Object Representation for Object Recognition

Jean Ponce
University of Illinois

Ruzena Bajcsy and Dimitri Metaxas
University of Pennsylvania

Thomas O. Binford
Stanford University

David A. Forsyth
University of Iowa

Martial Hebert and Katsushi Ikeuchi
Carnegie-Mellon University

Avinash C. Kak
Purdue University

Linda Shapiro
University of Washington

Stan Sclaroff and Alex Pentland
Massachusetts Institute of Technology

George C. Stockman
Michigan State University

Abstract: *This paper discusses some representation issues and challenges involved in object recognition. It is intended as a step toward assessing current object representation schemes and proposing design and evaluation criteria for future ones. The presentation is based on position statements by panelists at the 1994 IEEE Workshop on CAD-Based Vision and the 1994 IEEE Conference on Computer Vision and Pattern Recognition.*

1 Introduction

This paper is a collection of short position statements by researchers interested in object representation. Before leaving the stage to my colleagues, I (J. Ponce) would like to explain why and how it came about. The context is object recognition, that is, the identification in an image of a 3D object selected out of a collection of models. (This has to be distinguished from other model-based vision problems such as in-door robot navigation for example.)

Object recognition is without doubt a successful and healthy area of computer vision, with very impressive progress accomplished over thirty years or so of research: implemented systems now exist for recognizing objects modeled by polyhedra, superquadrics, generalized cylinders, algebraic surfaces, and even free-form surfaces in range and in video images. Despite these undeniable successes, most of today's recognition systems are still aimed at recognizing a handful of simple object models in clean, uncluttered images. Tomorrow's systems should demonstrate the recognition of complex man-made and natural objects among many –a thousand or more– candidates from noisy and cluttered images. This will require addressing extremely difficult –and I think fundamental– problems such as the automatic construction and indexing of large model databases of free-form objects, or the description of object classes.

I think that it is time to evaluate the progress achieved and try to anticipate what the future holds. A string of different object representations has been a key factor in past successes, and I believe that representation will remain a key factor, maybe *the* key, to future successes and to the wider application of computer vision techniques in industrial settings. (Clearly, representational and algorithmic issues are inter-mixed, as future recognition approaches will have to deal with the combinatorial problems due to clutter in both the model database and the image.)

This paper is intended as a first step toward several complementary goals: first, identifying the main issues and promising research directions in object representation for object recognition; second, evaluating the current representation schemes and identifying design and evaluation criteria for future ones; third, challenging the computer vision community to address the identified problems and to develop and implement tomorrow's representation paradigms.

The presentation is based on two panels, the first one held at the 1994 IEEE Workshop on CAD-Based Vision (WCBV), and the second one to be held at the 1994 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). The rest of the paper is a collection of statements by the panelists, with a few words of introduction and discussion by M. Hebert and myself.

2 The WCBV Panel

A first panel on Object Representation for Computer Vision was organized by J. Ponce at the second IEEE Workshop on CAD-based Vision in February 1994. The panelists were M. Hebert, K. Ikeuchi, A.C. Kak, D. Metaxas, S. Sclaroff, L. Shapiro, and G.C. Stockman. Two distinct approaches to the object representation problem emerged from a very lively discussion at the panel.

The first view holds that a common paradigm is necessary for making progress in this area, and that objective measures of success are needed independently of specific tasks. The need for such a unified approach comes from the observation that the applicability and performance of existing representations have not been established in a systematic manner. As a result, progress is hard to quantify and promising new directions are hard to identify. The second view, supported by the majority of the panel, holds that representations are defined only relative to specific

1063-6919/94 $3.00 © 1994 IEEE

tasks and applications, and that success can be defined only in the context of specific tasks. Examples that were given included, among others, the use of generalized cylinders for the recognition of tubular industrial parts and the use of polyhedral representations for bin-picking of industrial parts.

Independently of the overall approach, three main challenges were identified by the panel. First, representations that support indexing in a large database have be to developed. Second, current feature extraction and segmentation techniques do not have the level of robustness needed to support the most advanced representation and recognition schemes. Third, new techniques are needed for learning object models from actual data. Several solutions were suggested in order to learn object models from multiple observations. Finally, the need to evaluate and quantify the performance of specific object representation schemes was emphasized by all participants in order to move our field from an art to a more solid engineering discipline. Suggestions included the use of a benchmark library of objects and scenes, and the use of a small set of target tasks against which the performance of specific solutions can be measured.

