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

Report of the AAAI Fall Symposium on Machine Learning and Computer Vision: What, Why and How?

Kevin W. Bowyer
Lawrence O. Hall
Department of Computer Science & Engineering
University of South Florida
Tampa, FL 33620

Pat Langley
Institute for the Study of
Learning and Expertise
2451 High Street
Palo Alto, CA 94301

Bir Bhanu
College of Engineering
University of California
Riverside CA 92521

Bruce A. Draper
Department of Computer Science
University of Massachusetts
Amherst, MA 01003

1. Introduction

This report gives an overview of the 1993 AAAI Fall Symposium on *Machine Learning in Computer Vision: What, Why and How?* The level of interest in the symposium topic was indicated by the degree of participation. Over 70 researchers registered for the meeting, and 60 of these were still present at the end of the second full day of sessions. There was strong attendance from both the machine learning and the computer vision communities, although, perhaps predictably, some from each community felt that the other area had greater representation.

The symposium was divided into ten 90-minute sessions, with seven devoted to moderator/author coverage of contributed papers,[1] two consisting of invited talks, and one involving a panel discussion. The moderator/author format for the contributed paper sessions proved interesting and valuable. Each moderator summarized and commented on five papers and then let the authors respond. The moderators had done their homework, and their questions for authors were almost always right on the mark. Several authors even used the transparencies made by the moderator to guide their comments. Invited speakers included Tom Mitchell and Rich Sutton from machine learning and Chris Brown and Ramesh Jain from computer vision. Abe Waksman from the Air Force Office of Scientific Research organized the panel discussion.

When attempting to categorize work on machine learning for computer vision, one must decide whether to approach the problem from the perspective of learning or vision. In this report, we follow both paths, first describing the tasks for which machine learning holds potential for aiding vision research, and then describing challenges that vision presents for work in machine learning. In each case, we note relevant presentations from the symposium. In closing, we consider the state of research on machine learning for computer vision and recommend some steps that would lead toward a more mature discipline.

2. Roles for Machine Learning in Computer Vision

One common definition holds that learning involves the improvement of performance through the acquisition of knowledge from experience. In this view, it makes no sense to talk about learning in the absence of a well-defined performance task, and a recurring theme during workshop discussions was the development of end-to-end, 'task-oriented' vision systems. Some of the visual tasks to which machine learning might contribute are object recognition, surface reconstruction, pose determination, and change detection (monitoring).

In many cases, the basic performance task of a vision system can be viewed as mapping from sensory data (the input to the problem) to one or more of a set of possible decisions or actions (the output to the problem). For example, take the traditional problem of recognizing a 3-D object from a single arbitrary view. In this case, the input is an image and the desired output is recognition

[1] For those interested in the details of individual papers, the working notes of the meeting are available through AAAI as technical report FS-93-04. Send electronic mail to FSS@AAAI.ORG for details.

and localization of all instances of a known set of object models that appear in the image. The vision community would generally accept that the performance of an object recognition system had improved if there was reduction in either the error rate for recognition decisions or in the computation required to arrive at those decisions.

Although research in computer vision aims to develop complete systems, one can often decompose the overall performance task into three subtasks.[2] First, the system must transform its sensory input to a set of features, that is, early symbolic or qualitative abstractions of the input. Depending on the system, this process of *feature extraction* may infer edges, corners, texture energy, surface normals, surface patches, optic flow vectors, or many other structures. In the second subtask, the vision system must go from inferred features to a set of partially instantiated entities that it may discern in the sensory input. This *indexing* or retrieval process may deal with models of specific objects, entire classes of objects, patterns of motion, contexts of scenes, or yet other phenomena. Finally, the vision system must go from these candidate models to decisions about the best models for the given sensory input, using some *model evaluation* or *recognition* process. Different systems may place different relative emphasis on these subtasks, but all systems must in effect deal with them in some manner. Below we briefly consider how learning might improve performance for some of these subtasks.

