ELSEVIER

Image and Vision Computing 15 (1997) 277–292

image AND vision COMPUTING

Panel report: the potential of geons for generic 3-D object recognition

Sven J. Dickinson[a], Robert Bergevin[b], Irving Biederman[c], Jan-Olof Eklundh[d], Roger Munck-Fairwood[e], Anil K. Jain[f], Alex Pentland[g]

[a]*Department of Computer Science and Rutgers Center for Cognitive Science (RuCCS), Rutgers University, New Brunswick, NJ 08903 USA*
[b]*Computer Vision and Systems Laboratory, Department of Electrical and Computer Engineering, Laval University, Que. G1K 7PA, Canada*
[c]*Department of Psychology, University of Southern California, Los Angeles, CA 90089-2520 USA*
[d]*Computational Vision and Active Perception Laboratory, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, S-100 44 Stockholm, Sweden*
[e]*Department of Electronic and Electrical Engineering, University of Surrey, Guildford, Surrey, GU2 5XH UK*
[f]*Department of Computer Science, Michigan State University, East Lansing, MI 48824-1027 USA*
[g]*Vision and Modeling Group Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA*

Received 3 March 1995; accepted 3 September 1996

Abstract

Biederman's introduction of geons to the vision community has spawned considerable interest in building geon-based vision systems. However, numerous issues must be addressed before such systems can make a practical contribution to machine vision. At *IJCAI 1993*, a group of distinguished researchers, each of whom has worked with geon-based recognition, was brought together to form a panel whose goal was to identify and discuss these issues. This paper is based on that panel discussion.

Keywords: 3-D Object recognition; Geons; Object modelling

1. Introduction

In 1985 Biederman introduced to the computer vision community a theory of human visual object recognition called Recognition by Components (RBC) [1]. The main tenets of RBC, supported by psychophysical experiments [2,3], are that: (i) the basic-level representation recovered from an image during *primal access* is metrically invariant, i.e. invariant to scale, translation or orientation in depth, and (ii) the basic-level representation is composed of *parts* and their interrelations. In addressing these two principles, RBC also suggests that from nonaccidental contour relations in the image, a set of contrastive dichotomous (e.g. straight vs. curved axis) and trichotomous (e.g. constants vs. tapering vs. expanding/contracting cross-sectional sweep) 3-D volumetric primitive properties can be determined. The Cartesian product of the values of these properties gives rise to a set of volumetric primitives, or parts, which Biederman called *geons*. The notion of quickly recovering a set of object-centered, qualitative parts (geons) from an image, both without a priori knowledge of the scene contents and without expending any effort on depth recovery, quickly attracted the attention of the object recognition community. Since then, a number of researchers from around the world have engaged in building computational models based on Biederman's RBC theory.

Beginning in 1988, Bergevin and Levine [4,5] introduced their PARVO system, representing the first significant effort to build a vision system based on geons. From manually segmented line drawings (and from range data in 1991 [6]), they were able to recover and recognize a variety of multi-part objects. Then in 1989, Dickinson, Pentland and Rosenfeld [7] introduced a hybrid object representation integrating object-centered (volumes) and viewer-centered (aspects) models, with the object-centered models composed of parts closely resembling geons. This was to become the backbone of the OPTICA system which recognizes 3-D objects from real images [8,9].

A long-time advocate of recognition-by-parts, Pentland introduced a parametric part representation to computer vision, using deformable implicit functions such as superquadrics [10]. Similar to geons but with continuous rather than qualitative properties, this type of representation has been successfully used for recognition in segmented data [11], and in modeling complex natural objects [12].

PII S0262-8856(96)01134-1

In 1991, Munck-Fairwood investigated a logic programming approach to the recovery of geons from a manually segmented line drawing [13]. Recently, he has integrated causal probabilistic networks into his approach, assigning confidence measures to Biederman's dichotomous and trichotomous properties [14]. More recently, Du and Munck-Fairwood have applied their approach to geon recovery from real 2-D imagery [15]. Jacot-Descombes and Pun have explored probabilistic methods for distinguishing among metrically different instances belonging to a single geon class [16]. Diverging from the analysis of labeled contours, Raja and Jain explored geon recovery from range data [17]. Their technique, which recovers geons from superquadrics fitted to range data, has been successfully tested on range images of real geon-based objects with straight axes. Since then, Dickinson, Metaxas and Pentland [18] and Wu and Levine [19] have also looked at geon-like part extraction from range data.

Eklundh and his colleagues are building an active vision system whose goal is to recognize geon-based objects from images acquired by their stereo head [20]. In recent work, Eklundh and Olofsson were the first to study the problem of geon-based object maintenance (of position, orientation and appearance) over time [21], a critical problem facing the design and construction of geon-based active vision systems.

During these efforts by the vision community to build geon-based recognition systems, Biederman was exploring his own computational models for geon recovery. In 1988, Hummel, Biederman, Gerhardstein and Hilton began exploring connectionist models for geon recovery [22], while in recent work [23], Hummel and Biederman report a neural net model for shape recognition, using dynamic binding, that can recognize simple geon-based objects from line drawings.

Despite the growing interest in geons in the vision community, the above systems are still being applied to very constrained domains in which segmentation is often performed manually, lighting is tightly controlled, or the objects themselves are unrealistically simple. Nevertheless, the motivation for building generic recognition systems is strong. If a robotic vision system is to move about its environment, quickly recognizing new object exemplars from a large database of generic, articulated object models as well as learning new object descriptions, it must have the capability to recover and match image representations which capture the coarse shape of an object in terms of its parts. Geons have emerged as the most popular shape representation for generic object recognition. However, before successful geon-based recognition systems can be realized, a number of important issues must be addressed, including:

- Are geons an appropriate modeling vocabulary for 3-D generic object recognition systems? If so, for what classes of objects are they best suited? If not, can the representation be altered or enhanced to make them more usable?
- What are the major obstacles in recovering geons from real imagery? What image features should be extracted in order to recover geons?
- What kinds of architectures and control strategies are appropriate for geon-based recognition?
- Finally, with the growing interest in active vision, what issues face the design and construction of geon-based active recognition systems? And what role can geon-based recognition play in a larger context where, for example, object manipulation may be required?

At the 13th International Joint Conference on Artificial Intelligence (IJCAI), Chambery, France, 1993, Dickinson chaired a panel discussion which brought together a number of researchers, each with first-hand experience in building geon-based recognition systems, to address the above issues [24]. In the following sections, these researchers evaluate their progress and assess the potential of geon-based object recognition.

2. Testing the RBC theory: an approach respecting the general principles and the computational hypotheses

Robert Bergevin
Laval University

I see the original RBC proposal by Biederman as made up of two loosely connected sets of predictions: (i) the general principles; and (ii) the computational hypotheses. The general principles state that:

1. A line drawing which represents the discontinuities in an image is an efficient description of the image and is sufficient for the primal access task.
2. Objects are better represented and analyzed by decomposing them into their natural components (parts).
3. A qualitative description of the components is necessary and sufficient to permit fast access to a large database of object models.
4. Non-accidental instances of viewpoint-invariant features in the 2D line drawing are sufficient to permit fast access to the qualitative model of a 3D object.
5. Primal access for visual object recognition is obtained by matching a description of the spatial structure of components making up an object to an indexed database of models in a similar representation.

On the other hand, the computational hypotheses state that:

1. Five specific classes of 2D line groupings are sufficient to access the parts representation.
2. Segmentation should happen at concavities in the outline of an object.

3. The geons, a subset of the generalized cylinders, form an efficient qualitative shape representation for the parts which is suitable for the primal access task.
4. The symbol descriptions for objects and models should include geon labels, aspect ratios and relative sizes of parts.

While many psychological results directly support the general principles [25,2,26], the computational hypotheses are more like extrapolations or predictions regarding an actual implementation. This was meant to fill the gap between the levels of study in the psychological and computational perception fields (the 'what (is needed)' and the 'how (is it obtained?)' problems, respectively).

2.1. Line drawings

The main assumption of RBC is the suitability of single view line drawings for primal access. This, together with the small set of volumetric primitives making up the geons, represent, in my opinion, the atomic kernel of the theory without which the goal of rapidly recognizing arbitrary unexpected objects cannot be attained. Although in some circumstances it could make sense to diverge from these principles and, say, analyze multiple 2D views of an object, recover the object from range data, reconstruct quantitative surfaces, volumes, or exact models, etc., the resulting system would not solve the primal access task as defined in RBC.

