

A Short Survey: What Is a “Generic” Object Model for Computer Vision? ¹

Melanie Sutton, Louise Stark and Kevin Bowyer

Department of Computer Science and Engineering
University of South Florida
Tampa, Florida 33620 U.S.A.

sutton, stark or kwb @csee.usf.edu

ABSTRACT

An area of computer vision research which has recently become more active is the development of *generic* object models. We consider a model to be “generic” if it is meant to represent an object category rather than a single object instance. The motivation for work in this area is both to make recognition algorithms more efficient, but more importantly, to make them robust in real world applications. In this review, we attempt to bring many of the diverse efforts in the area of generic object representations into a common conceptual framework. This framework shows that there is a natural progression of increasingly sophisticated techniques available for creating generic object models.

1 Introduction

The predominate approaches to object representation all assume that the vision system will begin with a geometric model of each object that it is supposed to be able to recognize. The surveys by Besl and Jain [3] and by Chin and Dyer [9] give numerous examples of such research efforts in “model-based” and “CAD-based” vision. Use of a geometric model directly in the recognition process would constitute an *object-centered* approach. For example, in what is sometimes called *recognition by alignment*, the projection parameters of a 3-D geometric model onto the 2-D image are adjusted until the features in the projected image are properly aligned with features in the real image [19, 26]. In a *view-centered* approach, information about the collections of features that are visible from different viewpoints is derived from the geometric model (often as some form of aspect graph [6]) and this derived information is used in the recognition process. The information about the different *general views* of the object may then be used to compute an *interpretation tree* which becomes the control structure for the recognition process [15, 20].

¹This work is supported by Air Force Office of Scientific Research grant AFOSR-89-0036, and by National Science Foundation grants IRI-91-20895 and CDA-91-00898 (Research Experiences for Undergraduates).

While traditional model-based vision has been and still is a useful research paradigm, there is an increasing awareness that something more and different will be needed in order to achieve visual perception for “autonomous,” “real world” systems. The system must be able to recognize and deal with truly novel objects; that is, objects for which it has *no explicit prior model*, at least in the traditional sense of the word “model.” Recognition, in this sense, is taken to mean object categorization.

The remainder of the paper is organized as follows. Section two discusses parameterizations of 2-D feature configurations, as might result from generalizing some view-centered representation. Section three discusses different types of object-centered parameterizations. Function-based models in which the definition of the object category is some minimal set of constraints that serve to ensure that the object can be used for the appropriate function are discussed in Section four. It is impossible to individually mention every research effort that might be viewed as using one of these types of generic model. Instead, we have tried to sample the most recent work, the best known early work and unique or unusual examples.

2 Image Feature Configurations as Generic Models

The spirit of a view-centered representation is that it captures the fundamentally different configurations of features which may appear in an image. For some application areas, it may be possible to generalize such a representation by specifying a parameterized or qualitative arrangement of image features which adequately describes the appearance of object instances in a given category. For example, in the interpretation of aerial images one can assume an “overhead” view of the objects of interest.

Recent work by Fua and Hanson [13] provides an example of this type of approach. They define models of object categories in terms of parameterized and qualitative configurations of image features. Buildings are modeled as rectilinear networks in the edge-detected image. Similarly, the model of roads is based on curvi-linear parallel networks, and the model of vegetation is based on irregular edge outlines. Several researchers have developed more sophisticated generic models for buildings by exploiting the context provided by the sun and shadows [17, 22, 25]. In addition to rectilinear edge structure, it is possible to apply the constraint that buildings should cast shadows of a consistent shape in a consistent direction.

Huertas et al. [18] use similar “generic knowledge” about objects and context in the analysis of aerial images of airport scenes. Edge grouping operations of continuity, collinearity, parallelism, and symmetry are applied initially for the detection of runways. In a “hypothesize and verify” approach, this information is further used to aid in recognition of taxiways, and probable locations of other building structures (likely to be found at the edges of runways) and mobile objects (planes, which would be on the runway).

