Three-Dimensional Object Recognition Systems
A.K. Jain and P.J. Flynn (Editors)

3D Object Recognition: Inspirations and Lessons from Biological Vision

Terry Caelli, Michael Johnston and Terry Robison[a]

[a] Department of Computer Science
The University of Melbourne
Parkville, Vic. 3052
Australia

1. Introduction

Visual object recognition may be defined as the process by which image data is referenced to known descriptions of objects. Consequently, models of both machine and human object recognition involve the extraction and interpretation of image features which can index 3D world structures from what is sensed. However, "visual systems" are restricted to the sensing and processing of information which can be displayed as 2D projections, possibly varying over time, and also restricted by specific constraints such as the lack of transparency. Such restrictions on acceptable representations for vision systems prohibit, for example, higher dimensional analysis, seeing "inside" without a view, and specific topologies. It does not restrict vision to the study of images produced by passive means nor exclude images generated by artificial or active sensors (for example, ultra-sound, infra-red or range sensors). Any object recognition system (ORS) which claims either to describe human object recognition, or to be useful in machine object recognition applications must include the following properties:

- It must be possible to recognize objects with some degree of view invariance.
- The data structure for the stored description of objects should be such that it is possible to access it through partial information as, for example, with single views, degraded, or partially occluded object data.
- A description must be general enough to include any known stimulus within the object class to which it refers, and specific enough to exclude other stimuli.

Theories of machine object recognition are inherently algorithmic while physiological and psychophysical descriptions are typically more qualitative, often due to practical difficulties in the observation of brain structure and function but, sometimes, also due to what the biological vision (BV) scientific community is prepared to accept as "theory". That is, most theories in biological vision do not consist of *algorithms* which actually predict behavior in explicit detail but, rather, are *descriptions* of encoding and information processing.

Machine vision (MV) provides representations and algorithms for solving 3D vision problems (e.g. Jain and Hoffman [1], Fan, Medioni and Nevatia [2]). However, these theories are not aimed at satisfying biological constraints. Rather, they are typically developed as a result of current applications of engineering technologies and preconceived notions of just what constitutes an 'efficient' solution. Many ORS's developed in the MV tradition have not been adequately evaluated and often lack the generality of a biologically-based ORS. Further, there is no standards for quantifying the performance machine ORS's - at least up to recently.

Experimental research on human object recognition has focussed mainly on subprocesses and general qualitative descriptions rather than providing complete computational models. Those "models" which do exist are typically undetailed and lack algorithmic description (e.g. Triesman [3], Biederman [4]). Because of the complexity and abstractness of the psychological approach to object recognition, it has been easier to investigate it in a piecemeal fashion. To preempt the conclusion of this chapter, the MV conceptual analyses of how we *understand* the processes of recognition in machines, may provide BV with questions and tools which will allow for a truly explicit computational theory of BV to evolve. From a BV perspective, MV theories may provide a theoretical outline which must be translated into the language of experimental psychology and specified in ways which allow for psychophysical experimentation. The thesis of this chapter is that BV and MV can mutually learn from each other and, in particular, MV offers the theory necessary for BV to ask appropriate questions of biological information processing.

Indeed, one can well argue that in many areas of biological inquiry, successful languages and models are borrowed from more formal areas of science: Mathematics, Physics, Engineering and Artificial Intelligence. Perhaps the most relevant example of this, in BV, is that of the concept of a *receptive field* (RF). This is the notion that we can describe the information processing characteristics of a nerve cell in terms of the stimulus which best activates it. RF's were first delimited in the mammalian retina by Ratliff and Hartline [5] during a period when the formulation of filter theory, adaptive filters, and their underlying control system (network) formalization were taking place. That is, the formal definition and analyses of such structures gave BV scientists the "tools" for representing what **may** be processed at a given level of function. This notion of filtering became very popular in the 1960's to the present in most areas of BV since it fit well with the notion that the activity of individual neurons is sufficient to describe visual processing - the "Neuron Doctrine" [6]. The notion that BV systems do filtering has, in turn, motivated many engineers and computer scientists to construct pre-processing procedures which incorporate such mechanisms. However, the latter is done less critically than one would expect and so there are many misconceptions in the MV community about what we really know about BV. For example, it is by no means proven that there are only four, six or even a fixed number of frequency bands of the signal that the visual system is selectively sensitivity to. It is certainly not the case that the system is insensitive to phase nor even that the global Fourier transform is computed by the biological vision system [7,8].

In a similar vein, a number of BV scientists have borrowed representations and procedures from the MV literature without due critical analysis of the status of such models. For example, the work of Marr [9] has been taken too literally and definitively representative of developments in MV by the BV community and, here, we wish to clarify these misconceptions and demonstrate the both BV and MV have much to offer each other so long as each area understands the status of theory and results in the other. In particular, MV scientists should remember that BV is less formal and less is known objectively about how the system works: BV is driven by observation whereas MV is typically part of Information Science.