Regarding these assumptions, a well acknowledged fact in the computer vision field is the difficulty to extract a very good line drawing in an uncontrolled imaging situation, e.g. when the lighting and viewpoint are arbitrary. Two approaches have been taken around this problem in the geon research community. Most groups assumed the (usually ideal) line drawing as given from the outset. Others decided to extract line drawings from real data (usually from range images) but had to restrict themselves to very simple scene with ideally shaped custom-made objects. From the point of view of low-level vision processes, the main question to be addressed is how one goes from imperfect data from a non-ideally shaped object to the much simplified and compressed description in terms of an idealized line drawing. A number of ideas have been proposed in the computer vision community to abstract a non-ideal shape by fitting surface or volume models to depth data. Similarly, many two-dimensional line fitting and multi-scale edge extraction methods exist. However, not much is known about extracting appropriate edges of complex, free-form and textured three-dimensional objects from noisy two-dimensional intensity images. Among the reasons for this state of affairs is the non-availability of a general-purpose approximation model for three-dimensional shape. One interesting piece of research, in this respect, is the study of quantitative invariant properties of the projected contours of generalized cylinders. Unfortunately, the results of this research are also difficult to apply to intensity images of real scenes comprising complex objects [27].

2.2. PARVO

In my work with the PARVO (Primal Access Recognition of Visual Objects) system, two main goals were pursued [5,28]. First, I aimed to determine what are the underlying assumptions about the geometric properties of objects and their line drawings that make them acceptable to the kind of processing proposed in RBC. That is, what objects may be recognized with such an approach?, what is their complexity?, etc. Second, given such acceptable line drawings (ideal or noisy), I wanted to develop methods to actually perform the feature extraction, part segmentation, geon labeling, object description and model matching stages to determine whether unexpected object recognition is possible using the RBC approach under realistic conditions. In other words, given the huge difficulty of the generic object recognition problem, I decided instead to investigate the capabilities and limitations of an actual computer vision system to be built according to the principles and hypotheses of RBC, which are themselves based on a significant piece of research in human image understanding. From the obvious fact that geons cannot be an appropriate modeling vocabulary for all existing objects, I tried to determine what generic set of objects this representation is appropriate for, if any, while making more precise and explicit the architecture and control strategies of an object recognition system based on RBC.

My main conclusions from this work with PARVO are that:

1. Primal access geon-based recognition is possible for a large number of man-made objects. Objects selected in my experiments are of the same type as those used in Biederman's work, i.e. they are part-based by construction. Still, the modeling power of geons appears impressive, at least from a high-level point of view, given the large diversity of recognized objects.
2. Recovery of geon-based descriptions from natural objects is likely to be very difficult using the RBC approach. Besides obvious reasons such as the idealized nature of the geons and the coarseness of the attributes that are to be extracted from the intermediate line drawing representation of free-form textured objects, there is the fact that many natural objects simply cannot easily be described as an assembly of a few volumetric parts. On the positive side, the growing process of many natural objects gives rise to a limb structure reminiscent of an assembly of elongated parts, which is appropriate for RBC and PARVO.
3. Explicit part segmentation is needed and may still be

possible even if some elements of the object or line drawing do not respect the implicit assumptions on acceptable shapes. Experiments with line drawings having missing or spurious features (lines, faces, parts, etc.) were successful in demonstrating this inherent capability of RBC and PARVO.

4. Face extraction and labeling is necessary for proper geon labeling. The main reason being the necessity to address robustness and uniqueness issues which leads to the insertion, at this and other processing stages, of a verification or validation step.
5. Errors arising from blind labeling of the parts (i.e. without explicit verification of the hypotheses) are likely to destroy the indexing power of the geons. Fortunately, verification steps in PARVO are 'wired-in' and do not destroy the bottom-up nature of the system. Besides, no iterative feedback to lower levels is needed given the redundant nature of the part-based representation of objects by RBC and PARVO.
6. Parallel architectures and algorithms are necessary to cope with the combinatorial complexity arising from manipulating sets of features. The implemented processes in PARVO were all designed with this constraint in mind.

2.3. Grouping and MAGNO

My recent research [29] on these topics has focused on the problem of deriving more efficient and robust contour grouping processes. Such processes, operating at multiple scales, are meant to rapidly identify significant structured subsets of contour elements in an intensity image of a complex scene. The identification of these subsets is seen as a means towards individual object detection and segmentation, itself a prerequisite to the part decomposition and feature extraction steps as implemented in PARVO. As such, the grouping processes could help bridge the gap between noisy intensity images of arbitrarily viewed multiple-object scenes and 'PARVO-acceptable' object-based geometrical contour descriptions. In other words, grouping might help relax the constraints on the quality of the extracted line drawing. For instance, it makes possible the removal of spurious, non-geometric, contours and the restoration and completion of partially missing or noisy contour segments.

MAGNO (Multi-level Access to Generic Notable Objects), a front-end to PARVO, is presently under development. This module is going to extend the primal access model of unexpected object recognition to a more realistic situation where an active agent (e.g. mobile robot) controls the positioning and orientation of a camera inside a room; the goal being that interesting objects in the scene be detected and centered in the field of view to ease their proper recognition by PARVO. Multi-level grouping processes and qualitative depth cues are to be at the center of MAGNO's processing. MAGNO and PARVO represent the two complementary systems needed to address the 'where (are the objects?)' and 'what (are those objects?)' aspects of object recognition.

At this stage, MAGNO is going to limit itself to intensity images as input, respecting the principal assumption behind RBC and primal access object recognition. Moreover, the domain of interest of MAGNO and PARVO, as an integrated system, will be similar to the domain of PARVO, which I consider a success with respect to its stated goals. That is, it is not planned to invest much efforts in the near future to extend the domain of applicability of PARVO beyond what has been identified as appropriate from a geon-based point of view. Instead, the demonstration that this approach is applicable and useful in a realistic experimental setting, using active control for the acquisition of intensity images from complex scenes containing real man-made objects, will be aimed at.

2.4. Conclusion

The success with PARVO and the still incomplete understanding of the complex issues related to three-dimensional shape representation determined the present direction of my object recognition research. Application of the whole approach to a wider-scope domain of objects encompassing, for instance, man-made or natural textured and free-form objects, remains an ultimate goal to be kept in mind. However, any serious progress in this direction will depend on the availability of a new, and perhaps complementary, shape representation originating from progress in biological, psychological or computational vision research.

In the longer term, the basic-level type of recognition represented by MAGNO and PARVO could also be a starting point to both subordinate-type and superordinate-type recognition, in the terminology of human image understanding research. In the former case, a more specific recognition and pose suitable, say, for robotics manipulation is to be obtained using quantitative object and part models. In the latter case, a more generic recognition is to be obtained using higher-level abstraction and function-based reasoning.

3. Recognition-by-geons: 1997's current progress and current challenges

Irving Biederman
University of Southern California
Eric E. Cooper
Iowa State University
John E. Hummel
University of California, Los Angeles (UCLA)

In a fraction of a second humans are able to comprehend novel images of objects and scenes on the basis of

their shape. Geon theory [23,25,30,31] offered an account of this phenomenon characterized by four assumptions: (a) Objects are represented as an arrangement of simple parts (geons); (b) the geons can be distinguished by a small number of (typically binary) differences in viewpoint invariant properties (VIPs), such as straight vs. curved, rather than fine differences in metric properties, such as aspect ratio; (c) the relations among the geons are explicit, such as top-of or perpendicular-to, as part of a structural description, rather than implicit in a coordinate space; and (d) a relatively small number of geons is sufficient to distinguish objects at a basic- (or entry-) level, such as between a table and an elephant.

We review the current status of the theory and its challenges. Since the theory was first proposed almost 10 years ago, considerable evidence has accumulated supporting several of its basic assumptions and extending its scope to the vast majority of subordinate-level distinctions that people make. We will first consider two frequently made criticisms of geon theory: (a) that geons are inadequate for representing all objects, and (b) because geons are activated from orientation and depth discontinuities, and current systems in computer vision cannot extract such discontinuities with high accuracy, we should turn to another type of representation. On closer analysis, neither criticism appears to carry much weight in that (a) geons do not represent well what people do not represent well, and (b) human behavior offers an existence proof that orientation and depth discontinuities can readily be determined, bottom up. We will then review the empirical results relevant to geon theory, current theoretical work, and current challenges.

