Unsupervised Learning of Biologically Plausible Object Recognition Strategies

Bruce A. Draper and Kyungim Baek

Colorado State University, Fort Collins CO 80523, USA
{draper,baek}@cs.colostate.edu

Abstract. Recent psychological and neurological evidence suggests that biological object recognition is a process of matching sensed images to stored iconic memories. This paper presents a partial implementation of (our interpretation of) Kosslyn's biological vision model, with a control system added to it. We then show how reinforcement learning can be used to control and optimize recognition in an unsupervised learning mode, where the result of image matching is used as the reward signal to optimize earlier stages of processing.

1 Introduction

Traditionally, object recognition has been thought of as a multi-stage process, in which every stage produces successively more abstract representations of the image. For example, Marr proposed a sequence of representations with images, edges, $2\frac{1}{2}D$ sketch, and a 3D surfaces [7]. Recently, however, biological theories of human perception have suggested that objects are stored as iconic memories [6], implying that object recognition is a process of matching new images to previously stored images. If so, object recognition is a process of transformation and image matching rather than abstraction and model matching.

Even according to the iconic recognition theory, object recognition remains a multi-stage process. In the iconic approach, it is not reasonable to assume that the target object is alone in the field of view, or that the viewer has previously observed the object from every possible perspective. As a result, there is still a need for focus of attention (e.g. segmentation) and transformation/registration steps prior to matching.

At the same time, there is another psychologically-inspired tradition in computational models of biological vision that suggests that object recognition should not be viewed as a single, hard-wired circuit. Instead, there are many techniques for recognizing objects, and biological systems select among them based on context and properties of the target object. Arbib first described this using a slide-box metaphor in 1970's [1]. Since then, Ulllman's visual routines [17] and the "purposive vision" approach [2] could be viewed as updated versions of the same basic idea.

This paper tries to synthesize Kosslyn's iconic recognition theory with the purposive approach. In particular, it builds on the author's previous work on

using reinforcement learning to acquire purposive, multi-stage object recognition strategies [4]. Unlike in previous work, however, this time we assume that memory is iconic and that object recognition is therefore an image matching task. We then use the match score between the stored image (memory) and the sensed image (input) as a reward signal for optimizing the recognition process.

In this way, we build a prototype of an iconic recognition system that automatically develops specialized processes for recognizing common objects. In this way, we not only combine two biologically motivated theories of biological perception, we also avoid the need for hand-labeled training images that limiting our earlier work. Instead, we have an unsupervised rather than supervised system for learning object recognition strategies.

At the moment, our prototype system is extremely simple. This paper presents a demonstration in which the image match score is used as a reward signal and fed back to earlier stages of processing. The goal is to show that this reward signal can be used to make object recognition more efficient. More sophisticated versions, with hopefully higher over-all recognition rates, are under development.

2 Previous Work

Kosslyn has argued since at least 1977 that visual memories are stored essentially as images [5]. This idea received critical neurological support in 1982, when researchers were able to show a retinotopic map of a previously viewed stimulus stored in the striate cortex of a monkey [14]. Since then, the psychological and neurological evidence for iconic memories has grown (see chapter 1 of [6] for an opinionated overview). At the same time, SPECT and PET studies now show that these iconic memories are active during recognition as well as memory tasks [6]. The biological evidence for image matching as a component of biological recognition systems is therefore very strong.

More specifically, Kosslyn posits a two-stage recognition process for human perception, where the second stage performs image transformation and image matching. Although he does not call the first stage "focus of attention", this is essentially what he describes. He proposed pre-attentive mechanisms that extract nonaccidental properties and further suggests that these nonaccidental properties serve as cues to trigger image matching. Beyond hardwired, pre-attentive features, Kosslyn also suggests that biological systems learn object-specific features called signals (see [6] pp.114-115) to predict the appearance of object instances and that these cues are used as focus of attention mechanisms.

Kosslyn's description of image transformation and image matching is imprecise. Much of his discussion is concerned with image transformations, since the image of an object may appear at any position, scale or rotation angle on the retina. He argues that our stored memories of images can be adjusted "to cover different sizes, locations, and orientations" [6], although he never gives a mathematical description of the class of allowable transforms. Tootell's image [14] suggests, however, that at least 2D perspective transformations should be allowed, if not non-linear warping functions. With regard to the matching process

itself, Kosslyn doubts that a template-like image is fully generated and then compared to the input. Without further explanation, one is left with a vague commitment to an image matching process that is somehow more flexible than simple correlation.