This paper is focused on one central problem in vision: Object Recognition Systems (ORS) – the ability of systems to store and/or learn the descriptions of 3D objects and recognize them in 2D data structures invariant to their rigid motion, viewer position and, to some extent, environmental complexity. Common to all approaches to ORS, in man or machines, are the following two major questions:

- How do we generate/store 3D models including their explicit feature-based representation?
- How do we match 2D data to such models?

Associated with both questions are specific questions related to:

- How do we sense depth or shape from image data?
- How do we encode basic surface information?
- How do we segment models and/or data?
- What features do we use for model/data representations?
- How do we actually do 3D-ORS in BV, or, efficiently in machines?

In the following sections we briefly review each such process and, where possible, compare ideas and results from BV and MV.

2. Theory in Biological and Machine Vision

In MV, the term "theory" typically refers to a description of a process by algorithms which are implementable on digital computers. In the psychological literature the term is used more loosely to refer to a description of a process which is testable with psychophysical experimentation. To illustrate this difference, two psychological theories are examined in brief. These "theories" are the Recognition By Components model of Biederman [4], and the Feature Integration Theory of Treisman [3].

The Recognition by Components(RBC) model provides a non-algorithmic and sketchy account of "primal access: The first contact of a perceptual input from an isolated, unanticipated object, to a representation in memory" (Biederman [4], p. 32). Briefly, according to this theory, an image is segmented into regions and combined with information about non-accidental properties in order to describe

the 3D components in the scene – which are known as "geons". Geons have the highly advantageous property of being finite in number; they are defined in terms of non-accidental properties of generalized cones. In fact, apart from lacking explanatory validity, in that it includes no algorithms, this model is not even descriptively complete. No account is given of how geons and their spatial relationships may be encoded. However, there has been some psychological research following up the RBC theory but, as can be seen from the following examples, because its processes are not described in detail, any number of models, other than RBC, could explain the findings. For example, Biederman and Ginny [10] examined surface and edge information in recognition. They found that subjects recognized full color photographs and line drawings with the same degree of speed and accuracy. They concluded that line descriptions of visual images are what are used for object recognition, and that surface information such as color and texture gradient may be useful in edge extraction, but is not directly useful in object recognition tasks. This conclusion is claimed to be supportive of the RBC model, though, from the authors' perspective, it is difficult to imagine any theory that it wouldn't support!

On the other hand, Price and Humphreys [11] found that both color and surface detail information can enhance object recognition. They suggested that the degree to which surface detail is useful is proportional to the amount of within category discrimination required: features are relative to the classification problem – a characteristic of human pattern recognition well-documented in 2D vision (see, for example, Caelli, Bischof and Liu [12]). Other research, in the geon tradition, includes that of Biederman and Cooper [13], who primed subjects on images with either half the features (lines and vertices), or half the geon components removed. Priming facilitation was tested, using identical, complementary, and different class samples. For feature primed subjects, there was no difference in priming between identical and complimentary images, meaning the actual features present in the original image had nothing to do with the priming. Different exemplars showed somewhat less priming. For component primed subjects, the identical images showed more priming than the complementary ones: priming took place at the component level, rather than the feature level. This suggests that early preprocessing in biological ORS involves some form of segmentation and it was claimed that these results support RBC. The difficulty with this is that RBC is not well defined and, again, the results are so predictable from most ORS theories that one can hardly use them to strongly support RBC, per se.

Physiological evidence shows that in early stages of visual processing, color, orientation, motion, etc. are analyzed by separate but related channels (Livingstone and Hubel [14]). According the feature integration theory (FIT) of Treisman and Galade [15], a visual scene is processed first along these lines of separate dimension. Where attention is focussed, these features are combined into "unitary objects". They are then encoded in memory in this unitary form. However, with memory decay, the features may once again become dissociated. Without attention, these features cannot become associated in the first place. This model is interesting in that it brings visual attention into play, an aspect which is neglected in most models of ORS. However, it suffers from the same problems as Biederman's RBC model, in

that it is descriptive, but not explanatory, in terms of algorithms. Further, it should be noted that such features are not even conceptually independent in so far as color and motion must be indexed by space, pattern and form – unless they are derived from some perceptual "ether." Indeed, we have recently shown how this notion of modularity actually misses rich properties of the signal, at least with respect to spatio-chromatic information processing (Caelli and Reye [16]).