3.1. Two criticisms of geon theory

3.1.1. Criticism: geons cannot represent all objects

Geon theory offers an account of real-time entry-level shape classification by humans — what Biederman [25] referred to as 'primal access'. It will be helpful to consider what geon theory is not, as this has been the source of some misdirected commentary (e.g. [32,33]). Geon theory is not a full graphics system. Consequently, there will be many image distinctions that will not be easily depicted by geons, such as when an object or object region does not readily decompose into parts, as occurs with some sculptures and animal bodies, or when an object or region of an object is complex and/or highly irregular, as with trees or bushes. But, as predicted by geon theory, differences between such shape variations are not readily employed by people in real-time classification, nor are they treated as invariant under depth rotation, nor do they form the basis of any spontaneous classification in any culture [34,35]. Put another way, though we can see a particular oak tree in all its complexity, it is not at all clear that we require its exact shape to achieve classification, nor would we be able to readily distinguish that tree from others at a new orientation in depth, nor would we even think about the differences between that tree and other oak trees as the basis of a useful class [34,36]. The shape variations that are not easily represented by geon theory thus appear to be the same kinds of shape variations that people do not represent well. Whether the difficulty of expressing these kinds of image variations is a shortcoming of geon theory as a general purpose object recognizer may well depend on how one defines 'general purpose object recognizer'. Indeed, Jan-Olof Eklundh astutely raises the question as to whether such image variations (e.g. arising from an image of bushes) should even be considered as variations in shape.

Is it possible for geon theory, nonetheless, to provide a principled representation of highly complex irregular objects or objects that do not decompose into parts? Space precludes an extensive discussion of this issue, but it may be possible for the Hummel and Biederman network [23] to represent such objects by employing the same units that are successful for simple, regular objects. A highly complex, irregular object, such as a bush, will result in accidental synchrony preventing activation of a consistent geon structural description. The system will then 'saturate' into a texture description, defined in terms of low level (layer 2) contour and blob descriptors. Similarly, a simple but irregular object could be represented by whatever image attributes it would activate. In neither case would a geon structural description be activated, but the object could be easily distinguished from other objects only if they differed in their low level feature descriptors, such as straight or curved, an expectation consistent with the psychophysical data (e.g. [3,34]).

3.1.2. Criticism: edge extraction is difficult

Geon theory posits that the determination of the orientation and depth discontinuities in an image is sufficient for the vast majority of shape classifications that people achieve. Indeed, for multipart objects, there is virtually no difference in identification performance between full color slides and line drawings [26]. Although people evidence little difficulty in determining these discontinuities from a gray level image of unfamiliar objects, extraction of these discontinuities remains a daunting challenge to most computer vision systems. Biological vision thus furnishes an existence proof and, therefore, an inspiration to attempts at achieving edge extraction by machines. From this perspective, the appeal to stereo and motion by the 'active vision' community represents a premature admission of defeat.

3.2. Empirical results

3.2.1. Is recognition part-based?

Biederman [25] showed that recognition is impossible (median accuracy 0%) if contour is deleted from an

image of an object such that the part structure cannot be recovered. As long as the parts can be recovered, recognition can be highly accurate (median accuracy 100%) even with a considerably greater proportion of the contour removed. Results from priming experiments [2], in which the speed and accuracy of perceiving a briefly presented picture is facilitated by a prior presentation of a related picture, provide direct support for a simple, parts-based representation that is intermediate between (a) local features such as filter outputs or image edges, and (b) a global model (shape or concept) of the object. In these studies, complementary pairs of contour-deleted images of an object were created in which each image had half the vertices and edges of each geon. Presumably, the same geons could be activated from either image but through very different image features.

In a first (primary) block of trials, subjects named as quickly and as accurately as possible, one of the contour-deleted images. Each image was briefly presented (for a fraction of a second) followed by a mask. On a second block of trials, subjects named either the original or the complementary member. Even though there was no contour in common between the two members of a complementary pair, the remarkable result was that the speed and accuracy of naming the images was equivalent. Performance was much worse for objects with the same name but a different shape, e.g. an upright piano when a grand piano was initially shown, indicating that the priming was visual, rather than verbal or conceptual. A further control for specific concept priming was that complementary pairs of images composed of different (intact) geons showed no visual priming. That is, all the priming could be attributable to activation of a representation of the geons and relations – none to the activation of a specific object model, e.g. a grand piano or a round table, or the image features. A striking subjective impression is that the members of a contour-deleted complementary pair appear to be identical, as long as the same geons can be activated.

3.2.2. *Are the parts geons?*

Biederman [31] showed that variations in VIPs, such as changing the base of a lamp from a cylinder to a brick, have a more significant effect on an object's real-time classification than variations in the geon's aspect ratio. In that experiment, the magnitude of the aspect ratio differences were selected so as to be more detectable than the geon differences in a task where subjects had to judge whether two object images were physically identical. We have investigated the extent to which a two-layer recognition system [37] that has been highly successful at face recognition, could produce the advantage shown for recoverable vs. nonrecoverable images and the equivalence of complementary images. The input layer of the system consists of a lattice of Gabor-filters at various scales and orientations that can serve as an approximate model of V1 hypercolumns. The output of these kernels is mapped onto a representation layer that stores the activation values and performs standard matching functions in determining the similarity between a stored image and a probe image. The system does not explicitly represent any intermediate entities such as edges, surfaces or parts, nor does it explicitly distinguish VIPs or relations. The system showed very high accuracy in the recognition of two frame grabs of the line drawings in the previously described experiments, but its performance manifested none of the effects shown by human observers in those experiments. Its recognition performance was equivalent for recoverable and non-recoverable images, it was worse on complementary images than the originals, and it did not show the greater sensitivity to VIP compared to metric differences.

3.2.3. *Is recognition aspect invariant?*

In contrast to sets of objects that are not distinguished by geon differences, Biederman and Gerhardstein [3] showed that objects that differ in geons are recognized with little or no cost from rotation in depth, as long as the same part and relations were in view. In general, the availability of distinctive VIPs allows objects to be recognized from novel viewpoints, in contrast to what would be expected from certain 'view-based' accounts that posit no special status for VIPs. Without distinctive VIPs, enormous costs are evident in attempting to recognize objects from new orientations [31]. The invariance to viewpoint also holds for changes in size, position and mirror reflection [38].

3.3. *Modeling: Hummel and Biederman's JIM*

A neural net (NN) implementation of geon theory, JIM [23], takes a line drawing of an object as input and recruits a unit that is invariant over translation, size, and rotation-in-depth. The unit represents a structural description (geon + relations) of the object. The model's capacity for structural description derives from its original solution to the binding problem of neural networks. The input layer consists of a lattice of orientation-tuned cells. Fast Enabling Links (FELs) between cells that are collinear, coterminating, or closely parallel cause those cells that are currently activated by one geon to fire in synchrony but out of phase with the cells activated by another geon. The synchronous activity in a single time slice is used to bind independent units representing an object's geons, its attributes (e.g. vertical) and relations to other geons (e.g. above). For example, given an image of a vertical cylinder above a horizontal brick, the units for cylinder, vertical and above will be induced to fire in synchrony with one another but out of synchrony with the units for brick, below and horizontal. Limitations in the temporal bandwidth of firing results in accidental synchrony causing a

large number of elements to be perceived as texture unless irrelevant activity is suppressed by attention. In general, NN accounts have much to recommend them in this area. Their capacity to handle degraded and missing information and the availability of algorithms to exploit statistical learning are just two of the advantages of such formalisms.