It should be noted that Kosslyn's model is not the only model of biological object recognition. For example, Rao and Ballard propose a biological model in which images are transformed into vectors of filter responses, and then matched to vectors representing specific object classes [9]. Biederman presents a model based on pre-attentive figure completion [3]. Nonetheless, neither of these theories explain the strikingly iconic nature of Tootell's striate cortex image [14].

In addition, there has been a great deal of interest recently in PCA-based appearance matching [8,16]. While powerful, appearance matching should be understood as one possible technique for the image matching stage of object recognition. In fact, appearance matching is a computationally efficient technique for approximating the effect of correlating one test image to a large set of stored model images (see [15] Chapter 10 for a succinct mathematical review). Thus appearance matching is a potentially useful component of an object recognition system, but it is not by itself a model of biological object recognition.

Previously, we developed a system that learns control strategies for object recognition from training samples. The system, called ADORE, formalized the object recognition control problem as a Markov decision problem, and used reinforcement learning to develop nearly optimal control policies for recognizing houses in aerial images [4]. Unfortunately, the use of this system in practice has been hindered by the need to provide large numbers of hand-labeled training images.

Kosslyn's theory suggests that hand-labeled training images may not be necessary. By using the result of image matching as the training signal, we can move ADORE into an unsupervised learning mode. This removes the need to hand-label training instances, thereby creating an unsupervised system that learns and refines its recognition policies as it experiences the world.

3 The Proposed System

Fig. 1 shows our computational model of biological vision. It can be interpreted as adding a control and learning component to our instantiation of Kosslyn's model [6]. At an abstract level, it has three recognition modules: focus of attention, transformation/registration, and matching[1]. At a more detailed level, each module has multiple implementations and parameters that allow it to be tuned to particular object classes or contexts. For example, the final image matching stage can be implemented many ways. If the goal is to match an image of a specific object instance against a single template image in memory, then image correlation remains the simplest and most reliable comparison method. Alternatively, if the goal is to match an image against a set of closely related templates

[1] The focus of attention module has both a pre-attentive and attentive component, although we will not concentrate on this distinction in this paper.

in memory (e.g. instances of the same object under different lighting conditions), then principle components analysis (PCA) may be more efficient. In more extreme cases, if the goal is to find an object that may appear in many different colors (e.g. automobiles), then a mutual information measure may be more effective [18], while a chi-squared histogram comparison may be most appropriate for certain highly textured objects such as trees [12].

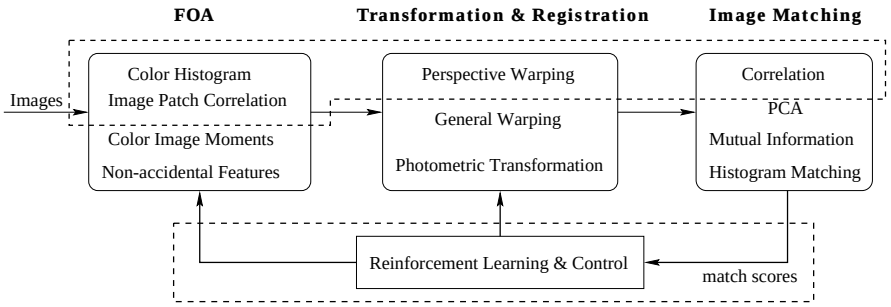

Fig. 1. The proposed system architecture.

In other words, the top-level modules in Fig. 1 represent roles in the recognition process that must be executed in sequence. Each role, however, can be filled by a variety of techniques, depending on the scene and object type. In this way, we address Kosslyn's need for a flexible image matching system, and at the same time incorporate all of Kosslyn's suggestion for focus of attention mechanisms [6].

However, the presence of options within the modules creates a control problem: which technique(s) should be applied to any given input? This is a dynamic control decision, since the choice of technique may be a function of properties of the input image. More importantly, it involves delayed rewards, since the consequences of using a particular FOA or transformation technique are not known until after image matching. We therefore use a reinforcement learning module to generate control policies for optimizing recognition strategies based on feedback from the image matching module (see Fig. 1).

4 The Implemented System

The dashed lines in Fig. 1 outline the parts of the system that have been implemented so far. It includes two technique for FOA module and one technique for each of the others: color histogram matching and image patch correlation for focus of attention; matching four image points to four image points for perspective image transformation and registration; and correlation for image matching. While this is clearly a very limited subset of the full model in terms of its object recognition ability, our goal at this point is to test the utility of the image match score as a reinforcement signal, not to recognize objects robustly.