Below are statements by the panelists summarizing their positions.

Martial Hebert

Object recognition research has led to a number of object representation schemes, among them polyhedra, parametric patches, generalized cylinders, superquadrics, algebraic surfaces, and discrete surfaces. A cursory look at this list shows two things. First, each of these representations has its own associated recognition algorithm and its own set of results, independent of the other representations. Second, although this partial list is approximately in chronological order, there is no obvious evolution from one solution to the next. Rather, each representation uses a completely different approach which is successful for a carefully selected set of objects.

These two observations illustrate the need for continuous progress leading to general results applicable to a wide range of problems in computer vision.

Some will argue that this is not an issue because the set of representations can be viewed as a toolkit from which specific tools can be used in practical tasks. This argument proceeds further by suggesting that representational tools may be developed in an independent manner so long as they cover the needs of the applications. This would be true enough if we had provided the "users" of computer vision with a clear picture of the performance and relevance of the tools. For example, the performance of a particular representation needs to be defined in terms of, among other things, the class of objects to which it applies, to which representations it can be converted, and the tools needed to build the representations. However, such criteria are not systematically used. Rather each representational scheme is developed independently using its own domain. An added benefit of a more systematic approach is that it may lead to more evolutionary research in which each new representational scheme builds upon the results from the previous ones. This would enable the community to focus its attention on the next level of capability needed in representations.

On the positive side, new representations have permitted major advances in the capabilities of recognition algorithms, for example, the recognition of more general 3D curved objects, or the recognition of 3D objects from 2D images. Although these problems are far from being solved, these examples show that the object representation capabilities are critical in making progress in object recognition and other areas.

Katsushi Ikeuchi

Is the current CAD technology useful and scalable?

The current technology is quite effective in task domains such as estimating the pose of man-made objects in in-door scenes using range data. Industrial applications include bin picking and material handling in assembly lines. However the current technology emphasizes geometry. By adding photometric properties to CAD models and using sensor simulators, we can extend the task domain, for example, to ATR (automatic target recognition) and the identification of man-made objects in out-door scenes using SAR or FLIR images. CAD technology can also be applied to sensor planning for object inspection, and by adding parameterization capabilities, to hierarchical object recognition.

Learning capability.

We propose the Learning CAD models (LCAD) paradigm that would observe a real object and modify its current representation. Such a capability is necessary because no matter how accurate a CAD model might be, there will always be a discrepancy between a real object and its model; however precise a sensor model might be, a real image will always be different from a simulated one. Thus, it is definitely necessary to develop a capability to adjust and to fill in this gap through real examples. Learning a CAD model from scratch may be very difficult but, fortunately LCAD can start with approximate solutions given by a programmer. Applications that will greatly benefit from the LCAD paradigm include learning object models for object recognition and scene models for virtual reality.

A new framework for object representation.

Vision systems should be designed in the task-oriented framework: vision theories should be developed with keeping in mind under what task condition the theory work. I do not believe that any single representation could serve for all vision tasks. Even if such a representation existed, a system using it would likely be extremely inefficient.

Thus, I believe the research directions on representations should be:

- To expand the collection of representations: up to now, we have reasonable representations for man-made ob-

ject recognition. However, other tasks may require other representations. For example, we have recently developed the Spherical Attribute Image representation to handle natural objects in NASA's Mars sample return mission.

• To investigate the relationship between the task specification and the functionality of representations, and to establish a systematic theory to select the optimal representation that satisfies such functionality required.

Avinash C. Kak

Tremendous progress has been made in model-based computer vision in the past thirty years. Automatic scene interpretation systems built recently have demonstrated capabilities beyond anything Roberts could have dreamed of back in the sixties. For example, a very successful idea that emerged in the eighties was the prediction/verification paradigm, that uses a few hypothesized matches between model and scene features to determine the object pose, then uses the estimated pose to efficiently verify the correctness of the matches. The main advantage of this approach is that it beats the combinatorial explosion associated with feature matching (matching all image features to all object features requires exponential time in the number of features).

