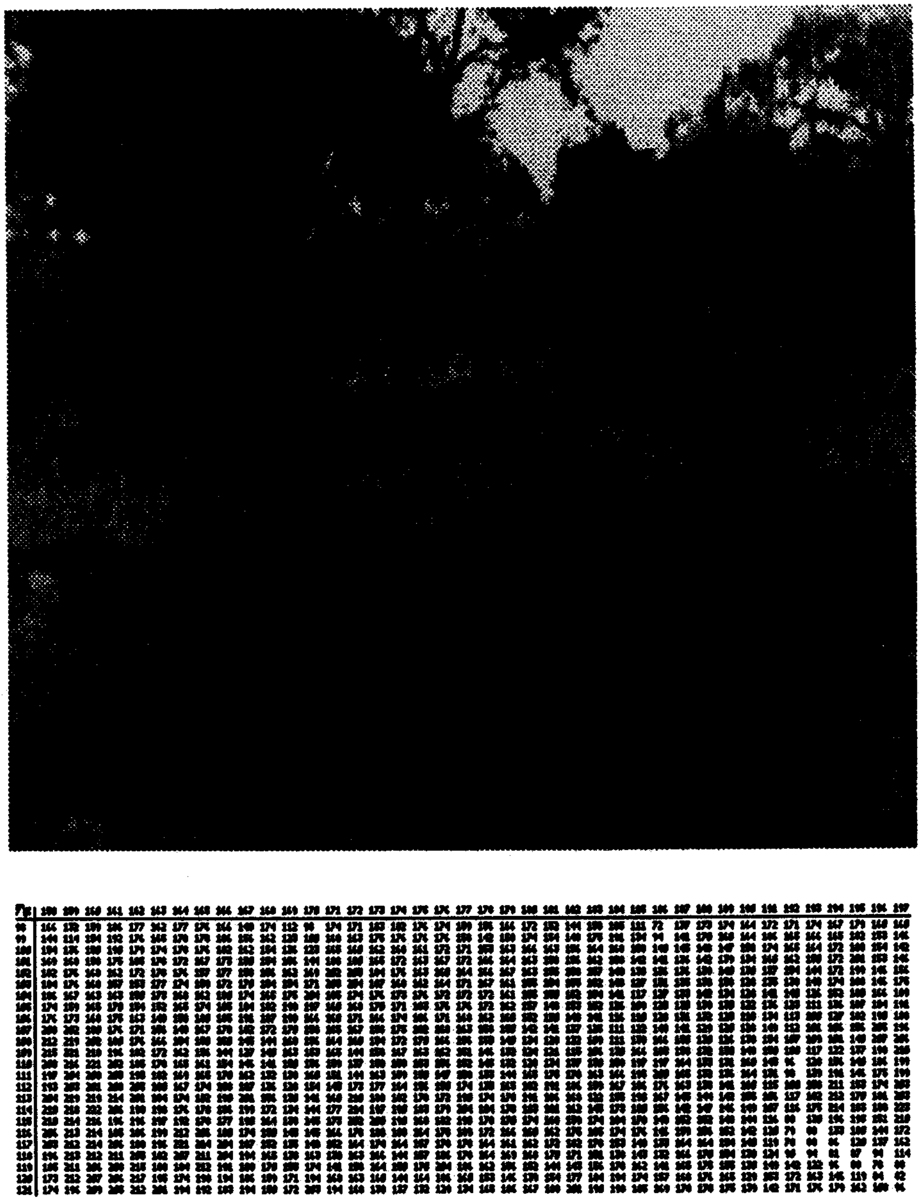

Figure 1: Top: a photograph of a natural scene, with two antelopes. Bottom: the array of the intensity values corresponding to the image of the antelope in the right part of the scene. These intensities are the raw data of vision, as they arrive in the camera or at the retina.

on this approach advanced our understanding of different aspects of vision, such as perceptual grouping, stereopsis and motion perception.

In areas in which physical constraints are hard to define, as in object recognition and scene understanding, the reconstructionist research program runs into difficulties. Recently implemented recognition methods ([87], [156], [145], [49], [70]) appear to circumvent the problem of 3D reconstruction, either by addressing other important issues in recognition, such as viewpoint normalization, while assuming that the 3D models of objects are built into the system, or by carrying out the entire recognition process in 2D.

In this chapter, we review the progress made in computational vision, as represented by Marr's approach, in the last fifteen years. Section 2 outlines computational theories developed for low, middle and high-level vision. Section 3 contains a more detailed discussion of solutions proposed to three representative problems in vision, each dealing with a different level of visual processing. The last section suggests modifications to the currently established computational paradigm that appear to be dictated by the recent developments in vision.

Since Marr's influence on vision research has been the strongest at MIT, we follow in our review the development of the MIT perspective on computational vision. Other recent reviews of computational and machine vision include ([21], [62], [41], [132] and [57]).

2 The components of vision

The analysis of images can be done at different levels. Low-level vision undertakes the reconstruction of the visual world, but only in a limited sense. Borrowing from Marr's terminology, its end product is a kind of $2\frac{1}{2}$D sketch of the surrounding scene: a representation in which many attributes of the visible surfaces are made explicit and used as labels in a scene-based map. Among the attributes computed at this stage are the relative motion and depth of the visible features, as well as luminance, color and texture based descriptions of different surfaces. Information from the different visual cues may be further processed by middle-level modules that compute more abstract properties defined over the extracted features. Combining surface patches and building descriptions of objects is within the scope of high-level vision.

The rest of this part is organized as follows: in sections 2.1, 2.2 and 2.3 we review research in low-level, middle and high-level vision, respectively, and in section 2.4 we review some of the aspects of neuronal modeling in the context of vision.

2.1 Low-level vision

One possible definition of the domain of low-level vision is through the notion of bottom-up, or data-driven, processing (as opposed to top-down, or model-driven processing). Low-level vision is concerned with those transformations of the input image that are common to most visual tasks and can be carried out as fast as the input changes, given the resources of the system.

To some extent, the analysis of the different visual cues can be done independently of each other and in parallel. Different behavioral tasks such as navigation, recognition or high precision manipulation may also require the extraction of different kinds of visual information. It is not clear, though, how independent the analysis of the different low-level modules can and should be and to what extent information should actually be integrated. The question of module interdependence may be approached both through the development and integration of low-level algorithms and by the study of biological vision systems.

Evidence from human visual psychophysics suggests there is at least some degree of independence among the modules. For example, one can easily understand black and white movies that lack color and stereo, and 3D structure can be perceived in extremely impoverished stimuli, such as the projections of rotating wire-frame objects [162], or random-dot stereograms [73]. Physiological evidence for module independence is provided by the existence of separate subcortical and cortical pathways carrying different kinds of visual information ([142], [102], [174]; see also [77], p.381). This independence is probably not complete: single units that respond to several different cues, such as stereo disparity and direction of motion, are found in some visual cortical areas ([35]; cf. [27]).

2.1.1 Stereopsis

Since the image on the retina of an eye, or in a camera, is obtained by projecting a three-dimensional world onto a two-dimensional surface, the information on the third dimension, depth, is lost. However, a perception of depth can still be achieved by stereopsis, i.e., by combining information from two or more images, taken simultaneously from slightly different positions, e.g., by the two eyes. In human vision, binocular stereo, together with other cues such as motion and perspective, is an important source of depth information.

The importance of seeing depth in manipulation tasks becomes clear if one attempts to thread a needle, or to insert a key into a keyhole, with one eye closed. The importance of stereo in recognition tasks seems to be more limited, as people can recognize objects in drawings and photographs that lack binocular information. Nevertheless, much attention has been given to the analysis of stereo vision, mainly because the goals of stereo seemed relatively simple and well understood, and because stereo has been demonstrated by Julesz to create a perception of 3D shape in the absence of any other cue ([73]; see figure 2).

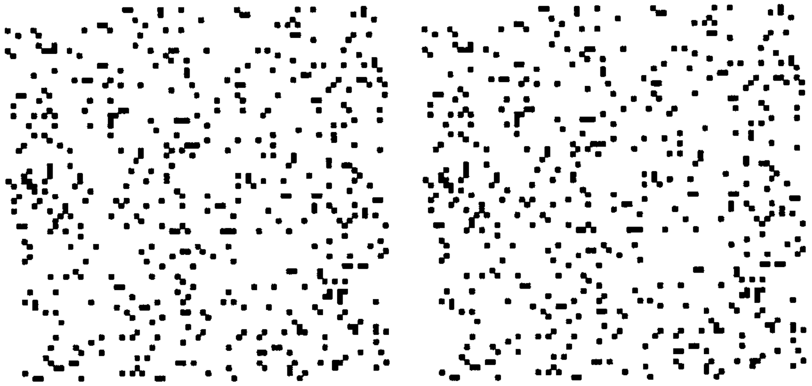

Figure 2: A random dot stereogram. The left and the right images should be viewed by the left and the right eyes, respectively. When the two images are fused, a square floating in depth above the background appears in the center. This is an example of shape perception when the only cue is stereopsis: no monocular shape cues are present in either image.

In computer vision algorithms, the extraction of depth from binocular stereo begins with the formation of a disparity map by matching the two images (the disparity of an object is defined as the difference between its positions in the two images). Thus, a disparity value is assigned to every location in the image. The representation of disparities in the brain is different: cells tuned to a relatively narrow range of disparity are found in the primary visual cortex along with cells that respond selectively to features whose disparities are higher or lower than certain values [118].

In a random dot stereogram, e.g. in figure 2, (and in many natural scenes) matching objects in the two images to compute disparity is difficult. For every object there are many potential target matches, all but one of which are false. Two constraints on matching were proposed to solve the false targets problem [95]. The first constraint, uniqueness, forces each object in one image to be matched to at most one object in the other image. The second constraint, smoothness, requires that nearby objects in one image be matched to nearby objects in the other image. Most of the existing stereo algorithms, as well as most of the theories of human stereo vision, incorporate these two constraints.

The matching of stereo pairs of complicated scenes remains difficult even after the application of these constraints. For example, objects may look different in the two images, due to different perspective distortions. For images that have enough texture, e.g., in random dot stereograms, matching can be done by cross-correlation of regions in the two images ([111], [171]). Many attempts have been made to obtain a general matching algorithm by the use of more sophisticated matching techniques ([9] is a review of early algorithms; see also [101], [59]). Two examples are multi-resolution matching (processing the images at several levels of detail; [97], [48], [100]) and adjusting the fixation point to control the range of disparities in the images [48]. After intensive development, a few of the algorithms appear capable of achieving better performance on a single stereo pair than that of the human visual system, albeit at the cost of a high computational complexity.