Currently, the system is "primed" to look for a specific object by a user who provides a sample image of the target. This sample image is then used as the template for image matching. The user also provides the locations of four distinctive intensity surface patches within the template image, for use by our (primitive) focus of attention mechanism. The FOA mechanism then extracts templates at nine different scales from each location, producing nine sets of four surface patches[2].

When an image is presented to the system at run-time, a local color histogram matching algorithm is applied for pre-attentive FOA. If the control module decides that the resulting hypothesis is good enough to proceed further, the FOA system uses a rotation-free correlation algorithm to match the surface patches to the new image. (As described in [10], the rotation-free correlation algorithm allows us to find an object at any orientation without doing multiple correlations.) The result is nine sets of four points each, one set for each scale. Thus the focus of attention mechanism produces nine point set hypotheses for the image transformation step to consider.

Under the control of the reinforcement learning module, the image transformation module selects one of these nine hypotheses to pursue. It then uses the four point correspondences between the input image and model template to compute the perspective transformation that registers the input image to the template. In principle, this compensates not only for changes in translation, rotation and scale, but also for small perspective distortions if the target object is approximately planar.

After computing the image transformation, the system has a third control decision to make. If the transformation is sensible, it will apply the transformation to the input image and proceed to image matching. On the other hand, if the selected set of point matches was in error the resulting image transformation may not make sense. In this case, the control system has the option to reject the current transformation hypothesis, rather than to proceed onto image matching.

The final image matching step is trivial. Since the transformed input image and the object template are aligned, simple image correlation generates the reward signal. In general, if we have found and transformed the object correctly, we expect a greater than 0.5 correlation.

4.1 Optimization through Unsupervised Learning

The proposed system casts object recognition as a unsupervised learning task. Users prime the system by providing a sample image of the target object and the location of unique appearance patches. The system then processes images, rewarding itself for high correlation scores and penalizing itself for low scores. In this way, the system optimizes performance for any given template.

To learn control strategies, the system models object recognition as a reinforcement learning problem. As shown in Fig. 2, the state space of the system has four major "states": two for pre-attentive/attentive FOA hypotheses, one

[2] Each set of four points includes patches at a single scale.

for image transformation hypotheses, and one for image matching hypotheses. The algorithms in Fig. 1 are the actions that move the system from one state to another and/or generate rewards, and the system learns control policies that select which action to apply in each state for the given task.

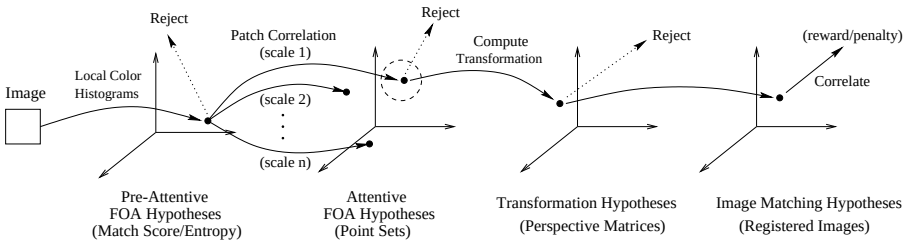

Fig. 2. The iconic view of state spaces and actions. At each state space, the control system decides whether to proceed further or reject current hypothesis.

Two necessary refinements complicate this simple system description. The first is that the four states are actually state *spaces*. For each type of hypothesis, the system has a set of features that measure properties of that type of hypothesis. For example, for the FOA point set hypotheses the system defines nine features: the correlation scores for each point (4 features); the difference between the best and second best correlation scores for each point (to test for uniqueness; 4 more features); and a binary geometric feature based on the positions of the points that tests for reflections[3].

The control policy therefore maps points in these four feature spaces onto actions. In particular, every action takes one type of hypothesis as input, and the control policy learns a Q function for every action that maps points in the feature space of the input data onto expected future rewards. At run-time, these Q functions are evaluated on the features associated with hypotheses, and the action with the highest expected reward is selected and run.

The second refinement is that some actions may return multiple hypotheses. For example, our current attentive FOA algorithm returns nine different point sets corresponding to nine different scales. When this happens, the control system must select the best state/action pair to execute next, where each hypothesis is one state. Again, this is done by selecting the maximum expected reward.