Part of the reason for this success has been the development of efficient mathematical and computational techniques for hypothesis generation. For example, the early work of Faugeras and Hebert and of Horn demonstrated that the pose of a polyhedral 3D object could be computed from range data by using only linear algebra. This made it possible to generate new hypotheses very efficiently. The other half of the story has been the development of efficient control structures and data structures for verifying the correctness of the matches. While the time complexity of the recognition algorithms developed in the early eighties was difficult to ascertain, progress has continued in this area, and the 1989 3D-POLY system by Chen and Kak was shown to have low-order polynomial time complexity in the number of features.

This is a sign of health for the model-based vision field, and more work and energy should be, and is, devoted to the development of efficient control structures for object recognition (see the work of Grimson and Huttenlocher for example). This should also apply to other problems such as the recognition of free-form surfaces.

Dimitri Metaxas

While the tasks of shape reconstruction and object manipulation require shape models that can represent shape very accurately, shape recognition often requires models that can offer shape abstraction. Therefore, the accuracy and form of shape representation depends on the task. Lately, we (Metaxas and Terzopoulos) have tried to bridge the gap between reconstruction and recognition by building models that offer both shape accuracy and hierarchical shape abstraction. Such models though, can only represent a certain class of existing objects. Clearly, more research needs to be done in the direction of developing models that can satisfy the needs of both fields (CAD models are obviously not sufficient) and at the same time offer a much broader shape coverage than currently existing models.

One way of generalizing the currently existing models is through the rigorous evaluation of their limitations both at the representational level and at the shape estimation level. Possible improvements in current mostly static shape estimation approaches, is the use of dynamic physics-based shape estimation techniques (Terzopoulos, Pentland, Metaxas) and the use of active strategies (Bajcsy) at the level of data processing and at the level of camera control. Furthermore, we need to develop the necessary control strategies and constraints which will allow the simultaneous object segmentation and shape estimation from raw data. These two problems which have been traditionally studied in isolation are definitely interdependent. One possible direction is through the integration of qualitative and quantitative shape estimation techniques (Dickinson and Metaxas), while another is the use of motion (Kakadiaris, Metaxas and Bajcsy) in order to segment and estimate the shape of articulated objects.

Finally, we believe that object recognition should not be based exclusively on geometric shape, an assumption made by most existing recognition systems. Research is needed towards the integration of cues like motion, color, texture and shape (qualitative and quantitative) in order to achieve robust object recognition. This will lead to the development of object representations that are not only based on geometry and are clearly necessary for recognition.

Stan Sclaroff and Alex Pentland

The first point is that the name "CAD-Based Vision Workshop" may be misleading, as much of the work presented here includes not only 3D models but also 2D aspects and appearances. Also, the fact that much of CAD-based vision is tightly focused on bin-picking means that we may be missing important, potentially solvable industrial problems. Consequently, we believe that a better focus might be "practical, useful vision".

The second point is that although the use of a 3D CAD model seems very attractive, there are indications that 2D techniques might be a better way to proceed. For instance, consider the 2D eigenspace techniques that have been developed by our group at MIT, Nayar at Columbia, and Thorpe at CMU. We have been able to obtain reliable recognition on a database of over 3,000 natural objects (faces), and pose/recognition with over 100 faces. Nayar has been able to obtain reliable pose/recognition with 20 vehicle models. Thorpe has used this technique to log miles of continuous, autonomous driving time in the NAVLAB. Moreover, each of these systems runs in real time on a standard workstation. Perhaps it is time to revisit our pattern-recognition roots?

Another promising coupling between 3D models and 2D views can be found in aspect graphs. Although it seems

that exact, global aspect graphs are computationally impractical, it appears that "local" and "qualitative" aspect graphs (graphs of subparts) are alive and well. Take for instance, the work done by Dickinson and Metaxas. Here, a qualitative, parts-based aspect graph is used to locate and categorize objects, followed by a physically-based superquadric modeling to recognize the object. Such an approach seems particularly promising since it can handle poorly segmented scenes and articulated objects, and it can deal with objects it has not encountered before.