Despite the impressive performance of some existing stereo algorithms, problems with the basic paradigm of applying physical constraints arise already in this relatively narrow domain. Specifically, imposing the constraints of uniqueness and smoothness on the matching process may be detrimental in some situations. For example, the smoothness assumption breaks down for scenes that include transparent objects, such as fences. Looking through a fence, one may see a tree to the left of a given bar in the left eye and the same tree to the right of the same bar in the right eye. (It should be noted that simple modifications of the continuity constraint can overcome this problem [124]). A more serious problem with smoothness arises at those locations in the images where there are depth discontinuities. Smoothing across discontinuities is bound to produce errors in stereo (and other low-level vision algorithms that employ this constraint, such as shape from shading and motion; see sections 2.1.2, 2.1.4 and 3.2). Among the recent studies that address this problem are [45], [18] and [144]. See also [122] for an overview of the use of smoothness constraints in low-level vision.

There is also evidence that human vision may not always follow the uniqueness assumption and that people may perceive simultaneously several surfaces corresponding to multiple matchings between elements in the two images [167]. In some cases the relationship between disparities derived from matching and perceived depth is not unequivocal. For example, the depth perceived in a dot pattern may correspond to an average disparity, rather than to one of the disparities derived from a possible matching [103].

A major open problem in stereo is what the representation at the output of a stereo system should be. The disparity field, once computed by a matching algorithm, can be used to extract relative depth information. Objects with large disparities are usually closer to the viewer than objects with small disparities (when the optical axes of the cameras are almost parallel). If the location, the orientation and the focal length of both cameras (or eyes) are known, a simple geometrical transformation can be used to compute the exact depth from the disparities. Typically, however, the precise camera parameters are not known. Computing the parameters from the disparity field is possible, though computationally expensive and sensitive to noise [62]. It is unclear whether this computation is necessary or whether relative depth is sufficient for all practical purposes ([7], [168]).

2.1.2 Motion

The analysis of motion is similar to stereo in that it involves the analysis of a set of images of a given scene. The differences are that the images are taken in succession rather than simultaneously, more than two images may be available for analysis, and the objects may undergo an arbitrary transformation between the images. Specifically, in stereo it is assumed that the two images are taken by two fixed cameras (eyes) whose relative location is known, at least approximately. When images are taken at different times, any object can translate in any direction, rotate and deform, so that the geometrical transformation between the images is not sufficiently defined, unless some assumptions are made. Among the goals of motion analysis are the extraction of depth information for shape perception and navigation (possibly in a less precise way than stereo), the segmentation of an image into distinct objects, and the identification of moving objects.

Psychophysical studies indicate that motion analysis alone is sufficient to extract shape information when the moving objects are rigid bodies (that is, the distance between any two points on an object remains constant with time; [162], [152]). The rigidity constraint is useful to make the problem computationally tractable ([152]; see [153] and [56] for a review). It is derived from a computational analysis of the problem and, similar to the constraints imposed in stereo analysis, it should be applied with care. There are many examples where humans perceive nonrigid motion (e.g., expansion and contraction), even though a rigid motion interpretation (such as a rotation in depth) is possible (see [4], [53]). In some cases, the motion of nonrigid objects leads to a clear perception of shape. For example, watching a person walking in the dark with small lights attached to the joints, one can perceive a human figure walking ([71], [60]). Although several recent studies address the issue of the interpretation of nonrigid motion, the algorithms they develop usually allow only certain classes of nonrigid transformations, such as bending (see, e.g., [81] and [29]).

Similar to the analysis of stereo, the analysis of motion can be divided into two stages: the matching, either of features, or of regions, in the successive images, followed by a computation that relies on the resulting disparity information. Another approach involves the computation of optical flow (the instantaneous 2D velocity field) instead of the motion disparity map. This computation can be presumably done by local intensity change detectors (e.g. [172]). A principal difficulty in the first approach is matching, whereas in the optical flow formulation it is relatively high sensitivity to noise, associated with numerical differentiation [62]. The distinction between the matching and the optical flow formulations parallels the difference between long-range and short-range motion perception processes observed in psychophysical studies [20].

The matching problem is more difficult in motion than in stereo, for example because of changing illumination and because of moving shadows, whose motion differs from the true motion of the objects. As in stereo, a smoothness constraint, equivalent to the assumption that the velocities of different features change slowly over the image, appears plausible and is used by many algorithms ([64], [53], [107]). Typically, computer vision algorithms enforce smoothness of velocity either over regions of the image [64], or along its contours [53]. Psychophysical findings suggest that both ways of imposing smoothness may coexist in the human visual system. For example, the perceived motion of points on a contour is similarly affected by the motion of neighboring points and of more distant points that lie on the same contour [108].

Given matched images, there are many ways to recover the three-dimensional structure of the objects. Early studies of this problem concentrated on the issue of minimal information needed to recover unique structure from motion ([152], [31], [123], [104], [150], [165]). One major problem common to the algorithms that recover structure from motion is their numerical instability [153]. To decrease the resulting sensitivity to errors, motion information may be pooled over many features (extended space; see [62], [23]), or over many time frames (extended time; see [154]). Human vision appears to use both extended time and space analysis ([71], [55]). Some psychophysical evidence suggests that, in disagreement with most structure from motion algorithms, the perception of 3D structure may not depend on prior perception of motion [157].

It is still an open question to what extent the exact structure of objects is necessary to fulfill the ultimate goals of vision. Some shape information, such as the classification of object surfaces as convex, concave, planar, or hyperbolic (saddle-like) can be obtained without the complete recovery of the exact locations in space of all the points in the image [166]. For specific narrowly defined tasks, such as obstacle avoidance in navigation, an even simpler analysis may suffice [110].

2.1.3 Edge detection

A problem common to the design of all low-level vision modules is the choice of the input representation. The raw input to the visual system, an array of intensity values (see figure 1), does not suit well tasks that require image matching, for example, because of its sensitivity to noise and to changes in the illumination. Intensity edges (image locations where the intensity changes significantly) have been proposed as a more stable initial representation for stereo and motion. Different definitions of significant change in intensity lead to different algorithms for edge detection (see [34], [6] and [54] for reviews). Finding computationally efficient algorithms that capture those edges that correspond to what we intuitively perceive as edges proved to be a difficult problem. If edge detection is meant to provide a cartoon of the image, i.e., the set of edges that has a physical origin, it effectively subsumes other difficult problems in vision, such as figure-ground separation and object recognition.

Intensity edges may be defined as those places in the image where the rate of change of intensity attains a local maximum. The derivative operation included in this definition makes it sensitive to noise. Noise amplification can be reduced by smoothing the image before subjecting it to the derivative operation. A computationally efficient, biologically plausible edge detection operator based on this approach is a linear filter that combines smoothing (by a convolution with a 2D Gaussian) with differentiation (by the application of a 2D Laplacian, a rotationally symmetric operator). Equivalently, the image may be convolved with the Laplacian of a Gaussian. The edges are then found by locating the zero-crossings in the output of the convolution [93]. Psychophysical evidence [164] suggests that zero-crossings may be involved in early processing in human vision, along with additional features, such as intensity extrema. (For an anatomical model of zero-crossing detection in the retina see

[133]).

Since the Laplacian of Gaussian operator is spatially symmetric, it ignores the asymmetry of edges, which are one-dimensional curves with a preferred direction. Some algorithms address this problem by computing second directional derivatives of the input [52]. The computation of the second derivative in the direction of the intensity gradient has been shown optimal for the detection of oriented edges [149]. As in stereo, a multi-resolution approach (computing the edges at several levels of smoothing) proved useful in edge detection. A widely used well-engineered implementation of edge detection that employs both approaches is due to Canny ([26]; see figure 3). Labeling edges according to their physical origin (shadows, occluding contours etc.) is a subject of current research (e.g. [44]). Edge labeling may be useful in tasks such as object recognition.

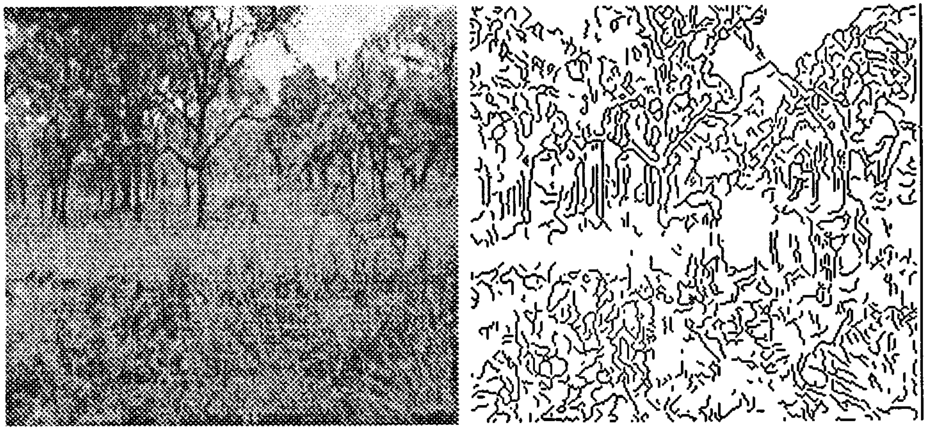

Figure 3: Left: an image of a natural scene (same as in figure 1). Right: intensity edges extracted from this image by the Canny algorithm [26].

2.1.4 Shape from shading and shadows

Shading information seems to be quite important in human shape perception [63]. In the perception of shape from shading, the visual system must separate intensity changes that are caused by changing orientation of object surfaces from those due to changing surface reflectance (including color) and illumination. As an example of the perception of shape from shading, consider the human face in figure 4. The basic information on the 3D shape of the face in this image is obtained from shading analysis (e.g., the nose appears to protrude from the face because its flanks are shaded darker than its tip).

When the illumination and the viewing direction are fixed and the color is constant, the shading of a surface depends solely on its local orientation. When posed quantitatively, the problem of inferring shape from shading turns out to be one of the most difficult in low-level vision. The main complication is ambiguity: a shaded image can be interpreted in many different ways. For example, a concave surface, a convex surface and a saddle-like surface all appear the same from certain viewpoints. If surface color is unknown and variable, computing shape from shading becomes even more difficult.

