Learning Object Models From Visual Observation and Background Knowledge

PAT LANGLEY
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

THOMAS O. BINFORD
TOD S. LEVITT
Robotics Laboratory, Computer Science Dept.
Stanford University, Stanford, CA 94305

ISLE

Institute for the Study of Learning and Expertise

Technical Report 94-4

November 15, 1994

Learning Object Models From Visual Observation and Background Knowledge

PAT LANGLEY[◇] (LANGLEY@CS.STANFORD.EDU)
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306

THOMAS O. BINFORD (BINFORD@CS.STANFORD.EDU)
TOD S. LEVITT (LEVITT@FLAMINGO.STANFORD.EDU)
Robotics Laboratory, Computer Science Department
Stanford University, Stanford, CA 94305

Abstract

This research project aims to use machine learning techniques to improve the performance of three-dimensional vision systems. Building on our earlier work, our approach represents and organizes models of object classes in a hierarchy of probabilistic concepts, and it uses Bayesian inference methods to focus attention, recognize objects in images, and make predictions about occluded parts. The learning process involves not only updating of the probabilistic descriptions in the concept hierarchy but also involves changes in the structure of memory, including the creation of novel categories, the merging of similar classes, and the elimination of unnecessary ones. An evaluation metric based on probability theory guides decisions about such structural changes, and background knowledge about function and generic object classes further constrains the learning process. We plan to carry out systematic experiments to determine the ability of this approach to improve both classification accuracy and predictive ability on novel images.

[◇] Also affiliated with the Robotics Laboratory, Computer Science Department, Stanford University, Stanford, CA 94305

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 15, 1994	Interim Report 8/1/94–11/1/94

4. TITLE AND SUBTITLE

Learning Object Models From Visual Observation and Background Knowledge

5. FUNDING NUMBERS

N00014-94-1-0746 (G)

6. AUTHOR(S)

Pat Langley, Thomas O. Binford, and Tod S. Levitt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for the Study of
Learning and Expertise
2164 Staunton Court
Palo Alto, CA 94306

8. PERFORMING ORGANIZATION REPORT NUMBER

ISLE Technical Report 94-4

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Intelligent Systems Program
800 North Quincy Street
Arlington, Virginia 22217

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Published in *Proceedings of the ARPA Image Understanding Workshop* (1994). Monterey, CA: Morgan Kaufmann.

12a. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This research project aims to use machine learning techniques to improve the performance of three-dimensional vision systems. Building on our earlier work, our approach represents and organizes models of object classes in a hierarchy of probabilistic concepts, and it uses Bayesian inference methods to focus attention, recognize objects in images, and make predictions about occluded parts. The learning process involves not only updating of the probabilistic descriptions in the concept hierarchy but also involves changes in the structure of memory, including the creation of novel categories, the merging of similar classes, and the elimination of unnecessary ones. An evaluation metric based on probability theory guides decisions about such structural changes, and background knowledge about function and generic object classes further constrains the learning process. We plan to carry out systematic experiments to determine the ability of this approach to improve both classification accuracy and predictive ability on novel images.

14. SUBJECT TERMS

computer vision, machine learning, background knowledge, concept formation, Bayesian networks, probabilistic inference

15. NUMBER OF PAGES
8

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

Learning Object Models From Visual Observation and Background Knowledge*

Pat Langley (LANGLEY@CS.STANFORD.EDU)
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA

Thomas O. Binford (BINFORD@CS.STANFORD.EDU)
Tod S. Levitt (LEVITT@FLAMINGO.STANFORD.EDU)
Robotics Laboratory, Computer Science Department
Stanford University, Stanford, CA 94305 USA

Abstract

This research project aims to use machine learning techniques to improve the performance of three-dimensional vision systems. Building on our earlier work, our approach represents and organizes models of object classes in a hierarchy of probabilistic concepts, and it uses Bayesian inference methods to focus attention, recognize objects in images, and make predictions about occluded parts. The learning process involves not only updating of the probabilistic descriptions in the concept hierarchy but also involves changes in the structure of memory, including the creation of novel categories, the merging of similar classes, and the elimination of unnecessary ones. An evaluation metric based on probability theory guides decisions about such structural changes, and background knowledge about function and generic object classes further constrains the learning process. We plan to carry out systematic experiments to determine the ability of this approach to improve both classification accuracy and predictive ability on novel images.

1. Introduction

The ability to perceive surroundings and recognize familiar objects is a basic capability of intelligent agents. In the past decade, research on vision has met with increasing success, but as in other areas of AI, the need for domain-specific knowledge has become increasingly apparent. In addition to knowledge of sensors, image formation, and image processing, a robust vision system requires a repertoire of geometric and material models for three-dimensional objects. Together with the physical laws that govern image formation, such models can be used to predict imaged features of objects, focus attention in the processing of images to recognize these objects, constrain the search for alternative interpretations of what objects are present in an image, and improve the ability to make inferences about occluded or obscured structures. However, the creation and tuning of object models is a painstaking and time-consuming process, making it a bottleneck in the implementation of practical AI vision systems.