Another recent example of this sort of representation being used in aerial image analysis is the “generic bridge finder” described by Vergnet et al. [33]. The definition of the concept bridge is that it is a roadway supported by pilings that passes over a river. The definition is implemented as a “deformable 2-D model” that first looks for a homogeneous gray-level region in the image that could represent a river. Using the

candidate river region as context, a search for a linear set of edge points that could represent a candidate roadway crossing the river and for a set of periodic linear edge segments which would be the pilings of the bridge is performed.

Parameterized image feature configurations have also been applied as generic models in the analysis of outdoor scenes at ground level. Several examples occur in the well-known VISIONS systems developed at the University of Massachusetts [16]. This system labels objects in an image to construct a symbolic representation of the three-dimensional world represented. Depending on the current object of analysis, the labeling may be based on similarities in color, size, shape, texture, location within the image and the possible labels of neighboring regions.

Additional research efforts in this area include the work generated by the *autonomous land vehicle* (ALV) project which used some qualitative 2-D feature model for the generic model of a road. As an example of an early effort in this area, Nagao and co-workers used similar techniques in an aerial photo-interpretation system well over a decade ago [29]. Binford's survey article of nearly a decade ago includes some other early examples [5].

The generic models discussed in this section are all similar in that they specify qualitative or parameterized configurations of essentially (2-D) image features. The appearance of a set of image features which satisfies the model is taken to indicate the presence of an instance of the object category. These models are also similar in that they depend heavily on contextual constraints and sequential identification of candidate regions in the image.

3 Representing 3-D Shape

Whereas the models described in the previous section were based on generalizations of the object appearance in a 2-D image, the models to be described in this section are based on generalizations of 3-D object shape. We distinguish between three level of increasing sophistication in these models: (1) *parameterized geometric models*, (2) *structural models* and (3) *parameterized structural models*.

3.1 Parameterized Geometric Models

A parameterized geometric model is created by replacing some of the constants in a geometric model with parameters that may be constrained to lie within defined ranges. (See Figure 1). The use of parameterized geometric models has been explored by a number of researchers.

Vayda and Kak [32] have used parameterized geometric models in their INGEN (INference engine for GENeric object recognition) system. This system considers three basic families of shapes: parallelepipeds, cylinders and "irregulars." Coplanarity and adjacency of surface patches in range data are used to capture widely varying shape instances.

Mulgaonkar et al. [28] used similar parameterized shape families in analysis of piles of 3-D objects in bin-picking systems. Knowledge about symmetry, stability, viewpoint

AUTOMOBILE FRAME							
IS_A	artificial_object						
SUBCLASSES	car		truck_1		truck2		
CONSISTS_OF	engine_area		cabin		hauling_area		
SIZE	LENGTH	min_L1	max_L1	min_L2	max_L2	min_L3	max_L3
	WIDTH	min_width		max_width			
	HEIGHT	min_H1	max_H1	min_H2	max_H2	min_H3	max_H3

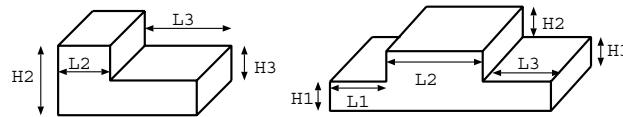


Figure 1: Example of a Parameterized Geometric Model

independence, and object solidity is used to generate object hypotheses and verify the consistency of hypothesized configurations.

More sophisticated parameterization of the geometric model can capture “object families” which allow substantial shape variations between object instances. Grimson [14] defines (2-D outlines of) object families which incorporate parameters for rotation, translation, scaling and stretching. Examples are discussed for scissors, which may be seen with different amounts of rotation between the scissor blades, and for a family of hammers that have the same head but whose handles may be stretched different amounts.

