A Hierarchical Representation for Efficiently Learning of Three-Dimensional Object Recognition and Description

Jihoon Yang & Vasant Honavar*
Department of Computer Science
226 Atanasoff Hall
Iowa State University
Ames, IA 50011-1040, USA

March 11, 1993

Abstract

Giving human-like visual capabilities to computers is an important goal in computer vision research. Recognition and description of 3-dimensional objects is a largely unsolved problem in this area despite the fact that many proposals have been put forth by a number of researchers in recent years (Chin & Dyer, 1986; Besl & Jain, 1985; Marill, 1991; Honavar, 1992b). We propose a framework for representation of complex three-dimensional objects which is motivated by the need for: parsimony (efficiency) in model construction; amenability to learning (incremental acquisition, assimilation, adaptation, and refinement of models of objects in the environment); and seamless integration as part of an intelligent agent architecture (by keeping the representation and learning structures multi-purpose and flexible).

This paper argues for a hierarchical representational scheme motivated by considerations such as the ones enumerated above; presents such a representational framework; and outlines an approach for the design of learning algorithms for task-driven, parsimonious, incremental acquisition and refinement of the necessary representations within the proposed framework.

1 Description of the Representational Framework

1.1 Background Assumptions

We will proceed under the assumption that complex 3-dimensional objects are made of 3-dimensional primitive components each of which can be identified in the 2-dimensional image(s) of the scene using known techniques. A number of such techniques have been developed to date in the literature (Wang & Freeman, 1990; Gigus et al., 1991; Kim & Kak, 1991; etc.).

It is also assumed that the model base stores the representation of objects from a standard view (that is, we will not store different views of the object resulting from rotation). Our aim here is to focus on the design of the hierarchical representation scheme and the accompanying learning algorithms. The limitations imposed by this assumption can be revoked later exploiting advances in viewpoint-invariant representations such as those based on aspect graphs (Wang & Freeman, 1990; Gigus et al., 1991; Koenderink & van Doorn, 1979; Plantinga & Dyer, 1990; Korn & Dyer, 1987; Watts, 1988; Kriegman & Ponce, 1990).

We will also assume that the scene contains only a single object or the system has some attentional mechanisms (Tsotsos, 1990) that can limit its field of active visual analysis to a single object in the scene.

1.2 Design Considerations

Our aim is to design a representational framework that facilitates the descriptions of complex 3-dimensional objects to be incrementally constructed and refined as necessary for the task of object recognition, description, (and eventually, manipulation by effectors). This is accomplished using an ordered sequence of extraction, abstraction, differentiation, and generalization operators that operate on instances of 3-dimensional objects and the existing model base. The design of the representational framework and the operators that effect changes in representation is motivated by the need for building parsimonious representations with just the right amount of detail necessary for adequate recognition and description of the objects encountered by the

*Partially supported by Iowa State University College of Liberal Arts and Sciences

system in its environment (Honavar, 1992b). For instance, some complex objects might be recognized by the system based on the mere occurrence of the constituent parts in the 2-dimensional image. When such a coarse-level representation is found inadequate the system adds additional detail necessary for successful recognition (e.g., relative spatial arrangement of parts, further decomposition of parts into subparts and so on) to its internal representation of objects in its environment. Within this framework, different objects might be encoded at different levels of detail (in a manner that is loosely analogous to multi-resolution encoding (Rosenfeld, 1984; Uhr, 1987) of two-dimensional images).

The level of detail in the representation is a function of the recognition and description tasks demanded of the system. The refinement of internal representations (model base) maintained by the system is accomplished through feedback-guided learning.