Another basic feature of intelligent agents is the ability to learn – to transform experience into knowledge that improves performance. In the past decade, research on machine learning has also made significant strides, with the emergence of new algorithms and new methods for evaluating those techniques. However, most AI work on learning remains focused on symbolic domains such as medical diagnosis or abstract planning. Such machine learning research usually assumes logical representations of knowledge, rather than ones based on measurements extracted from sensor data about the environment. For example, a typical machine learning system might describe an object as 'red', or diseased cells as 'notched ovals', but these features are not directly available in the world. Color is a multidimensional phenomenon, as revealed when one measures the light frequencies over an object's surface, part of which may be in shadow. Similarly, shapes differ considerably in their detail, and recognizing a 'notch' requires representing its physical structure, then matching this description against observed curves in the presence of significant noise around boundaries. In contrast to purely symbolic AI systems, intelligent agents must represent information about the physical world, often in terms of numeric and probabilistic descriptions. Machine learning researchers would benefit from working in domains that force them to address issues of perceptual representation and processing.

Research on both machine learning and computer vision emphasizes structure, but the appropriate categorization for an object is often determined by its function. For example, chairs as a class have an enormous variation in appearance, but humans can recognize them by the functionality implied by their generic 3D structure. A chair must have a platform approximately parallel to

* This research was was supported by Grant No. N00014-94-1-0746 from the Office of Naval Research, with partial funding from the Advanced Research Projects Agency.

the ground with supporting structure below, at a height convenient to most humans. It must also have a surface elevated above the platform and approximately perpendicular to that surface to form a back support, with arm and foot rests being optional. Thus, function and structure constrain one another, often making it possible to infer one from the other. An unconventionally shaped door is recognizable as an articulated object between two spaces even when the particular shape model is not represented in the knowledge base. Functional knowledge plays an important role in both vision and learning.

In this paper we outline a research program that we hope will unify these aspects of intelligence. We plan to use methods from machine learning to automatically refine object models for use in machine vision, starting with knowledge of function. At the same time, we will use vision as a challenging domain for the testing and development of methods for machine learning. We begin by reviewing our approach to representing experience, object models, and function, along with our scheme for organizing this knowledge in long-term memory. We then turn to mechanisms for recognizing new cases of object classes in visual data, including techniques for probabilistic inference and focus of attention. After this we discuss an incremental approach to learning object models, drawing on both training instances and domain knowledge, then consider the contributions of this approach to our understanding of vision and learning. Finally, we present our plans for evaluating the resulting system in visual domains and discuss related work on vision and learning.

2. Representation and Organization of Object Models

In order to develop programs that can learn from visual observations, we must first select some representation of physical objects. As in our previous work (Binford, Levitt, & Mann, 1989), we represent object models in long-term memory at different levels of part/subpart aggregation. The lower levels of these 'part-of' hierarchies include constructs such as edges and regions, edge/region relations, projections of volume primitives into 2D images called *ribbons*, relations among ribbons, surfaces in 3D space, and primitive volumes called *generalized cylinders*. The latter are the basic building blocks of three-dimensional object models; these models are described as logical (and/or/not) combinations of generalized cylinders, spatial relations among local coordinate systems attached to model components, and relations among these combinations. Beliefs about objects in the world are represented using probabilistic distributions over events expressed in the same hierarchical model representation.

Object classes are also organized hierarchically in 'type' or 'is-a' hierarchies. This scheme reflects the notion of object specialization and represents classes at different levels of abstraction. For instance, most automobiles have a hood, a trunk, a chassis, and four wheels, but they differ in their relative sizes and locations. Thus, one can organize knowledge about classes of cars into an is-a hierarchy that is partially ordered according to generality. Concepts high in this hierarchy denote abstract classes (sports cars, sedans), lower ones correspond to particular makes (Jaguar XKE's and Mazda RX7's), and terminal nodes indicate specific cars (John's XKE). One can organize the components of cars (represented as generalized cylinders) in a similar is-a hierarchy. Because more general categories cover a wider range of objects, their descriptions will typically have higher variances and thus provide fewer constraints on recognition and inference than more specific ones. For instance, hypothesizing that an object is a car makes less detailed predictions than hypothesizing it is a Jaguar XKE. We will return to this issue later, when we discuss learning.

The function of an object is directly related to the forces or motions that it exerts on other objects. In our framework, functions are represented in terms of spatial relationships among objects and sequential transformations on those objects. Such sequences can be continuously indexed, as in the parametrized motion of a robot arm as a function of time. Other functions involve persistence over time, as in chair legs that support a seat. This view of function lets our framework represent, compute, and infer functions from 3D object structure, and to infer 3D object structure from knowledge of function.