3. Sensing Objects

During a time when most vision research was related to understanding how to encode and process images as signals, per se – in man and machines – one scientist made a stand which was to become the basis for a change in vision research: the "Gibsonian" position (see Gibson [17]). Paraphrasing and consolidating what was really a very qualitative and introspective inquiry into visual perception, we can simply point out that Gibson's essential "insight" was that the various algorithms which constitute the "act of perception" must have evolved to reference **world** structures and not just image features, per se. This perspective on vision was taken up by the Artificial Intelligence (AI) community while both Image Processing and BV scientists were still mainly focussed on how the image is encoded, etc. However, the early AI approaches to such a view were mainly symbolic, as in "Block World" interpreters [18] and not until the late 1970's was there any formal attempt to integrate known image processing technologies of visual systems with properties of objects and structures within the world around us. This is not to say that BV avoided such issues. Indeed, "ecological optics" (a term used by Gibson to describe this perspective) was of central interest to many BV researchers from motion, spatial encoding, through to color vision research programmes. The main difference, however, was that experiments were conducted which really did not enable us to unambiguously infer how given processing systems actually encode the signal to infer world structures. This is particularly true of most of the work done on the threshold detection of grating patterns which not only have not provided of consistent theory of threshold spatial vision but has not been able to predict how humans process above threshold images, in general, nor how such information is used to solve problems of inferring depth or shape. On the other hand, exceptions to this situation are the studies of Reichardt [19] on understanding the encoding of motion by flies in flight, and the experimental work by Johannson [20] on how observers solve motion correspondence problems with the motion of realistic shapes. However, fundamental research into the variety of passive sensing sources and just how they are used in inferring shape from image intensity information is still evolving in the BV literature.

Though most recent techniques for ORS in MV use range data, there is still need for intensity-based ORS not only because of its relevance to BV but also because of the robustness of passive sensing for industrial and other applications. Hence there is still strong interest (see, for example, Seibert and Waxman [21]) in both MV and BV circles in problems of inferring depth from passive sensing resources such as focus, stereo, motion, perspective/texture. What follows is a brief resume of

the known relationships between MV solutions to Shape-from-X and the associated results, where possible, from BV.

- **Shape-from-Shading.** Here, with appropriate assumptions and one image, or, with photometric stereo, etc., it is possible to generate shape and depth from intensity information with knowledge of the rendering model. The algorithms are generally of the relaxation type, propagating depth from boundaries - as has been also shown to be the case with human vision [22,23].

- **Shape-from-Focus.** Here, at least 2 images are required and the degree of blur about a pixel (measured, say, by variance) is used to fit a blur [24] or even point spread function [25] from which depth is determined. Since the human eye, under photopic vision, functions with a small aperture, focus is not a strong cue, per se. However, accommodation of the lens is, implicitly, a response and/or source for depth though the exact degree of afferent-to-efferent processing with this is still yet to be fully determined. That is, the visual system does adjust the power of the lenses with respect to distance but the degree to which this can serve as a depth cue is not resolved.

- **Shape-from-Stereo.** Whether it be with two or multiple images, the use of image disparity information to infer depth is well established in photogrammetry or stereo procedures. However, the problems with this resource's insensitivity to small disparities and the "correspondence problem" with large ones, still leaves this as a source needing additional analyses [26]. However, human vision does use binocular disparity to infer depth and cells have been found within the vertebrate visual cortex which directly encode this and so produce a signal which can be be used by higher-order neurons to infer depth [27].

- **Shape-from-Motion.** Motion flow has been studied in both areas as a strong resource for depth and, again, without large disparities the process is quite insensitive. With large displacements however, there is, again, a "correspondence problem". Recent neurophysiological data shows that vertebrates also have neurons selectively sensitive to various types of relative motions including looming and linear motion components [28]. Also, recent evidence suggests that the same type of constraint satisfaction algorithm used to solve motion correspondence problems in MV may well explain and predict similar problems in BV [29].

- **Shape-from-Perspective/Texture.** This resource for MV is not that popular due to the fact that many imaging devices have little perspective. In the human visual system, depth may be inferred from edges, which shows that the preprocessed image, in terms of lines, bars and edges, described by Marr [9] as the "primal sketch", for example, is enough to provide sparse depth information. The psychological literature has many reports of how observers use this cue for depth – so long as these types of features are present [17].

However, for passive sensing there is one major limitation: no depth or shape can be inferred from pixels whose neighborhood variances are zero: where there is no

variation in light or "features". This shows that the inference of full range from intensity information cannot be obtained without prior or "top-down" knowledge or constraints. Indeed, recent solutions to this "interpolation" problem [30,31] demonstrate how this knowledge can be introduced either locally and algorithmically, or, globally via knowledge of the actual models involved.

To this stage, however, only broad relations between the known physiology of intensity processing and the extraction of depth are available in the sense that it is well established that the visual system is capable of differentiating the intensity image via the hierarchy of orientation-specific receptive fields. However, at this stage the only depth resources with direct physiological substrate are stereo and motion.