The discussion of the representational scheme presented here is somewhat tentative (in terms of some of the details which we may have to modify in the light of the experience gained by implementing and evaluating a prototype system). The general idea is to have an ordered sequence of levels of representation in which successive levels can incorporate details that were ignored by preceding levels. To keep the system as general as possible and to facilitate modular design, we have chosen to use *generalized distance measures* (GDM) (Honavar, 1992a; Honavar, 92c) for matching objects represented within this framework. GDM is a generalization of the concept of *Levenshtein edit distance*. The *distance* between an instance and a model is given by the *cost* of the *minimum cost* sequence of *edit operations* needed to transform the instance into the model. The costs of individual edit operations is allowed to be different for each model and these costs are determined by a process of inductive learning. GDM offer a natural generalization of the notion of distance or measure of mismatch between two objects as used in a variety of pattern recognition and artificial intelligence techniques (e.g., k-nearest neighbor classifiers, artificial neural networks using radial basis functions, and structured or syntactic methods that use strings, trees, pyramids, attributed relational graphs, as well as artificial intelligence techniques that employ conceptual graphs, frames, and schema structures). A class of GDM-based inductive learning algorithms is currently under development (Honavar,1992a). Such algorithms offer a natural extension of generative or constructive algorithms that enable the adaptive and parsimonious determination of necessary artificial neural network topologies through learning (Honavar & Uhr, 1992).

1.3 Hierarchy of representational levels

Primitive parts are identified using known methods and relations among parts in the scene are computed. Then the scene is matched against the hierarchically represented set of models (which themselves are learned). Matching proceeds from the coarsest level toward the most detailed level until a unique model is identified as the best-match or the match fails (in the event no existing model gives a sufficiently close match — in this case, if the system is in a learning mode, it will construct and store a new model from the scene for future use). The identification of the best match at each level is performed using a generalized distance measure (GDM-based matching criterion is used at each level for finding the best match.

Our current design incorporates the following ordered sequence of levels (at present, the designs for levels 1-3 is more or less final whereas levels 3-6 are somewhat tentative).

- Level 1: This level treats the models as well as the object in the scene as if they were made of unordered sets of primitive objects. The best matching model (roughly speaking), is the one that has the largest number of components in common with the object in the scene (this is not quite true because GDM allows weighting of the mismatches and the weights are set by the learning algorithm).

- Level 2: This level incorporates spatial relationships between primitives. The distance reflects the similarity between the scene and models in terms of the number of pairs of primitives which has the same relationship. Relationship between primitives are encoded with respect to the standard view. The distance between the model and the scene is again computed using the GDM which in this case measures the weighted cost of converting the scene to the model by performing a sequence of addition and deletion operations. The matching algorithm for this level is an adaptation of existing edit-distance computation algorithms for strings and trees (Wagner & Fischer, 1974; Zhang & Shasha, 1989; Wong et al., 1990). We have devised a mapping from graphs to strings that relies on a predetermined ordering over the vertex set that enables efficient matching of relational graphs. A natural extension of this paradigm involves the use of *random graphs* to encode the models, the specification of a generalized distance measure for random graphs, and the development of the corresponding inductive learning algorithms. Recently, we have made some progress in this direction by defining entropy/information-theoretic cost func-

tions for random graphs which can be minimized using connectionist-like learning algorithms which modify the parameters of models represented by random graphs. Many of the details of this approach remain to be worked out.

- Level 3: This level incorporates additional information on the the adjacency of surfaces for each pair of primitives connected by a relation at level 2. That is, it counts the number of mismatched (i.e., has different adjoining) surfaces. It is assumed that unique number is assigned to each surface in each primitive before the identification occurs. Therefore, the distance reflects the similarity in terms of the number of correct adjoining surfaces of the scene and models.
- Level 4: This level might consider the relative size of primitive pairs (scaling). The GDM at this level yields roughly speaking, the normalized summed difference in relative sizes of primitives in the scene with respect to the model.
- Level 5: This level might consider the distance between the centers of primitive pairs are considered.
- Level 6: This level might consider the difference between the solid angles subtended between pairs of primitives in the scene and the model.

2 Learning

One of the primary objectives of this work is to incorporate effective learning mechanisms that enable the system to incrementally acquire and refine its model base to meet the task-requirements in a given environment. The general idea is to have an ordered sequence of learning operators that correspond roughly to the different levels in the representational hierarchy. These operators extract, encode and assimilate segments of the object in the scene into the model base at the appropriate level of detail. Entirely new GDM-based inductive learning algorithms as well as adaptations of existing symbol processing, connectionist, as well as statistical learning methods are being developed for this task (for a discussion of such methods, see Honavar, 1992a; Honavar, 1993).