Kadono et al. [23] describe a fairly sophisticated system oriented toward interpretation of man-made objects which appear in outdoor scenes. Parameterized geometric models are used to define the class of “automobile” objects and the class of “house” objects, each with several subclasses.

3.2 Structural Models

A structural model is distinguished from a simple geometric model by specifying an explicit construction of the object as a set of parts (a “part-whole” model) The model may be hierarchical, with several stages in which a “part” is specified by another structural model before the definition is eventually reduced to primitive geometric descriptions.

A representative example of the use of structural models has been described by Mulgaonkar et al. [27]. In this work there are just three types of qualitatively-defined part types: sticks (linear features), plates (flat parts), and blobs (large volumes) (see Figure 2 [27]). Recognition is based upon matching the list of parts, together with the connections between them. The generality of these parts allows the system to recognize object instances which vary greatly in shape.

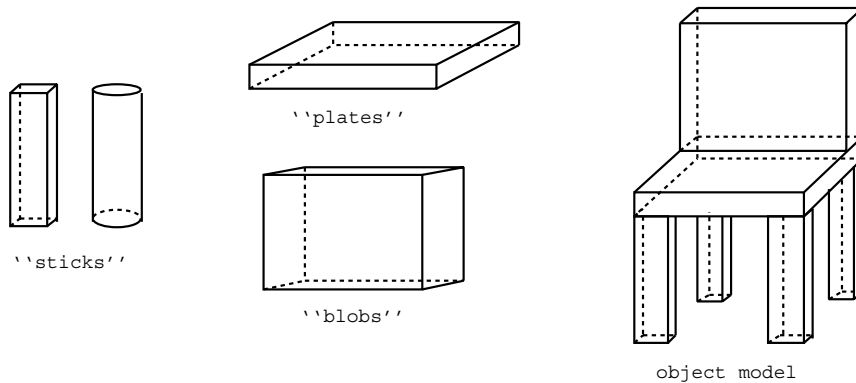


Figure 2: Examples of Sticks, Plates and Blobs

Qualitative structural models are fundamental to Biederman's *recognition by components* (RBC) theory [4]. In RBC, an object category is defined as a particular set of qualitatively-described primitive 3-D shapes, called "geons," and the qualitative relationships between them ("on top of," "larger than," ...). Biederman suggests that immediate recognition is achieved, at the object category level, through a form of indexing based on the structural composition of the object. Biederman's RBC theory has inspired several computer vision research efforts to implement a recognition system using some of the RBC concepts.

Bergevin's PARVO system uses RBC-like assumptions of coarse qualitative models to represent classes of objects [2]. PARVO analyzes a line drawing to determine individual faces of the object, from which it infers the geon types and their connections. It then performs recognition by matching to a graph model of the object in which a node represents a geon type and an arc represents a connection between two geons. Dickinson et al. describe a different RBC-inspired system [11]. An interesting element of their approach is the use of a precomputed *aspect hierarchy* to index from groups of edge segments to geon faces, then from geon faces to geon aspects and finally from geon aspects to geons.

Probably the earliest work using structural models related to computer vision is Winston's classic "arch-learning" program [35]. This program was able to learn structural descriptions of object families, such as *arch*. The system used ideal 2-D line drawings as input. The underlying representation of an object category is a semantic network which indicates particular qualitative relationships between qualitatively specified parts.