With these refinements, we have a reinforcement learning system defined over continuous state spaces. This implies that we need function approximation techniques to learn the Q functions. Currently we are using backpropagation neural networks for this, although there is some evidence that memory-based function approximation techniques may provide faster convergence [11] and we will experiment with other techniques in the future. Given a function approxi-

[3] An image reflection would imply looking at the back of the surface, which we assume to be impossible.

Fig. 3. A *cgw_tree* magazine (left), a *cgw_watch* magazine (middle), and a *wired* magazine image (right).

mation technique, the system can be trained as usual using either $TD(\lambda)$ [13] or Q-Learning [19].

5 Experimental Results

As mentioned earlier, we do not claim at this point to have a robust object recognition system, since much of Fig. 1 remains unimplemented. Instead, the goal of the experiments is to test whether the image match score can be used as a reinforcement signal to make object recognition more efficient. To this end, we consider a "control-free" baseline system that exhaustively applies every routines in Fig. 1, and then selects the maximum correlation score. We then compare this to the strategy learned by reinforcement learning. While the reinforcement learning system obviously cannot create a higher correlation score than exhaustive search, the ideal is that the controlled system would produce nearly as high correlation scores while executing far fewer procedures.

The experiment was performed with a set of color images of magazines against a cluttered background. The dataset has 50 original images with different viewing angles and variations in illumination and scale. There are three types of magazine images in the dataset: 30 *cgw_tree* images, 10 *cgw_watch* images, and 10 images containing the *wired* magazine. Fig. 3 shows these three types of magazines. Each original image was scaled to 14 different resolutions at scales ranging from 0.6 to 1.5 times the original. As a result, the dataset contains 700 images. We define a good sample to be any image that produces a maximum correlation score of 0.5 or higher under exhaustive search.

The first row of Table 1 shows the number of good and bad samples in each dataset. Table 1 also contains several measures of the performance of our control system. The number of rejected samples indicates the number of false negatives for good samples and the number of true negatives for bad samples. Therefore, it is better to have small values for the former and large values for the latter. According to the values in the table, the control system works well on rejecting bad samples for all three cases. The true negative samples classified by the system include all the images without target object and all the false positive samples have final matching scores less that 0.5, which means that, eventually,

Table 1. Results obtained by the policy learned without backtracking.

	CGW_TREE		CGW_WATCH		WIRED	
	good	bad	good	bad	good	bad
# of samples	129	571	31	669	59	641
# of rejected samples	25(19.4%)	550(96.3%)	11(35.5%)	668(99.9%)	20(33.9%)	640(99.8%)
operation_count/sample	2.92	1.97	2.94	1.33	2.85	1.54
#of optimal prediction	53(41.1%)		15(48.4%)		22(37.3%)	
average prediction/optimal	0.936343		0.954375		0.956907	

the system will not consider those samples as positive samples. Therefore, we can say that the control system generates no false positives.

When it comes to the false negatives, however, the control system had a somewhat harder time. One of the reasons is inaccurate predictions by the neural network trained on pre-attentive FOA hypothesis. Once the mistake is made, there is no way to recover from it. Also, even one weak image patch out of four can easily confuse the selection of point matches and the resulting transformed image gets low match score. This is happened most of the times in *cgw_watch* and *wired* cases. These concerns lead us to the need for defining more distinctive feature set.

As a point of reference for efficiency of object recognition, there are few enough control options in the current system to exhaustively pursue all options on all hypotheses for any given image, and then select the maximum match score. This approach executes 28 procedures per image, as opposed to less than three procedures per image on average for *cgw_tree* and even less that two on average for *cgw_watch* and *wired* under the control of the reinforcement learning system. On the other hand, the average reward (i.e. match score) generated by the reinforcement learning system is 94% of the optimal reward generated by exhaustive search for *cgw_tree*, 95% of optimal for *cgw_watch*, and 96% of optimal for *wired*. Thus there is a trade-off: slightly lower rewards in return for more than 93% savings in cost.

We also noticed that the neural networks trained to select point matches are less accurate than the networks trained to decide whether to reject or match transformation hypothesis. This creates an opportunity to introduce another refinement to the reinforcement learning system: since we are controlling a computational (rather than physical) process, it is possible to backtrack and "undo" previous decisions. In particular, if the system chooses the wrong point set, it may notice its mistake as soon as it measures the features of the resulting transformation hypothesis. The system can then abandon the transformation hypothesis and backtrack to select another point set.

