Adaptive User Interfaces: Papers from the 2000 AAAI Spring Symposium, 74–79
Technical Report SS-00-01, Menlo Park, CA: AAAI Press

An Adaptive Architecture for Hierarchical Vision Systems

Marcus A. Maloof
Department of Computer Science
Georgetown University
Washington, DC 20057-1232
maloof@cs.georgetown.edu
http://www.cs.georgetown.edu/~maloof

Abstract

We describe a machine learning architecture for hierarchical vision systems. These vision systems work by successively grouping visual constructs at one level, selecting the most promising, and passing them up to higher levels of processing. This continues from the pixel-level of the image to the object-model level. Traditionally, researchers have used static heuristics at each level to select the best constructs. In practice, this approach is brittle, because people have not been successful at surveying the evidence necessary for robust performance, and is static, because designers have not incorporated learning mechanisms that would let the system improve its performance with the aid of user feedback. The machine learning architecture proposed herein is an attempt to address both of these issues.

Introduction

We have devised a novel machine learning architecture that shows promise of significantly improving the accuracy of hierarchical machine vision systems. The proposed approach organizes learning and recognition modules into a hierarchy. Modules at the lower image-level survey local information, such as pixel intensities and shape features, and form simple visual constructs, which they pass to higher construct- and object-level modules. These modules higher in the hierarchy take into account more global types of information, such as spatial or geometric properties. In general, modules in the hierarchy take lower-level constructs and use perceptual grouping operators to form new constructs. To avoid computational bottlenecks, the modules evaluate new constructs and select only the most promising to pass to the next level of processing.

Traditionally, people have programmed heuristics for construct selection, and, as a result, these heuristics often lack robustness because of our inability to survey large amounts of evidence. A greater problem is that these mechanisms are static, meaning that they cannot adapt and improve once deployed in a system. People often interact directly with such systems, or at least supervise their operation, which suggests a role for interaction on the part of the user, and a role for adaptation and learning on the part of the intelligent vision system.

The proposed approach uses on-line learning algorithms, instead of static heuristics, to induce concept descriptions for selecting the most promising visual constructs at each level in the hierarchy. If a system based on this approach makes a mistake, then an operator or technician can provide feedback interactively, which is used to update the concept descriptions for selecting constructs associated with each module. This presents a challenge because we must develop strategies for correctly and efficiently propagating feedback to the modules in the hierarchy that produced the error.

Ours is the first learning approach to take advantage of the decomposition of visual objects into constituent parts that is inherent to model-based and hierarchical vision systems. We anticipate that our approach will yield systems that attain higher accuracy than systems that use a single learning process at the top of a recognition hierarchy to map features to object classes, which is characteristic of many current approaches. This work also promises a methodology for managing user feedback in hierarchies of heterogeneous learning algorithms.

Our plan is to build an experimental vision system based on the proposed learning approach for the domain of detecting blasting caps in X-ray images of airport luggage. We intend to conduct experimental studies designed to measure the improvement in accuracy of our approach versus existing approaches, and to measure over time the affects of interaction and feedback on performance. The experimental studies will include cost-sensitive learning methods, since our data sets invariably will be skewed toward the class of lesser importance, and since we do not have a precise cost analysis of errors for our domain. To assess the accuracy of the cost-sensitive learning methods, we will use Receiver Operating Characteristic (ROC) analysis and area under ROC curves as our performance metric. Statistical tests, such as Analysis of Variance and Duncan's test, will indicate whether the experimental results are statistically significant. We have preliminary results on the task of blob detection that illustrate our experimental

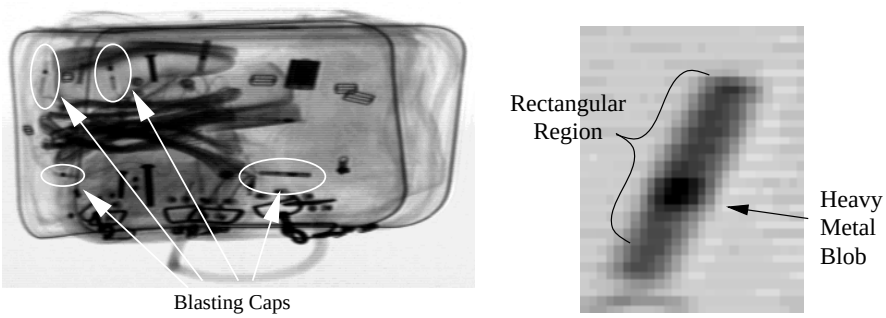

Figure 1: X-ray images of blasting caps in airport luggage. Left: A representative image from a collection of thirty. Right: A closeup of a region containing a blasting cap.

methodology, but, due to space constraints, we present these results elsewhere (Maloof 1999).