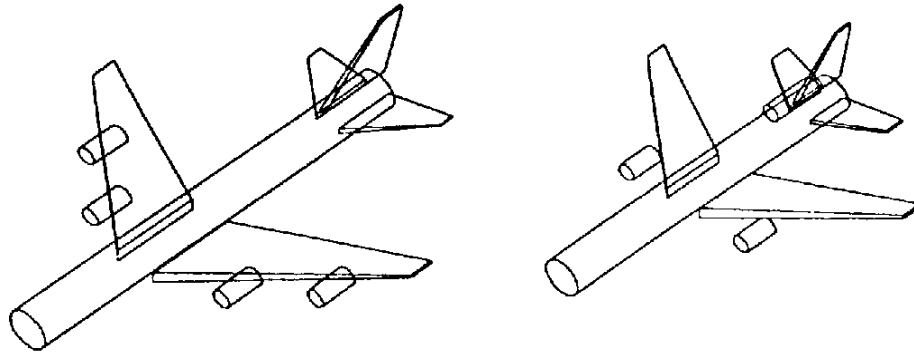


Figure 3: Instances of Aircraft in the Same Generic Family Drawn by ACRONYM

3.3 Parameterized Structural Models

As noted earlier, one motivation for using a structural model is to make it convenient to specify sophisticated parameterized geometry. However, it also becomes possible to parameterize the structure itself. In a *parameterized structural model*, variations in the essential structure of the model may be modeled by parameterizing the number and location of some of the component parts of the object.

One of the earliest uses of parameterized structural models occurs in Brooks' ACRONYM system [8], where families of aircraft are represented by a composition of a variable number of parts (see Figure 3).

Kriegman has used what can be viewed as a parameterized structural model in a somewhat different application context [24]. In this work, the “prototypical model” of a building is uninstantiated. Loose dimensional constraints, combined with sensory data, provide the parameterization, allowing the model to be even more robust to gross changes in object geometry and topology. The acquired model could then be used in path planning for robot navigation.

4 Function-Based Models

A *function-based model* differs from the types of models described so far in that it does not specify any explicit geometric or structural plan for instances of the object category. Instead, an object category is defined in terms of knowledge about what is necessary in order for an object to *function as* an instance of the object category. Primitive chunks of knowledge about shape, physics and causation are used to build minimally sufficient definitions of required function. For this reason, function-based models seem to provide better support for “purposive” (in the sense of Aloimonos [1]) and “task-oriented” (in the sense of Ikeuchi [21]) computer vision.

Stark and Bowyer [31] have described a function-based object recognition system **GRUFF** (**G**eneric **R**epresentation Using **F**orm and **F**unction) that uses a set of five “knowledge primitives” to construct a generic object definition. The knowledge primitives deal with the concepts of stability of a 3-D shape in a given orientation, clearance

of a specified volume of 3-D space, dimensions of a specified element, relative orientation of two surfaces, and proximity of two elements. These primitives are used to define a “category tree” which specifies the *functional plan* for the object category.

DiManzo [12] proposes a recognition system design that utilizes knowledge about function within an expert system framework. The design is limited in knowledge about different subcategories. Primitives are defined in the form of individual expert systems that evaluate the 3-D information using constraints defined by the user. A prototype system is being implemented which receives a 3-D description of a scene generated by an octree solid-modeler.

Brady and colleagues discussed the relation between geometric structure and functional significance in their design of the “Mechanic’s Mate” system [7, 10]. In part of this work, semantic network descriptions are computed from 2-D shapes, and a generalized structural description is learned from a sequence of positive examples.

Stansfield discussed the modeling of generic object categories, in the context of combined visual/tactile sensing [30]. The concept of a *spatial polyhedron* is introduced to represent a class of objects in terms of the spatial configuration of required parts and function-based features that must be able to be sensed. This provides a means of classifying objects with wide structural variations of a particular feature, provided that the feature is sensed in the spatial location indicated by the model [30]).

Some well-known early work in the area of function-based definitions of object categories was done by Winston, Binford and co-workers [34]. They point out that there can be an infinite number of different shape descriptions for objects in a category as simple as *cup*, but that a single functional description can be used to represent all cups in a concise manner.