Assuming a simplified surface reflectance model, and given some a priori knowledge about

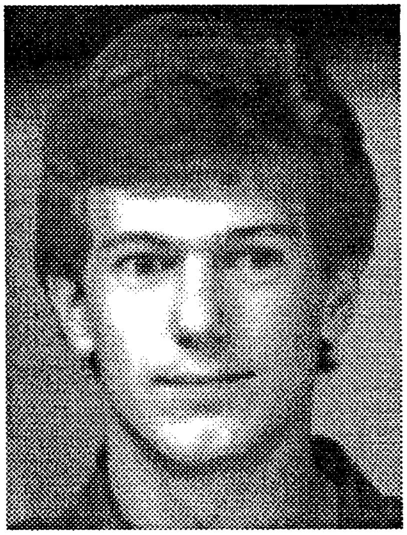

Figure 4: A shaded image of a face. The basic information on the 3D shape of the face in this image is obtained from shading analysis (e.g., the nose appears to protrude from the face because its flanks are shaded darker than its tip).

the objects in the image (e.g., the depth along the outlines of the objects), the exact shape from shading problem becomes solvable, though still computationally difficult ([170], [69], [62]). A further assumption of oblique illumination may substantially reduce the computational complexity [115], although the problem of self-shadowing, neglected by most algorithms including [115], becomes significant. In such cases, special shape from shadows algorithms [138] may be employed to infer shape from the distribution of shadows in the image. One complicating factor, mutual illumination (a secondary illumination due to the reflectance of light by other objects in the scene) is neglected by all of the above algorithms. Recent work has shown that, despite the simplified conditions assumed by most of the shape from shading algorithms, mutual illumination can make the problem more ambiguous than had been previously appreciated [43]. On the other hand, experiments with computer graphics systems indicate that simulating mutual illumination is important for the creation of realistic-looking images.

Recent psychophysical data suggest that humans use shading information in a limited way ([10], [147], [25]). People seem to use shading information to build qualitative descriptions of object surface (e.g., in terms of convex and concave regions), rather than to compute exact shape from shading (in light of the complexity of extracting shape from shading, these findings may not come as a surprise). The computation of qualitative shape from shading may be facilitated by using highlights [17], which are a nuisance in most exact approaches ([16], [131]). Finally, humans appear to employ high-level heuristics, such as the assumption that the scene is illuminated from above, to disambiguate the shape from shading problem [131].

2.1.5 Color

In human vision, color contributes important information for object recognition, although its contribution is by no means necessary for most everyday tasks. The human visual system exhibits an impressive ability to infer the correct color of objects under illumination that may vary in direction, intensity and spectral content. Thus, while light arriving at the retina from a given surface patch under different illuminations may have different spectral compositions, the patch would normally appear to have the same color in both cases. This phenomenon is called color constancy. It has been suggested that color constancy can ensue if we assume that the color of illuminant changes slowly and smoothly, whereas surface color changes abruptly. Many algorithms rely on this assumption to assess the true color of the surface ([84], [90]; [66] show that a linear color operator can be learned from examples). Under the assumption that all colors appear in the image with equal probabilities, the color of the illuminant can be estimated by averaging color over large neighborhoods of the image [83]. Another approach uses the color of highlights to compute the color of the illuminant [85]. Psychophysical evidence suggests, however, that the human visual system does not use this specific technique [67].

Humans perceive color by having three different types of receptors in the retina whose peak sensitivities are in the red, green and blue regions of the spectrum. Thus the dimension of the color space perceived by humans is only two, not considering the overall brightness. This means that there are many different reflectance functions that appear to humans to have the same color. A principal component analysis of the distribution of surface reflectance functions of natural objects and of ambient daylight reveals, however, that the number of degrees of freedom needed is actually small. More specifically, the color of many natural objects can be largely approximated in terms of only three base functions (see [89], [24]).

2.1.6 Texture

The analysis of texture deals with statistical properties of collections of features (textons or elementary units of texture [74]) that cover object surfaces. Statistics of texture may contain important perspective information and may enable the segmentation of the image into distinct objects. For example, in the image of a tilted uniformly textured plane the density of elements diminishes in the direction away from the viewer, providing a hint to the orientation of the plane in depth. For non-planar objects, the distribution of texture elements is more complex: a sphere sprayed with paint projects to a circle filled with dots, in which the density of the dots is low around the center and increases towards the periphery.

Differences in the statistical distribution of texture primitives may help segment the image into distinct objects (see [13] for a review; [160], [28]). If the analysis is based on complex texture elements [12], it is first necessary to identify the elements in a hierarchical manner and then compute their distribution in the image. The characterization of useful textons (in particular, those involved in human perception of texture) has been a subject of extensive research [75]. One relevant question is that of the relationship between texture-based segmentation and the structure of the constituent textons. Another theory relies on a statistical analysis of the raw intensity values in the image [74]. Gradients of the size distribution and the density of textons can be interpreted as depth cues for receding surfaces ([46]; cf. [141]), through Fourier analysis [5] or statistical assumptions about the scene [169].

2.1.7 Occluding contours

A line drawing of a complex 3D object, containing no shading, stereo or textural cues, may provide information sufficient for its identification and for the perception of its shape. One major source of information in a line drawing is the shape of the object's outline, or the occluding contour. Marr [91] argued that by themselves the occluding contours are not sufficient for shape classification. He therefore concluded that our ability to perceive the shape of an object from its outline is achieved through the imposition of strong constraints on possible shapes of objects, i.e., that most shapes can be described in terms of volumetric primitives called generalized cones [15]. Koenderink [80] showed, however, that such restrictions are not necessary, and that occluding contours carry important information about shape. Specifically, he showed that a concave segment of the object's boundary implies that the object's surface is locally hyperbolic (saddle-like), while a convex occluding contour implies a locally elliptic (convex or concave) surface. It should be noted that saddle-like regions are good candidates for object segmentation, and that human vision appears to use simple heuristic rules of this type in achieving descriptions of objects in terms of their parts [61].

2.1.8 Low-level vision: an interim summary

In distinction from other theoretical approaches, computational study of vision combines an analysis of the computational strategies employed by biological systems with psychological and physiological investigations, and with building artificial vision systems. Thinking in terms of representations and their transformations and subjecting the resulting theories to empirical tests proved especially fruitful in the domain of low-level vision, or the bottom-up recovery of visual properties of the environment. The interaction between computational, biological and machine studies led to a better understanding of the difficulties involved in low-level vision, and to a reopening of basic questions concerning the output representation, the physical constraints and the modularity of low-level processing. These issues become even more important at the higher levels of visual processing, where at present little feedback is available from empirical investigations.

2.2 Middle vision

Two classes of visual operations appear to fall outside the scope of both the input-driven low-level processes and the high-level goal-oriented ones. The first class includes processes that operate on the "primal sketch" representation of the input, obtained in the first stage of vision [92]. The purpose of these processes is, in general, the completion and the enhancement of the primal sketch. These operations are lateral in the sense that they do not necessarily rely on either a more detailed input than already available, or on higher-level cues. Examples of phenomena that may involve such processes are boundary completion (figure 5, left), interpolation of depth and motion information to regions that lie in between features used to compute this information, spatial grouping of features (figure 5, right) and increased perceived saliency of some contours relative to others ([92], [27], [137]).

Intermediate-level processes of the second type may be described computationally as multipurpose visual routines [155], invoked at need by higher-level modules. These routines may enhance the intermediate representation by making explicit spatial properties and relations such as contiguity (of a contour) and insideness (of a feature with respect to the contour). This type of information is useful in many visual tasks, such as recognition and navigation,

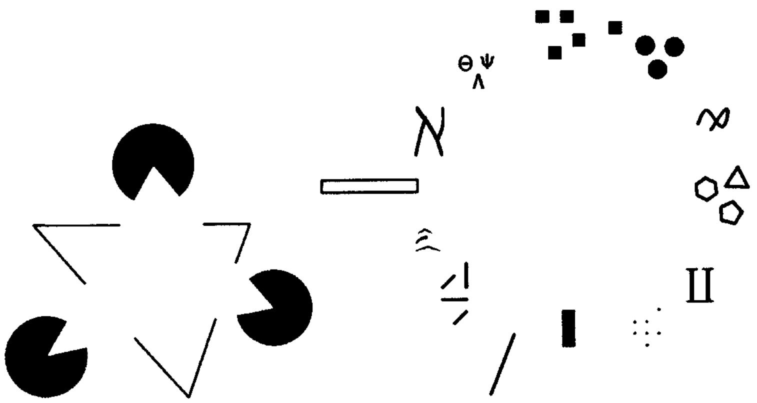

Figure 5: Left: Kanisza's triangles, an example of boundary completion. Illusory contours that form a triangle are seen to occlude three black disks. Right: an example of hierarchical grouping by shape similarity and physical proximity. At the highest level, the small shapes form a circle.

but is too abstract and resource-intensive to warrant automatic bottom-up computation. An additional kind of routines may be involved in the computational substrate of visual attention, where the basic operations are indexing (marking a location in the visual field) and the shifting of the processing focus to the marked location [155].

Grouping of edges receives increased attention in recent studies of middle vision. Lowe [87] proposed to start the grouping by detecting image properties that are likely to convey useful information about the 3D world. Among such properties are collinearity and parallelism: the chance that edges that appear collinear or parallel in the image do so only by accident (due to the particular viewpoint) is small. Another example of a middle vision operation on edges is the computation of the saliency of curve segments by a simple local mechanism [137]. In many cases, this operation can isolate perceptually important curves in a noisy edge map (produced, e.g., by the Canny method). Our last example, which resembles visual routines in that it operates "laterally" on the intermediate representations, is the integration of different low-level cues (e.g., [120], [3]).

The advances in the computational theory of integrated low and intermediate level vision are beginning to draw the attention of psychologists and physiologists. Psychophysical evidence indicating that the human visual system may indeed combine different cues into an integrated representation can be found in [27]. Some of the recent experiments that address the issue of intermediate level vision and support the existence of boundary tracing and indexing as visual routines are described in [72] and [128]. An example of a neurophysiological study of attention in monkey is [105], which describes the modification of the receptive field of an extrastriate cell by attention shifts. Cells that respond to abstract visual entities such as illusory contours were found in the striate cortex of the monkey [159].