Preliminaries

To focus and ground discussion, we will use the vision domain of blasting cap detection in X-ray images since blasting caps are representative of the class of visual objects we wish to recognize. They have compositional structure and recognition requires both local and global processing. Furthermore, we have chosen to build upon Nevatia's *perceptual grouping* approach (Mohan and Nevatia 1989) for three-dimensional object recognition because it is mature and has been applied successfully to complex recognition problems, such as building detection in overhead images (Huertas and Nevatia 1988; Lin and Nevatia 1996). More importantly, the hierarchical nature of Nevatia's approach lends itself to a novel architecture for visual learning, which is the topic of this article. In the next two sections, we provide relevant details of blasting cap detection and of Nevatia's perceptual grouping approach.

Blasting Cap Detection

When we X-ray blasting caps as they might occur in luggage in an airport scenario, they appear similar to the images in figure 1. One feature useful for recognition is the low-intensity blob near the center of the object, which is produced by a concentration of heavy metal explosive near the center of the blasting cap. Another is the rectangular region surrounding the blob, which is produced by the blasting cap's metal tube. However, the mere appearance of these two regions of interest is usually not sufficient for detection, so we must also ensure that the proper spatial relationship exists between a given blob and rectangular region (or "rectangle"). As we will see in the next section, Nevatia's perceptual grouping approach provides an elegant framework for solving this recognition problem.

Hierarchical Vision Systems

The main tenet of Nevatia's perceptual grouping approach is that machines can recognize objects by repeatedly grouping low-level constructs (e.g., lines) into higher-level ones (e.g., rectangles). At the lower levels of processing, the machine surveys local information, such as a region's area or pixel intensities. As recognition proceeds up the hierarchy, higher-level reasoning processes take into account global information, such as geometric and spatial relationships. The left diagram in figure 2 shows an architecture of a hierarchical vision system for recognizing blasting caps.

Inherent to this recognition process are mechanisms at each level that select the most promising visual constructs for further processing at the higher levels, as shown in the right diagram of figure 2. Traditionally, people have manually programmed these heuristics for construct selection using methods such as constraint satisfaction networks (Mohan and Nevatia 1989) and linear classifiers (Lin and Nevatia 1996). We contend that using on-line machine learning techniques and user feedback to acquire the criteria for selecting the most promising constructs at each level will yield more robust vision systems capable of improving their accuracy[1] over time. Experimental results on a rooftop detection task support this claim (Maloof *et al.* 1998).

Statement of the Problem

Modern hierarchical machine vision systems are often brittle because, due to cognitive limitations, humans cannot adequately survey the evidence required for these systems to cope in their intended environments. Furthermore, the fact that these systems, once deployed, cannot adapt to changes in their environment contributes to their lack of robustness. Visual learning approaches (i.e., approaches that combine machine learning with computer vision) hold the potential for addressing both of these problems.

[1] The term *accuracy* is often used to mean "percent correct." However, there are different measurements of accuracy, percent correct being one (Swets 1988). Hence, we will use the term *accuracy* in the general sense, and, as we will describe, use area under an ROC curve instead of percent correct as our measure of accuracy.