5 Discussion

We have distinguished between examples of several quite different types of “generic” object models for computer vision. Parameterized image feature configurations are most useful in applications which involve the identification of man-made objects in aerial images or outdoor images within some known context. This is because the possible poses of the man-made objects can be assumed to be reasonably constrained (seen “from above,” for example) and they are likely to have features (strong rectilinear edge segments, for example) which easily distinguish them from the background. Parameterized geometric models are most useful when the application allows the assumption that all instances of an object category will fall into an easily parameterized range of 3-D object shapes. As the range of different 3-D shapes to be captured becomes more varied, structural and parameterized structural models are needed. In the most general case, a function-based model is required in order to deal with truly novel shapes which may fulfill the function required for an instance of the object category, but do so in an unanticipated manner.

The development of representations which will support generic object recognition is clearly of fundamental importance if the goal of autonomous real-world systems is to be achieved. Just as clearly, this area of research is still in its infancy. Object representations which are eventually developed to support this goal are likely to incorporate

elements of all the types of models described here, as well as information for additional sensing modalities (such as tactile sensing) and for making inferences about the material composition from properties of surface appearance.

References

- [1] Aloimonos, J. 1990. Purposive and qualitative action vision, *DARPA Image Understanding Workshop*, pp. 816-828.
- [2] Bergevin, R. and M.D. Levine. 1989. Generic object recognition: building coarse 3D descriptions from line drawings, *IEEE Workshop on Interpretation of 3-D Scenes*, pp. 68-74.
- [3] Besl, P.J., and Jain, R.C. 1985. Three-dimensional object recognition, *Computing Surveys*, **17**, **1**, pp. 75-145.
- [4] Biederman, I. 1987. Recognition-by-components: a theory of human image understanding, *Psychological Review* *94*, **2**, pp. 115-147.
- [5] Binford, T.O. 1982. Survey of model-based image analysis systems, *Int. J. of Robotics Research* **1**, pp. 18-64.
- [6] Bowyer, K.W. and Dyer, C.R. 1991. Aspect graphs: an introduction and survey of recent results, *Int. J. of Imaging Systems and Technologies*, to appear.
- [7] Brady, M., Agre, P.E., Braunegg, D.J., and Connell, J.H. 1985. The mechanics mate, in **Advances in Artificial Intelligence**, T. O'Shea (ed.), Elsevier Science Publishers, B.V., pp. 79-94.
- [8] Brooks, R.A. 1984. **Model-based computer vision**, UMI Research Press, Ann Arbor Michigan. See also Symbolic reasoning among 3-D models and 2-D images, *Artificial Intelligence*, **17**, 285-348.
- [9] Chin, R.T. and Dyer, C.R. 1986. Model-based recognition in robot vision, *ACM Computing Surveys*, **18**, **1**, pp. 67-108.
- [10] Connell, J.H. and Brady, M. 1987. Generating and generalizing models of visual objects, *Artificial Intelligence*, **31**, pp. 159-183.
- [11] Dickinson, S.J., A.P. Pentland, A. Rosenfeld. 1990. Qualitative 3-D shape reconstruction using distributed aspect graph matching, *Third Int. Conf. on Computer Vision*, pp. 257-262.
- [12] Di Manzo, M., E. Trucco, F. Giunchiglia, F. Ricci. 1989 FUR: Understanding FUncional Reasoning, *Int. J. of Intelligent Systems*, **4**, pp. 431-457.
- [13] Fua, P. and A. Hanson. 1987 Using generic geometric models for intelligent shape extraction, *AAAI*, pp. 706-711.
- [14] Grimson, W.E.L. 1988. On the recognition of parameterized 2D objects, *Int. J. of Computer Vision*, **3**, pp. 353-372.