4. Parts and Features

It is widely accepted in the both the psychological (e.g. Biederman [4], Hoffman and Richards [32], Braunstein, Hoffman and Saidpour [33]), and in the machine literature (e.g. Jain and Hoffman [1], Fan, Medioni and Nevatia [2]), that some form of image segmentation must occur before recognition can take place.

Current practice in the MV literature is to segment surface data into parts as a function of various degrees of prior knowledge or constraints. The purpose of segmentation in this literature is to reduce the matching problem and obtain descriptions which are more robust and less dependent of specific pixel-based matching criteria. However, the segmentation problem is underconstrained without additional knowledge and, equally, segmentation does not necessarily imply that the complete surface needs to be partitioned into surface "parts". That is, one form of segmentation is the location of surface "edges", corners, etc., without determining what is non-edge, etc.

The issue of segmentation for ORS's, and for range data, specifically, has received a good deal of attention in recent years [35,36]. Common to most approaches is the development of surface part clustering in terms of similarities in surface point position, normals, or curvature information or surface curve fitting parameters [35,36]. Actual techniques vary from simply grouping via curvature sign (-,0,+) values or by complex clustering algorithms with hybrid constraints [36] to actually merging and splitting initial clustered regions to be consistent with known part properties in the model database [1].

Incorporating the notion of image segmentation into a model of human ORS is a very good way of providing the flexibility required for recognition from novel views, and recognition from partially occluded or degraded images. Segmentation reduces geometric information about an object into discrete, manageable chunks. In this way, parts may be recognized in isolation, making it unnecessary for all object parts to be visible for object recognition to take place. It is necessary, however, to include information not only about the segmented parts themselves, but also about the spatial relationships between these parts. One common view of the segmentation problem is that of defining where the boundaries between parts should occur. Herein lies the strong contrast between segmentation in MV and BV.

Hoffman and Richards proposed an algorithm for how humans segment surfaces that many others have already used in MV – in one form or another [2], and they termed it the minima rule: "Divide a surface into parts at loci of negative minima of each principal curvature, along its associated lines of curvature" (Hoffman and Richards [32], p.275). In intuitive terms, this means that part boundaries will occur at regions of concavity on the surface of the object, so that parts would consist of protruding surfaces. Segmentation using the minima rule has two obvious benefits: It yields a segmentation schema which is view invariant, and it requires no knowledge about the object itself; it is dependent only on the geometry of the object surface. Braunstein, Hoffman and Saidpour [33] found that when presented with object parts partitioned either at negative minima, or at positive maxima, and asked to choose a part that matched an object, subjects usually chose one partitioned at negative minima. Furthermore, when asked to mark part boundaries of novel objects manually, all subjects partitioned the objects at negative minima. This research shows evidence for the minima rule, although it does have some problems. The stimuli used were random dot patterns, which were rotationally symmetrical in the vertical plane. It is also clear that simple curvature extrema are not sufficient for all types of segmentations - as is clear from the MV segmentation algorithms where, for example, it is important to identify different surface types, etc. [2]. Another important consideration for segmentation is the role of jump boundaries, which commonly mark regions at which adjacent objects meet. If a model of object segmentation is to have high ecological validity, it should account for these additional requirements for part type descriptions.

For MV, surface features are usually of three generic forms [1]. One, Morphological(M): features derived from the complete object or model; two, unary(U): features extracted from individual parts, and three, binary(B): features derived from part relationships. Unary features can refer to typical surface patch pixels (local, as in curvatures), global patch properties (such as areas) and can also refer to patch boundary properties (such as perimeter). Binary features typically capture part relationships such as distances, angles, and also include boundary relationships. Those used in the literature are shown in Table 1(right column). These features, again, fall into seven types: Morphological (M) unary curvatures(U.C), unary distance (U.D), unary boundary (U.B) and binary boundary (B.B), binary distance (B.D), and binary angles (B.A) [37]. Those computational forms, as shown in the right column of Table 1, correspond to ways of "measuring" such features as real-valued functions whose values, wherever possible are invariant to rigid motions.

5. Encoding Objects

MV model representations are typically surface-based, either obtained via active sensing, CAD, or the use of Shape-from-X methods. Given the constraints of surface "visibility," most approaches represent models via "aspect graphs" where the fully view-independent representation is defined by having adequate numbers of views. This, we believe is consistent with what we understand as 3D model knowledge in so far as it is presumably impossible to have fully view-independent models