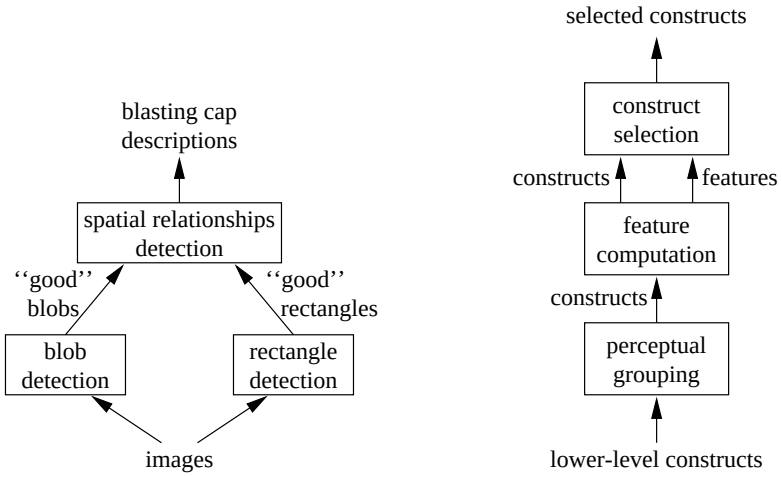

Figure 2: Hierarchical vision systems. Left: A hierarchy for recognizing blasting caps. Right: One level of the hierarchy.

Recently, there has been considerable work on visual learning approaches; however, much of this research has concentrated on single-step learning and recognition schemes in which a learning technique is applied at one point in a recognition process to, for example, map features to object classes.

Although using a learning process at the end of a vision process for the task of mapping features to object classes may improve accuracy over traditional, hand-constructed classification methods, we contend that such an approach is undesirable for one primary reason: Current work in visual learning does not take advantage of the fact that the hierarchical approach decomposes objects into simpler parts, which may make learning easier and may result in vision systems with higher accuracy.

Consequently, we propose a new visual learning architecture that tightly integrates vision and learning processes, organizing them in a hierarchy, and addresses our criticism of existing approaches. As for our research hypothesis, we anticipate that the proposed visual learning architecture, which we discuss in the next section, will significantly improve accuracy as compared to existing architectures that use a single learning step at the top level of a recognition hierarchy.

Proposed Approach

Our proposed approach involves organizing a set of learning and recognition modules into a hierarchy. The low-level modules in the hierarchy are responsible for detecting local features, such as blobs or straight lines, and for then passing these constructs to the next level in the hierarchy. As we proceed to the topmost level of the hierarchy, the object level, the modules take into account more global types of information, such as spatial or geometric information.

Since each module of the hierarchy is similar in design, although each will be configured differently depending on its task, we will ground discussion by concentrating on one module in the hierarchy, which we present in figure 3. In the next section, we present the design of an experimental system for the domain of blasting cap detection that is based on our proposed approach.

The first step is a perceptual grouping process that takes constructs from modules lower in the hierarchy and groups them to form new constructs. For example, in one level of a building detection system, such a module would group linear features into parallelograms, which potentially correspond to rooftops. In our blasting cap domain, a module may group blob regions and rectangular regions based on spatial constraints.

These newly formed constructs are then passed to a feature computation component that computes a predetermined set of attribute values for the construct. Depending on where the module resides in the hierarchy, these could be low-level attributes, such as statistics computed using the intensities of a region's pixels, or attributes that characterize the shape of the region, including area or measures of compactness. Conversely, higher-level modules would take into account more global types of information, such as geometric constraints. In our blasting cap domain, a module that groups the blob and rectangular regions might use as an attribute the distance between the centroids of the two regions.

Next, the construct selection process takes the current set of concept descriptions and uses the features computed in the previous step to classify each construct. Those constructs that are labeled as positive (and thereby selected) are then passed to the next level of processing in the hierarchy. However, we assume that at some point in the recognition process, the user, or the environment, will provide feedback on the objects that were or were not correctly identified.

When provided, the positive or negative feedback will start on-line learning processes that will use the set of misclassified constructs and their features to modify each level's current set of concept descriptions. Learn-

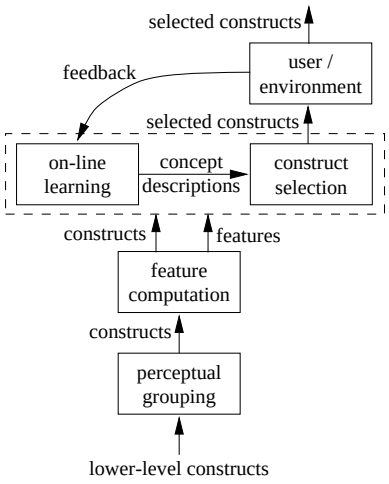

Figure 3: One level of the proposed approach to visual learning.

ing and adaptation will occur throughout the life of the vision system.