3.4. Challenges

3.4.1. Are there a small number of visual primitives?

We (Sven Dickinson, Sandy Pentland, David Wilkes and myself) are investigating this issue in the context of aspect graphs. Are there a limited number of part types that would be distinguishable from all other part types over large regions of the viewing sphere generating the aspect graph? For example, a brick can be identified as such from almost any viewpoint. If the areas of the viewing sphere corresponding to each interpretation are converted to probabilities, then geons have the property of low uncertainty in the Shannon sense. Is this a good criterion for identifying part types for human recognition? One attractive feature of this conceptualization is that one can define geons to be the set of volumes that have minimal viewpoint uncertainty. Low viewpoint uncertainty, moreover, may yield optimal conditions for a self-organizing network to develop hidden units that function as geon detectors. However, aspect invariance is not the only constraint that is likely to prove important. A given aspect, such as a rectangle, might be a possible aspect for a large number of other simple volumes, e.g. a brick, a wedge, a cylinder, a curved cylinder, among others. Other aspects might be a possible projection of a small number of volumes. We thus have to consider not only the uncertainty of the mapping from volumes to aspects but, if the aspects are to serve as the basis of recognition, the uncertainty of the mapping from aspects to volumes.

3.4.2. Can the binding processes assumed by JIM be fast enough to account for real-time recognition?

Can the mechanisms of contour binding and geon activation be completed quickly enough to account for 100 msec recognition times? An elaboration of JIM by Hummel and Stankiewicz [39] posits that an unbound representation of the image might be classified in 100 msec, but the binding necessary for invariance and memory might require 300–700 msec; Biederman et al. [40] present evidence for this possibility.

3.4.3. Irregular objects

As noted earlier, geon representations are not suitable for representing objects that do not readily decompose into simple parts or are highly irregular. Differences among such objects, however, tend not to define boundaries between basic level objects nor do they evidence the viewpoint invariance shown by objects that are readily described as an arrangement of geons [35]. A possible benefit from employing a neural network representation is that it can offer some capacity for achieving recognition (through partial activation) in the face of the ambiguity often inherent in images of such objects. Nonetheless, such objects are ultimately classified and a complete account of object recognition would include how such objects are represented. In many cases, when classification cannot be achieved through a geon structural description, it is accomplished through a classification of texture, position in the scene, or other characteristics.

3.5. Middle down constraints

One advantage of assuming a geon-type of representation is that one can exploit 'middle down' properties of generalized cylinders (GCs, of which the geons are a partition) as reported by Zerroug and Nevatia [27]. There are a number of regularizing assumptions that people employ in interpreting their visual world. These assumptions include symmetry, parallelism, orthogonality of angles and the orthogonality of cross sections and axes. Unless these assumptions are made, Zerroug and Nevatia conclude that the pose, and therefore the shape, of an object may be indeterminate.

3.5.1. Structural description vs. a coordinate space

Geon theory posits a structural description in which units for explicit relations (e.g. above, perpendicular to) are bound to geons and attributes of geons (e.g. vertical axis much longer than diameter of cross section). Other requirements assume a coordinate space in which there is no distinction between parts and relations. Some evidence that is consistent with a structural description derives from Kobatake and Tanaka [41], who showed that cells in the Inferior Temporal Cortex (where, presumably, high-level object representations are stored) are tuned to complex features that are largely translationally and size invariant and appear to be distinguishable by VIPs. If it is shown that the tuning of these cells is invariant with their local context, it will be further evidence that these cells participate in an 'anding' operation in their representation of objects in a manner suggesting a structural description. It remains to be seen, however, whether cells can be found that explicitly code relations, independent of shape.

3.5.2. Development of object recognition capabilities

There is ample evidence that adult neural connectivity could not have been genetically determined. Although genetics provides a rough scaffolding for determining what statistics of images are going to affect connectivity, the actual organization must be activity dependent. Assuming that object recognition capacity develops

from viewing objects and scenes, then this experience, by definition, is viewpoint dependent. From robust statistics of such images, general aspect-invariant capacities might develop. How these aspect-invariant capacities for object recognition develop from the statistics of viewpoint-dependent experiences is one of the great challenges in the developmental neurobiology of object recognition.

3.6. Extensions to subordinate level concepts and scene perception

Although geon theory was initially proposed to account for basic-level classification, Biederman and Gerhardstein [34] discuss its extension to almost all the subordinate level classifications that people readily make. One kind of these classifications is where large geon differences exist, such as that between a round table and a square table. A second type is where the subordinate classification depends on small viewpoint invariant properties, the location of which is determined by the initial classification. Thus, people look for the name or logo in distinguishing a Toyota Camry from a Lexus, but the search for that information is dependent on an initial classification that the object is a car. Biederman [30] includes a discussion of how 'geon clusters' can form the basis of quick access to the interpretation of a scene.

3.7. Conclusion

Although geon theory is still evolving, it offers a unified account of a broad range of empirical results of real-time human entry — and most instances of subordinate-level classification performance and certain aspects of scene recognition. These processes are the fundamental routes by which we gain access to knowledge about our visual world.

4. Geons: a viable approach for computer recognition of generic objects?

Jan-Olof Eklundh
The Royal Institute of Technology

Visual object recognition is a wide and unwieldy subject on which computer vision has made but limited progress. One reason for this may be that problems range from recognizing known objects to general categorization, and that a common framework therefore may be elusive. Considering recognition of generic objects using volumetric primitives, like geons, can in this perspective form an attractive middle ground. Although it does not encompass all thinkable types of objects, since some natural objects, such as bushes, may not even be characterized by their shapes, it does provide a rich enough world to be of interest. Moreover, it is also known that geon representations allow indexing and hence rapid recognition. Finally, even though categorization problems in their most general form can hardly be addressed if only shape is used, recognition by shape primitives does apply to some of them. Nothwithstanding important advances in the past few years, there are several critical issues to be addressed before we know how far the geon-based approach will take us. I will discuss some of these and how I think they can be treated at least in a general sense.

The first issue concerns how geons can be recovered from imagery. Most high level work has in this field, like in other recognition work, been based on perfect or near-perfect line-drawings. This is certainly not a realistic way to go. However, geons are qualitative in nature, and could be derived by some shape abstraction principle, which fits or associates primitive shapes to data. Several schemes in this vein have been suggested in the literature. However, a crucial requirement is that the typical geon properties should be made explicit, e.g. properties like being flat or curved, growing or tapering, etc. This is not generally the case for approaches based on fitting methods, and may require non-linear techniques or cause uniqueness problems. Another important aspect is the treatment of structures at different scales. Generally speaking, I don't think we know how to handle these problems, but I do believe that there exist promising methods.

The second issue is about the use of depth information. In his work on human recognition, Biederman is using line-drawings only. I think one must be careful in drawing conclusions about machine vision approaches from that. The display situations are such that the subject sees the object clearly separated from other objects in the fore- and background. Unless qualitative depth is provided either from accommodation cues or from binocular or other cues, a machine vision system using images with a large depth of field is faced with difficulties. Truly, Lowe and others have shown that recognition can be done in static monocular images. However, I think that then the possibilities of performing rather complex types of reasoning are needed, which goes against the general advantages of geon-based recognition. In principle, I feel that the problem should and can be posed in the presence of qualitative depth information, which can be used for segmentation and disambiguation. On the other hand I think that having dense depth information from stereo reconstructions or range sensing defines a completely different problem, and also a problem that is mainly of a technical character.

The third issue concerns whether the full geons need to be recovered or if observed subparts can have enough indexing power. Biederman has convincingly demonstrated that partial information is sufficient. However, this is again in a context where you don't see a large

scene collapsed in depth. If those conditions are valid, partial information may not be constraining enough and can lead to combinatorial problems. On the other hand, if it seems as if there is some other process selecting a region of interest, or generally guiding the attention, then there are viable computational techniques which could be based on typical partial information that we can obtain from real images.

Generally, I think that geon-based recognition of generic objects can be a powerful approach, even though it has limitations. However, I also think that the technique must be applied in a context where the issues I've listed, and also other problems, can be addressed appropriately.

5. Geon-based recognition of 3D objects

Anil K. Jain
Michigan State University

Recognition by components (RBC) theory, proposed by Biederman [25], offers an explanation of human object recognition supported by several psychophysical studies. It maintains that a 3D object is decomposed into qualitative volumetric part-based primitives, called *geons*, at the concavities of an object. The *primal access* to geons and the extraction of their structural relationship accounts for the quick and reliable recognition of objects by humans. The theory obviates an explicit 3D reconstruction for the recognition; it maintains that geon structure of an object could be constructed from the line drawing extracted from its (2D) image. Both the extraction of the perfect line drawing from an intensity image as well as the subsequent extraction of geons from their silhouettes are challenging and open research problems.