Table 1
Typical Surface Features used in Machine ORS's

Feature	*ORS1-measure*
Morphological	M.1 Perimeter M.2 Number of Parts M.3 Total Area M.4 Genus (Mean K)
Sense	U.C.1 Mean H U.C.2 Mean K
Size	U.D.1 Area U.D.2 3D Spanning Distance (max)
Boundary Features	U.B.1 Perimeter U.B.2 Mean Curvature U.B.3 Mean Torsion
Boundary Type	B.B.1 Length of Jumps B.B.2 Length of Creases
Part Distance Relations	B.D.1 Bounding Distance B.D.2 Centroid Distance B.D.3 Max Distance
Part Angle Relations	B.A.1 Normal Angle Differences (average) B.A.2 Bounding Angle between surfaces (average) B.A.3 Normal Angle Differences (average)

without having viewed all surface points or without having equivalent alternative knowledge. This is not to be confused with the issue of how such data is symbolically encoded. For example, one can have enough data to have a full description of a given 3D object but if surface descriptors are used which are not invariant to rigid motions then the model cannot be said to be "view-independent". For this reason, most MV models consist of range surface mean and Gaussian curvature descriptors which are invariant to rigid motions whereas their original (x, y, z) surface coordinates are not. As will be seen in the following section, for MV, objects are fundamentally represented via "shape descriptors" which are "positionless" in the sense that the curvatures are fully independent of the 3D location of the object. As seen in Table 1, such descriptors are then used to segment and, in general, extract model features which can be found in data – invariant to position and pose.

The internal representation of objects in humans is a dynamic memory process, in that it involves a model of objects which may be both accessed and modified by perceptual information. Current models of internal representation are rooted in early work on 2D shape recognition. For example, the work of Deutsch [38] and Sutherland [39] found that shapes could be recognized independently of location,

size and brightness, but not orientation. Theories of internal representation in human object recognition may be grouped into three divisions (see Pinker [40] for a review): view-independent versus object centered models, single-view-plus-transformation models, and multi-view models.

According to the view-independent models, objects are represented as a collection of spatially independent features, such as intersections, angles, curves and surfaces. View-independent theories assign an object a representation that is the same regardless of its orientation, location, or size. Opposing this are object-centered theories, under which an object representation may be described as a data-base consisting of a store of descriptions, from multiple view-points, with which an image may be directly compared. Rock, DiVita, and Barbeito [41] found that novel wire objects shown in one position, are not often recognized when they were rotated about a vertical axis and presented later. Rock, DiVita, and Barbeito [42] also found that mirror images and left-right reversals are difficult to discriminate. Rock, Wheeler, and Tudor [43] asked subjects to imagine how these 3D wire objects would appear from positions other than the one they were in. They found that subjects were unable to perform this task unless they made use of strategies that circumvent the process of visualisation. This suggests that objects are not simply rotated about in space as seen on a CAD computer screen, but, in fact, draw on several of the available heuristics (as, for example, 2D projected feature similarities) available to humans when performing such a task.

According to single-view-plus-transformation models, object recognition is achieved via transformation, typically mental rotation, of an input stimulus into either a perspective, or an orthogonal (canonical orientation) view. In these models, mental representations are defined in terms of the end products of the transformation process. Shepard and Metzler [44] found that reaction time in a "same-different" (binary classification) matching task was a linear function of the angular difference between two geometrically identical figures. This was true for rotations in the plane and for rotations in depth. Countering this research, experiments investigating orientation-independence have provided arguments against the notion of reaction time being a linear function of the angular difference in rotation. Corballis and Nagourney [45] suggested that the time required to name normal versions of letters and digits was largely independent of the orientation of the characters. Orientation-independence in recognition time seems to occur only for highly familiar combinations of shapes and orientations; when unfamiliar stimuli must be recognized, orientation effects reminiscent of mental rotation appear. This suggests that humans may sometimes use mental rotation to recognize unfamiliar shapes or examples of shapes. However, the actual computational procedure including internal data structures and matching criteria have not usually been specified.

Because there is no theory of form perception that explains the necessity of mental rotation, Takano [46] became interested in why mental rotation appears to happen only in certain instances but not in others. If mental rotation does not occur, then simple template matching theories of form recognition would have difficulty coping with the storage and computational requirements in performing their task adequately. To deal with this problem, feature extraction theories have been

proposed. These theories suggest that recognition is based on those features that are not affected by rotation (see, for example, Sutherland [39]). In instances where these features are unavailable, or not relevant, objects may need to be aligned by mental rotation. Takano used a mental rotation paradigm to see whether the distinction between orientation-free information and orientation-bound information played a significant role in human form perception. The conclusions drawn from this experiment suggest that mental rotation is unnecessary if the forms differ in either type of orientation-free information, provided that the difference is actually encoded as such.