Because visual domains are inherently physical and perceptual, we represent all levels of aggregation, type, and function in terms of uncertain physical and geometric constraints. Many of these constraints come from general physical and geometric knowledge combined with sensor models. For instance, we represent the probability that certain types of edges and regions will be observed, given certain 3D object and sensor relations. Similarly, the occurrence of a certain generalized cylinder probabilistically 'implies' the observation of certain ribbon relations, and the occurrence of specific complex objects predicts the observation of specific cylinders. Different levels of visual knowledge may assume different probability distributions, but all share a need to represent the uncertain relation between observations of the environment and its actual state.

There exists a simple and direct mapping from hierarchies of probabilistic concepts into Bayesian inference networks (Lauritzen & Speigelhalter, 1988). Briefly, each object class in an is-a hierarchy can correspond to a single node in the Bayesian network. In addition, each part-of component that occurs in the object hierarchy can map onto a node in the Bayesian network. Information about conditional probabilities, which is stored with the concepts in the hierarchy, is stored on the influence links in the Bayesian network. Our approach to visual recognition relies directly on this correspondence (Binford, Levitt, & Mann, 1989).

3. Recognition of Object Classes

In vision, one of the basic tasks is to recognize and infer the structure of three-dimensional objects from observational information that is present in imagery. This task requires processing at the multiple levels described in the previous section, each of which introduces its own

forms of uncertainty. In tackling the recognition problem we will draw on the methods we have used in our earlier work (Binford et al., 1989), which incorporate both bottom-up processing (from pixels to edges to ribbons to cylinders to objects) and top-down processing. At each level, one maintains competing hypotheses about the proper interpretation for portions of the image. One may believe with 0.95 probability that an inferred region is associated with surface A and believe with 0.05 probability it belongs with surface B. Similarly, one may believe that an inferred generalized cylinder is a tire with 0.8 probability, a boulder with 0.15 probability, and a manhole cover with 0.05 probability. These probabilities typically change over time, as additional evidence emerges through further processing.

In our previous vision research, we have used Bayesian networks to propagate evidence across different levels of the part-of hierarchy. Conditional probabilities relate part-of relations on the arcs of the hierarchy, whereas belief in object components is represented at the nodes of the hierarchy. Leaves in the graph correspond to observed measurements, whereas interior nodes specify derived measurements. Inference is seeded by the output of edge operators that provide evidence for boundaries in the imaged objects. Potential boundary segments are grouped to match models of ribbons and projected surfaces. Matches are instantiated as hypotheses for instances of models occurring in the world supported by the match evidence. Beliefs in the truth of model matches are represented as conditional probability distributions over competing interpretations.

Partial matches along the part-of hierarchy are compared against the object models to predict locations of imagery features for other imaged object components. Decision-theoretic control algorithms order the actions with the highest predicted payoff to gather additional evidence in support or denial of hypothesized imagery interpretations (Levitt, Binford, & Ettinger, 1990), thus directing attention to useful regions of the image. The instantiated network of matches forms a hierarchical Bayesian network in which nodes represent competing local interpretations and arcs represent hypothesized part-of relationships. As new nodes are instantiated from matches, and as additional image processing of 2D/3D geometric and material evidence is attached to existing nodes, algorithms for updating Bayesian networks propagate beliefs over the entire network.

By contrast, the recognition module we have used in our work on machine learning (Gennari, Langley, & Fisher, 1989) starts with the most abstract concept in an is-hierarchy at a given level of aggregation, then estimates the probability for each specialized child of that concept. Briefly, recognition involves sorting an observed instance downward through the probabilistic concept hierarchy, selecting the most probable alternative at each is-a level. However, based on the mapping from a hierarchy of probabilistic concepts to a Bayesian network (described above), our future work will use the propagation techniques associated with the latter to classify objects instead. This approach should give the effects of sorting through a hierarchy of probabilistic concepts but provide both a cleaner semantics and consistency with our previous work on visual recognition. The scheme also provides a more coherent way to handle objects that have part-of structure, which our previous learning work (Thompson & Langley, 1991) addressed in a somewhat ad hoc manner.

We will use the same probabilistic inference procedure to draw on functional knowledge during recognition. The random variables at these nodes will specify time-indexed parametric mappings among sets of random variables that represent objects. Bayesian networks support inference in both directions, from functions to objects or vice versa. This unified approach to representing part-of, is-a, and functional knowledge will simplify the processes of both visual recognition and learning.

4. Refining Object Classes through Machine Learning

A central insight of machine learning is that background knowledge can significantly constrain the acquisition of new knowledge, and we are taking advantage of this idea in our work on learning in visual domains. In particular, we assume the system already has accurate probabilistic knowledge about lower levels of aggregation, from edges to generalized cylinders. Processing at this level involves specialized optimal estimation problems that have been addressed elsewhere; thus, we are not dealing with learning at these levels. Moreover, we assume the system already has an initial set of generic object models, including the generic pieces of which they are composed. For instance, in the domain of vehicles, we would assume initial knowledge of sedans, sports cars, station wagons, vans, and trucks, as well as knowledge of tires, hoods, doors, trunks, and the like.

