Transformation-invariant indexing and machine discovery for computer vision

Darrell Conklin
Department of Computing and Information Science
Queen's University, Kingston, Ontario, Canada K7L 3N6
conklin@qucis.queensu.ca

Abstract

Future computer vision systems must have the ability to discover new object models. This problem can be addressed by relational concept formation systems, which structure a stream of observations into a taxonomy of discovered concepts. This paper presents a representation for images which is invariant under arbitrary groups of transformations. The discovered models, also being invariant, can be used as indices for 3D images. The methodology is illustrated on a small problem in molecular scene analysis, where discovered models, invariant under Euclidean transformations, are efficiently recognized in a cluttered molecular scene.

Introduction

An ultimate goal of a computer vision system is to produce symbolic descriptions for scenes, identifying objects and their locations. The approach which has achieved the greatest success in this area is *model–based vision*, where object models drawn from a preconstructed library are matched against segmented parts of an image.

The classical model–based paradigm has two limitations which seriously affect its use as a general–purpose vision system. First, many systems explicitly search for a transformation mapping the model coordinate frame to the image coordinate frame. The model coordinate frame may be object–centered (Marr, 1982). This search can be very expensive, and is complicated by the fact that not all features of an object may appear in a 3D reconstruction of the 2D image (some object parts may be occluded in a particular viewpoint). It also requires that the "correct" model be chosen rapidly. Model–based vision systems usually work with rigid models and cannot cope with the "contortion" (Kosslyn and Koenig, 1992) or "articulated object" (Wolfson and Lamdan, 1992) problem, where overall object identity is preserved but object parts are positioned in novel ways.

The second limitation of current model–based vision systems is that they lack a facility for learning or discovering new models. A general–purpose vision system must be able to encounter new types of objects, relate them to existing types, and thereafter recognize them (Pentland, 1986; Edelman and Poggio, 1992). The ultimate goal of computer vision is unlikely to proceed until this limitation is overcome. Supervised learning systems for object models (e.g., (Connell and Brady, 1987; Winston, 1975)) only partially address the model formation problem since they assume that a model is generalized/specialized by positive/negative examples of a particular class of object. General model formation, though also relying on similar learning operators, is an unsupervised learning process.

In this short paper I argue that techniques from machine discovery and concept formation research can address many of the above problems. First I present a knowledge representation for models which facilitates transformation-invariant descriptions of objects and their parts. Second, I show how recurrent substructures in scenes can be discovered, abstracted, and incorporated into a knowledge base of models organized by subsumption. Finally, I discuss how these principles are being applied to a project in the analysis of 3D molecular scenes.

Transformation-invariant indexing

The task of establishing a correspondence between object model and scene is highly underconstrained. The mapping from model parts to image parts may not be uniquely determined; there may be occlusions in the image (not all parts of the model can be mapped), and there may be multiple objects and/or spurious parts in the image (not all parts of the image are mapped onto) (Grimson, 1990). These problems, combined with the further problem of model selection and database search, have led to other approaches to the model–based vision problem.

Many researchers have departed from the 2D image – 3D model "reconstructionist" paradigm of model–based vision. One such technique that is gaining popularity is the use of 2D characteristic view classes of objects (Wang and Freeman, 1990). A characteristic

view is a "stable" arrangement of an object, in that the view is invariant as the vantage point changes within a neighborhood. There may be several characteristic views of an object: each is a node in that object's *aspect graph*. During object recognition the characteristic views of the selected model are tested for graph isomorphism with parts from the segmented image. Aspect graphs are a promising framework, but efficient algorithms need to be developed for their matching and indexing.

Another promising approach to the viewpoint problem is encompassed by the field of *geometric invariance* (Mundy and Zisserman, 1992). This is the study of geometric properties and relations that remain invariant under transformations. Simply put, an invariant of a representation, with respect to a group of transformations, is a property or relation whose values remain unchanged after application of a transformation from the group. Example transformation groups are the Euclidean, Similarity and Affine groups; transformations can also be dimension-reducing (see (Mundy and Zisserman, 1992) for a good collection of research on invariance). For example, the relation "nearer" between three parts (van Benthem, 1991, Appendix A: "On Space") is invariant under the Similarity group of transformations (translation, rotation, scaling). Invariant features of models can be measured directly from images, have the same value regardless of the image coordinate frame, and can subsequently be used to index a model directly.

