From Knowledge Bases to Markov Models to PCA

Bruce A. Draper
Department of Computer Science
Colorado State University
Fort Collins, CO 80523 U.S.A.
draper@cs.colostate.edu

Introduction

The study of control in computer vision is necessitated by the simple observation that few if any significant vision systems are implemented as one-step mappings. Instead, most systems are multi-stage processes that map images onto high-level interpretations, either in the form of labels, 3D models, or both. The steps involved change from system to system. Common examples include selective attention, stereo and motion analysis, feature extraction (edges, corners, lines, Gabor responses), segmentation, image registration, camera resection (pose determination), appearance matching, geometric matching, graph matching, and constraint matching. The control problem is to put these steps together, either concurrently, sequentially or recurrently, to create an effective and efficient vision system.

What makes control in computer vision more than a scheduling and information fusion problem is the *multiple vision systems* (MVS) hypothesis. MVS is based on theories of human vision that suggest it is not a single, monolithic system, but is instead composed of multiple, anatomically distinct pathways. It hypothesizes that artificial vision systems should be built the same way; as a collection of interacting subsystems, each serving a unique visual task and interpreting images with its own mechanisms.

From a control perspective, MVS implies that there is not one control problem in computer vision, but many. Every subsystem selects basic visual processes and integrates them to achieve a specific goal. For example, geometric techniques might be used to recognize rigid objects such as buildings, while appearance-based techniques are used to draw subtle distinctions among human faces, and color and motion are exploited to recognize natural objects such as trees or rivers.

The theory and practice of how these control problems are solved is the one common thread among recognition tasks. Although each visual task is solved uniquely, the principles for finding optimal (or at least good) control policies are common to all of them.

This talk quickly summarizes the multiple vision system hypothesis in humans. It then reviews twenty years of research in the control of computer vision, with an admitted bias toward the author's own work. It argues that there has been a progression from model-free approaches to probabilistic models in low dimensional spaces, toward current

research in high-dimensional probabilistic control models.

Cognitive Models

The multiple vision systems hypothesis has deep roots in the psychology and cybernetics literatures. Since complex computer vision systems did not appear until the 1970's, however, we will leave older references for another forum.

In the 1970's and early 80's, many computer vision researchers were influenced by Arbib's schema theory [3, 4]. Schema theory had three major components. First, it emphasized that the role of perception is to enable action through the *action-perception cycle*. Second, it hypothesized an underlying set of core visual abilities. Third and most importantly, it hypothesized that these core abilities were combined into special-purpose visual schemas, each of which enabled a particular (physical or cognitive) action.

In the 1980's, scheme theory as a whole lost popularity, although most of its components were bolstered. For example, Ungeleider and Mishkin provided the first anatomical evidence for distinct pathways with their description of the ventral ("what") and dorsal ("where") pathways [25]. Ullman further developed the idea of core visual abilities in his *visual routines* [24]. Finally, in the late 1980's, the purposive and animate vision paradigms reinforced the ideas behind the action-perception cycle [1, 5].

In the 1990's, Milner and Goodale collected and organized evidence supporting the thesis that primate vision is composed of multiple special-purpose pathways [17]. Their argument has subsequently been supported by brain imaging studies, showing different loci of activation for different tasks [19, 20]. Just as important, Milner and Goodale show that the evidence from behavioral and lesion studies suggests that pathways are differentiated to serve (physical or cognitive) actions.

Control in Computer Vision

Schema theory and its subsequent refinements form the background for models of control in computer vision. In the 1970's and 80's, vision researchers followed the lead of their colleagues in AI and built expert systems for computer vision. Some of these were rule-based systems [16, 21]; others were blackboard systems [2, 18] or semantic networks [12, 13]. These systems shared a methodology in which designers encoded common sense knowledge about objects and their relations in a form that could be used to control the vision process. My own work on the Schema System (named after Arbib's theory) clearly fits this description [10].