The specific learning task we will examine can be stated in terms of its inputs and outputs:

- Given a set of generic object models and knowledge of their function;
- Given a set of images for instances of those object classes;
- Acquire specializations of the object classes that improve accuracy of future object recognition.

For instance, given generic models for various vehicle types and images of particular vehicles, acquire object models for specific makes of sedans and vans, as well as models for individual vehicles. Similarly, given generic models for vehicle components and images of particular components, acquire models for specific types of tires and doors, as well as models of individual components.

To address this problem of learning specialized models, we are borrowing directly from our previous work on incremental, unsupervised concept learning (Gennari, Langley, & Fisher, 1989). One can view generic object models as nodes at the top (most general) level of an is-a hierarchy. In this view, the learning process involves the gradual addition of specialized concepts lower in this hierarchy, each making finer distinctions than its parent, and thus containing more information that the system

can use to direct its attention and evaluate competing hypotheses. This is the sense in which learning should gradually improve the visual recognition of object models as the system gains experience in a domain.

However, before one can specialize an existing generic object model, it is necessary to first decide that an image contains an instance of that object class and hypothesize a specific three-dimensional description of that instance, including sub-parts, attributes, and specializing features. This is where background knowledge plays an essential role. We believe that the initial knowledge about generic object models, combined with knowledge about the lower levels of aggregation, is sufficient to produce reasonably accurate descriptions of instances that occur in images. Functional knowledge also plays an important role in this process, constraining both the types and parts of training objects. The resulting descriptions are then passed to the learning module, which uses them to modify knowledge in the is-a hierarchy. This specialized knowledge in turn constrains future recognition and learning.

Given a mechanism for assigning training instances to object classes, and thus sorting the instance through an is-a hierarchy, learning can occur in a number of ways. The simplest mechanisms merely update the probabilistic descriptions for each model to which the instance is assigned, but others actually modify the structure of the hierarchy. Figure 1 illustrates the four learning operators that can lead to such changes:

- *extending downward*, which occurs when a training case reaches a terminal node in memory; under these circumstances, the learner creates a new node N that is a probabilistic summary of the case and the terminal node, making both children of N;
- *creating a sibling*, which occurs if a training case is sufficiently different from all children of a node N; in this situation, the learner creates a new child of N based on the case;
- *merging two concepts*, which occurs if a case is similar enough to two children of node N that the learner judges all three should be combined into a single child;
- *splitting a concept*, which occurs when a case is different enough from a child C of node N that the learner decides C should be removed and its children moved up to become children of N.

The last three of these actions are considered at each level of the hierarchy, as the system sorts the new training instance downward through memory. If none of these are deemed appropriate, the induction algorithm simply averages the case into the probabilistic structural description for the best-matching category, then recurses to the next level.

The most important issue here involves deciding whether, at a given level in the hierarchy, the instance should be incorporated into one of the existing subclasses, or whether it is sufficiently different from them to justify creation of an entirely new subclass. Gennari, Langley, and Fisher (1989) describe an evaluation function based on information theory that can be used to make these decisions. However, experimental studies with this method have revealed significant sensitivity to the ordering of training instances. The merge and split operators mitigate this effect, but much of the problem seems due to overcommitment to an unfavorable hierarchy structure early in the learning process. Fortunately, McKusick and Langley (1991) have shown that priming the concept hierarchy with reasonable top-level categories reduces order effects, producing more rapid improvement in predictive accuracy and giving more understandable concept hierarchies. Although their results focused on attribute-value formalisms, they should also hold for more complex representations and thus further recommend our introduction of background knowledge in the form of generic object models.

Like all induction algorithms, Gennari et al.'s learning method is negatively affected by the presence of irrelevant attributes in the instance descriptions, and we expect this issue will be exacerbated in visual domains, where complex objects can involve hundreds of features. Our response to this problem relies on knowledge of objects' functions, embodied in a *utility metric* (Edwards, 1993) that influences the application of the four learning operators. The basic probabilities passed to this utility metric would still come from the reconstruction of 3D object parts and their spatial relations, and if we based learning decisions on this factor alone, we would obtain classes and subclasses based entirely on similarities of 3D structure. However, the utility metric can encode task-specific knowledge about the relative importance of different factors, whether these involve the structure of parts and their configuration, their surface appearance, or their overall function.

For example, in some domains one may care more about the orientation of an object's parts than the sweeping function used to describe them. Thus, chairs with round seats and backs (which provide vertical and lateral support) would be more similar chairs with square seats and backs (which provide the same support) than they would to objects with round components oriented in different ways (and thus provide no support). Likewise, some differences in surface appearance (e.g., a country designator on an aircraft) may be more important in some contexts than others that are equally large (e.g., mud on an aircraft). Even though door knobs and door handles have very different 3D structures, they serve the same purpose, which is reflected in their similar heights on doors and their graspability by the human hand.