Wolfson and Lamdan (1992) describe the Geometric Hashing paradigm for transformation-invariant indexing. For a given model object, an *affine basis* is a triplet of non-collinear points. All other points in the model can be expressed in the reference frame determined by an affine basis. The GH paradigm uses a hash table of (model, affine basis) pairs. During a *training stage*, for every affine basis of a model object with n points the model is transformed, and all other $n-3$ coordinates in the model are indexed in the hash table. During the *matching stage* an affine basis for the image is chosen, the hash table is consulted for all points in the transformed image, and the number of votes for each model are tabulated. Models with many votes are promising candidates for further examination.

The GH transformation-invariant indexing method can be costly; for a database of m models with an average of n points in each, $O(mn^4)$ hash table entries are needed. The method is also sensitive to noise and minor variation in an input image, and has difficulty with the contortion problem.

Representation and Machine Discovery

In this section I propose a new approach to transformation-invariant indexing, where a subsumption taxonomy of transformation-invariant indices is used to retrieve candidate models. These indices are not hand-crafted, rather they are automatically produced by a machine discovery procedure. In addition to guiding the selection of models based on the structure of the image, this subsumption taxonomy can provide a method for *model-driven segmentation* of a scene, that is, integrating the selection and indexing phases of object recognition.

The approach is based on knowledge representation principles of *description logics* (Nebel, 1990). The central mode of reasoning in description logics is *subsumption*; reasoning about the relative generality of concepts. In our extension to description logics images, like all concept terms, are given an extensional semantics and subsumption is sound with respect to this semantics. The result is $\mathcal{SDL}$, a *spatial* description logic. More formally, an *image term* is a pair (I, R), where:

- I is a symbolic image. This is a multidimensional point-data spatial data structure (Samet, 1990): concept terms are placed at coordinates in the structure. The canonical form of a symbolic image is a set of *components*, which are (concept term / coordinate) pairs. The components can inductively refer to image terms.
- R is a set of transformation-invariant relations preserved by this image. The relations are of arbitrary arity and are computed by functions which detect the validity of a relation by directly manipulating the symbolic image data structure.

This representation for object models provides a *diagrammatic* rather than a *sentential* representation (Larkin and Simon, 1987) of invariant spatial relations.

Subsumption between two image terms is a (directed) relational monomorphism (Haralick and Shapiro, 1993) between them; this is a standard graph-theoretic technique in computer vision. Similarity between images is related to the size of a maximal morphism between them — see (Conklin and Glasgow, 1992) for details. Subsumption morphisms between image terms are cached in the subsumption taxonomy (in effect, implementing a mathematical category structure (Conklin and Jenkins, 1993)), providing a compilation mechanism for efficient image indexing. The subsumption taxonomy will have general models close to the root, and instances of models at the leaves. The general models may occur in many different scenes.

One of the main facilities of a description logic is a *classifier*: this procedure places a new concept in its correct location in the taxonomy, just below all most specific subsumers, and just above all most specific subsumees (Woods, 1991). As a new concept progresses down the taxonomy during the classification process, models which are increasingly specific and similar to the concept are encountered.

An important aspect of the knowledge representation scheme outlined above is that image terms can be generalized like all concept terms. It is not clear what the generalization of symbolic images alone should be

(i.e., what is the "generalization" of an image?); however when they are associated with relations they take on a semantics in which generalization must be logically sound. Generalization in $\mathcal{SDL}$ is based on nondeterministic rules applied to image terms. For any symbolic image I and relation set R, the image term (I, R) can be generalized in various ways, for example: ($\succeq$ denotes the subsumption relation):

1. by replacing a part p of a component (p, c) in I by a more general concept term q:

$$\frac{q \succeq p}{(\texttt{replace}(I, (p, c), (q, c)), R) \succeq (I, R)}$$

2. by deleting a component C from the image I:

$$\frac{}{(\texttt{delete}(I, C), R) \succeq (I, R)}$$

3. by removing relation identifiers from the relation set R:

$$\frac{R' \subseteq R}{(I, R') \succeq (I, R)}$$

The contributions of $\mathcal{SDL}$ are that relational models are structured into a subsumption taxonomy and can be created dynamically by generalization as new scenes or objects are indexed by the system (although image terms can also appear as background knowledge). The IMEM (Image MEMory) algorithm (Conklin and Glasgow, 1992) is a structured concept formation algorithm which incrementally develops a knowledge base by expanding it with new image terms. Recurrent spatial arrangements of parts are found by classifying an interpreted image with respect to a current knowledge base, performing matching and generalization with similar image terms having the same classification (see Figure 1). During scene analysis these discovered image terms, being recurrent patterns, are good indices into the knowledge base of models.