Improving model evaluation. The model-based recognition process requires some models of the entities that one must match against inferred features. Traditional vision systems have dealt with small sets of models (typically in the tens), because developers have been forced to enter their models by hand. However, the vision field aspires to systems that can recognize thousands of different objects, and it desires to build them in a reasonable time and at reasonable costs. Perhaps the most obvious use of machine learning for vision involves the automated acquisition and refinement of models from training images. Learning entirely new models can improve recognition accuracy by increasing the number of objects or classes covered by the system. Refining existing models can improve accuracy by reducing confusions about similar entities. Most research along these lines has focused on the acquisition of models of specific objects that incorporate characteristic views; the learning methods used have differed widely, but many have dealt with the selection or weighting of relevant image features. Symposium talks in this vein were presented by Pope and Lowe, Gros, Murase and Nayar, Cook et al., and others.

Improving feature extraction. The vision community has already developed well-defined algorithms for computing many low-level features (e.g., edges, texture energy, optic flow) from images. Thus, at first glance there appears to be little benefit in using learning techniques to improve performance on this task. However, the computation of all possible features can be an expensive process, and one can use learning methods to determine *which* of many low-level features to compute (feature selection process) in given situations, and thus improve the efficiency of the inference process. At the symposium, Viola and Bhandaru et al. each presented work of this general nature. The knowledge acquired during learning can make the computation of one feature conditional on the results of other feature extractions or on more global factors, such as whether the image was taken on a cloudy or clear day. Murphy presented an approach that incorporated this latter idea.

Improving indexing. Given a set of inferred features and a set of stored models, one could in principle find all instantiations of each model, evaluate each of them in turn, and select the best ones. However, computational considerations make this impractical even for small numbers of models. Typically, vision systems use some scheme to index models in terms of low-level features, letting them generate a set of instantiated candidate models with relatively little cost. One can create such indices manually, which can be a time-consuming process, or one can use machine learning methods to generate them automatically. Better indices can lead to either reduction in retrieval costs, as in the work described by Draper, or more accurate retrieval of candidate models, as in Beis and Lowe's work on indexing for occluded objects. Other presentations on this topic were given by Mann and Jepson and by Remagnino, Bober, and Kittler.

Not all symposium talks focused on learning knowledge for use in vision. A few researchers took a different approach, using the output of a vision system as training data for learning on an entirely distinct performance task. Ikeuchi, Mitchell, and Salganicoff presented work along these lines that used visual feedback as the source of information in learning for robotic planning and control. For example, Ikeuchi's technique acquires assembly plans by observing a video sequence of a human operator performing the assembly task. These efforts serve to illustrate that, although learning clearly holds promise for improving the performance of vision systems, we should also remember that vision can provide useful input for machine learning.

3. Challenges to Machine Learning from Computer Vision

The goal of improving the performance of computer vision systems presents a number of challenges to the field of machine learning. Here we outline the more prominent issues and note examples of progress represented at the symposium.

[2] Some systems, such as the eye-tracking program that Pomerleau and Baluja presented at the meeting, may be difficult to decompose along these lines.

Structured representations. Algorithms for machine learning are typically designed to operate with flat attribute-value formalisms. Yet most research on computer vision assumes that knowledge about an image has inherent structure, and thus represents information at multiple levels of aggregation. For instance, many vision systems make inferences about edges, corners, surface regions, object components, and the spatial relations among those components. Recent work on inductive logic programming within the machine learning community only partly addresses this issue, and such methods are not yet robust. In the symposium, papers by Sengupta and Boyer, Pope and Lowe, and Conklin dealt directly with learning over structured descriptions, and additional attempts to adapt learning techniques to visual domains should produce more work in this area.

Handling uncertainty. Many learning algorithms represent acquired knowledge in logical terms that either match or mismatch a given instance, and even more assume that the features of instances are certain. However, many aspects of visual domains are inherently uncertain; edge detectors can give quite different results for very similar images, and variations in perspective and lighting can also introduce considerable ambiguity. Some induction algorithms operate with probabilistic descriptions, and others achieve similar effects by other means, but we predict that serious attempts to use learning in vision will produce more work along these lines. Symposium papers by Segen, Pope and Lowe, and Sengupta and Boyer provided examples of this approach.