Multi-view theories are hybrids of object centered and transformation models. Representations consist of pools of object views in familiar orientations. Recognition will occur rapidly when an image is oriented according to a stored, familiar view. When an image does not match one of these views, transformation is necessary. Jolicoeur and Kosslyn [47] investigated long term memory representation of three dimensional shapes. Their study provides converging evidence that people can store three dimensional shapes in long term memory using both object-centered and viewer-centered recognition, and that these dimensions were very stable appearing in every subjects data for every family of stimuli. A second experiment showed that the introduction of memory requirements did not seem to mitigate the subjects tendency to use both sorts of coordinate systems to compare the stimuli within each family. This contradicts Marr and Nishihara's [48] claim that recognition proceeds solely through the use of object-centered representations. Tarr and Pinker [49] presented participants with several objects each at a single rotation. They were given extensive practice at naming and classifying them as normal or mirrored- reversed at various orientations. Their preliminary findings were consistent with the early 3D mental rotation studies in that response times increased with departure from the study orientation. This suggests that subjects mentally transform the orientation of the input shape to one they had initially study or familiar with. An interesting finding from this research was that whenever mirror images of trained shapes were presented for naming, subjects required the same amount of time at all orientations. This finding suggests that mental rotation transformations of orientation can take the shortest route of rotation that allows the alignment of the input shape with its memorized counterpart.

6. Recognition Processes

One of the most difficult problems in ORS's is that of grouping object (model) parts and feature states into a form which can optimize recognition. This is because different objects share similar feature values on different parts of their surfaces. That is, given adequate features and the situation where, ***from any view***, object part features are clearly differentiated in feature space, then well-known optimal classification procedures can be employed - precisely the scenario which does not occur in generic ORS environments. This situation is similar to problems in concept learning where different concepts share common states and the learning procedure, whether in modeling human function or machine applications, has to be capable

of forming a structure which captures common properties within class (object) examples and highlight features which differentiate between classes. Techniques in MV which attain these goals are of two basic types: Feature Indexing (FI) and Evidenced-Based Systems (EBS's). In the former case object parts, and their feature states are ordered according to their discrimination power – along the lines of being representative of given classes of objects and discriminating others. The search procedure usually is in the form of a decision/strategy tree or some form of graph matching (where binary constraints prune the state space) initiated from more critical parts - indexed for their complexity (see, for example, McLean [50], Grimson [51] and Lowe [52]).

Evidence-Based Systems (EBS) solutions to this problem revolve around the generation of clusters of different object samples in feature space which, to various degrees, "evidence" different objects [1]. A feature space is simply an n-dimensional Euclidean space, on which each dimensional axis corresponds to some property of the data. For example, object part features such as perimeter and average Gaussian curvature would constitute two such feature dimensions. These predicates are normally either unary (describing properties of single features), or binary (describing relations between pairs of features). Evidence-based systems use bounded regions of feature space as probabilistic evidence rules for the occurrence of objects which have features represented within that region. An object in an evidence-based system is represented in terms of a series of these rules, which are triggered by the occurrence of the features of that object. Once triggered, a rule provides a certain amount of evidence for each object in the database, according to the likelihood of any given object having features which fall within the bounded region defining that rule. An object is recognized on the basis of accumulated evidence over all triggered rules.

The role of world knowledge and semantic association is an important issue for human object recognition, which has not been well addressed in the literature. Most researchers who put forward models of object recognition tend to ignore this issue. For example, Biederman [4], in proposing the recognition by components model defined object recognition in terms of matching geons to non-accidental image features without specifying the search process. Without fear of over-generalizing it appears that there is no search model for human object recognition – certainly one which makes specific predictions about the complexity of the matching process for given classification problems.

Human ORS must make sense of complex scenes containing multiple objects, some of which may be occluded to a very high degree. The evidence-based approach is able to account for both perceptual and semantic considerations in object recognition, with explanatory efficiency. When evidence rules (which may be abstractly described as bounded regions of feature space) are triggered, they provide a level of activation to objects represented within that region of feature space, proportional to the amount of evidence provided by that rule for each object. Activation may be mediated by either perceptual or a semantic processes. Perceptual processes build up the visual percept of a scene. Semantic processes use perceptual information to invest perceptual entities, including objects in the scene, with meaning, includ-

ing names and properties. These semantic processes feed back into the perceptual channel to provide contextual information as an aid to the visual organization of the scene. Whenever an evidence rule is triggered by perceptual information, all representations which are semantically associated with each of the objects represented on the area of feature space designated by the triggered rule are also activated. The degree of activation is proportional to the strength of memory association, as well as the amount of evidence for any particular object, provided by the triggered rule. This allows for rapid recognition of objects in familiar environments, as well as recognition of objects which are highly occluded. For example, in an office scene, a telephone may be entirely occluded by papers, apart from the cord. As an isolated cue, a cord alone would probably provide insufficient evidence for the recognition of the telephone. Nevertheless, recognition of the cord will provide considerable semantic activation for the telephone representation, which together with semantic activation due to the triggering of evidence rules for various other items of office paraphernalia, may well be enough to facilitate the recognition of the telephone. This concept of semantic facilitation is drawn from an extensive literature in psycholinguistics (e.g. Meyer and Schvanaveldt [53]).