We can encode knowledge about the relative importance of such factors – whether they involve structure, appearance, or function – in the utility metric. In this scheme, one multiplies the importance values by the probabilities inferred during 3D reconstruction, thus biasing decisions made during the selection of learning operators. The incorporation of task-oriented knowledge about utility leads to a much richer theoretical framework for learning than one that deals only with differences in 2D or 3D structure, and we predict that it will significantly speed the rate of learning by reducing the influence of irrelevant features.

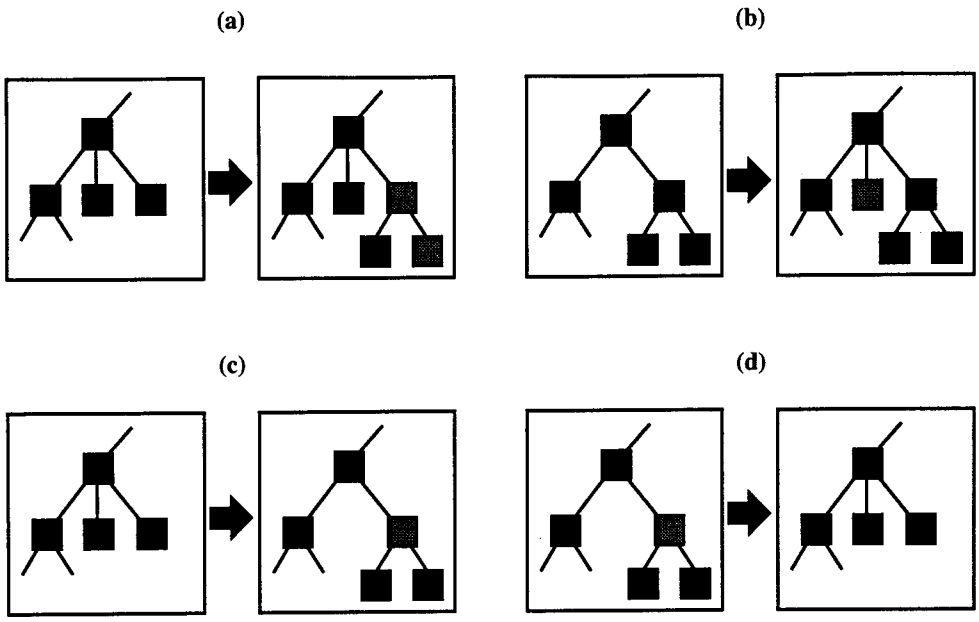

Figure 1: Learning operators used to modify the structure of a hierarchy of probabilistic concepts: (a) extending the hierarchy downward; (b) creating a new sibling at the current level; (c) merging two existing concept; and (d) splitting an existing concept. Newly created nodes are shown in gray.

5. Contributions of the Research

Although the proposed research will build on our previous work on computer vision (Binford et al., 1989) and machine learning (Gennari at al., 1989), it will address issues that go beyond the simple merging of these two traditionally separate paradigms. Our earlier efforts shared some basic assumptions, such as the importance of handling uncertainty and a reliance on probabilistic methods, but the full integration of the two approaches still requires some important advances. In particular, we must:

- Identify robust methods for unsupervised learning over structural descriptions and multiple levels of aggregation. Previous work on structural induction, including our own (Thompson & Langley, 1991) has used impoverished representations, and Binford et al.'s (1989) formalism for representing physical objects provides a much richer description language but also a greater challenge than addressed in previous learning research.
- Explore the use of background knowledge, including functional information, to constrain the learning process. Most research on learning in the presence of functional knowledge has used representations designed for logical reasoning (e.g., Horn clauses), rather than formalisms that support recognition in uncertain domains like vision. Our goal of storing background knowledge in the same hierarchy as learned knowledge, and our scheme for mapping function onto physical structure, provides a more coherent framework but also takes us into unexplored waters.
- Clarify the relation between probabilistic concept hierarchies and Bayesian inference networks. Most recent work on probabilistic induction has used concept hierarchies, but Bayesian networks, which have played a central role in our vision work, have drawn considerable attention in other circles. Our mapping between these approaches to organizing, using, and learning probabilistic knowledge should help both communities understand the contributions of the other.

In addition, we must also find ways to evaluate the learning ability of the system we are developing. Although the field of machine learning has a reputation for careful experimental studies, it has focused on simple learning tasks involving attribute-value representations, it has seldom addressed the role of background knowledge, and it has avoided complex tasks like 3D vision. In contrast, the computer vision community typically works with quite complex images, but sometimes lacks careful empirical studies involving many test cases and well-defined metrics. The empirical studies of our learning algorithm, to which we now turn, must take the best from both of these worlds.