2.3 Object recognition

An intelligent visual system is expected to allow its host to navigate through the environment and possibly to manipulate any objects present in it. Keeping a visual library of models of potentially important objects enables such a system to recognize an object and to select an appropriate behavior based upon its past experience. Consequently, computational theories of object recognition postulate that there exist in the visual system representations of familiar objects and scenes. To recognize an object, the system compares it with each of the stored models, or templates [112]. An estimate of the goodness of fit between an object and a template can be obtained, e.g., by a correlation-like operation, in which the two are superimposed and the proportion of pixels that agree in their values is computed.

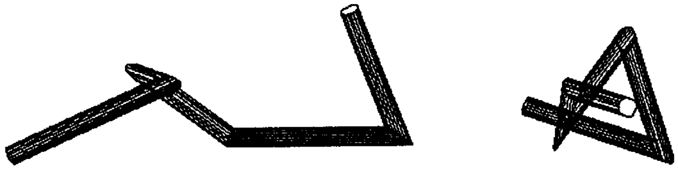

Figure 6: The appearance of a 3D object can depend strongly on the viewpoint. The image on the right is of the same object as the image on the left, rotated in depth by 90°. The difference between the two images illustrates the difficulties encountered by any straightforward template matching approach to 3D object recognition.

Recognition of 3D objects seen from arbitrary viewpoints is difficult because an object's appearance may vary considerably depending on its pose relative to the observer (see figure 6). Thus, while straightforward template matching [2] may be useful in the recognition of 2D objects in a controlled environment, it will not work for 3D object recognition, unless a template is stored for each view that will ever have to be recognized. Although the extent to which people can recognize novel, radically different, views of 3D objects is not clear ([134], [135]), we obviously do have some ability to generalize recognition to novel views. This ability is termed visual object constancy (see e.g. [65]).

Most of the schemes for object recognition proposed to date can be divided into three main classes, according to their approach to the problem of object constancy [156]. The first approach assumes that objects have certain invariant properties that are common to all their views and different between object classes (in practice, this twofold assumption proved difficult, if not impossible, to satisfy). Under this approach, objects are represented by vectors of property values, or, equivalently, by points in a multidimensional space. Recognition then becomes a problem of clustering in this space (e.g. [36]).

A second approach to object recognition relies on the decomposition of objects into simple generic constituents. The components are generic in the sense that all objects can be described in their terms. For a structural approach to succeed, the components must be rapidly and reliably detected in any given view of an object, otherwise the problem of component identification tends to become as complex as object recognition itself. An analogous situation results if the breakdown of objects into components is too fine-grained, e.g. when the components are individual edge elements or lines. In that case, component detection is relatively easy, but their relationships become complicated.

Some older object recognition systems (e.g. [22]), and at least one psychological theory [14], describe objects in terms of elongated volumes called generalized cylinders ([15], [94]). The use of generalized cylinder primitives may be regarded as a compromise between the conflicting requirements of generality and detectability of object components. The description of an object under this representation scheme is said to be object-centered, in the sense that it involves the relations among the object's parts that do not depend on the viewer's position.

Variants of the decomposition approach that use contour-based primitive descriptions are relatively successful in certain domains, such as industrial part recognition. The two major limitations of this approach are that in many cases the precise shape of the object matters more than its decomposition into parts, and that many objects have no natural structural description.

A third major approach to object recognition, which may be termed normalization [114] or alignment [156], addresses the requirement for precise characterization of object shapes by employing a modified form of template matching. This approach may be illustrated using the simple example of the recognition of a 2D shape such as the letter A, whose size is larger than that of A's template stored in the system [109]. In this case, template matching can be still used, provided the input shape is first "normalized" by scaling it down to the standard size.

The normalization approach can be extended to the domain of 3D object recognition. To recognize a rigid 3D object, one may first normalize its appearance by undoing the transformations (such as 3D rotation) by which it differs from a stored model. Combined with algorithms that compute the necessary transformations, the normalization method has been recently applied to the recognition of natural objects from unconstrained viewpoints ([87], [156]). As the main idea of normalization is the ultimate use of template matching, 3D recognition schemes based on normalization typically involve 3D object-centered templates or models.

State of the art object recognition systems are typically based either on sophisticated search techniques ([47], [49]), or on variants of the normalization approach ([19], [88], [145], [68], [82]). These systems are mainly useful in tasks such as industrial part recognition, because of their reliance on strict geometrical models. In addition, search-based systems tend to perform poorly in cluttered environments, where segmenting the object from the background is especially difficult. A partial solution to the first shortcoming has been offered in the form of parametrized object models. For example, a pair of scissors can be recognized irrespective of the angle between the two blades, if this angle is considered as an additional dimension of the search space, along with the viewpoint parameters (this, of course, aggravates the complexity of the search). The problem of sensitivity to clutter and noise may in principle be approached through the use of distinctive labels for recognition primitives (edges, corners etc.). It is not clear, however, what kind of labeling would be easy enough to compute and at the same time informative enough to be useful in recognition. Another possibility here is to use middle vision techniques (section 2.2) to enhance the image before recognition is attempted. An additional problem, common to most recognition methods, is that of indexing: choosing a subset of models that is likely to include a potential match to the input, rather than trying to match the input to all known models. Finally, we remark that automatic acquisition of object models appears to be the main difficulty associated with those approaches that rely on 3D geometrical representation.

The two related questions, which recognition scheme is employed by the human visual system and what is the nature of visual object representation, are at present unresolved [65]. In this respect, our understanding of recognition lags considerably behind our understanding of low-level tasks such as motion detection and stereopsis. We discuss possible reasons for this situation in section 3.

2.4 Biophysics of computation

The possibility to separate abstract, or computational, from concrete, or implementational, aspects of visual perception is a central feature of the approach to the study of vision advocated by Marr and Poggio [96]. Experience of the last fifteen years suggests that this view is too idealized, and that the complexity of most problems in vision renders impractical algorithms that ignore constraints imposed by the available hardware. Moreover, the choice of low-level visual modules for which a computational theory is sought depends on the hardware. For example, depth can be recovered either through binocular stereo or through a laser rangefinder. The computational problems that need to be addressed in these two cases are clearly different.

One important source of differences between solution classes that are available to biological vision and those that best suit machine vision is the relevant computational primitives. Brains process large amounts of low-precision data at a relatively slow rate and in a highly parallel fashion, while most digital computers are serial, fast and can carry out high-precision arithmetic operations (see, e.g., [66], [136]). A gradual realization of the importance of hardware-related constraints led to an increasing cooperation between the fields of neuroscience and computational vision. One research goal in this field is to find out "what are the biophysical mechanisms underlying information processing and how are these mechanisms used to perform specific tasks" ([78], p.640).

Although the first detailed models of a neuron were proposed in the fifties [58], computational biophysics is still at an early stage of development. The study of neuronal function is now approached at different levels, from that of the biophysics and biochemistry of membranes to the level of ensembles of neurons. In particular, functional understanding is sought through anatomical and physiological investigation, tightly coupled with computer simulation (see [32], [79] for recent collections of papers in the field). Biochemical and electrical mechanisms have been invoked as explanations of phenomena such as retinal adaptation and spatial and temporal filtering. Models that involve networks of neurons with simple excitatory and inhibitory connections have been proposed, among other computational tasks in vision, for directional selectivity in the retina ([8], [148]), for the computation of a smooth velocity field [53], and for the winner-take-all, or maximum, operation [173].

3 Case studies

From the preceding section, it appears that at the low levels of the visual processing the developments in machine vision closely paralleled biological and psychological findings. In high level tasks such as recognition, most machine vision approaches bear less resemblance to their putative counterparts in biological systems. We identify radical differences between the nature of representation at the low and high levels of the human visual system as a possible cause of this distinction. We support this view by an analysis of three characteristic cases. The first case is the measurement of visual motion, a low-level vision process whose objective, the computation of a projected 2D velocity field, is well-defined in representational terms and has support in biological studies. The second case deals with the integration of low-level visual cues, a middle vision process whose biological reality has not yet been demonstrated. The discussion of integration illustrates some of the engineering aspects of vision research. The third case, recognition of three-dimensional objects, is an example of a field still in search for a computationally efficient, biologically plausible representation paradigm, and in which current engineering solutions do not seem to converge on human-like performance.

3.1 The measurement of visual motion: well-defined representations allow a principled solution

A standard formulation of the computation of visual motion distinguishes between two processing stages [57]. In the first stage, the movement in the changing 2D image is measured. In the second stage, motion estimates obtained through this measurement are used in different ways, e.g., in navigation, or in the recovery of the 3D layout of the environment.

The problem confronted by the visual system in the first stage of motion perception may be posed as the computation of a projected 2D velocity field, i.e., the assignment of a velocity vector to each feature in the image (alternative, more qualitative formulations are conceivable, but will not be discussed here). In this computation, the visual system must rely solely on the changes in the light intensity patterns projected on the retina. In general, many possible movements may give rise to the same changes in the retinal illumination. The situation is further aggravated if the measurement of visual motion is carried out by a mechanism which examines only a limited area of the image.

The ability of a limited-area motion detection mechanism to extract only partial information about the real 2D velocity field, called the aperture problem ([161], [98], [64]), may be illustrated by the following example. Consider an extended oriented pattern in the image, such as an intensity edge, moving behind a relatively small aperture, representing the limited area of the image analyzed by a motion detection mechanism. Because of the aperture, it is only possible to perceive the movement of the edge in the direction perpendicular to its orientation (figure 7). Some algorithms that incorporate smoothing (equivalent to using information from extended portions of the image) can solve this problem. An example of such an algorithm is given below. It should be noted that differential formulations of velocity field computation that yield a single equation per image point (e.g. [64]) are inherently sensitive to the aperture problem. A complete 2D velocity may be recovered, at the expense of an increased sensitivity to noise, by writing down second-order differential equations [158].