Lastly, it seems to us that model construction is increasingly important, and that rapid progress is being made in this area. There has even been mention of recovering CAD models from image and range data for reverse engineering. It seems unlikely that we will be able to automatically recover engineering-precision solid models; however, it does seem plausible that we could recover models which are good enough that high precision can be obtained with relatively little human intervention. And, of course, it is quite possible that the models will be good enough for machine vision!

Linda Shapiro

There are two major types of existing representations for model-based vision: the relational matching representation and the point matching representation. Relational matching requires primitive features, their attributes, and their relationships. Point matching requires a set of identifiable points. Relational matching is a well-defined, mathematical formalism that uses as much information as possible to hypothesize correspondences intelligently. Point matching algorithms (such as alignment and geometric hashing) do not need to extract complex features and can be performed rapidly with proper (large) preprocessing. Point matching fails to use relationships that could suggest correspondences to try, and it depends heavily on the method for point detection.

CAD-based vision was designed for and is intended for industrial use. In industrial vision, the task is well-defined. Common tasks are pose estimation for robot guidance and inspection. The environment for the task is controllable and can be duplicated or simulated for training. Bin-picking is rare, because parts from bins generally come down chutes, one at a time. Multi-object scenes do not come about randomly; they are part of some process that has a known initial state and is controlled at every step. In this environment, CAD models, along with lighting, sensor, and feature detection models can be used to predict the detectability of features. The important thing is to discover what kinds of features are most reliable for a given object or set of objects and a given task.

George C. Stockman

There is no single representation for all problems.

One thing learned from AI is that problem representation is a very important step in problem solution; and, that it is very unlikely that a single representation will serve all purposes. From CAD, we know that no single modeling system serves all design needs; a system for the design of sheet metal parts will be fundamentally different from a system to design castings. In the general case, computer vision also shares the "frame problem" with AI. On what features should our representations be built and indexed? How can relevant frames be accessed during recognition while irrelevant frames are ignored? There is also an intractable scale problem in the general case. Do we need to represent the grain of wood in the wall or do we just need to represent the plane of the wall?

When a general 3D problem is well-defined, then an abstract data type (ADT) can be created for it.

A representation useful for 3D computer vision should be an ADT – a mathematical model together with the operations defined for operating on that model. Parametric models are well suited for indexing and recognition since a few parameters encode shape and location. but a few parameters will not be sufficient in general for inspection or for tasks requiring a fairly fine interpretation of local features. On the other hand, mesh models allow local control and support dynamic modeling via FEM, and support other applications such as rendering, but do not by themselves represent features for recognition or indexing. A module which creates general mesh models for higher-level visual processes would have to provide sufficient detail for an unknown set of these processes. Such models would be poor for recognition purposes, as are the polygonal models of curved objects used in computer graphics. This is another aspect of the scale problem mentioned above.

Where specific problems have been addressed, machine vision has produced success.

Peanuts can be inspected as they flow through a tube so fast that they look like a stream of fluid to the human eye. With small changes, the same machines inspect grains of rice! There is no hope of adapting this method to the inspection of foundry castings, however. At MSU we have successfully used specific tube models to recognize normal blood vessels in MRI images. Model-matching tends to fail in cases of stenosis or aneurism; while this failure gives good information to a specific recognition task, it is because the representation itself breaks down. Machine vision has been successfully used in electronic inspection and in robotic applications in the automotive industry. We should be optimistic that many vision problems can be covered by a reasonably small set of paradigms and representations. It is the job of computer vision researchers to sort out the classes and create new ones.

3 The CVPR Panel

At the time of this writing, the CVPR panel has not taken place yet. It is co-organized by J. Ponce and M. Hebert, and the panelists are T.O. Binford, R. Bajcsy, and D.A. Forsyth.

A number of concerns, most of them common to the three panelists, emerge from the statements enclosed be-

low: the need for multi-level, part-whole representations, based on appropriate geometric primitives; the importance of learning models from image data; the difficulty of segmentation; the importance of invariant and quasi-invariant representations.