Ideally, we would like for a user to provide feedback at the highest level, the object level. Assuming that the system draws a graphical representation of recognized objects on the image, to provide feedback for false positives, we would like for the user to simply click on the graphical representation of the misidentified object. Similarly, to provide feedback for false negatives, we would like for the user to select the image region containing the unidentified object.

This approach to feedback leads to difficulties regarding credit assignment because it is difficult at the top level of the hierarchy to pinpoint the construct(s) at lower levels that led to the misclassification. The primary problem is that a low-level construct may be part of both a positive and negative object description. For example, a line corresponding to the edge of a building is a part of the building, but it is also a part of the region adjacent to the building. Presently, we are assuming that the user will identify the specific constructs at each level that led to the misclassification. However, we will postpone the specific details of this scheme until we describe our design for an experimental vision system, which we discuss in the next section.

Evaluation

To evaluate our approach, we have begun to implement an experimental visual learning system for detecting blasting caps in X-ray images of luggage, which are representative of a class of objects that we wish to recognize. Furthermore, during and after its construction, we will conduct experiments in an effort to support our research hypothesis, which we stated previously. The following sections provide details about the image set and the implementation of the experimental visual learning system.

Description of the Image Data

The image data that we will use for our inquiry consists of thirty X-ray images of suitcases containing blasting caps that appear much as they would in an airport scenario: flat with respect to the X-ray source, but rotated in the image plane. The images vary in the amount of clutter (e.g., clothes, shoes, batteries, pens) and in their planar orientation. There is enough variability in the position of the blasting caps within the bags to provide a range of difficult recognition problems. In some instances, the long axis of the blasting cap is perpendicular to the X-ray source. In others, the cap is behind opaque objects, partially occluded by various metal objects, and rotated out of the imaging plane. The left image in figure 1 shows a representative image from the collection. Each of the 8-bit grayscale images has dimensions of roughly 565×340.

As we have discussed previously, blasting caps are appropriate objects for this study because they require a hierarchical approach for recognition. The system must detect the blob and the rectangular region of the X-rayed cap, and it must then analyze the spatial relationship between these two regions, thus requiring, in some sense, a minimal hierarchy.[2]

Implementation of an Experimental Visual Learning System

To validate our proposed approach for visual learning, we plan to use the approach to construct an experimental learning system for detecting blasting caps in X-ray images of luggage. The design for this system appears in figure 4.

The experimental system will consist of three primary modules: one for detecting the blob region in the center of the blasting cap, one for detecting the rectangular region, and one for detecting the appropriate spatial relationships between the blob and the rectangular region. We have already begun work on the blob detection module and present preliminary results elsewhere (Maloof 1999).

Each module in the hierarchy will consist of the components pictured in figure 3: perceptual grouping, feature computation, on-line learning, and construct selection. The user will provide feedback at the top level of the hierarchy and, if necessary, at the lower levels as to the correctness of the selected constructs.

To implement feedback mechanisms, we envision the system presenting its results of processing as a graphical representation of the object detected in the image. To provide feedback for a false negative (i.e., an object erroneously identified as a blasting cap), the user, or some other qualified person, would select the graphical representation of the mistaken object and indicate, by

[2]Detecting rectangular regions will require processing that involves linear feature detection and grouping, but, for this study, we do not plan to use learning at these lower levels. Consequently, we are treating the detection of the rectangular region as a single vision process.