Modeling human ORS in terms of the evidence based system accounts well for the issues of view-independence, partial occlusions, variation between object within object classes, and novel exemplars of object classes. View independence is achieved simply by including all the segmented features of an object in the database. No matter which direction an object is viewed from, it will trigger rules which give evidence for the presence of that object. Preferred views are also accounted for, in terms of those views which provide the most evidence for the object in question. Partial occlusion is accounted for in much the same way. Provided at least some of the features of an object are displayed, some rules will be triggered, giving evidence for the object. Variation within object classes, and novel exemplars of object classes are also accounted for, because each rule is a bounded region in feature space, and each feature is simply a point. Therefore, any given feature predicate may vary in its precise position within a bounded region, and still provide the same evidence for its object class. As long as an object has enough similarity to the other objects in its class, that it triggers approximately the same set of evidence rules, it will be recognized as a member of that object class.

A final point should be made about the underlying physiology of object recognition. First, it should be noted that vision, as a sense, is passive and so any theory of biological object recognition that does not address the problem of inferring depth or shape from intensity – as an integral part of the processing system – is not complete. For these reasons it is difficult to evaluate results from neurophysiological studies showing certain sensitivities to photographs of faces, etc.(see Perrett [54], as the results confound the perception of *patterns* (as 2-D structures) in comparison to *objects* (as 3-D structures). Unfortunately, to this stage, we know very little about how the visual system solves, and uses solutions to, Shape-from-X in the *act of object recognition* though experiments are underway to investigate such issues. The evidence-based paradigm should prove useful in the study of these processes as it lays out a blueprint for how BV may learn relationships between surface types

and expected image intensities, or, vice-versa.

REFERENCES

1 Jain, A., and Hoffman, R. (1988). Evidence-Based Recognition of Objects. *IEEE:PAMI*, 10, 6, 783-802, 1988.

2 Fan, T., Medioni, G., and Nevatia, R. (1989) Recognizing 3-D Objects Using Surface Descriptions *IEEE:PAMI*, 11, 11, 1140-1157, 1989.

3 Treisman, A. (1986) Features and objects in visual processing. *Scientific American*, 255(5),114-126.

4 Biederman, I. (1985). Human image understanding: recent research and a theory *Computer Vision, Graphics, and Image Processing*, 32, 29-73.

5 Ratliff, F., and Hartline, H. (1959) The response of *Limulus* optic nerve fibres to patterns of illumination on the retina mosaic *Journal of General Physiology*, 42, 1241-1255.

6 Barlow, H. (1972) Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, 1, 371, 394.

7 Caelli, T., and Oguztoreli, N. (1988) Some task and signal dependent rules for spatial vision *Spatial Vision*, 2, 4, 295-315.

8 Caelli, T., and Moraglia, G. (1987). Is pattern matching predicted by the cross-correlation between signal and mask? *Vision Research*, 27(8), 1319-1326.

9 Marr, D. (1982) *Vision*. San Francisco: Freeman.

10 Biederman, I., and Ginny, J. (1988). Surface versus edge-based determinants of visual recognition *Cognitive Psychology*,20, 38-64.

11 Price, C., and Humphreys, G. (1989). The Effects of Surface Detail on Object Categorization and Naming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 15, 797-825.

12 Caelli, T., Bischof, W. and Liu, Z. (1988) Filter-based models for pattern classification *Pattern Recognition*, 21, 6, 639-650.

13 Biederman, I., and Cooper, E. E. (1991). Priming contour deleted images: evidence for intermediate representation in visual object recognition Cognitive Psychology, 23, 393-41.

14 Livingstone, M., and Hubel, D. (1988). Segregation of form, color, movement and depth: anatomy, physiology and perception *Science*, 240, 740-750.

15 Triesman, A., and Gelade, G. (1980). A feature integration theory of attention *Cognitive Psychology*, 12, 97-136.

16 Caelli, T., and Reye, D. (1992) Classification of Images by Color, Texture and Shape *Pattern Recognition* (in press).

17 Gibson, J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton-Mifflin.

18 Waltz, D. Understanding Line Drawings of Scenes with Shadows. In Patrick Winston (ed.) *The Psychology of Computer Vision*, McGraw-Hill, New York.

19 Reichardt, W. (1987) Evaluation of optical flow information by movement detectors *Journal of Comparative Physiology A*, 161, 533-547.