Thomas O. Binford

A hierarchy of shape representation is essential from the lowest levels of computational vision through the highest levels. My concerns in representation are given in a survey of model-based image analysis systems in 1982, in a paper on Bayesian Inference in 1987, in a review on shape representation in 1991.

At the lowest level, segmentation of the image intensity surface (2D data) or range surface (3D data) are identical. Segmentation of 3D images is similar (MRI or CT). The essential representation is a covering of the image with diameter-limited neighborhoods. On each neighborhood, a semi-differential geometry representation of piecewise 2nd-order patches with locally straight boundary curves is relevant. This is a local representation.

The second-level is a surface spline, a web of C^2 surface patches bounded by extended curves that terminate at vertices. Surfaces are not necessarily C^2 or C^1 across curve boundaries. The basis is not necessarily polynomial. This is an SCP graph (surface-curve-point). The SCP is a quasi-local representation including incidence and connectivity. Although there is local surface information (e.g. principal curvatures) this is a weak form of shape representation.

At the next level, 3D data are represented by a VSCP graph, i.e. CSG. There is a part-whole volume representation with volume primitives. A surface representation is not equivalent to a volume representation, i.e. the boundary is not equivalent to the interior. This is obviously true for volumes that are not uniform. Most representations used in vision are surface representations. They are deficient in that volume representations have relations between surfaces that can only be derived from surface representations with computational effort. E.g. the diameters of volumes are opposite relations between surfaces. Elongation is a relation between diameters.

It is at the volume part level that we can talk about representing the shape of objects, as opposed to the local shape of a surface. Provocative statement: we cannot talk about object shape with surface representations. Generalized cylinder volume primitives (GCs) provide a useful basis for shape representation of objects. We regard GCs as a volume spline basis. Completeness requires only approximation at a suitable continuity level. GCs also enable segmentation into parts. Parts are volume concepts; i.e. continuity of surfaces is not meaningful for defining parts; continuity of volumes is. Volume representation is essential for segmentation into volume parts. Incidentally, the same arguments hold for the relation between surface representations and their curve boundary representations. We represent surfaces by ribbons, i.e. a part-whole web of GCs in a lower dimension. Voronoi diagrams are a class of volume representations based on a distance measure, similar to GCs. The structural part-whole decomposition is the most important issue. The set of basis functions for part shape is relevant.

GCs are useful for generating invariants and quasi-invariants among surfaces for recognition.

The discussion here is related to two issues. The first issue is building shape descriptions from data. These representations are volume splines, i.e. well suited to building up local descriptions from data. The second issue is comprehensive representation in a complex world, e.g. outdoors with vegetation, terrain and man-made structures, or indoors with offices, clutter, etc.

In brief: Provocative Statement 2: any representation for shape of objects that is useful with wide variation should be a volume representation by parts; spline-like.

Ruzena Bajcsy

Shape representation: the chicken and egg problem.

How Form or Shapes are represented/modeled has a long history. The problems has been investigated by many scientific disciplines: mathematics, engineering-design, manufacturing, arts, psychology, zoology, philosophy and others.

In this presentation, we shall take a rather narrow point of view, that is: issues related to form representation of man-made solid objects. One underlying assumption is that such man-made objects have some purpose or use and therefore the use, and the user has something to do with their form.

Another assumption we make is that there is some finite (hopefully not too large) set of geometric (in this case mathematically well-defined) primitives from which one can compose the rest of the man-made shapes. This assumption stems from our scientific training, such as physics and mathematics which favors reductionism as opposed to holism.

Then our problem is:

- to find the set of geometric primitives that are mutually exclusive and cover as large a set of man-made objects as possible;
- to find computational procedures that will extract/estimate these primitives in a consistent fashion.

In our past work we have selected the parametric representation of superquadrics and algebraic surfaces for such primitives, but of course this is not the only possibility. The reason why I put into my title the chicken and egg problem is that there is a tradeoff between selecting simple primitives and then having more complex part-whole descriptions and having more complex primitive descriptions and simpler and fewer part/whole descriptions. It is this issue of *complexity* of selection of the representation which makes it unclear which way to go and remains one of the more difficult open problems.