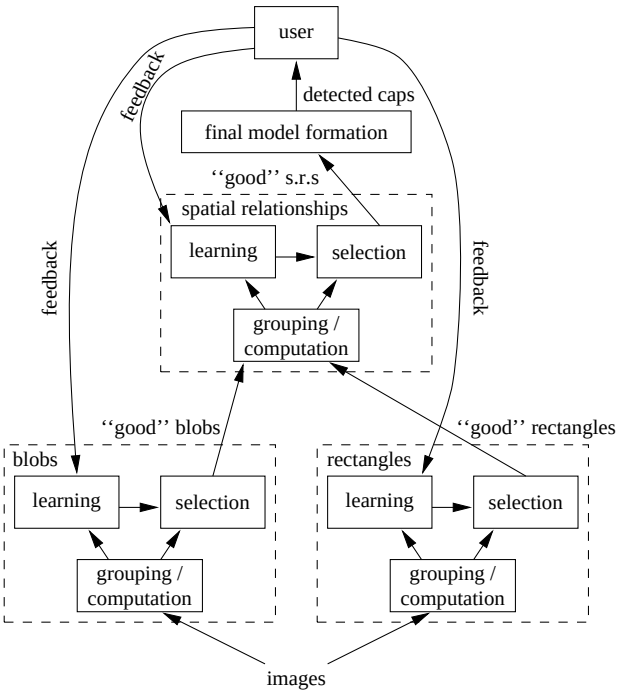

Figure 4: Design of an experimental visual learning system for blasting cap detection based on the proposed architecture.

means of a menu or a button, that the selected object was misclassified.

The system would then attempt to label all of the constructs involved in the formation of the object description as negative. However, as we have discussed, a problem arises when a construct is used in both a true positive and a false negative description. For example, a linear construct could be part of a valid blasting cap description, and it could be part of an invalid description that corresponds to a rectangular region adjacent to the real object. In this situation, we cannot simply label the linear construct as a negative example because this action could reduce the system's ability to correctly identify blasting caps.

Therefore, as the system attempts to label constructs as negative, if it discovers conflicts, then it would present each construct in question to the user. The user would decide which constructs should retain positive labels and which should get new negative labels. The process of presenting this information would start at the highest level of the hierarchy and descend through the lower levels. At each node of the hierarchy, the system would highlight all of the questionable constructs that are part of the object's final description. The user would review these constructs, identify any misclassified constructs, and proceed to the next module in the hierarchy. When the user completes this identification process, an on-line learning step will ensue that will use the newly labeled constructs as training examples to update the concept descriptions for selecting visual constructs at each level of the hierarchy. If the number of constructs is large, then the system could cluster the examples and present the most representative to the user for labeling. (Methods for further reducing the amount of required interaction will be a fascinating area of future work.)

To provide feedback for false positives (i.e., blasting caps that were not identified), the user will use the mouse to identify the image region surrounding an unidentified blasting cap. The system will retrieve all of the visual constructs that fall within this region and present them to the user. As before, the system will present the constructs at each level in turn, and the user will identify the misclassified constructs. After completing this process, the newly labeled constructs will serve as training examples to the on-line learning process which will update the concept descriptions for selecting the most promising constructs.

The user would use these procedures any time the vision system makes a mistake. And as we have indicated previously, this process of feedback and adaptation continues throughout the life of the vision system.

Bootstrapping the System

A problem with the preceding discussion is that it assumes that the system already possesses a set of reliable concept descriptions for selecting visual constructs at each of the levels. When constructing the vision system, we will configure each module in the hierarchy by starting at the image level and moving up the hierarchy to the object level. Once we have created the routines at a given level for vision processing, perceptual grouping, feature computation, we will use batch learning to induce the concept descriptions necessary for selecting the most promising constructs.

When we have configured the modules at one level and empirically validated performance, we can begin work at the next level, using the lower-level modules as input. After completing the hierarchy, we may need to further refine the system's accuracy before deployment in the intended environment. To accomplish this, we will employ the methodology described in the previous section that uses on-line learning and feedback to incrementally refine the concept descriptions for selecting promising visual constructs at each of the levels of the hierarchy until the system achieves an acceptable level of performance.

Related Work

Our hierarchies share some interesting similarities and dissimilarities with other hierarchical or tree-structured approaches. For example, our hierarchies are structurally similar to decision trees (Quinlan 1990), which have been used for visual learning tasks (e.g., Draper 1997), but our approach processes information bottom-up rather than top-down.