20 Johansson, G. (1974) Vector analysis in visual perception of rolling motion *Psychological Forschung*, 36, 311-319.
21 Seibert, M., amd Waxman, A. (1992) Adaptive 3-D Object Recognition from Multiple Views *IEEE:PAMI*, 14, 2, 107-124.
22 Bulthoff, H., and Mallot, H. (1988) Integration of depth modules: stereo and shading *Journal of the Optical Society of America*, 5, 10, 1749-1758.
23 Todd, J., and Mingolla, E. (1983) Perception of surface curvature and direction of illumination from patterns of shading *Journal of Experimental Psychology: Human Perception and Performance*, 9, 4, 583-595.
24 Caelli, T., and Xu, S. (1990) A new method for Shape-from-Focus. In C P. Tsang(Ed.) *AI'90*, Singapore: World Scientific.
25 Pentland, A. (1987) A new sense of depth of field.*IEEE:PAMI*, 9, 4, 523-531.
26 Horn, B. (1986) *Robot Vision*. Cambridge, Mass.: MIT Press.
27 Pettigrew, J. (1973) Binocular neurons which signal change of disparity in area 18 of cat visual cortex *Nature(London)*,241, 123-124.
28 Frost, B., and Nakayama, K. (1983) Single visual neurons code opposing motion independent of direction *Science*, 220, 744-745.
29 Caelli, T., Manning, M. and Finlay, D. (1992) A general correspondence approach to apparent motion *Perception* (in press).
30 Dillon, C., and Caelli, T. (1992) Inferring Shape from Multiple Images using Focus and Correspondence Measures(In Submission).
31 Dillon, C., and Caelli, T. (1992) Generating Complete Depth Maps in Passive Vision Systems *IAPR-92 Proceedings, Hague, September.*
32 Hoffman, D., and Richards, W. (1986) Parts of Recognition. In A. Pentland(Ed) *From Pixels to Predicates*. New Jersey: Ablex, 268-294.
33 Braunstein, M., Hoffman, D., and Saidpour, A. (1989) Parts of Visual Objects: an Experimental test of the Minima Rule *Perception*, 18, 817-826.
34 Besl, P., and Jain, R. (1988) Segmentation through variable-order surface fitting*IEEE:PAMI*, 10, 2, 167-192.
35 Fan, T., Medioni, G., and Nevatia, R. (1987) Segmented Descriptions of 3-D Surfaces *IEEE Journal of Robotics and Automation*, vol RA-3, 6, pp527-538.
36 Yokoya, N., and Levine, M. (1989) Range Image Segmentation Based on Differential Geometry: A Hybrid Approach *IEEE:PAMI*, 11, 6, 643-649.
37 Caelli, T., and Dreier, A. (1992) Some New Techniques for Evidenced-Based Object Recognition *IAPR-92 Proceedings, Hague, September*.
38 Deutsch, J. (1955) A theory of shape recognition. *British Journal of Psychology*, 46, 30-37.
39 Sutherland, N. (1968) Outliners of a theory of visual pattern recognition in animals and man *Proceedings of the Royal Society of London(Series B)*, 171, 297-317.
40 Pinker, S. (1984) Visual cognition: An introduction. *Cognition*, 18, 1-63.
41 Rock, I., DiVita, J., and Barbeito, R. (1981) The effect on form perception of change of orientation in the third dimension *Journal of Experimental Psychology: Human Perception and Performance*, 7, 719-732.
42 Rock, I., DiVita, J. (1987) A case of viewer-centered object perception *Cognitive*

Psychology, 19, 280-293.

43 Rock, I., Wheeler, D., and Tutor, L. (1989) Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, 21 185-210.

44 Shepard, R., and Metzler, J.(1971) Mental rotation of three-dimensional objects *Science*, 171(3972), 701-703.

45 Corballis, M., and Nagourney, B. (1978) Latency to categorize disoriented alphanumeric characters as letters or digits *Canadian Journal of Psychology*, 32, 186-188.

46 Takano, T. (1989) Perception of rotated forms: A theory of information types *Cognitive Psychology*, 21, 1-59.

47 Jolicoeur, P., and Kosslyn, S. (1983) Coordinate systems in the long-term memory representation of three-dimensional shapes *Cognitive Psychology*, 15, 301-345.

48 Marr, D., and Nishihara, H. (1978) representation and recognition of the spatial organization of three-dimensional shapes *Proc. Royal Soc. (Lond)*, 200, 269-294.

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

50 McLean, S., Horan, P., and Caelli, T. (1992) A Data-Driven Indexing Mechanism for the Recognition of Polyhedral Objects *SPIE Proceedings: Advances in Intelligent Robotic Systems*, 1609.

51 Grimson, W., and Lozano-Pérez, T. (1985) Recognition and Localization of overlapping parts from sparse data *MIT A.I. Memo No. 841.*

52 Lowe, D. (1987) Three-dimensional object recognition from single two-dimensional images *Artificial Intelligence*, 355-395.

53 Meyer, D., and Schvanaveldt, R. (1971) Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90, 2, 227-234.

54 Perrett, D., Mistlin, A., and Chitty, A. (1987) Visual neurons responsive to faces *Trends in Neuroscience*, 10, 9, 358-363.