The IMEM discovery algorithm is similar to Levinson's (1985) graph discovery algorithm, however it differs in that 1) a broader class of spatial relations can be expressed using the image term representation, 2) the category–theoretic structure of the concept taxonomy leads to a very efficient classifier implementation, and 3) hierarchical models can be expressed in our formalism. Also, we use a similarity threshold to guide the creation of new concepts, whereas Levinson uses a match cost heuristic. IMEM is similar in spirit to Thompson and Langley's (1991) LABYRINTH structured concept formation algorithm — see (Conklin and Glasgow, 1992) for a discussion of key differences.

In contrast to the IMEM approach, where inter–image similarities trigger concept formation, the process described by Holder et al. (1992) analyzes a single image to discover substructures. Heuristic criteria such as compression, compactness, and connectivity are used to guide the discovery. It would be an interesting exercise to combine the inter– and intra–image discovery methods used by these two approaches.

```
incorporate(image)
  let S be the most specific subsumers of image.
  for each concept C in S do
    (a) place a subsumption link between C and
    image.
    (b) merge(image,C).

merge(image,C)
  if there exists an immediate subsumee D of C
   which is compellingly similar to image, then
    (a) form a new concept which is a least
    common subsumer of image and D, and give it
    a unique name U.
    (b) classify(U).

classify(concept)
  if concept is not equivalent to an existing
   concept, place concept in the current concept
   taxonomy just below all most specific
   subsumers, and just above all most general
   subsumees.
```

Figure 1: The IMEM algorithm. Top–level call: incorporate(*image*).

The success of the $\mathcal{SDL}$ approach to knowledge representation and learning depends intimately on the relations used to describe images. Merely stating that they must be transformation-invariant is not enough — the "universal" relation and many others are transformation-invariant, but clearly differ in utility for discrimination during concept formation or for filtering out invalid object models. On the other hand, if very specific relations are used then little similarity between instances can be detected. Overly general and overly specific relations will therefore violate, respectively, the *separability* and *stability* requirements of object models (Wolfson and Lamdan, 1992). The transformation-invariant relations we have used for the molecular scene analysis domain (next section) include proximity, planarity, qualitative orientation, and connectivity.

Application

Our main application of computer vision techniques is *molecular scene analysis* (Glasgow et al., 1992; Fortier et al., 1993): the reconstruction of an interpreted symbolic image from an electron density map of a molecule. The initial image data is complex and noisy, but various image processing techniques are being developed to reduce this complexity. In particular, we are interested in extracting parts corresponding to peaks of electron density; peaks may will correspond to meaningful parts of the molecule in question. This image preprocessing will transform the image data into a *peak map* which has the same form as the discovered models. The technique of matching model to image using the concept taxonomy, called *abductive subsump-*

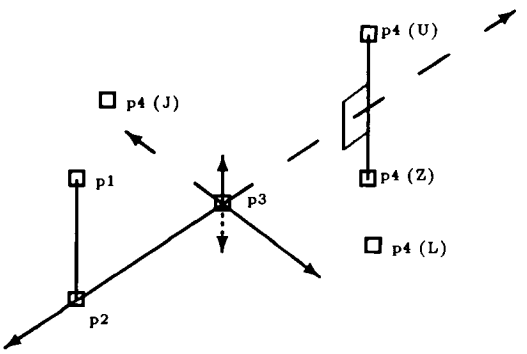

Figure 2: The Δ relation. p1–p2–p3–p4 form a connected chain.

tion, will be discussed and demonstrated below.

The molecular scene analysis problem has several interesting properties from a machine vision standpoint. First, the electron density images are 3D and no 2D-3D reconstruction is necessary (hence no parts are occluded). Second, at the outset of a scene analysis we usually know the covalent structure of the molecule, that is, its atomic constitution and bonding topology. This prior knowledge provides strong constraints on possible scene interpretations. Third, there is much spurious data in the peak map. Finally, molecules are not rigid objects, rather, they assume a number of energetically favourable conformations (contortions).

Machine discovery and transformation-invariant indexing is critical to the molecular scene analysis process. Very large databases of 3D molecular structures exist, and it is highly likely that substructures from the databases can be used to accurately construct a significant portion, if not all, of the new molecule. Defining these substructures and their various conformations is an excellent problem for machine discovery. The substructures should be transformation-invariant so that they subsume any image containing them, regardless of the coordinate frame of the image.