Another related issue is segmentation or estimation of parts from the data. The segmentation in general is *not*

unique! Hence one needs more constraints. One source of such constraints can come from the functionality and/or usage of the object. This area again is still not researched properly.

In my presentation, I will demonstrate some of the above problems in real examples from our recent work.

David A. Forsyth

Effective object representation is one of the grand challenges of computer vision; a challenge to date largely unsolved, because of its amorphous nature - effective representations have, to date, depended intimately on applications - and the ease with which workers are distracted by the intrinsics of the problem - it is all too easy to define a representation that looks good, but cannot be used for anything. Because object recognition involves extracting an object representation from an image and then using that representation to determine an object's identity, it presents challenges both to the *distinctiveness* and the *recoverability* of a representation.

Organizing a modelbase to cope with such variables as viewing position, illumination properties and camera parameters results in an explosive growth of the modelbase; insensitivity to these variables is, therefore, a most desirable property for an object representation. For plane objects, constructing effective, recoverable representations that are invariant to all major variables is very largely a solved problem. For 3D objects, the situation is far more interesting; much is known about polyhedra, but curved surfaces remain rather mysterious.

Invariant representations for simple surfaces.

Although outline curves bear a special, rigid relationship to surface geometry, it is hard to infer a distinctive surface description from a single image curve. However, when a surface is known to belong to a clearly defined, globally constrained, surface class, it is often possible to recover an invariant representation. For many such cases the result is plausible because the surface is, in some sense, a curve in disguise - examples such as rotationally symmetric or canal surfaces spring to mind. Algebraic surfaces are remarkable, because a representation can, in principle, be recovered that determines the surface up to a projective transformation of space. There are, however, tremendous computational difficulties, so that rigid surface classes present little hope for a robust, long term solution to the problem of recovering representations for curved surfaces.

More promising research directions include:

- segmentation techniques to decompose image outlines so that representations for combinations of simple surfaces can be recovered usefully - this well established topic should be revived by the recent mass of information covering distinctive geometric properties of surface outlines for various types of surface;
- combining surface markings and, possibly, colour, with outlines to obtain more distinctive representations.

Near free-form surfaces and quasi-invariance.

General or free-form surfaces have too great a potential for geometric peculiarity to be of interest in vision. The issue then becomes: what class of surfaces is both sufficiently inclusive to cover a wide range of objects, and at the same time sufficiently constrained that it is possible to use an outline to make statements about the surface? A number of answers, none entirely satisfactory at the time of writing, have been proposed; examples include various forms of generalized cylinder and cone, superquadrics, and algebraic surfaces. As the class of surface becomes less strongly constrained, it becomes more difficult to recover geometric invariants from the outline.

There is a continuum from invariance to uselessness; if invariants are not available, quasi-invariants[1], image properties that are stable over a range of viewing geometries, must do. These properties are used to assemble a surface description that is likely to be reasonably accurate. Quasi-invariants are poorly understood at present - for example, it is not known how to assemble all (or many) of the quasi-invariants pertaining to a particular geometry.

Quasi-invariants present fruitful research avenues. Such effects as limited field of view, limited image resolution, and frame to frame consistency are essential in determining the usefulness of quasi-invariants - for example, a symmetry spine constructed from an image of an object of revolution, although clearly not the image of the axis of the object, lies extremely close to the axis for all practical camera views. In fact, it is possible to show that as the camera moves, the contour generator generally moves around the object in a way that keeps the error small. There are very few examples known to have such properties and it is not yet known how to construct more.

Because their application is broader, I believe that quasi-invariants of some form will come to be central in representations for near free-form surfaces. Their present application is limited because little is known about how an armory of such cues might be constructed for a large class of surfaces, or about how to use them effectively. Finally, the problem is complicated by continuing uncertainty about the data necessary to effectively identify a curved, 3D object.

4 Conclusion

As stated in the introduction, this paper and the associated panels are only a very first step toward a better assessment of current object representations and the design of future ones. Our hope is that they will steer renewed interest for shape representation in the computer vision community. The next step calls for a synthesis of the opinions expressed here and of those of a broader sample of our community. Further discussion is required.

[1] By this term I mean properties whose value is within a given, small, range for all possible views less a set of small measure; there is a local definition that is current, but not equivalent.