Partial information. Most work on machine learning assumes that all features are present during both training and testing. In contrast, images seldom contain all the information that would be useful in vision. For example, the object of interest in a scene may be partly occluded, and even when this does not occur, an image reveals only one side of an object. Learning researchers have adapted many of their algorithms to handle some missing information, but they seldom examine the effect of removing half of the available features, as vision tasks will force them to do. Contributions to the symposium from Beis and Lowe and from Gros began to deal with this issue, but much more work remains to be done.

Focusing attention. The performance components associated with most learning algorithms assume that information about instances falls outside the system's control, and that no costs are involved in collecting such information. One recent body of work in computer vision assumes exactly the opposite, that focusing attention is central to the processes of visual inference and recognition. Some approaches focus computational resources on useful parts of an image; others actually direct the collection of images over time. A few efforts in machine induction have attempted to learn strategies for focusing attention, but we can expect many more examples to emerge as work progresses at the intersection of machine learning and computer vision. The papers by Draper and by Remagnino et al. dealt with learning in this context.

Incremental learning. Typical machine learning techniques process training instances in a nonincremental manner, using statistical regularities to direct search through the space of hypotheses. Although one can collect images for processing of this sort, a more natural approach attempts to learn incrementally from images as they are encountered. For instance, the vision system for an autonomous vehicle would encounter images over time, and it might attempt to learn from each one as it becomes available. There exists some work on incremental induction, but we predict that serious attention to vision tasks will increase efforts in this direction. Pope and Lowe, Conklin, and Segen presented papers at the symposium on this topic.

Learning with many classes. The majority of supervised induction techniques have been designed to handle only a few classes, and even unsupervised methods are seldom tested on domains with many different categories. Most existing vision systems also deal with small numbers of classes, but the field's long-term goal requires the ability to discriminate among thousands of different object classes. Before machine learning can contribute to achieving this goal, it must develop algorithms that scale well along this dimension. Unfortunately, none of the symposium papers presented significant progress on this front.

Dealing with large spaces. Vision systems often depend on parameters that one must tune to obtain reasonable performance, and the size of the resulting parameter space can be very large. Although methods for parameter tuning have a long history within machine learning, computer vision requires more robust techniques that scale well to high-dimensional spaces. One symposium paper, by Bhanu, Lee, and Das, focused on such a parameter-tuning task.

In summary, machine learning must address a variety of issues before it can make a significant contribution to computer vision. Each of these problems has received some attention within the learning community, but a focus on visual domains would force researchers to develop more robust algorithms and evaluate them in more realistic settings.

4. Future Research on Vision and Learning

The Raleigh meeting revealed an emerging research community that has considerable energy and that has produced many promising ideas. However, it also showed an area with little common terminology, poor knowledge of related work, and not enough concern for careful evaluation. To be fair, such characteristics are typical of most young disciplines, and one should not expect significant collaborations from scientists in traditionally separate areas to blossom overnight. Nevertheless, research at the intersection of machine learning and computer vision must significantly improve the quality of its work before it can be recognized as a mature field.

One approach to raising standards is to establish requirements for an acceptable research paper on learning in vision. We believe that the typical paper should include five main features:

- *Specify the performance and learning tasks* that are the focus of the research, clearly distinguishing between the two aspects and stating each in terms of inputs and outputs.
- *Describe the representation* for both the data given to the learning system and the acquired knowledge that it generates.
- *Explain the performance and learning algorithms* in enough detail to allow reimplementation. If space allows, include pseudocode and an extended example of the system in operation.
- *Evaluate the learning algorithm* in terms of improvement on the performance task, giving experimental evidence that the system gets better with experience.
- *Place the approach in context*, discussing its relation to other work (including non-learning approaches to vision) and noting its limitations.

We believe that papers containing such features will constitute clear contributions to research on vision and learning, and that those working in this area should strive to meet these criteria.