In many vision applications, 3D information about the scene is necessary (e.g. navigation, grasping, manipulation) and can be sensed (e.g. stereo, range finders). While the RBC theory is primarily advanced in the context of extraction of geons from 2D intensity images, similar strategies could be used for extracting geons from 3D (depth, range) data. These representations, whether derived from intensity or range images, could then be used for 3D object recognition [8,28,42].

The information offered by geons is only qualitative in nature. Therefore, geons are not adequate for describing and discriminating objects which are similar in their coarse part structure (e.g. sculpted or free-form objects). In such situations, geon structure of an object is mostly useful for an indexing of a given object into the model database. Quantitative models (e.g. superquadrics, algebraic surfaces, generalized cylinders) can then be used for further refining and verifying the initial hypotheses about the input object. Therefore, it is advantageous and necessary to augment the qualitative information offered by the geon-based representation by a quantitative representation strategy. In such situations, the recovery of geon structure might proceed from other quantitative representations[1]. Our research in geon-based recognition addresses several issues, including segmentation of the 3D data into volumetric parts, recovering quantitative models from the data belonging to each of the parts, and estimating geons, given the quantitative representation primitives.

Our early work in geon-based recognition studied the recovery of geons from the superquadric models [17]. The results of the study indicated that recovery of the geons from superquadric models is indeed feasible for synthetic data and 'smooth' real range data. We are currently working on improving the accuracy of such a recovery from real data by use of surface shape information. We are also investigating the problem of recovery of geons from algebraic surfaces and juxtaposing merits of such a recovery with the use of superquadrics [43].

The part segmentation of a range image requires identification of loci of local minima of concavities. In reality, the extracted local minima of the concavities are noisy and cannot be directly used for part segmentation. We have proposed a method for obtaining a reliable part segmentation of the depth data by integrating principles of *gestalt* and transversality [44]. Our approach can successfully segment a large number of range images into component parts and is a viable alternative to the existing multiscale approaches to reliable part decomposition.

Most of the research studies focusing on geon-based object recognition work with 2D intensity images. Geon-based recognition from 3D depth data has not received as much attention. We have developed a recognition system which uses geons as the primary representation primitive. Recognizing the fallibility of the principle of transversality to recover the part structure of an object [45], our system reliably recovers the geon structure of a 3D object from (i) a representation describing surface configurations of the input object (surface adjacency graph, or SAG), and (ii) the intensity information associated with the range image. The resultant representations could then be matched using a graph matching algorithm [42].

Representation of an arbitrary object in terms of a few *generic* primitives is an attractive and intuitive idea. However, it appears that much more research is required to establish its feasibility in solving real-world recognition problems. Given our lack of understanding of the computational issues in low- and intermediate-level vision processes, a reliable recovery of the qualitative features of *parts* as well as their interrelationships

[1] Although, qualitative information could also be used to *direct* the recovery of quantitative information, quantitative information cannot be recovered from the qualitative models.

is difficult. In the short term, we could do much better by adopting an engineering approach to the recognition problem.

6. Geons and the 'what'-function

Roger Munck-Fairwood
Li Du
University of Surrey

Generic object recognition may be regarded as the attempt to fulfill the 'what'-function of machine vision. In addressing the first of the four issues in the introduction, we identify a fundamental question which must be answered if generic object recognition (as it is currently understood) is to be posed as a scientific pursuit. The question is: How can the 'what'-function be *defined* for both human and machine vision? We present a formal definition of this function. We address the second issue by presenting a working implementation for extracting geons from real images. We propose a system architecture which addresses the third issue, and discuss the fourth issue — about active vision — with respect to the 'what'-function.

6.1. The 'what'-function and the purpose concept

We define the 'what' function as a function which maps a 3D entity into a predefined bin of classification on the basis of some particular recovered geometrical properties. Let us assume that W is the vocabulary for describing the geometry of 3D entities (as deliverable by a collection of lower level vision processes), C is the set of predefined classification bins, and M is the classification or mapping process. A single task of recognition can then be formally expressed as $c = M(w)$, where $c \in C$ and $w \in W$.

This formalism helps to critically examine the common statement that it is desirable for a generic object recognition system to work 'reliably'. Informally, one may mean by this that when the system is presented with images of different objects, all of which we would naturally call, say, 'cup', it should deliver the label 'cup'. But the above formalism shows that this would require a universal (all-purpose) C. Clearly, this is an unreasonable assumption because even given a perfect w (geometrical description of an object), the way the bins are designed strongly depends on either engineering criteria or cultural customs.

If C is defined according to engineering criteria, i.e. pragmatic geometric classification, then we have moved away from the generally accepted meaning of 'generic' object recognition and closer to multi-model-based recognition. On the other hand, if we aim at defining C to represent 'natural' categories we have a hopeless task, as even within one culture we cannot accurately define W, C or M, i.e. precisely what geometric characteristics are used in distinguishing named objects from each other. Moreover, many overlapping Cs exist.

It may be argued that one can always resort to W (object geometry) in order to achieve a universal unambiguous object description. However, even if low-level processes could provide any desired degree of accuracy in W, one could only achieve a description and not classification of an object, i.e. no 'recognition' would have occurred.

Hence, we argue that a C must be defined and that it cannot be universal or totally 'generic'. In fact, the above formalism points the way to being able to *measure* the generality ('genericness') of any generic object recognition system. It shows that one cannot use the cardinality of W or C (i.e. the number of geometrically distinct descriptions, or the number of object categories) as a measure of genericness, as these can always be made arbitrarily large (e.g. by introducing finer distinctions in geon classes). In that case, we would not necessarily call such systems 'highly generic' because many or most of the distinctions within W or C may be practically useless, i.e. have no significance for the intended *use* of the recognition system.

Instead, we suggest that the way towards measuring genericness is to assess (a) the richness of the set of purposes P which a given C supports and (b) the versatility of the W (shape recovery) system, the outputs of which can be mapped to C. Thus, we see that the pursuit of generic object recognition can never result in a self-contained, universal system but must be bounded by purposes at the top end and richness of geometric description and recovery ability at the bottom end.

6.2. A framework for an object recognition system

The above formalism also leads to an outline of how we see the design of generic object recognition systems could usefully proceed. The key idea is to keep the shape recovery subsystem (W-subsystem) and object classification subsystem (C-subsystem) conceptually separate. It is not always clear where this distinction lies in the reported work on generic object recognition systems, and it can therefore be very difficult to judge how 'generic' a given system is likely to be.

We have outlined [46] a W-subsystem which would provide a versatile repertoire of image labeling tools. The basic principles behind the system design include the following: (a) it is a *servant* to a C-subsystem and has a high degree of transparency in the sense of making process parameters (scale, thresholds, etc.) explicit; (b) the architecture has the potential to incorporate any existing approaches of 3D geometry recovery; (c) the image labeling vocabulary is divided both into levels of abstraction and levels of cost; and (d) reasoning occurs

in three directions: bottom-up (b-u), top-down (t-d), and horizontally. When b-u and t-d are in conflict, either b-u can dominate (corresponding to visual illusion in human vision) or t-d can dominate (corresponding to subjective contours). We illustrate the system using 2D edge-based features at the lower end and geon clusters at the higher end.

The C-subsystem has two-way communication with the W-subsystem. It not only takes b-u input from the latter but gives cognitive (predictive) information to it when possible. Its basic function is to take the recovered geometry at the geon-cluster level of the W-subsystem and to use M to map each cluster into the appropriate bin of C. In a given engineering task, the human operator may use the definition of a particular set of purposes P to manually define C and M, thus simplifying the task of this sub-system. But in general, the sub-system should automatically create C and M according to a given P.

6.3. Geon extraction by robust feature grouping

We refer now [15] to a working implementation of a small and specialized W-subsystem capable of extracting geon-type descriptions from real images. Progress in geon-based generic object recognition has been hindered by the lack of effective means to handle noisy and cluttered images. In fact, many schemes are aimed at handling line-drawing input or data of very good quality. A classical problem is involved: to find an effective method for grouping (i.e. determining which edge features correspond to which geon) in the presence of noise, clutter and occlusion. It seems that the lack of such a strong 'perceptual group' has hindered the ability of existing systems to cope with real images.