Table 2 shows the results with backtracking. The additional operation per sample is less than one on average, and it greatly reduced the false negative rates with no significant increase in the false positive rates. With backtracking, the false negative rate drops from 19.4% to 7.0% on *cgw_tree*, from 35.5% to 19.4% on *cgw_watch*, and from 33.9% to 10.2% on *wired* case, while the largest increase

in the false positive rates is only 0.9%. It is open to interpretation, however, whether backtracking could be part of a model of biological vision.

Table 2. Results obtained by the policy learned with backtracking.

	CGW_TREE		CGW_WATCH		WIRED	
	good	bad	good	bad	good	bad
# of samples	129	571	31	669	59	641
# of rejected samples	9(7.0%)	545(95.4%)	6(19.4%)	668(99.9%)	6(10.2%)	640(99.8%)
operation_count/sample	3.72	2.15	3.35	2.11	3.53	1.55
#of optimal prediction	69(53.5%)		17(54.8%)		33(58.9%)	
average prediction/optimal	0.963196		0.939729		0.975222	

6 Conclusions and Future Work

We have described a system that learns control strategies for object recognition tasks through unsupervised learning. The basic recognition process of the system is biologically motivated, and uses reinforcement learning to refine the system's overall performance. We tested the system for three different target objects on a set of color images with differences in viewing angle, object location and scale, background, and the amount of perspective distortion. The learned control strategies were within 94% of optimal for the worst case and 98% of optimal for the best case.

In the future, we would like to complete our implementation of the broader system outlined in Fig. 1. For example, histogram correlation and mutual information can also be used to match images, thereby allowing the system to recognize a broader range of objects. Even more improvements can be made in the focus of attention module. As the set of procedures grows, there will be more actions that can be applied to every type of intermediate data. How accurately can we predict the expected rewards of these actions is, therefore, one of the most important issues in the system implementation. To increase the prediction power, we need to define more distinctive features than those used at the moment.

References

1. M. Arbib. *The Metaphorical Brain: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory.* Wiley Interscience, New York, 1972.
2. J. Aloimonos. Purposive and Qualitative Active Vision. *IUW,* pp.816-828, Sept. 1990.
3. I. Biederman and E. Cooper. Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology,* **23**:393-419

4. B. A. Draper, J. Bins, and K. Baek. ADORE: Adaptive Object Recognition. *International Conference on Vision Systems*, Las Palmas de Gran Canaria, Spain, 1999.
5. S. K. Kosslyn and S. P. Schwartz. A Simulation of Visual Imagery. *Cognitive Science*, **1**:265–295, 1977.
6. S. K. Kosslyn. *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge, MA, 1994.
7. D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information,* W.H. Freeman & Co., San Francisco. 1982.
8. H. Murase and S. K. Nayar. Visual Learning and Recognition of 3-D Objects from Appearance. *International Journal of Computer Vision*, **14**:5–24, 1995.
9. R. P. N. Rao and D. Ballard. An Active Vision Architecture based on Iconic Representations. *Artificial Intelligence*, 78:461–505, 1995.
10. S. Ravela, *et al.*. Tracking Object Motion Across Aspect Changes for Augmented Reality. *Image Understanding Workshop*, Palm Springs, CA, 1996.
11. J. Santamaria, R. Sutton, A. Ram. "Experiments with Reinforcement Learning in Problems with Continuous State and Action Spaces", *Adaptive Behavior* **6**(2):163-217, 1998.
12. B. Schiele and J. L. Crowley. *Recognition without Correspondence using Multidimensional Receptive Field Histograms*. MIT Media Laboratory, Cambridge, MA, 1997.
13. R. Sutton. Learning to Predict by the Models of Temporal Differences. *Machine Learning*, **3**(9):9–44, 1988.
14. R. B. H. Tootell, *et al.*. Deoxyglucose Analysis of Retinotopic Organization in Primate Striate Cortex. *Science*, **218**:902–904, 1982.
15. E. Trucco and A. Verri.*Introductory Techniques for 3-D Computer Vision,* Prentice Hall, Upper Saddle River, NJ., 1998.
16. M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, **3**(1):71–86, 1991.
17. S. Ullman. Visual Routines. *Cognition*, **18**:97–156, 1984.
18. P. Viola and W. M. Wells. Alignment by Maximization of Mutual Information. *ICCV*, Cambridge, MA, 1995.
19. C. Watkins. *Learning from Delayed Rewards*. Ph.D. thesis, Cambridge University, 1989.