As indicated above, we believe that experimental studies must play a central role in the evaluation of visual learning. Such studies should include one or more well-defined measures of performance that serve as dependent variables. Note that it is easy to define performance measures at the system level. The challenge for the researchers is to define measures at the algorithm level. The most obvious measures are recognition accuracy and processing time, but others are possible. Equally important, studies must measure performance on a set of test cases that are distinct from the instances used during training. Otherwise, the experiment evaluates nothing more than a system's ability to memorize the training set. Moreover, to ensure against fortuitous splitting of images into training and test sets, results should be averaged over different random partitions.

An experiment must also vary one or more independent variables to determine its effect on the dependent measures. In studies of learning and vision, there are three main types of independent factors that affect performance. The first involves the number of training cases available to the learner; plotting performance as a function of this variable gives a *learning curve*, which shows whether the system improves with experience and, if so, the rate of such improvement. A second type of independent variable concerns the learning system itself; experiments that vary this factor, sometimes called *comparative studies*, relate the learning behavior of different algorithms or measure the contribution of specific components or parameters within a given method. Finally, one can vary aspects of the domain, such as the amount of occlusion, the number of objects in images, and the number of classes being learned. Such *domain studies* are important in evaluating an algorithm's ability to scale along dimensions that make vision tasks difficult.

Clearly, one cannot study vision or learning without focusing on particular domains. One factor that has encouraged experimentation within the machine learning community has been the collection of public data sets that are available by ftp from a central site. The availability of visual data should have the same impact on the study of visual learning. Images are the most obvious type of data, but other forms of shared information are also possible. For example, the ARPA-sponsored Image Understanding Environment will provide a common protocol for exchanging and distributing vision data at many different levels of abstraction.[3] This should reduce the overhead required for comparative studies of different vision learning techniques.

In summary, it is clear that computer vision and machine learning have much to contribute to each other. The Fall Symposium on Machine Learning and Computer Vision brought together a community of researchers who are excited about the great potential of the area, but it also revealed that the area has a long road to travel before realizing that potential. Nevertheless, the symposium laid a good foundation for future work on this promising topic, and we hope that future meetings will produce more significant results on vision and learning.

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: AAAI Press.

Bhandaru, M., Draper, B., & Lesser, V. (1993). Learning image to symbol conversion. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 6–9). Raleigh, NC: AAAI Press.

Bhanu, B., Lee, S., & Das, S. (1993). Adaptive image segmentation using multi-objective evaluation and hybrid search methods. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 30–34). Raleigh, NC: AAAI Press.

[3] Readers interested in the Image Understanding Environment should send electronic mail to Joe Mundy at MUNDY@CRD.GE.COM.

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, NC: 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.

Draper, B. (1993). Learning from the schema learning system. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 75–79). Raleigh, NC: AAAI Press.

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, NC: AAAI Press.

Ikeuchi, K., & Kang, S. B. (1993). Assembly plan from observation. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 115–119). Raleigh, NC: AAAI Press.

Mann, R., & Jepson, A. (1993). Non-accidental features in learning. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 55–59). Raleigh, NC: AAAI Press.

Murase, H., & Nayar, S. K. (1993). Learning and recognition of 3D objects from brightness images. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 25–29). Raleigh, NC: AAAI Press.

Murphy, R. R. (1993). Learning to eliminate background effects in object recognition. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 144–147). Raleigh, NC: AAAI Press.

Pomerleau, D. A., & Baluja, S. (1993). Nonintrusive gaze tracking using artificial neural networks. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 153–156). Raleigh, NC: AAAI Press.

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.

Remagnino, P., Bober, M., & Kittler, J. (1993). Learning about a scene using an active vision system. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 45–49). Raleigh, NC: AAAI Press.

Salganicoff, M. (1993). A vision-based learning method for pushing manipulation. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 90–94). Raleigh, NC: 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: 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.

Viola, P. A. (1993). Feature-based recognition of objects. *Working Notes of the AAAI Fall Symposium on Machine Learning in Computer Vision* (pp. 60–64). Raleigh, NC: AAAI Press.