3 Summary

In this paper, we have proposed a framework for task and data-driven inductive and incremental construction of parsimonious structured hierarchical representations of 3-dimensional objects using connectionist-like learning algorithms that extract, abstract, encode and tune information-rich substructures from the environment. Ongoing research is aimed at the development and evaluation of prototype systems for 3-dimensional object recognition and description. The insights gained from the study will be used to further refine and extend the framework. This is part of a broad research agenda whose long-term objective is the development of powerful learning mechanisms that exploit the strengths of multiple representations as appropriate for the task at hand.

References

[BJ85] P.J. Besl & R.C. Jain. Three-dimensional object recognition. *ACM Computing Surveys*, 17(1):75–145, 1985.

[CD86] R.T. Chin & C.R. Dyer. Model-based recognition in robot vision. *ACM Computing Surveys*, 18(1):67–108, 1986.

[GCS91] Z. Gigus, J. Canny, & R. Seidel. Efficiently computing and representing aspect graphs of polyhedral objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(6):542–551, June 1991.

[Hon92a] V. Honavar. Inductive learning using generalized distance measures. In *1992 SPIE Conference on Adaptive and Learning Systems*, 1992. In press.

[Hon92b] V. Honavar. Learning parsimonious representations of three-dimensional shapes. In *NATO Advanced Research Workshop on Mathematical Representation of Shape*, Dribergen, Netherlands, 1992. To appear.

[HU92] V. Honavar & L. Uhr. Generative learning structures and processes for generalized connectionist networks. *Information Sciences*, 1992. Special Issue on Neural Networks and Artificial Intelligence (In press).

[Hon93] V. Honavar. Connectionist learning algorithms for structured symbolic representations. In: Honavar, V. and Uhr, L. (Ed). Integrating Symbol Processors and Connectionist Networks in Artificial Intelligence and Cognitive Modelling. New York: Academic Press. To appear.

[KD87] M.R. Korn & C.R. Dyer. 3-d multiview object representations for model-based object recognition. *Pattern Recognition*, 20:91–103, 1987.

[KK91] W. Kim & A.C. Kak. 3-d object recognition using bipartite matching embedded in discrete relaxation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(3):224–251, March 1991.

[KP90] D.J. Kriegman & J. Ponce. Computing exact aspect graphs of curved objects: Solids of revolution. *International Journal of Computer Vision*, 5(2):119–135, 1990.

[KvD79] J.J. Koenderink & A.J. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216, 1979.

[Mar91] T. Marill. Emulating the human interpretation of line-drawings as three-dimensional objects. *International Journal of Computer Vision*, 6(2):147–161, 1991.

[PD90] H. Plantinga & C.R. Dyer. Visibility, occlusion, and the aspect graph. *International Journal of Computer Vision*, 5(2):137–160, 1990.

[Ros84] A.Rosenfeld. Multiresolution image representation. In S. Levialdi, editor, *Digital Image Analysis (Selva di Fasano, Italy, November 15-18, 1982)*, pages 18–28, London, 1984. Pitman.

[SFH92] P. Suetens, P. Fua, & A.J. Hanson. Computational strategies for object recognition. *ACM Computing Surveys*, 24(1):5–61, 1992.

[Tso90] J.K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13:423–469, 1990.

[Wat88] N.A. Watts. Calculating the principal views of a polyhedron. In *Ninth International Conference on Pattern Recognition (Rome, Italy, November 14–17, 1988)*, pages 316–322, Washington, DC, 1988. Computer Society Press.

[WF74] R.A. Wagner & M.J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, January 1974.

[WF90] R. Wang & H. Freeman. The use of characteristic-view classes for 3d object recognition. In Herbert Freeman, editor, *Machine Vision for 3D Scenes*, pages 109–161. Academic Press, 1990.

[WYC90] A.K.C. Wong, M.You, & S.C. Chan. An algorithm for graph optimal monomorphism. *IEEE Trans. Systems, Man, and Cybernetics*, 20(3):628–636, June 1990.

[ZS89] K.Zhang & D.Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM Journal on Computing*, 18:1245–1262, 1989.