6. Plans for Experimental Evaluation

In evaluating our system's ability to learn, we will draw upon experimental methods that are now commonly used within the machine learning community (Kibler & Langley, 1988). In this framework, the goal of learning is to improve performance on some task, in this case the recognition and description of objects in images. We will divide images into training and test cases, present

the training images to the learning system sequentially, and measure its performance on the test instances after every N training instances. We will average the resulting learning curves over many runs based on different random orders of the training cases. We will also compare these curves against the performance of a nonlearning system that uses only generic object models.

We will use three main measures of performance, which will serve as the dependent variables in our experiments. The first is simply the accuracy of classification, i.e., the percentage of instances correctly labeled by object class. Although our learning algorithm will be unsupervised, one can provide the system with class information in the training data, provided it is not used in learning, and then ask the system to predict the class on test instances. This provides a reasonable measure of the learning algorithm's ability to acquire classes that were present in the data. A more difficult task involves predicting the three-dimensional structure of objects in test images: given an image of an object, the system must use its acquired knowledge to generate a three-dimensional description of the object. Here the performance measure is the average difference between the actual and inferred structure. A third metric involves the time required for recognition, measured both in CPU seconds and in terms of basic inference steps taken by the performance system.

We are considering a number of domains for use in our experimental studies. One such domain involves ten machined parts that make up a benchmark suite developed by researchers at the University of Utah. Five parts are cover plates and five are steering arms; one of each type is correctly machined, whereas others are defective parts that have missing, extra, or misplaced features. These images would let us avoid some issues, such as figure-ground separation, but the object shapes are sufficiently complex to challenge our representation, performance method, and learning algorithm. For this domain, we would provide background knowledge about the shape of correct objects and their function, and we expect that our induction method would acquire subclasses that correspond to different types of defective parts.

A second domain we are considering involves the interpretation of aerial images for the purpose of surveillance or inventory. In this case, the objects of interest are buildings, roads, vehicles, and related cultural artifacts such as fences and parking lots. The aim here goes beyond the recognition of individual objects to the detection of significant object configurations. For example, an important distinction in this domain is between airfields that are preparing to launch attacks and ones that are not. Such concerns relate directly to our ideas about the importance of function in recognition and learning, and they suggest obvious forms of background knowledge to give the system. The availability of high-resolution aerial images (10,000 × 10,000 pixels), through the RADIUS program's Fort Hood data set, would aid testing in this domain.

Our approach to learning object models suggests a number of explicit hypotheses that we plan to test in our experiments. The most obvious of these is that, as the system encounters more training images, both its recognition and prediction accuracy should improve on novel test images. However, we also expect that recognition time will decrease, due to an improved ability to focus attention on objects of particular subclasses. Given limited recognition time, we naturally expect a tradeoff between speed and accuracy, but we also predict this tradeoff will become flatter with experience, so that (after enough learning) the vision system will recognize objects very rapidly with little loss in accuracy.

We will also use our experiments to assign credit to different components of the learning system. For example, we will compare the full algorithm, which refines the initial concept hierarchy by creating new subclasses, with a lesioned version that only alters the probability distributions associated with the original object models. We hypothesize that the refinement process will lead, asymptotically, to higher accuracy and faster recognition in domains where subclasses exist. We also expect that background knowledge, whether in the form of generic object models or a function-motivated utility metric, will reduce the number of training cases needed to reach asymptotic performance, minimize the effects of training order, and mitigate the influence of irrelevant features. However, all of these predictions are subject to experimental tests, and we must know the results before drawing conclusions about the usefulness of the various facets of our approach.

7. Related Work on Vision and Learning

Recently, a number of other researchers have also explored techniques for inducing object models from training images, and we should briefly compare their approaches to our own. In each case, we discuss the representation and organization of learned knowledge, along with the performance and learning components that use and acquire that knowledge.

For example, Pope and Lowe (1993) represent a particular object as a set of characteristic views, each described as a set of features at multiple levels of aggregation. Associated with each feature is a probability of occurrence and a probability distribution for its numeric attributes. These descriptions are exclusively 2D, which contrasts with our emphasis on 3D models. Recognition of new images involves using Bayes' rule to compute the probability of each view given the features found in the image, then selecting the most likely one. Pope and Lowe's learning scheme incrementally assigns each image description to the most likely view and updates the probability distributions for that view, but creates new characteristic views for sufficiently novel instances. This method is very similar to our induction algorithm, except that it creates only one level of clusters and uses a different evaluation metric. In related work, Beis and Lowe (1993) have focused on creating concept hierarchies and indexing object models, which comes closer to our hierarchical approach.