- [15] Hansen, C., and T. Henderson. 1988. Towards the automatic generation of recognition strategies, *Second Int. Conf. on Computer Vision*, pp. 275-279.
- [16] Hanson, A. and E. Riseman 1988. The VISIONS image-understanding system, in **Advances in Computer Vision I**, C. Brown (ed.), Lawrence Erlbaum Associates, Publishers, New Jersey, pp. 1-114.
- [17] Huertas, A. and R. Nevatia. 1983. Detection of buildings in aerial images using shape and shadows, *Eighth Int. Joint Conf. on AI*, pp. 1099-1951.
- [18] Huertas, A. and R. Nevatia. 1988. Detecting buildings in aerial images, *Computer Vision, Graphics, and Image Processing*, **41**, pp. 131-152.
- [19] Huttenlocher, D.P. and S. Ullman. 1987. Object recognition using alignment, *DARPA Image Understanding Workshop*, pp. 370-380.
- [20] Ikeuchi, K. 1987. Generating an interpretation tree from a CAD model for 3D-object recognition in bin-picking tasks, *Int. J. Computer Vision*, pp. 145-165.
- [21] Ikeuchi, K. and Hebert, M. 1990 Task-oriented Vision, *DARPA Image Understanding Workshop*, pp. 497-507.
- [22] Irvin, R.B. and D.M. McKeown, Jr. 1989. Methods for exploiting the relationship between buildings and their shadows in aerial imagery, *IEEE Trans. on Systems, Man and Cybernetics*, **19**, pp. 1564-1575.
- [23] Kadono, K., M. Asada and Y. Shirai. 1991. Context-constrained matching of hierarchical CAD-Based models for outdoor scene interpretation, *IEEE Workshop on Directions in Automated CAD-Based Vision*, (June, 1991).
- [24] Kriegman, D.J., T.O. Binford, and T. Sumanaweera. 1988. Generic models for robot navigation, *DARPA Image Understanding Workshop*, pp. 453-460.
- [25] Liow, Y. and T. Pavlidis. 1990. Use of shadows for extracting buildings in aerial images, *Computer Vision, Graphics, and Image Processing*, **49**, pp. 242-277.
- [26] Lowe, D. 1987. The viewpoint consistency constraint, *Int. J. of Computer Vision*, **1**, **1**, pp. 57-72.
- [27] Mulgaonkar, P.G., L.G. Shapiro, and R.M. Haralick. 1984 Matching 'sticks plates, and blobs' objects using geometric and relational constraints, *Image and Vision Computing*, **2**, May, pp. 85-98.
- [28] Mulgaonkar, P.G., C.K. Cowan, and J. DeCurtins. Scene description using range data, *IEEE Workshop on Interpretation of 3-D Scenes*, pp. 138-144.
- [29] Nagao, M., T. Matsuyama, and Y. Ikeda. 1979. Region extraction and shape analysis in aerial photographs, *Computer Graphics and Image Processing*, **10**, pp. 195-223.
- [30] Stansfield, S.A. 1988. Representing generic objects for exploration and recognition, *IEEE Int. Conf. on Robotics and Automation*, pp. 1090-1096.

- [31] Stark, L., and K.W. Bowyer. 1991. Generic recognition through qualitative reasoning about 3-D shape and object function, *IEEE Conf. on Computer Vision and Pattern Recognition*, 251-256. See also Achieving generalized object recognition through reasoning about association of function to structure, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **13**, 10, pp. 1097-1104.
- [32] Vayda, A. and A.C. Kak. A robot vision system for recognition of generic shaped objects, *CVGIP: Image Understanding*, to appear.
- [33] Vergnet, R.L., P. Saint-Marc, J.L. Jezouin. 1991 A generic bridge finder, *IEEE Workshop on Direction in Automated CAD-Based Vision*, (June, 1991), to appear.
- [34] Winston, P.H., T.O. Binford, B. Katz, and M. Lowry. 1984. Learning physical description from functional definitions, examples, and precedents, **Proc. of the Int. Symp. on Robotics Research:1**, Michael Brady and Richard Paul, (eds.), MIT Press.
- [35] Winston, P.H. 1975. Learning structural descriptions from examples, in **The Psychology of Computer Vision**, P.H. Winston, (ed.), McGraw-Hill Book Company, New York, pp. 157-209.