We present the multilateral (at least six sides) convex polygon as the key perceptual group to facilitate a robust feature grouping method. The grouping power of such convex closures comes from the fact that they have much greater excluding power than open perceptual groups (such as parallel and co-terminating groups). In contrast to minimum closures used in earlier systems, multilateral closures are also far less likely to arise as a result of accidental configuration of surface markings. To exploit this perceptual group, we introduce a new coarse level of geon representation so that the silhouette of such a geon can always be identified with a convex polygon. Subsequent focussed reexamination of the image data can determine the traditional (e.g. curved/straight) geon attributes, once main geon features (sides, ends, etc.) have been localized.

The initial input is a set of lines extracted from an image. The recognition process is organized into iterations. The first iteration is responsible for recognizing all unoccluded geons. It finds convex polygons, generates and verifies geon hypotheses and removes explained data from the rest. All subsequent iterations aim to extract the remaining geons which are occluded by previously recognized geons. These iterations are similar to the first iteration except that they use partially occluded convex polylines to initiate extraction of occluded convex polygons (i.e. polygons broken by previously recognized geons, implying partially occluded geons) which in turn lead to hypothesis and verification of occluded geons. The whole process terminates when no more geons can be found. Experiments have shown that this system handles images which are much more noisy and cluttered than those used as input to previously reported systems.

6.4. About active vision

Our definition of generic object recognition shares the 'purpose' concept with that invoked in the study of active vision, but from an entirely different point of view. Currently in active vision research, the emphasis is shifted towards sensor-motor interaction, and the 'purpose' concept is used to circumvent the need for the 'what'-function (e.g. by gazing on a bright spot or a color pattern). In contrast, we employ the 'purpose' concept to directly address the issue of object recognition by linking P with C. We consider object recognition as a key component for a general active vision system, where appropriate actions can only be taken *after* 'what' is answered.

Elements of P may be, for example, 'pick up the container', and 'put this object on another which can support it'. Each element of P can be used to define one or more elements of C which can serve it. A sufficiently rich W is then required to support the necessary distinctions between elements of C (e.g. containment, object support).

Viewed in this way, the W which geons provide is unlikely to be able to support the C (and in turn P) for an active vision system because object category concepts such as ability to contain or support — and even absolute object size — are not part of the model. However, if strong ad hoc assumptions are made about such properties of particular geon clusters (e.g. 'the geon cluster named "cup" is usually of this size and can contain liquid...') then geons may provide a convenient way of indexing into a large model database of objects which must be manipulated.

6.5. Summary

The question of whether a geon vocabulary (W) is appropriate for generic object recognition cannot be answered in isolation without reference to a set of purposes P. The question of what classes of objects C they are best suited for also cannot be answered without P. In fact, the question of whether the geon representation can be improved to be 'more usable' implies that an

intended use or purpose exists. It is not productive to search for the 'best vocabulary' (geons or otherwise) for generic object recognition as a universal (all-purpose) system. However, once P is defined, one can begin to quantitatively compare the 'genericness' of different systems.

The major obstacles in recovering geons from real imagery have their roots in the lack of effective image feature groups. The multilateral convex polygon is a primary group which can give greater robustness than groups employed hitherto.

An appropriate system architecture is where the shape recovery and object classification subsystems are clearly distinct, but communicating two-way. Both subsystems can benefit from development efforts: producing better general-purpose shape recovery systems, and automated ways of deriving shape classification systems given a set of purposes.

For active vision systems where object manipulation may be required, the above P clearly becomes explicit: the class of the object determines whether or how it should be manipulated. Hence, geon-based recognition can play a role if assumptions are made about the absolute size and other properties of specific geon clusters (object classes) — otherwise, an object representation which involves absolute size must be used.

7. Parts representations

Alex Pentland
Massachusetts Institute of Technology

The representation of objects by their parts has a long tradition in computer-aided design, simulation, and cognitive psychology. Indeed, in these areas it is the dominant strategy for representing complex 3-D objects. It is absolutely clear, therefore, that part representations are excellent for many computational and cognitive tasks. What is not so clear is how they might be useful in computer vision.

The first parts representation was suggested by Binford [47]; this is the idea of generalized cylinders (GC). Because of the relatively unconstrained nature of this representation, the recovery process seems to require elaborate line grouping and reasoning. Consequently there are relatively few reports of recovering such descriptions from real imagery. Moreover, because such descriptions are often not unique, it is unclear how they aid in object recognition.

About 1985, these problems with GC representations began to be taken seriously, and two different proposals arose to deal with them. One variation on the GC idea is due to Biederman [1], who suggested using the Cartesian product of qualitative properties such as tapering, cross-section, etc., to create a qualitative taxonomy of GC. One advantage of this type of representation, called geons, is that the properties can be chosen to be ones that are more easily recovered from imagery. Another is that it provides a way to define qualitative shape classes, an important problem in general-purpose vision.

To date, however, only Dickinson, Pentland and Rosenfeld [9], and more recently, Du and Munck-Fairwood [15], have successfully used this approach to recognize objects in real 2-D imagery and Raja and Jain and Levine et al. [6] from real range imagery, although Bergevin and Levine [5], Hummel and Biederman [23], and Munck-Fairwood [14] have reported good success in interpreting line drawings. The principal problems with this approach seem to be the difficulty in extracting sufficiently good line drawings, and the idealized nature of the geon representation.

A different variation of the GC idea was suggested by Pentland [10], who proposed a parametric version of GC based on deformable implicit functions. The original version of this proposal used superquadric functions with multiplicative deformations; later versions used spheres with additive ('Modal') deformations. Use of a parameterized implicit function (deformed superquadrics or spheres) converts the problem of recovering a description into a relatively simple numerical optimization. Moreover, if the parameterization is orthogonal then the description is unique, making the recognition problem much easier.

In recent years, several authors have reported success at recovering this type of description and using it for recognition, e.g. Pentland and Sclaroff [11] were able to recognize people's head shape with 96% accuracy. The major limitations of this approach seem to be the requirement for pre-existing part segmentation, and an initial estimate of orientation.

It is interesting to compare the geon and parametric approaches. Both describe a similar range of shapes; perhaps the major difference is that one is qualitative while the other is quantitative. This similarity has made it possible to generate geon statistics using parametric models [8], and allowed Raja and Jain [17] to use superquadrics to recover geons.

It seems to me that the qualitative geon representation is best suited to understanding intersections and groupings of parts, while the parametric representation is best for classifying individual part shapes. I speculate that the most successful systems will employ first a qualitative pruning/grouping stage to perform segmentation, and then a parametric fitting stage to extract detailed part shape. This approach was recently successfully employed by Dickinson and Metaxas [48].

A final question is representational adequacy. It seems clear that to model natural objects the GC representation must be augmented by a displacement map, as in Sclaroff and Pentland [12]. Moreover, the GC representations

currently used in vision are much simpler than those used in CAD. Consequently, I predict that the more expressive GC representations [11,12] will become increasingly popular.

8. The obstacles facing practical geon-based recognition

Sven Dickinson
Rutgers University

I see three major obstacles facing the success of geon-based object recognition systems: (i) the recovery of geons from real imagery; (ii) the difficulty in *explicitly* modeling real objects using geons; and (iii) the lack of representational power provided by geons for the task of interacting with the world.

8.1. Recovery

In our work on geon-based recognition, we have attempted to address the issues in building geon-based systems that work on real imagery [8,9]. We chose an approach to volumetric primitive recovery based on part-based aspect matching which, in fact, can be applied not only to geons, but to other classes of volumetric primitives as well. Unlike traditional aspect graph representations which use aspects to model the various appearances of a *complete* object, we use aspects to model a small, finite part vocabulary from which objects are constructed. By mapping aspects to parts, we avoid the intractability of traditional aspect graphs (particularly when the object's parts articulate). By indexing with object-centered volumetric parts, we retain the compactness of an object-centered model representation.