The output of a collection of limited-aperture motion detectors must be further processed to recover a better approximation to the true velocity field. According to Marr's approach, this can be done by constraining the solution to the recovery problem to comply with prior assumptions that reflect the physical nature of the problem. The assumption that the velocity field must be smooth ([98], [64]) proved to be a good compromise between physical reality and computational convenience. Hildreth [53] incorporated this assumption into the measurement of motion by formulating the computational problem as constrained minimization. The true velocity $\mathbf{V}$ was estimated by minimizing an error functional Θ that expressed a compromise between the requirement of smoothness and the compliance of the velocity component normal to the image contours with the measured data:

$$\Theta = \int \left[\left(\frac{\partial \mathbf{V}_x{}^2}{\partial s} \right) + \left(\frac{\partial \mathbf{V}_y{}^2}{\partial s} \right) \right] ds + \beta \int \left[\mathbf{V} \cdot \mathbf{u}^{\perp} - v^{\perp} \right]^2 ds$$

where $v^{\perp}$ is the measured velocity in the direction $\mathbf{u}^{\perp}$ perpendicular to the contour, β is a weighting factor that expresses the confidence in the measured velocity constraints, and the integral is taken along the image contours (in comparison, other formulations, e.g., [64], used

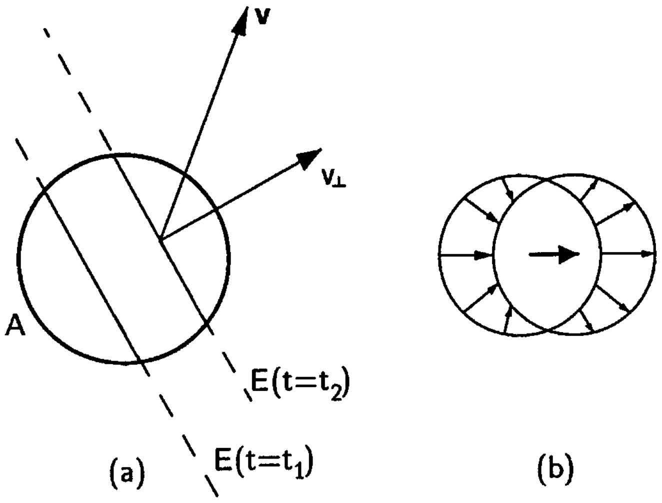

Figure 7: An illustration of the aperture problem. Left: a bar E is moving behind a small aperture A, so that its ends are not visible. Only the component $v_{\perp}$ of the bar's velocity that is perpendicular to the bar's orientation can be measured through the aperture. Right: an example of a circle translating to the right. The component perpendicular to the circle's contour, which would be computed by aperture-limited operators, does not correspond to the true motion of the circle.

area-based constraints). The first term in the above expression corresponds to an explicit imposition of the smoothness constraint, used to resolve the ambiguity illustrated in Figure 7.

Computing the maximally smooth velocity field consistent with the available evidence was shown by Hildreth [53] to model a number of aspects of the human motion perception mechanism, including visual illusions in which the algorithm and human vision fail in similar ways. (Algorithms that use area-based formulations, e.g., [172], are also compatible with psychophysical data.) Hildreth's algorithm was subsequently shown to belong to a general framework for approaching the so called inverse problems, of which the reconstruction of velocity field by aperture-limited detectors is a special case [122]. Following the theoretical analysis of the problem, Movshon et al. [106] discovered what could be the physiological correlates of the two successive stages of motion measurement: cells in cortical area V1 that are sensitive to the direction of motion in such a way that they can provide only the component perpendicular to image contours, and cells in area MT that appear to combine the outputs of simpler detectors to provide the true direction of motion in 2D.

3.2 Integration of low-level vision cues: refining low-level representations

Biological vision systems are more robust and flexible than any existing computer vision system, even in the relatively simple domain of low-level vision. One reason for this may be that biological vision combines information from different low-level cues to form richer and more robust intermediate representations. Since little is known about the nature of intermediate representation in human vision, attempts to integrate low-level modules tend to rely on engineering common sense and on mathematical tools such as probability theory.

The task of integrating low-level modules in computer vision encounters difficulties such as noisy data and differences between the output representations used by the individual modules. Most algorithms that attempt integration tend to deal only with modules that process related cues (e.g., stereo, vergence and focus [1]). As as example of a more comprehensive integration effort, we shall describe the MIT Vision Machine, a system, built around a parallel supercomputer, whose purpose is the integration of stereo, motion, color, texture, and intensity edge data ([120], [44]).

The two basic observations behind this particular approach to integration are (i) that physical edges, such as surface orientation discontinuities and object boundaries in the scene, cause the appearance of discontinuities in the output of at least some of the low-level visual modules, and (ii) that a generalized edge map would constitute a highly useful representation of a scene (consider, as an illustration, how informative can a mere cartoon of a scene be). Thus, combining information from different cues along edges, instead of over regions, may result in a computationally tractable and useful integrated representation. In many cases, the physical origin of the edges in this representation may also be deduced. For example, if it is known that along an edge there is no depth discontinuity (i.e., the stereo disparities are continuous) and no color discontinuity, it may be assumed that the edge is due to a shadow. A physical classification of the edges obtained through the integration of different cues may later be used in recognition and navigation.

Edge-based integration should start with a reliable map of discontinuities in the individual low-level cues. Since the output of most low-level modules by themselves is unreliable, one should choose the module whose performance is the most stable and consistent and use its output as the basis for the integration process. A good candidate for the basic representation is an intensity edge map, computed by the Canny algorithm [26]. Assuming that some of the other edges, corresponding to discontinuities in depth, color and texture, are also found by this edge detector, one can improve the reliability of the depth, color and texture maps by forcing discontinuities in these maps to align with the intensity edges found by the more reliable edge detector.

Since edge detection leads to noise amplification, unless the process includes sufficient smoothing, the above approach to integration faces a dilemma. If the smoothing is too strong, many of the discontinuities necessary for integration will be lost. Stereo disparities, for example, should in general be smoothed almost everywhere in the image, except at the boundaries between objects. The location of these is, of course, not known at the time the disparity map is computed, that is, before integration. Different techniques have been suggested to circumvent this problem, by performing a restricted smoothing ([144], [18]). The MIT Vision Machine uses for this purpose a statistical method, originally developed for image reconstruction ([45], [99]). This method employs Markov Random Fields (MRFs), a mathematical structure that specifies the probability for a certain parameter to attain a specific value at an image point in terms of the values of the parameter at the neighboring points. The value at a given site is allowed to affect its neighbor's value only if there is no discontinuity between the two. The placement of the discontinuities in this formalism is in itself statistical: the state of the system is allowed to fluctuate, with edges being turned on and off, until a stable configuration is obtained. This configuration is usually a less noisy version of the original image. Since the MRF technique uses an explicit discontinuity map, it appears well-suited for an edge-based integration system. The integration algorithm in such a system may include the following steps:

1. Apply different low-level vision algorithms, in parallel.

2. Smooth the output of each module with an MRF, constraining discontinuities to the locations of intensity edges, in parallel.
3. Use the discontinuities in the MRF of each module to label edges according to their physical origin, i.e., occluding boundaries, corners, shadow edges, specularity edges, color edges.
4. If necessary, return to step 2 and use the labeled edges to improve the quality of the output.

An example of the output of this system using real images is given in figure 8.

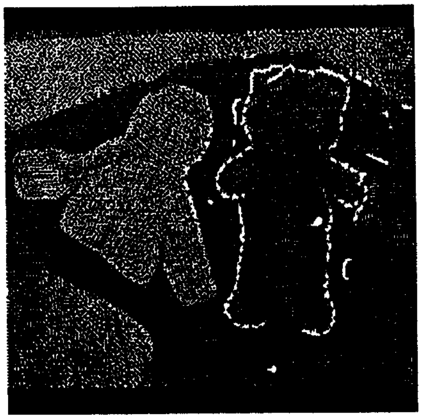

Figure 8: A cartoon-like representation of two objects segmented from the background using stereo, motion, color and texture cues.

At present, this approach still falls far short of achieving human-like performance in scene segmentation and in the classification of image edges. The major open problems in the MRF-based approach to integration are the poor initial output of some of the low-level modules and the need for a manual tuning of the MRF parameters. Furthermore, it is not clear whether a simple and qualitative rule-based approach would not perform better than formal approach of combining quantitative data.

3.3 Object recognition: looking for the right representation

It is difficult to find in the study of high-level vision a parallel to the achievements of the combined mathematical, psychological and physiological approach that advanced our understanding of low-level visual tasks such as motion and stereo. In low-level vision many physical assumptions have been found useful in dealing with the ill-posedness of inverse optics or the reconstruction problem. In comparison, in high-level problems such as object recognition relevant physical constraints are scarce and key computational issues, in particular the nature of representation, are highly controversial.

Consider the task of recognizing objects that belong to any fixed set, such as a collection of typefaces. The major problem confronted by the visual system in this case is that of noise. For example, machines can be trained to recognize complicated and diverse 2D objects such as printed characters that come in a variety of sizes and typefaces [76], with an efficiency approaching that of a human. In another example of a restricted domain, the interpretation of polyhedral scenes with shadows, relying on a complete enumeration of possible types of polyhedral junctions, has been demonstrated by Waltz as early as 1975 [163]. In both these cases, the finite domain places an effective constraint on the possible solutions.

When the shape of the objects is allowed to change continuously, but in a principled manner and without affecting their basically discrete and finite classification, some physical assumptions still apply and may be used to constrain the solution. In handwritten character recognition, for example, phenomenological knowledge of the physics of motor control in handwriting can be easily formulated in mathematical terms and leads to a simplification of the recognition process [39]. In comparison, when the problem is to recognize an arbitrary 3D shape from any possible viewpoint, physical constraints can be applied to compensate for only one source of image variability — the imaging process. The other part of the problem, the variability of the intrinsic 3D shape of the objects, remains unsolved.

In many recognition problems, the nearest conceivable thing that resembles a physical constraint seems too implausible or too impractical to be of any use. Consider, for example, face recognition, an important visual task in primates, and a challenging application of machine vision. A face recognition system that employs geometric models is bound to fail when confronted with a familiar face bearing a novel expression, even if a great number of templates corresponding to previously encountered expressions are stored. Applying the physical constraint method in this case amounts to the inclusion of the knowledge of facial and cranial anatomy in the recognition algorithm, an improbable requirement. It just might turn out that a feasible face recognition system would have to rely on stored examples of familiar faces bearing prototypical expressions more than on an algorithmic approach of computing some standard face representation using anatomical constraints.