Hierarchical mixtures of experts (Jordan and Jacobs 1994) is tree-structured approach for supervised learning that processes information bottom-up, but each

training example is presented to each node in the hierarchy. In our approach, each node in the hierarchy learns from its own set of training examples.

Hierarchical reinforcement learning (e.g., Dietterich To appear) is also similar, but the semantics of the hierarchies are different. A reinforcement learning task involves learning a sequence of actions. Hierarchical reinforcement learning composes this sequence of actions into subtasks. The temporal order of the actions and subtasks is important, and time in these graphs moves from left to right.

With our hierarchies, actions on the same level have no temporal order and could be executed in parallel. Further, all of a node's children must make their decisions (i.e., identify their visual constructs) before it can make its decision. Time, therefore, moves from bottom to top. In this regard, they are similar to tree-structured Bayesian inference networks (e.g., Neapolitan 1990).

However, as with reinforcement hierarchies, the semantics of our networks is different from that of Bayesian inference networks. Links in the latter type of network denote causal relationships, whereas in our network, links represent an evidentiary relationship (i.e., a good blob is evidence that a proper spatial relationship may exist).

Conclusion

The proposed work is the next logical step in a long-term investigation of mechanisms for interactive, adaptive software systems. Ours is a departure from current approaches that use a single learning algorithm at one point in a recognition process. As we have discussed, we propose to study multiple, possibly heterogeneous learning algorithms organized in a hierarchical manner. The need for such learning architectures arises in machine vision, especially for situations in which we must build systems to recognize 3-D objects that have compositional structure and that require both low-level, local processing and high-level, global processing for recognition. Building detection and blasting cap detection, our chosen domain, are two examples of this class of problems. And, as we saw in the previous sections, we have proposed a novel visual learning approach that stands to significantly improve the accuracy of hierarchical vision systems.

We anticipate that the success of this study will impact three fronts. Scientifically, this project seeks to tightly integrate vision and learning processes. If successful, the proposed research will yield vision systems with higher recognition rates and, in the longer term, will provide opportunities to adapt and improve the vision processes themselves. On a broader level, studying feedback mechanisms for learning algorithms organized in a hierarchy has scientific merit that extends beyond the context of vision problems and is important for a wide range of applications, such as computer intrusion detection and intelligent agents (e.g., agents that prioritize one's e-mail queue).

Acknowledgements. Earlier work with Pat Langley, Tom Binford, and Ram Nevatia provided the basis for many of the ideas developed in this paper. This research was conducted in the Department of Computer Science at Georgetown University in Washington, DC. This work was supported by the department and by a grant from the Graduate School of Arts and Sciences at Georgetown.

References

Dietterich, T. To appear. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research.*

Draper, B. 1997. Learning control strategies for object recognition. In Ikeuchi, K., and Veloso, M., eds., *Symbolic visual learning.* New York, NY: Oxford University Press. 49–76.

Huertas, A., and Nevatia, R. 1988. Detecting buildings in aerial images. *Computer Vision, Graphics, and Image Processing* 41(2):131–152.

Jordan, M., and Jacobs, R. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6:181–214.

Lin, C., and Nevatia, R. 1996. Building detection and description from monocular aerial images. In *Proceedings of the Image Understanding Workshop*, 461–468. San Francisco, CA: Morgan Kaufmann.

Maloof, M.; Langley, P.; Binford, T.; and Nevatia, R. 1998. Generalizing over aspect and location for rooftop detection. In *Proceedings of the Fourth IEEE Workshop on Applications of Computer Vision (WACV '98)*, 194–199. Los Alamitos, CA: IEEE Press.

Maloof, M. 1999. Design of a machine learning architecture for hierarchical vision systems. Technical Report CS-99-1, Department of Computer Science, Georgetown University, Wasington, DC.

Mohan, R., and Nevatia, R. 1989. Using perceptual organization to extract 3-D structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(11):1121–1139.

Neapolitan, R. 1990. *Probabilistic reasoning in expert systems: theory and algorithms.* New York, NY: John Wiley.

Quinlan, J. 1990. Learning logical definitions from relations. *Machine Learning* 5:239–266.

Swets, J. 1988. Measuring the accuracy of diagnostic systems. *Science* 240:1285–1293.