Given our goal of recovering the aspects corresponding to an object's volumetric parts, how do we proceed to recover these aspects? From our analysis, we discovered that image regions, not lines or line groups, were the most appropriate features to extract from an image [8,48]. Image regions not only provide a natural grouping of image contours, i.e. the bounding contours of a region, but region shape provides a powerful inferencing mechanism for aspect, and hence volume recovery. Our *aspect hierarchy*, a multi-level, viewer-centered representation of a set of volumetric parts, allows us to recover volumetric parts in the presence of noise, occlusion, and 'degenerate' views, by exploiting the viewing probabilities of a volume's aspects.[2]

Despite the utility of regions in recovering volumetric parts, we are still faced with the problem of quickly extracting regions from an image. In addition, how do we deal with the inevitable under- and over-segmentation of image regions? And finally, at what stage in the recovery of geons, do we attempt to rectify such segmentation problems: immediately following segmentation? at the geon verification stage? or, at the object verification stage? In working with real images, it becomes very clear that the segmentation cannot be decoupled from the recognition. I believe that a dynamic feedback scheme between the various stages in the recognition process must be explored before a geon-based object recognition system can be successfully applied to real images. We are currently exploring feedback schemes applied to segmentation and recognition.

8.2. Modeling

One of the most challenging tasks in building a geon-based recognition system was finding geon-based objects to test it on! In fact, in early experiments, we ended up constructing our own geons in order to build our own objects. That is not to say that there are not geon-based objects in the world. However, what we found was that although many of the real objects we did look at could be coarsely described in terms of geons, there was a plethora of detail in the objects that did not contribute to their coarse shape. Even ignoring surface markings and texture, there would be a lip on a cup, a small ridge around the edge of a table, or other fine object structure whose projected contours in the image were quite salient. Although Biederman suggested the use of nonaccidental relations between contours in recovering geons, neither he nor the line-drawing based geon-recovery systems which he inspired addressed the issue of structural scale.

Most recognition approaches deal with this problem by explicitly modeling the structural detail of the object. In such CAD-based vision systems, which model an object's exact geometry, each salient image contour has a corresponding model feature (surface discontinuity or occluding boundary). Although such vision systems are effective in manufacturing environments where object databases are small, objects are rigid, and objects can be explicitly modeled, they offer little hope for recognition in less constrained environments. I believe that geons are an appropriate modeling primitive for certain recognition domains where objects can be modeled as constructions of distinct volumetric parts. However, we must be able to recover abstract geon (or other volumetric) descriptions from images of objects with structural detail. Once geons can be abstracted from real objects, then further modeling abstraction can be pursued. For example, modeling an object's functionality instead of its geometry provides an even more powerful and robust means for modeling particular classes of objects. Reasoning about the shape and interaction of an object's volumetric parts can provide 'functional' indices into a database of objects defined by their function [50].

[2] For a discussion on the significant likelihood of degenerate views, see Wilkes et al. [49].

8.3. Representational power

For a vision system to interact with the world requires that it not only identify or recognize the objects that it wishes to interact with, but that it sufficiently localize (in depth) and characterize (in dimension) the objects in order to manipulate them. Moreover, a vision system should be active and be able to guide itself to advantageous viewpoints of an object. Finally, as the vision system moves through its world, it must be able to track the (possibly moving) objects in its field of view. In proposing geons, Biederman addressed only the problem of bottom-up, or unexpected, object recognition. Is there a representational framework for geons or, more generally, generic volumetric part descriptions, that will support the recognition-related behaviors of unexpected or bottom-up object recognition, expected or top-down object recognition, object localization, metric shape recovery, active object recognition (viewpoint control), and object tracking? I feel that the problem of generic shape representation must be viewed in this wider context of object recognition.

As a controlled experiment to examine the interaction between these behaviors, we are building a vision-based mobile robot whose goal is to allow handicapped children to interact with their environment through the use of a robotic manipulator. The child interacts with the system by instructing it to find and manipulate toys on a table (top-down recognition) [48]; hence, we need object identification, object localization, and metric shape recovery (for gripping) [18,47,51]. Alternatively, when the child is idle, the system can build up its database of objects in its domain (bottom-up recognition) [8,9]. If objects appear to be ambiguous or are heavily occluded, the system must move to a better viewpoint [52] and, while moving, the system must track the objects [53,54]. We have successfully addressed each of these problems within the framework of our hybrid object-centered/viewer-centered representation.

9. Conclusions

The field of object recognition by computer has been dominated by approaches assuming knowledge of the exact geometry of the objects to be recognized. Typically, any salient image feature, such as a line, a curve or a corner, has a corresponding feature in the object's model. Insisting on this direct correspondence has led to model-based vision systems which are sensitive to minor changes in the shape of the object, making them wholly inappropriate for generic or prototypical object recognition.

When a human looks at an object, they can easily describe its shape in terms of a set of abstracted parts. What is most impressive is the human's ability to extract the coarse shape of the object's parts, even when the parts contain superfluous structural detail that may give rise to very salient image structures. Biederman has proposed a set of volumetric shape abstractions, called geons, that offer a representation for this coarse shape. Granted, not all objects can be conveniently decomposed into volumetric parts; for example, amorphous objects and objects such as faces do not lend themselves to geon-based description. However, a great many objects, both natural and artificial, exhibit a definable part structure. The key contribution of Biederman's RBC theory to the computer vision community is the idea that one can define a set of volumetric part classes by permuting the values of qualitative, viewpoint-invariant properties of lines.

Since Biederman's introduction of geons to the computer vision community, much progress has been made in geon recovery from line drawings and images of simple objects. Such approaches have proposed different representations, architectures and control strategies, and have been applied to both intensity and range images. However, there is a strong consensus that one of the major obstacles to success is the problem of geon extraction from both intensity and range images of real objects. As in the traditional CAD-based vision community, the geon community has also avoided the problem of shape abstraction by choosing simple, smooth objects in which every salient image feature represents either a shape-defining surface discontinuity or occluding boundary on a geon. The matching of image detail to qualitative descriptions remains largely an unsolved problem.

Since an intelligent agent will move about and interact with its environment, there is additional support among some panel members for widening the context of geon-like vision to include, for example, active vision and metric shape recovery. How do we track geon-based objects in 2-D images? How do we recover quantitative shape information when we want to grasp an object? How do we change viewpoint to disambiguate an object? What role do geons play in these tasks?

The goal of generic object recognition is both important and ambitious. Until geon practitioners can demonstrate effective techniques for recovering geons or other qualitative shape primitives from images of real objects, the computer vision community and the human vision community (see [32,33]) will understandably remain skeptical. However, until coarse shape can be quickly extracted from scenes of real objects, 3-D object recognition will be restricted to overconstrained, industrial environments.

References

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

[2] I. Biederman and E. Cooper, Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. Cognitive Psychology, 23 (1991) 393–419.

[3] I. Biederman and P. Gerhardstein, Recognizing depth-rotated objects: evidence for 3d viewpoint invariance. J. Experimental Psychology: Human Perception and Performance, 19 (1993) 1162–1182.

[4] R. Bergevin and M.D. Levine, Recognition of 3-D objects in 2-D line drawings: An approach based on geons. Technical Report TR-CIM-88, Center for Intelligent Machines, McGill University, November 1988.

[5] R. Bergevin and M.D. Levine, Generic object recognition: building coarse 3D descriptions from line drawings. Proc. IEEE Workshop on Interpretation of 3D Scenes, pp. 68–74, Austin, TX, 1989.

[6] M. Levine, R. Bergevin and Q.L. Nguyen, Shape description using geons as 3D primitives. Proc. International Workshop on Visual Form (IWVF), Capri, Italy, May 1991.

[7] S. Dickinson, A. Pentland and A. Rosenfeld, A representation for qualitative 3-D object recognition integrating object-centered and viewer-centered models. Technical Report CAR-TR-453, Center for Automation Research, University of Maryland, 1989. (Also appears with the same title in: K.N. Leibovic (ed.), Vision: A Convergence of Disciplines, Springer-Verlag, NY, 1990, pp. 398–421.)

[8] S. Dickinson, A. Pentland and A. Rosenfeld, 3-D shape recovery using distributed aspect matching. IEEE Trans. Pattern Analysis and Machine Intelligence, 14(2) (1992) 174–198.

[9] S. Dickinson, A Pentland and A. Rosenfeld, From volumes to views: An approach to 3-D object recognition. CVGIP: Image Understanding, 55(2) (1992) 130–154.

[10] A. Pentland, Perceptual organization and the representation of natural form. Artificial Intelligence, 28 (1986) 293–331.