Even if standard representations, or models, of 3D objects are involved in recognition, it is not clear what method of modeling is computationally most efficient, and what method (or methods) is employed in human vision. A major point of controversy as to the nature of 3D object representation regards the question of whether it should be object-centered (symmetric with respect to the object itself), as advocated by Marr, or viewer-centered (dependent upon the viewer's position relative to the object). Although in principle a 3D object-centered representation can be readily constructed from a collection of $2\frac{1}{2}$D viewer-centered ones, in practice building a 3D object-centered model only to use it later by comparing its 2D projections with input images ([87], [145], [156]) seems to be redundant. Similarly, it appears equally unlikely that the human visual system goes through the effort of constructing computationally unwarranted 3D models, or, alternatively, that it necessarily reconstructs the third dimension of an input object before recognizing it (after all, we do recognize everyday objects in visually impoverished line drawings).

Humans construct their library of object models through experience. It appears that 3D machine recognition systems that use object-centered models fail to match this crucial component of the human performance in recognition. The few 3D machine recognition systems that are designed to learn object representations from examples ([70], [151], [119]) employ viewer-based rather than object-based representation. Some evidence to the effect that the ultimate representation need not be object-centered is already available. This evidence falls into three classes, psychophysical, physiological and computational, discussed separately below.

3.3.1 Hints from psychophysics

The most straightforward way to investigate the nature of object representation in long-term memory is to test it in a recognition task. An object-centered representation that does not allow the system to infer what the object would look like from an arbitrary viewpoint is, for all practical purposes, equivalent to a collection of viewer-centered view-specific representations. Consequently, if the human visual system effectively employs object-centered representations, a person should be able to recognize an object, previously seen from a limited range of viewpoints, from a novel viewpoint.

It should be noted that the structure from motion theorems ([152], [150], [86]) indicate that in principle the effect of having a full object-centered description of a 3D rigid body may be produced by actually storing only a few of its 2D views. The ability of the human visual system to infer the 3D structure of an object from a small number of its projections, termed the kinetic depth effect, has been known for a long time ([162], [71]). An alternative way to formulate the object-centered representation question is, therefore, to ask whether the 3D structure perceived in the kinetic depth effect is retained by the long-term memory, or whether this effect is a mere by-product of some other, more basic, perceptual operation.

Until recently, the ability of the human visual system to recognize objects from novel viewpoints has been taken for granted. Palmer, Rosch and Chase were the first to demonstrate that even for familiar objects speed and accuracy of recognition varies with viewpoint [113]. A more direct test would involve novel objects, shown to the subjects under controlled conditions which guarantee that some views remain unseen until the test time. Such experiments, involving novel 3D wire-like stimuli, have been carried out by Rock and his collaborators ([134], [135]). In one series of experiments [134], subjects saw novel 3D wires, each from a single fixed viewpoint. The subjects were subsequently shown other, similar, objects, one at a time, and required to decide whether these were the familiar wires, displayed at new attitudes, or unfamiliar wires. In these experiments, the subjects' performance approached chance level when the in-depth rotational distance between training and test views was about 30°. In another experiment [135], the test called for deciding whether one of two simultaneously displayed wire objects was a replica of the other object, shown at a different attitude. Again, subjects performed poorly, even when given the opportunity to reason explicitly about the relative positions of different features of the two objects. As during training the subjects perceived the stimuli in 3D (due to binocular stereopsis), the lack of generalization to novel views in this experiment could be attributed to the subjects' failure either to retain 3D information, or to translate it into a format that would permit later generalization, e.g., into an object-centered description. The likelihood of the first interpretation, namely, that subjects do not include 3D information in a long-term memory representation of wire-like objects, is diminished by the finding that the presence of binocular stereo cues significantly reduces recognition error rate ([38]; see also [65], p.81).

Additional evidence in favor of the hypothesis that objects are normally represented in long-term memory by collections of $2\frac{1}{2}$D viewpoint-specific descriptions comes from the experiments of Tarr and Pinker [143], who found that naming time for 3D objects similar to the ones used by Shepard and Metzler [139] increased linearly with the distance between the presented view and the closest learned view. This linear dependence, present in a wide variety of recognition tasks (cf. [113], [37]), is predicted by a theory that posits multiple-view representations, but is difficult to explain within the framework of a theory of recognition of the viewpoint normalization variety ([87], [156]). For example, Ullman's alignment scheme

([156]; see section 2) specifies an algorithm that computes the hypothesized viewpoint of an object model, given the positions of a few key model features in the input image. When combined with a 3D object-centered representation of the model, this scheme should result in recognition time that is independent of the viewpoint (unless implementational constraints bring about such dependence, in which case the representation effectively ceases to be symmetric with respect to the object, i.e. object-centered).

3.3.2 Hints from physiology

The notion that the primate visual system employs view-specific representations of objects is supported by the findings of Gross, Perrett and others ([51], [117], [50], [116]) that cells in the visual area IT in monkey respond preferentially to complex stimuli such as hands and faces. Perrett et al. [116] reported that some cells responded maximally to full views of a face, others to the face tilted upwards or downwards by 45°, and still others to profile and back views of the head. The cells' response remained strong when the preferred view of the face was displayed at different scales, or rotated within the image plane by as much as 90°. If indeed these cells are at the top of the representation hierarchy in the visual system, then their response pattern is most easily explained in terms of viewer-centered rather than object-centered representations.

3.3.3 Computational considerations

Recent computational studies provide further support for the hypothesis that a recognition strategy based on multiple-view representation can in principle account for much of the human performance in object recognition ([40], [11], [119]). In particular, if the recognition problem is formulated in terms of the approximation of a mapping that associates a standard view of an object with any of its other views, powerful mathematical tools from function approximation theory are available that can construct such a mapping from examples [121]. In other words, a system can learn to recognize an object from any viewpoint merely by being exposed to a random set of a few tens of the object's views.

4 Discussion and prognosis

During the last decade, the combined computational and biological study of vision, pioneered by Marr, has resulted in progress in some areas of vision research, notably, in understanding low-level vision. The computational approach proved most fruitful when the problem at hand could be formulated as a transformation between clearly defined representations. In other words, principled computational solutions were first offered to those problems for which a notion of competence (in Chomsky's sense; see [93], p.28) was readily available. One such problem was binocular stereo, for which both the input representation (edge maps of the two images) and the output representation (a dense depth map) seemed to be obvious.

The computational paradigm made an important contribution to the study of vision by encouraging the generation of concrete and testable theories of performance. In many cases, these theories eventually led to a revision of basic assumptions of the underlying theory of competence. For example, when even the sophisticated stereo algorithms appeared to be incomplete as models of human performance, the nature of the input and the output representations in stereo came under questioning.

The computational approach has also increased the understanding of the difficulties associated with high-level tasks, such as object recognition. Despite early enthusiasm, it now becomes increasingly clear that no adequate competence-level theory is yet available for these tasks. In particular, computational feasibility and biological (psychological and neurophysiological) plausibility of 3D object-centered models as the ultimate representations in object recognition is now being questioned. From the preceding review, it appears that a satisfactory reconstruction of the visual world is not feasible, unless its aim is a viewer-relative and qualitative (as opposed to absolute and quantitative) representation. Fortunately, it also appears that carrying out an ideal reconstruction is not a prerequisite for seeing, insofar as it does not seem that biological vision relies on such reconstruction.

The roots of the present situation in computational vision may be traced to the philosophical foundations of the currently accepted computational paradigm in cognitive science. An important part of Marr's argument in favor of the computational approach to vision has been functionalist. The functionalist program in the study of mind [125] has been prompted by the "suspicion that there are empirical generalizations about mental states that cannot be formulated in the vocabulary of neurological or physical theories" ([42], p.25). Glossing over technical details, a common version of the functionalist approach ([126], [127]) holds that two intelligent agents are in the same psychological state (e.g., both believe that they see a cat) if they are in the same computational state. In this sense, the two agents function similarly, although their physical realizations may be different. Similarly, Marr and Poggio suggested that functional, or computational, understanding of vision may be achieved in separation from the understanding of its physical substrate, just as it is sufficient to study aerodynamics to understand flight in both birds and airplanes ([93], p.345).

The advantages (and the feasibility) of the postulate of level separation as a research strategy in vision seem now less substantial than a few years ago. At the same time, functionalist theories of cognition are increasingly criticized from at least two points of view. One line of argument starts with the claim that the functional level is not sufficiently abstract to allow interesting generalizations about mental states ([127], p.74). Others claim that any such generalizations are, in principle, unwarranted ([33], p.224; see also [129]), or depend on a better understanding of the physical substrate of cognition [30]. A speedy resolution of the uncertainty concerning the philosophical foundations of the functionalist program appears at the present time unlikely.

Whereas a discussion of level separation in computational vision leads one to question the basic premises of the functionalist program, another problematic issue mentioned above, that of visual reconstruction as a goal of vision, can be addressed within the present paradigm. Persistent difficulties with solving the inverse optics problem (reconstructing the geometry of visible objects) prompt the search for a different competence theory of vision, one that is formulated in terms of more plausible representations than geometric models. Many of the emerging research directions can be described as an attempt to make the most of the rich intermediate representation provided by the low-level modules (see the preface to [132]). These approaches focus on qualitative rather than quantitative representations (e.g. [146], [168]), on active vision ([110], [7]), on the application of high-level rules to obtain fast approximate solutions [131] and on massive use of parallelism, memory and learning ([120], [151], [119]).

Acknowledgements

We are grateful to Ellen Hildreth, Norberto Grzywacz, Tomaso Poggio, Manfred Fahle and Anya Hurlbert for many useful comments and discussions.

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] Y. S. Abu-Mostafa and D. Psaltis. Optical neural computing. *Scientific American*, 256:66–73, 1987.

[3] J. Y. Aloimonos. Unification and integration of visual modules. In *Proceedings Image Understanding Workshop*, pages 507–551, San Mateo, CA, 1989. Morgan Kaufmann Publishers, Inc.

[4] A. Ames. Visual perception and the rotating trapezoid window. *Psychological Monographs*, 65(7), 1951.

[5] R. Bajcsy and L. Lieberman. Texture gradient as a depth cue. *Computer Graphics and Image Processing*, 5:52–67, 1976.

[6] D. H. Ballard and C. M. Brown. *Computer Vision.* Prentice-Hall, Englewood Cliffs, NJ, 1982.

[7] D. H. Ballard, R. C. Nelson, and B. Yamauchi. Animate vision. *Optic News*, 15:9–25, 1989.