Sengupta and Boyer (1993) have taken a similar approach that also represents objects models as probabilistic summaries at different levels of aggregation. Their work emphasizes the organization of models in an is-a hierarchy, through which they sort new descriptions during

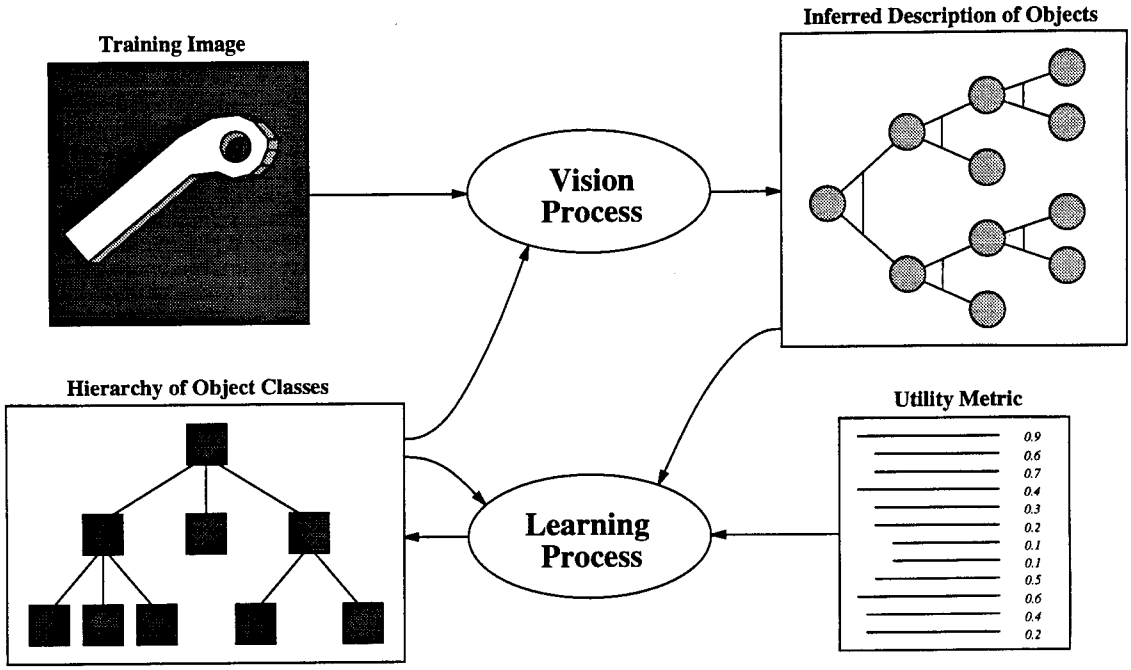

Figure 2: Overview of the framework for combining computer vision and machine learning. The vision process uses a hierarchy of generic models to transform a training image into a 3D description of objects in the image. The learning process then uses this description to modify the hierarchy, using a utility metric to bias its decisions.

recognition. This sorting process leads to updates in the probabilistic summaries through which the description passes, and it produces a new subclass when it reaches a level at which the description is sufficiently different from existing class summaries. Thus, the scheme is very similar to our earlier work on unsupervised concept learning, differing primarily in its evaluation metric and its use of beam search for sorting rather than a greedy method. Sengupta and Boyer have tested their approach using 3D object descriptions, but have taken their training cases from a CAD library rather than actual images.

Segen (1993) has also used probabilistic summaries, at different levels of aggregation, to represent object models. He describes each class of objects as a 'stochastic graph', which consists of a set of components, each described using a discrete probability distribution over a set of nodes that are themselves stochastic graphs. The recognition process assigns an image to the most likely top-level graph, and the incremental learning algorithm either updates the probabilistic summaries for the selected class or creates a new class if the image is different enough from existing ones. The evaluation metric used in both recognition and learning is closely related to notions of minimum description length. Segen has tested this approach in the domain of gesture recognition.

Conklin (1993) has taken a different approach to representation that relies on logical descriptions. His system organizes memory into an is-a hierarchy, but each nonterminal node contains not a complete summary of training cases but conjunctions of features held in common by all of its children. These descriptions, which are transformation invariant, serve primarily as indices for retrieving individual training cases, but also aid in parsing images as they are sorted through the hierarchy. As in our earlier work, learning is incremental and interleaved with the sorting process, with training cases being stored as new terminal nodes but also leading to more general descriptions along the paths they traverse. Conklin has tested his approach on molecular scene analysis given electron density maps.

Gros (1993) has taken a nonincremental clustering approach to the induction of 2D object models. His approach represents a model as a set of characteristic views, each having a set of logical features, such as line segments and their points of intersection, all occurring at a single level of aggregation. Although Gros does not describe a performance component, one might use such descriptions for object recognition to assign images to the characteristic view with the most matched features, using some version of a nearest neighbor algorithm. The learning stage draws on a nonincremental agglomerative clustering algorithm, which successively merges the two clusters of images that are nearest in the feature space, then employs a threshold to determine the top-level classes.