[11] A. Pentland and S. Sclaroff, Closed-form solutions for physically based shape modeling and recognition. IEEE Trans. Pattern Analysis and Machine Intelligence, 13(7) (1991) 715–729.

[12] S. Sclaroff and A. Pentland, Generalized implicit functions for computer graphics. ACM Computer Graphics, 25(2) (1991) 247–250.

[13] R. Fairwood, Recognition of generic components using logic-program relations of image contours. Image and Vision Computing, 9(2) (April 1991) 113–122.

[14] R. Fairwood and G. Barreau, A belief network for the recognition of 3-d geometric primitives. Proc. 6th International Conference on Image Analysis and Processing, Como, Italy, September 1991.

[15] L. Du and R. Munck-Fairwood, Geon recognition through robust feature grouping. Proc. 9th Scandinavian Conference on Image Analysis, Uppsala, Sweden, June 1995.

[16] A. Jacot-Descombes and T. Pun, A probabilistic approach to 3-D inference of geons from a 2-D view. Proc. SPIE Applications of Artificial Intelligence X: Machine Vision and Robotics, pp. 579–588, Orlando, FL, 1992.

[17] N. Raja and A. Jain, Recognizing geons from superquadrics fitted to range data. Image and Vision Computing, 10(3) (April 1992) 179–190.

[18] S. Dickinson, D. Metaxas and A. Pentland, Constrained recovery of deformable models from range data. Proc. 2nd International Workshop on Visual Form, Capril, Italy, May 1994.

[19] K. Wu and M. Levine, Recovering parametric geons from multi-view range data. Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 159–166, Seattle, WA, June 1994.

[20] J.-O. Eklundh, M. Li and G. Olofsson, Representations for 3-D geometric modeling. Technical Report IR.C.1.1., ESPRIT BRA-3038, Vision as Process, May 1990.

[21] J.O. Eklundh and G. Olofsson, Geon-based recognition in an active vision system. In ESPRIT-BRA 3038, Vision as Process, Springer-Verlag ESPRIT Series, 1992.

[22] J. Hummel, I. Biederman, P. Gerhardstein and H. Hilton, From edges to geons: A connectionist approach. Proc. Connectionist Summer School, pp. 462–471, Carnegie Mellon University, June 1988.

[23] J. Hummel and I. Biederman, Dynamic binding in a neural net model for shape recognition. Psychological Review, 99 (1992) 480–517.

[24] S. Dickinson, I. Biederman, A. Pentland, J.-O. Eklundh, R. Bergevin and R. Munck-Fairwood, The use of geons for generic 3-D object recognition. Proc. International Joint Conference on Artificial Intelligence (IJCAI), pp. 1693–1699, Chambery, France, August 1993.

[25] I. Biederman, Recognition-by-components: A theory of human image understanding. Psychological Review, 94 (1987) 115–147.

[26] I. Biederman and G. Ju, Surface vs. edge-based determinants of visual recognition. Cognitive Psychology, 20 (1988) 38–64.

[27] M. Zerroug and R. Nevatia, Volumetric descriptions from a single intensity image. International J. Computer Vision, 20 (1996) 11–42

[28] R. Bergevin and M.D. Levine, Generic object recognition: Building and matching coarse 3d descriptions from line drawings. IEEE Trans. Pattern Analysis and Machine Intelligence, 15 (January 1993) 19–36.

[29] J.-L. Arseneault, R. Bergevin and D. Laurendeau, Grouping of straight line segments and circular arcs for scene analysis. Proc. Vision Interface '94, pp. 137–144, Banff, Alberta, May 1994.

[30] I. Biederman, Aspects and extensions of a theory of human image understanding. In: Z. Pylyshyn (ed), Computational Processes in Human Vision, pp. 370–428. Ablex, New York, 1988.

[31] I. Biederman, Visual object recognition. In: S. Kosslyn and D. Osherson (eds), An Invitation to Cognitive Science, 2nd Ed, pp. 121–165, MIT Press, 1995.

[32] M. Kurbat, Structural description theories: Is rbc/jim a general purpose theory of human entry-level object recognition? Perception, 23 (1994) 1339–1368.

[33] E. Leeuwenberg, P. van der Helm and R. van Lier, From geons to structure: A note on object representation. Perception, 23 (1994) 505–515.

[34] I. Biederman and P. Gerhardstein, Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff. J. Experimental Psychology: Human Perception and Performance, 21 (1995) 1506–1514.

[35] E. Cooper, S. Subramaniam and I. Biederman, Recognizing objects with an irregular part. Poster, Meetings of the Association for Vision and Opthalmology, Ft. Lauderldale, FL, 1995.

[36] I. Biederman and E. Cooper, Viewpoint invariant differences during object recognition are more salient than metric differences. Submitted.

[37] M. Lades, J. Vorbruggen, J. Buhmann, C. von der Malsburg, R. Wurtz and R. Konen, Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. Computers, 42 (1993).

[38] E. Cooper, I. Biederman and J. Hummel, Metric invariance in object recognition: A review and further evidence. Canadian J. Psychology, 46 (1992) 191–214.

[39] J. Hummel and B. Stankiewicz, An architecture for rapid, hierarchical structural description. In: T. Inui and J. McClelland (eds), Attention and Performance XVI: Information Integration in Perception and Communication, pp. 93–121, MIT Press, Cambridge, MA, 1996.

[40] I. Biederman, S. Subramaniam and S. Madigan, Chance forced choice recognition memory for identifiable rsvp object pictures. Meetings of the Psychonomics Society (presentation), St. Louis, MO, November 1994.

[41] E. Kobatake and K. Tanaka, Neuronal selectivities to complex object features in the ventral pathway of the macaque cerebral cortex. J. Neurophysiology, 71 (1994) 856–867.

[42] N. Raja and A. Jain, Obtaining generic parts from range images using a multi-view representation. CVGIP: Image Understanding, 60(1) (July 1994) 44–64.

[43] S. Pankanti, A. Jain and G. Taubin, An evaluation of 3D object recognition system using algebraic surface primitives and superquadric primitives. Technical Report TR-CS-95-66, Michigan State University, January 1995.

[44] S. Pankanti, C. Dorai and A. Jain, Robust feature detection for 3d object recognition. Proc. SPIE Conference on Geometric Methods In Computer Vision II (Vol. 2031), pp. 366–377, San Diego, CA, July 1993.

[45] D. Hoffman and W. Richards, Parts of recognition, Cognition, 18 (1985) 65–96.

[46] L. Du and R. Munck-Fairwood, A formal definition and framework for generic object recognition. Proc. 8th Scandinavian Conference on Image Analysis, University of Tromsø, Norway, May 1993.

[47] T. Binford, Visual perception by computer. Proc. IEEE Conference on Systems and Control, Miami, FL, 1971.

[47] S. Dickinson and D. Metaxas, Integrating qualitative and quantitative shape recovery. International J. Computer Vision, 13(3) (1994) 1–20.

[48] S. Dickinson, H. Christensen, J. Tsotsos and G. Olofsson, Active object recognition integrating attention and viewpoint control. Proc. ECCV '94, Stockholm, Sweden, May 1994.

[49] D. Wilkes, S. Dickinson and J. Tsotsos, A quantitative analysis of view degeneracy and its application to active focal length control. Proc. International Conference on Computer Vision, Cambridge, MA, June 1995.

[50] E. Rivlin, S. Dickinson and A. Rosenfeld, Recognition by functional parts. Computer Vision and Image Understanding, 62(2) (1995) 164–176.

[51] S. Dickinson, D. Metaxas and A. Pentland, The role of model-based segmentation in the recovery of volumetric parts from range data. IEEE Trans. Pattern Analysis and Machine Intelligence (to appear).

[52] S. Dickinson, H. Christensen, J. Tsotsos and G. Olofsson, Active object recognition integrating attention and viewpoint control. Computer Vision and Image Understanding (CVIU) (to appear).

[53] S. Dickinson, P. Jasiobedzki, H. Christensen and G. Olofsson, Qualitative tracking of 3-D objects using active contour networks. Proc. IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, June 1994.

[54] M. Chan, D. Metaxas and S. Dickinson, A new approach to tracking 3-D objects in 2-D image sequences. Proc. AAAI '94, Seattle, WA, August 1994.