[8] H. B. Barlow and R. W. Levick. The mechanism of directional selectivity in the rabbit's retina. *J. Physiol.*, 173:477–504, 1965.

[9] S. T. Barnard and M. A. Fischler. Computational stereo. *ACM Comput. Surveys*, 143:553–572, 1982.

[10] H. G. Barrow and J. M. Tenenbaum. Computational vision. *Proc. IEEE*, 69:572–595, 1981.

[11] R. Basri and S. Ullman. Recognition by linear combinations of models, June 1989. forthcoming MIT AI Memo.

[12] J. Beck. Textural segmentation. In J. Beck, editor, *Organization and representation in perception*, chapter 15. Erlbaum, Hillsdale, NJ, 1982.

[13] J. Beck, K. Prazdny, and A. Rosenfeld. A theory of textural segmentation. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision*, pages 1–38. Academic Press, New York, NY, 1983.

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

[15] T. O. Binford. Visual perception by computer. In *IEEE Conference on Systems and Control*, Miami Beach, FL, Dec. 1971.

[16] A. Blake and G. Brelstaff. Geometry from specularities. In *Proceedings of the 2nd International Conference on Computer Vision*, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[17] A. Blake and H. H. Bülthoff. Does the brain know physics? perception of shape from specularity, 1989. to appear.

[18] A. Blake and A. Zisserman. *Visual reconstruction.* MIT Press, Cambridge, MA, 1988.

[19] R. C. Bolles and P. Horaud. 3DPO: A three-dimensional part orientation system. *International Journal of Robotics Research*, 5:3–26, 1986.

[20] O. J. Braddick. Low-level and high-level processes in apparent motion. *Phil. Trans. R. Soc. London B*, 290:137–151, 1980.

[21] M. J. Brady. Computational approaches to image understanding. *ACM Computing Surveys*, 14:3–71, 1982.

[22] R. A. Brooks. Symbolic reasoning among 3D models and 2D images. *Artificial Intelligence*, 17:285–348, 1981.

[23] A. Bruss and B. K. P. Horn. Passive navigation. *Computer Vision, Graphics, and Image Processing*, 21:3–20, 1983.

[24] G. Buchsbaum and A. Gottschalk. Chromaticity coordinates of frequency-limited functions. *Journal of the Optical Society of America*, 1:885–887, 1984.

[25] H. H. Bülthoff and H. A. Mallot. Interaction of different modules in depth perception. In *Proceedings of the 1st International Conference on Computer Vision*, pages 295–305, June 1987.

[26] J. F. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:679–698, 1986.

[27] P. Cavanagh. Reconstructing the third dimension: interactions between color, texture, motion, binocular disparity and shape. *Computer Vision, Graphics, and Image Processing*, 37:171–195, 1987.

[28] R. Chellappa, R. Chatterjee, and R. Baghdazian. Texture synthesis and coding using Gaussian Markov field models. *IEEE Trans. SMC*, 15:298–303, 1985.

[29] S. S. Chen and M. Penna. Shape and motion of nonrigid bodies. *Computer Vision, Graphics, and Image Processing*, 36:175–207, 1986.

[30] P. S. Churchland. *Neurophilosophy.* MIT Press, Cambridge, MA, 1987.

[31] W. F. Clocksin. Perception of surface slant and edge labels from optical flow: a computational approach. *Perception*, 9:253–269, 1980.

[32] R. M. J. Cotterill, editor. *Computer simulation in brain science.* Cambridge Univ. Press, Cambridge, 1988.

[33] D. Davidson. *Essays on actions and events.* Clarendon Press, Oxford, 1980.

[34] L. S. Davis. A survey of edge detection techniques. *Computer Graphics and Image Processing*, 4:248–270, 1975.

[35] R. Desimone, S. J. Schein, J. Moran, and L. G. Ungerleider. Contour, color and shape analysis beyond the striate cortex. *Vision Research*, 25:441–452, 1985.

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

[37] S. Edelman, H. Bülthoff, and D. Weinshall. Stimulus familiarity determines recognition strategy for novel 3d objects. A.I. Memo No. 1138, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, July 1989.

[38] S. Edelman and H. H. Bülthoff. Recognition of novel 3D objects in human vision, 1989. submitted.

[39] S. Edelman, S. Ullman, and T. Flash. Reading cursive handwriting by alignment of letter prototypes, 1989. submitted for publication.

[40] S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3d objects. A.I. Memo No. 1146, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, August 1989.

[41] M. A. Fischler and O. Firschein, editors. *Readings in computer vision: issues, problems, principles and paradigms.* Morgan Kaufmann, Los Altos, CA, 1987.

[42] J. A. Fodor. *RePresentations.* MIT Press, Cambridge, MA, 1981.

[43] D. Forsyth and A. Zisserman. Mutual illumination. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 466–473, San-Diego, CA, 1989.

[44] E. Gamble, D. Geiger, T. Poggio, and D. Weinshall. Labeling edges and the integration of low-level visual modules. *IEEE Trans. SMC*, 19(6), 1989.

[45] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[46] J. J. Gibson. *The perception of the visual world.* Houghton Mifflin, Boston, MA, 1950.

[47] C. Goad. Fast 3D model-based vision. In A. P. Pentland, editor, *From pixels to predicates*, pages 371–391. Ablex, Norwood, NJ, 1986.

[48] W. E. L. Grimson. *From Images to Surfaces.* MIT Press, Cambridge, MA, 1981.

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

[50] C. G. Gross and M. Mishkin. The neural basis of stimulus equivalence across retinal translation. In S. Harnad, R. W. Doty, L. Goldstein, J. Jaynes, and G. Krauthamer, editors, *Lateralization in the nervous system.* Academic Press, New York, NY, 1977.

[51] C. G. Gross, C. E. Rocha-Miranda, and D. B. Bender. Visual properties of cells in inferotemporal cortex of the macaque. *J. Neurophysiol.*, 35:96–111, 1972.

[52] R. M. Haralick. Digital step edges from zero crossings of second directional derivatives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:58–68, 1984.

[53] E. C. Hildreth. *The measurement of visual motion.* MIT Press, Cambridge, MA, 1984.

[54] E. C. Hildreth. Edge detection. In S. Shapiro, editor, *Encyclopedia of artificial intelligence*, pages 257–267. John Wiley, New-York, NY, 1987.

[55] E. C. Hildreth, N. M. Grzywacz, E. H. Adelson, and V. K. Inada. The perceptual buildup of three-dimensional structure from motion. A.I. Memo No. 1141, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[56] E. C. Hildreth and C. Koch. The analysis of visual motion: from computational theory to neuronal mechanisms. *Ann. Rev. Neurosci.*, 10:477–533, 1987.

[57] E. C. Hildreth and S. Ullman. The computational study of vision. A.I. Memo No. 1038, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1988.

[58] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. Lond.*, 116:500–544, 1952.

[59] W. Hoff and N. Ahuja. Extracting surfaces from stereo images: An integrated approach. In *Proceedings of the 1st International Conference on Computer Vision*, pages 284–294, June 1987.

[60] D. D. Hoffman and B. E. Flinchbaugh. The interpretation of biological motion. *Biological Cybernetics*, 42:195–204, 1982.

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

[62] B. K. P. Horn. *Robot vision.* MIT Press, Cambridge, Mass., 1986.

[63] B. K. P. Horn and M. Brooks. *Seeing shape from shading.* MIT Press, Cambridge, Mass., 1989.

[64] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[65] G. W. Humphreys and P. Quinlan. Normal and pathological processes in visual object constancy. In G. W. Humphreys and M. J. Riddoch, editors, *Visual object processing: a cognitive neuropsychological approach*, pages 43–106. Erlbaum, Hillsdale, NJ, 1987.

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

[67] A. C. Hurlbert, H.-C. Lee, and H. H. Bülthoff. Cues to the color of the illuminant. *Invest. Ophthalm. Vis. Science Suppl.*, 30:221, 1989.

[68] 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.

[69] K. Ikeuchi and B. K. P. Horn. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, 15:141–184, 1981.

[70] 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.

[71] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.

[72] P. Jolicoeur, S. Ullman, and M. Mackay. Curve tracing: a possible basic operation in the perception of spatial relations. *Memory and Cognition*, 14:129–140, 1986.

[73] B. Julesz. *Foundations of Cyclopean perception.* University of Chicago Press, Chicago, IL, 1971.

[74] B. Julesz. Experiments in the visual perception of texture. *Scientific American*, 232:34–43, 1975.

[75] B. Julesz. A brief outline of the texton theory of human vision. *Trends in Neurosciences*, 7:41–45, 1984.

[76] S. Kahan, T. Pavlidis, and H. S. Baird. On the recognition of printed characters of any font and size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9:274–287, 1987.

[77] E. R. Kandel and J. H. Schwartz. *Principles of neural science.* Elsevier, New York, 1985.

[78] C. Koch and T. Poggio. Biophysics of computational systems: Neurons, synapses, and membranes. In G. M. Edelman, W. E. Gall, and W. M. Cowan, editors, *Synaptic Function*, pages 637–697. Wiley, New York, NY, 1987.

[79] C. Koch and I. Segev. *Methods in neuronal modeling.* MIT Press, Cambridge, MA, 1989.

[80] J. J. Koenderink. What does the occluding contour tell us about solid shape? *Perception*, 13:321–330, 1984.

[81] J. J. Koenderink and A. J. van Doorn. Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America*, 3:242–249, 1986.

[82] 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.

[83] E. H. Land. An alternative technique for the computation of the designator in the retinex theory of color vision. *Proceedings of the National Academy of Science*, 83:3078–3080, 1986.

[84] E. H. Land and J. J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971.

[85] H.-C. Lee. Method for computing the scene-illuminant chromaticity from specular highlights. *Journal of the Optical Society of America*, 3:1694–1699, 1986.

[86] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

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

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

[89] L. T. Maloney. *Computational approaches to color vision.* PhD thesis, Stanford Univ., Stanford, CA, 1984.

[90] L. T. Maloney and B. Wandell. A computational model of color constancy. *Journal of the Optical Society of America*, 1:29–33, 1986.

[91] D. Marr. Analysis of occluding contour. *Proc. R. Soc. Lond. B*, 197:441–475, 1976.

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

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

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

[95] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.