For example, a six–member ring exists in various contortions; what remains constant is the covalent structure. IMEM was applied to a small set of training instances of six–member rings, drawn from the Cambridge Structural Database (Allen et al., 1991). Each instance was described using the relations of connectivity (two parts within a distance of 1.7 Å), and the Δ relation. The quaternary Δ relation measures the "qualitative orientation" of a fourth part relative to a plane formed by three other connected parts. It takes on four values: U, L, Z, and J (see Figure 2). Both connectivity and Δ are invariant under Euclidean transformations. Figure 3 depicts two of the models discovered by IMEM; these correspond to the accepted conformational classes "chair" and "phenyl" used by chemists.

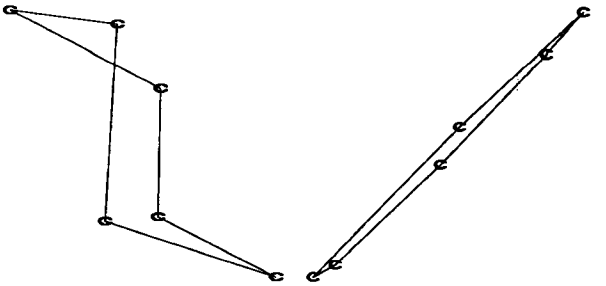

Figure 3: Two discovered models: Left: the chair, Right: the phenyl ring.

Abductive subsumption

I stated earlier in this paper that a subsumption taxonomy of concepts provides a method for model–driven segmentation of a scene. This includes finding the most specific subsumer of the input image with respect to the taxonomy. However, the parts of an image will initially be completely uninterpreted, and therefore their identity must be assumed by abduction during the subsumption test with a concept. When an image reaches a most specific subsumer, each morphism from concept to image gives a plausible interpretation of its parts. No explicit segmentation of the image is necessary.

The inference process of abductive subsumption can be viewed as one step in the *spatial analogy* process (Conklin and Glasgow, 1992); transferring structure and/or part identity from similar instances to an image.

The principles of abductive subsumption can be illustrated on a peak map interpretation problem. Figure 4 depicts a peak map for a small molecule (CSD refcode ACLYCA10). Virtual bonds are drawn between peaks within 1.7 Å. Hand interpretation of such a map requires expert knowledge of chemistry. By abductive classification of the peak map, using the taxonomy described in the previous section, we were able to find a phenyl ring (lower right corner of Figure 4). To be able to interpret other six–member rings present in the molecule and the peak map would require that additional conformations be encountered in a larger training set.

Conclusions

This paper has outlined a framework for the discovery of transformation–invariant object models. These models are organized into a subsumption taxonomy, and in addition to being useful abstractions, they accelerate the model indexing task. The framework has been illustrated on a small problem in the automated interpretation of a high resolution electron density peak map.

The $\mathcal{SDL}$ knowledge representation language is applicable to domains where the *identity of object parts* and their *interrelationships* determines to a large ex-

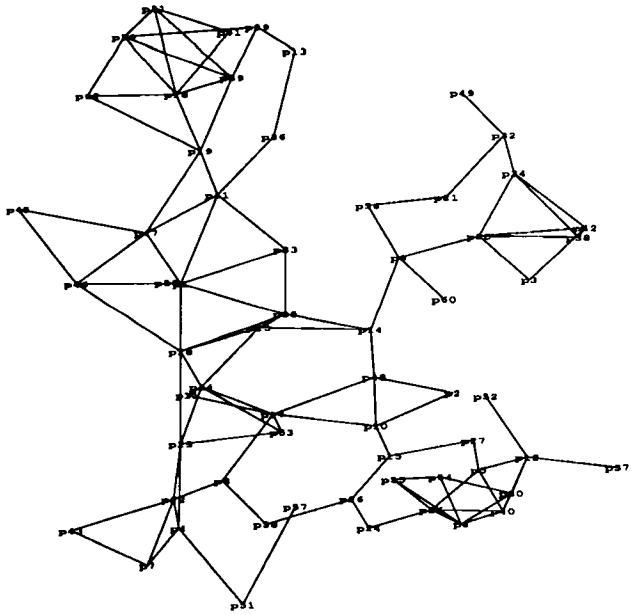

Figure 4: A peak map for the ACLYCA10 molecule.

tent the *identity of the object.* In this paper we did not consider volumetric or surface representations of objects. For the peak map interpretation problem this was a reasonable restriction, although it may not be so for general vision, where shape matching may also be necessary to determine part or object identity. The knowledge representation language $\mathcal{SDL}$ is used by the IMEM machine discovery procedure which, performing relational concept formation, discovers recurrent patterns in scenes and hypothesize them as new transformation–invariant object models.