Our emphasis on the role of background knowledge and function distinguishes our framework from most research on vision and learning, but Cook, Hall, Stark, and Bowyer (1993) describe an alternative method that also incorporates these ideas. They provide their learning system with models for an object class (e.g., chairs) and a set of 'fuzzy' inference rules for predicting the degree to which an object satisfies its function. A tutor provides training images with associated functionality scores, on which the learning algorithm bases its revision of the inference rules' conditions, using a technique similar to backpropagation. Background knowledge lets

the system infer a 3D description from the image, and it also constrains the functional learning process. This approach differs from ours in its use of fuzzy inference rather than probabilistic reasoning, and in its emphasis on predicting functionality rather than on recognition. The earlier work of Winston, Binford, Katz, and Lowry (1983) on learning recognition rules from functional knowledge comes closer to our approach, but here the learned process generated logical descriptions.

As we have seen, the recent literature includes a number of research efforts that address many of the same issues as our ongoing work. Some deal with the incremental, unsupervised induction of object models, others incorporates probabilistic representations and recognition mechanisms, a few focus on the organization of knowledge in long-term memory and the importance of abstract object classes, and one draws on functional background knowledge. However, none have attempted to combine all of these ideas into a coherent framework.

8. Concluding Remarks

In summary, we are developing a theoretical framework that unifies our previous work on computer vision and machine learning. As Figure 2 depicts, this framework assumes that the vision system uses background knowledge, in the form of generic models of object classes, to reconstruct three-dimensional descriptions of objects in a training image. The learning system then uses uses these descriptions, along with functional knowledge encoded in a utility metric, to revise the background knowledge by forming more specialized descriptions of the objects' classes, which the vision system uses in turn on the next image. Techniques for handling uncertainty are central to both the vision and learning modules.

Although our project remains in its early stages, we are confident that the basic approach will increase our understanding of both computer vision and machine learning. We have not yet integrated the two components of our system, but we have decided on a common representation language that will let them communicate, we have examined the domain of machined parts in some detail, and we have designed specific experiments to test our hypotheses about the impact of learning and background knowledge on the speed and accuracy of the vision process. Whether these hypotheses are borne out, or whether we must modify our system design to achieve the desired effects, is the central question that we hope to answer in our future research.

References

Beis, J. S., & Lowe, D. G. (1993). Learning indexing functions for 3D model-based object recognition. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 50–54). Raleigh, NC: AAAI Press.

Binford, T. O., Levitt, T. S., & Mann, W. B. (1989). Bayesian inference in model-based machine vision. In L. N. Kanal, T. S. Levitt, & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (Vol. 3). North Holland.

Conklin, D. (1993). Transformation-invariant indexing and machine discovery for computer vision. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 10–14). Raleigh: AAAI Press.

Cook, D., Hall, L., Stark, L., & Bowyer, K. (1993). Learning combination of evidence functions in object recognition. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 139–143). Raleigh, NC: AAAI Press.

Edwards, W. (1993). *Utility theories: Measurements and applications.* Boston: Kluwer.

Gros, P. (1993). Matching and clustering: Two steps towards automatic object model generation in computer vision. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 40–44). Raleigh: AAAI Press.

Levitt, T. S., Binford, T. O., & Ettinger, G. J. (1990). Utility-based control for computer vision. In R. D. Schacter, T. S. Levitt, L. N. Kanal, & J. F. Lemmer (Eds.), *Uncertainty in artificial intelligence* (Vol. 4). North Holland.

Gennari, J. H., Langley, P., & Fisher, D. H. (1989). Models of incremental concept formation. *Artificial Intelligence, 40*, 11–61.

Kibler, D., & Langley, P. (1988). Machine learning as an experimental science. *Proceedings of the Third European Working Session on Learning* (pp. 81–92). Glasgow: Pittman. Reprinted in J. W. Shavlik & T. G. Dietterich (Eds.), *Readings in Machine Learning.* Morgan Kaufmann.

McKusick, K. B., & Langley, P. (1991). Constraints on tree structure in concept formation. *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 810–816). Sydney: Morgan Kaufmann.

Pope, A. R., & Lowe, D. G. (1993). Learning 3D object recognition models from 2D images. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 35–39). Raleigh: AAAI Press.

Segen, J. (1993). Learning shape models for a vision-based human-computer interface. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 120–124). Raleigh, NC: AAAI Press.

Sengupta, K., & Boyer, K. L. (1993). Incremental model base updating: Learning new model sites. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 1–5). Raleigh, NC: AAAI Press.

Thompson, K., & Langley, P. (1991). Concept formation in structured domains. In D. Fisher, M. Pazzani, & P. Langley (Eds.), *Computational approaches to concept formation.* San Mateo, CA: Morgan Kaufmann.

Winston, P. H., Binford, T. O., Katz, B., & Lowry, M. (1983). Learning physical descriptions from functional descriptions. *Proceedings of the Third National Conference on Artificial Intelligence* (pp. 433–439). Washington, DC: AAAI Press.