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

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

[98] D. Marr and S. Ullman. Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London B*, 211:151–180, 1981.

[99] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82:76–89, 1987.

[100] J. E. W. Mayhew and J. P. Frisby. Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence*, 17:349–386, 1981.

[101] G. G. Medioni and R. Nevatia. Segment-based stereo matching. *Computer Vision, Graphics, and Image Processing*, 31:2–18, 1985.

[102] M. Mishkin, L. G. Ungerleider, and K. A. Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 4:414–417, 1983.

[103] G. J. Mitchison and S. P. McKee. Interpolation in stereoscopic matching. *Nature*, 315:402–404, 1985.

[104] A. Mitiche. On kineopsis and computation of structure and motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:109–112, 1986.

[105] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.

[106] J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome. The analysis of moving visual patterns. In C. Chagas, R. Gattas, and C. G. Gross, editors, *Pattern Recognition Mechanisms.* Vatican Press, Rome, 1985.

[107] H.-H. Nagel and W. Enkelmann. Towards the estimation of displacement vector fields by 'oriented smoothness' constraints. In *Proceedings Int. Conf. on Pattern Recognition*, pages 6–8, Montreal, Canada, July 1984.

[108] K. Nakayama and G. H. Silverman. The aperture problem ii: spatial integration of velocity information along contours. *Vision Research*, 28:739–746, 1988.

[109] U. Neisser. *Cognitive Psychology.* Appleton-Century-Crofts, New York, NY, 1967.

[110] 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.

[111] H. K. Nishihara. Practical real-time imaging stereo matcher. *Optical Engineering*, 23(5):536–545, 1984.

[112] A. Paivio. The relationship between verbal and perceptual codes. In E. C. Carterette and M. P. Friedman, editors, *Handbook of Perception*, pages 375–397. Academic Press, New York, NY, 1978.

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

[114] S. E. Palmer. The psychology of perceptual organization: a transformational approach. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and machine vision*, pages 269–340. Academic Press, New York, 1983.

[115] A. Pentland. Shape information from shading: a theory about human perception. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 404–413, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[116] D. I. Perrett, A. J. Mistlin, and A. J. Chitty. Visual neurones responsive to faces. *Trends in Neurosciences*, 10:358–364, 1989.

[117] D. I. Perrett, E. T. Rolls, and W. Caan. Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.*, 47:329–342, 1982.

[118] G. F. Poggio and T. Poggio. The analysis of stereopsis. *Ann. Rev. Neurosci.*, 7:379–412, 1984.

[119] T. Poggio and S. Edelman. A network that learns to recognize 3d objects, 1989. submitted for publication.

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

[121] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

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

[123] K. Prazdny. On the information in optical flow. *Computer Vision, Graphics, and Image Processing*, 22:239–259, 1983.

[124] K. Prazdny. Detection of binocular disparities. *Biological Cybernetics*, 52:93–99, 1985.

[125] H. Putnam. Minds and machines. In S. Hook, editor, *Dimensions of mind.* New York University Press, New York, NY, 1960.

[126] H. Putnam. *Mind, language and reality.* Cambridge University Press, Cambridge, 1975.

[127] H. Putnam. *Representation and reality.* MIT Press, Cambridge, MA, 1988.

[128] Z. Pylyshyn. The role of location indexes in spatial perception: a sketch of the finst spatial-index model. *Cognition*, 32:65–97, 1989.

[129] W. V. O. Quine. *Word and object.* MIT Press, Cambridge, MA, 1960.

[130] P. Quinlan. Visual object recognition reconsidered, 1989. submitted for publication.

[131] V. S. Ramachandran. Perception of shape from shading. *Nature*, 331:163–166, 1988.

[132] W. Richards, editor. *Natural computation.* MIT Press, Cambridge, MA, 1988.

[133] J. Richter and S. Ullman. A model for the temporal organization of X– and Y–type receptive fields in the primate retina. *Biological Cybernetics*, 43:127–145, 1985.

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

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

[136] J. Schwartz. The new connectionism. *Proc. AAAS*, 117:123–141, 1988.

[137] A. Sha'ashua 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.

[138] S. A. Shafer and T. Kanade. Using shadows in finding surface orientation. *Computer Vision, Graphics, and Image Processing*, 22:145–176, 1983.

[139] R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

[140] A. Sloman. What are the purposes of vision? CSRP 066, University of Sussex, 1987.

[141] K. Stevens. The visual interpretation of surface contours. *Artificial Intelligence*, 17:47–75, 1981.

[142] J. Stone and B. Dreher. Parallel processing of information in the visual pathways. *Trends in Neurosciences*, 3:441–446, 1982.

[143] M. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21, 1989.

[144] D. Terzopoulos. Regularization of inverse visual problems involving discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:413–424, 1986.

[145] 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.

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

[147] J. T. Todd and E. Mingolla. Perception of surface curvature and direction of illumination from patterns of shading. *J. Exp. Psychol.: HPP*, 9:583–595, 1983.

[148] V. Torre and T. Poggio. A synaptic mechanism possibly underlying directional selectivity to motion. *Proc. R. Soc. Lond. B*, 202:409–416, 1978.

[149] V. Torre and T. Poggio. On edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:147–163, 1986.

[150] R. Tsai and T. Huang. Uniqueness and estimation of three dimensional motion parameters of rigid objects with curved surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:13–27, 1984.

[151] L. W. Tucker, C. R. Feynman, and D. M. Fritzsche. Object recognition using the Connection Machine. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 871–878, Ann Arbor, MI, 1988.

[152] S. Ullman. *The interpretation of visual motion.* MIT Press, Cambridge, MA, 1979.

[153] S. Ullman. Computational studies in the interpretation of structure and motion: summary and extension. In J. Beck, B. Hope, and A. Rosenfeld, editors, *Human and Machine Vision.* Academic Press, New York, 1983.

[154] S. Ullman. Maximizing rigidity: the incremental recovery of 3D structure from rigid and rubbery motion. *Perception*, 13:255–274, 1984.

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

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

[157] L. Vaina and N. M. Grzywacz. Structure from motion with impaired local-speed and global motion-field computations, 1989. submitted.

[158] 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.

[159] R. von der Heydt, E. Peterhans, and G. Baumgartner. Illusory contours and cortical neurons' responses. *Science*, 224:1260–1262, 1984.

[160] H. Voorhees and T. Poggio. Computing texture boundaries from images. *Nature*, 333:364–367, 1988.

[161] H. Wallach. On perceived identity: 1. the direction of motion of straight lines. In H. Wallach, editor, *On Perception*. Quadrangle, New York, 1976.

[162] H. Wallach and D. N. O'Connell. The kinetic depth effect. *J. Exp. Psychol.*, 45:205–217, 1953.

[163] D. L. Waltz. Understanding line drawings of scenes with shadows. In P. Winston, editor, *The Psychology of Computer Vision*. McGraw-Hill, New York, 1975.

[164] R. J. Watt and M. J. Morgan. Spatial filters and the localization of luminance changes in human vision. *Vision Research*, 24:1387–1397, 1984.

[165] A. M. Waxman and S. Ullman. Surface structure and 3D motion from image flow: a kinematic analysis. *International Journal of Robotics Research*, 4:72–94, 1985.

[166] D. Weinshall. Direct computation of 3d shape and motion invariants. A.I. Memo No. 1131, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, May 1989.

[167] D. Weinshall. Seeing 'ghost' solutions in stereo vision. *Nature*, 1989. submitted for publication.

[168] D. Weinshall. Qualitative depth from stereo, with applications. *Computer Vision, Graphics, and Image Processing*, in press, January 1990.

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

[170] R. J. Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19:139–144, 1980.

[171] Y. Yeshurun and E. L. Schwartz. Cepstral filtering on a columnar image architecture: a fast algorithm for binocular stereo segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, July 1989.

[172] A. L. Yuille and N. M. Grzywacz. A computational theory for the perception of coherent visual motion. *Nature*, 333:71–74, 1988.

[173] A. L. Yuille and N. M. Grzywasz. A winner-take-all mechanism based on presynaptic inhibition feedback, 1989. in press.

[174] S. Zeki and S. Shipp. The functional logic of cortical connections. *Nature*, 335:311–317, 1989.

CS-TR Scanning Project
Document Control Form

Date : 12/20/94

Report # AIM-1158

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
□ Laboratory for Computer Science (LCS)

Document Type:

□ Technical Report (TR) ☒ Technical Memo (TM)
□ Other:

Document Information

Number of pages: 35 (41-IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
□ Double-sided

Intended to be printed as :
□ Single-sided or
☒ Double-sided

Print type:
□ Typewriter □ Offset Press ☒ Laser Print
□ InkJet Printer □ Unknown □ Other:

Check each if included with document:

☒ DOD Form 2 (PGS) □ Funding Agent Form □ Cover Page
□ Spine □ Printers Notes □ Photo negatives
□ Other:

Page Data:

Blank Pages (by page number):

Photographs/Tonal Material (by page number): 2, 8, 9, 19

Other (note description/page number):

Description : Page Number:

(A) XEROX MARKS IN LEFT MARGINS
(B) IMAGE-MAP (1) UNNUMBERED TITLE PAGE
(2-35) PAGES #'ED 1-34
(36) SCANCONT
(37-39) TRGT'S
(40-41) DOD'S

Scanning Agent Signoff:
Date Received: 12/20/94 Date Scanned: 1/11/95 Date Returned: 1/12/95

Scanning Agent Signature: Michael W. Cook

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
AIM 1158

2. GOVT ACCESSION NO.
AD-A216713

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Computational vision: a critical review.

5. TYPE OF REPORT & PERIOD COVERED
memorandum

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Shimon Edelman
Daphna Weinshall

8. CONTRACT OR GRANT NUMBER(s)
DACA76-85-C-0010
N00014-85-K-0124

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

12. REPORT DATE
Oct. 89

13. NUMBER OF PAGES
34

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Office of Naval Research
Information Systems
Arlington, VA 22217

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

computational vision
review

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We review the progress made in computational vision, as represented by Marr's approach, in the last fifteen years. First, we briefly outline computational theories developed for low, middle and high vision. We then discuss in more detail solutions proposed to three representative problems in vision, each dealing with a different level of visual processing. Finally, we discuss modifications to the currently established computational paradigm that appear to be dictated by the recent developments

(cont. on back)

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 cont.

in vision.