References

Allen, F. H.; Davies, J.; Galloy, J.; Johnson, O.; Kennard, O.; Macrae, C.; Mitchell, E.; Mitchell, G.; Smith, J.; and Watson, D. 1991. The development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comp. Sci.* 31.

Conklin, D. and Glasgow, J. 1992. Spatial analogy and subsumption. In Sleeman, D. and Edwards, P., editors 1992, *Machine Learning: Proceedings of the Ninth International Conference (ML92).* Morgan Kaufmann. 111–116.

Conklin, D. and Jenkins, M. A. 1993. Compilation of description logics. Submitted for publication.

Connell, J. and Brady, M. 1987. Generating and generalising models of visual objects. *Artificial Intelligence* 31:159–183.

Edelman, S. and Poggio, T. 1992. Bringing the grandmother back into the picture: a memory-based view of object recognition. *Int. J. Pattern Recognition and Artificial Intelligence* 6(1):37–61.

Fortier, S.; Castleden, I.; Glasgow, J.; Conklin, D.; Walmsley, C.; Leherte, L.; and Allen, F. 1993. Molecular scene analysis: The integration of direct–methods and artificial–intelligence strategies for solving protein crystal structures. *Acta Crystallographica* D49:168–178.

Glasgow, J. I.; Fortier, S.; and Allen, F. H. 1992. Molecular scene analysis: Crystal structure recognition through imagery. In Hunter, L., editor 1992, *Artificial Intelligence and Molecular Biology.* AAAI Press.

Grimson, W. E. L. 1990. Object recognition by constrained search. In Freeman, H., editor 1990, *Machine Vision for Three-Dimensional Scenes.* Academic Press. 73–108.

Haralick, R. M. and Shapiro, L. G. 1993. *Computer and Robot Vision*, volume 2. Addison-Wesley.

Holder, L. B.; Cook, D. J.; and Bunke, H. 1992. Fuzzy substructure discovery. In Sleeman, D. and Edwards, P., editors 1992, *Machine Learning: Proceedings of the Ninth International Conference (ML92).* Morgan Kaufmann. 218–223.

Kosslyn, S. M. and Koenig, O. 1992. *Wet Mind: The New Cognitive Neuroscience.* Free Press.

Larkin, J. and Simon, H. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science* 11:65–99.

Levinson, R. A. 1985. *A Self Organizing Retrieval System for Graphs.* Ph.D. Dissertation, University of Texas at Austin.

Marr, D. 1982. *Vision.* Freeman.

Mundy, J. L. and Zisserman, A., editors 1992. *Geometric Invariance in Computer Vision.* The MIT Press.

Nebel, B. 1990. *Reasoning and revision in hybrid representation systems.* Springer–Verlag.

Pentland, A. 1986. Perceptual organization and the representation of natural form. *Artificial Intelligence* 28:293–331.

Samet, H. 1990. *The Design and Analysis of Spatial Data Structures.* Addison-Wesley.

Thompson, K. and Langley, P. 1991. Concept formation in structured domains. In Fisher, D. H. and Pazzani, M., editors 1991, *Concept Formation: Knowledge and experience in unsupervised learning.* Morgan Kaufmann. 127–161.

van Benthem, J. 1991. *The Logic of Time.* Kluwer.

Wang, R. and Freeman, H. 1990. The use of characteristic-view classes for 3D object recognition. In Freeman, H., editor 1990, *Machine Vision for Three-Dimensional Scenes.* Academic Press. 109–162.

Winston, P. H. 1975. Learning structural descriptions from examples. In Winston, P. H., editor 1975, *The Psychology of Computer Vision.* McGraw-Hill.

Wolfson, H. J. and Lamdan, Y. 1992. Transformation invariant indexing. In Mundy, J. L. and Zisserman, A., editors 1992, *Geometric Invariance in Computer Vision.* The MIT Press. chapter 17.

Woods, W. 1991. Understanding subsumption and taxonomy: A framework for progress. In Sowa, J. F., editor 1991, *Principles of Semantic Networks.* Morgan–Kaufmann. 45–94.