Although not referenced much today, these systems had some positive points. They were flexible enough to make use of many kinds of contextual and relational information, and to integrate it with other information sources (e.g. color, shape, texture, and *a priori* knowledge). This made them broadly applicable. At a time when the available visual routines were still very weak, knowledge-directed systems as a class were able to interpret many types of images, including aerial images, outdoor ground level images, and indoor office

scenes[1]. Furthermore, because they acted according to rules from a knowledge base, there actions on any given image could be easily explained to even a non-technical user.

At the same time, knowledge-directed systems had serious flaws that eventually halted their development (although see [7]). The most fundamental is that they were *ad-hoc*. Probably as a result, knowledge-directed systems were brittle. Systems that were developed using a particular set of images would often fail on sets of similar (but not identical) images. They were also expensive to build, because of the time and effort required to construct the knowledge bases, and difficult to port from one domain to another. For a more detailed discussion of these and related failings, see [11].

In retrospect, knowledge-directed vision systems generally applied fine-grained but model-free approaches to visual control. By fine grained, I mean that they used a large number of relatively simple visual routines. In some cases, visual routines were simple enough to be embedded in if-then rules. More often, they were embedded in knowledge sources (to use the blackboard terminology), but these knowledge sources were rarely complex. Edge detection and image segmentation routines were among the most sophisticated knowledge sources.

Control was model-free in the sense that the control process itself was not formally modeled (although it was often monitored and informally reasoned about). For example, the Schema System included hand-crafted procedures that exploited what we knew about the domain. They also accessed the current state of the interpretation, for example to resolve conflicts. The control process itself, however, was not modeled as a Markov process or a Bayesian evidential process or by any other formal mechanism.

Another limitation, not considered at the time, was that these systems reasoned in low-dimensional spaces. They made their control decisions based on a small number of features computed over images or other intermediate representations (e.g. edge images or image regions). At the time, however, it was implicitly assumed that the goal of visual processing was to reduce complex images to comparatively simple features and labels, so this limitation was rarely if ever noted.

After these early efforts, most researchers shied away from building systems that interpreted images in unconstrained or loosely constrained domains. Instead, most research focused on narrower topics, such as improving specific visual routines or recognizing objects in very narrow domains. (Obviously, there were exceptions.) Research in control and computer vision became less popular.

Nonetheless, in the 1990's, a second generation of visual control systems emerged. These were generally better motivated if less ambitious than their predecessors. In particular, these systems explicitly modeled the control process itself, typically as either a Bayes net or Markov model, and used domain independent learning mechanisms to infer control strategies.

[1] No single system had this breadth, however.

Bayesian networks model image interpretation as an evidential process, in which visual routines gather observable measurements about a scene. A Bayesian net combines these measurements to infer probabilities for various (non-observable) interpretations of a scene. Control is implemented as a greedy algorithm that selects the visual routine with the highest utility at every time step, where utility is a trade-off between cost and the expected change in hypothesis probabilities. For a good example of a Bayes net system, see Rimey and Brown [23].

Bayes nets alleviate many of the problems with earlier knowledge-based systems. They are well grounded in probability theory. They should be robust to minor changes in imaging conditions (although this has not been well tested). The algorithms that infer control decisions can be easily ported across domains. On the down side, they are still expensive to build because the network must be carefully constructed from domain knowledge, and geometric and temporal information can be difficult to encode.

I took the other approach and modeled vision as a Markov process [9] (so did others; see [22]). In this view, image interpretation is not unlike an automatic programming problem. At any given moment, there is a state of the existing interpretation. The goal is to select the visual routine with the greatest expected future reward (as before, trading off certainty and cost). Control is again a greedy process, where this time the choice is based on the expected long-term reward.

Markov models have many of the same advantages and disadvantages as Bayes nets. Both are well grounded in probability theory. Both should be robust to minor changes in imaging conditions (an untested hypothesis). Both should be portable. To my knowledge, ADORE[2] is still the only knowledge-based visual control system to be ported across domains, and it is based on Markov models (see [8]). Markov models have the advantage that they are easier and therefore cheaper to build, because they do not rely on a hand-crafted dependency network. On the other hand, they require a training set of labeled images to learn the expected future reward functions.

Future Direction: High-dimensional Control Systems

Although better than their predecessors, second generation control systems like Bayes nets and Markov models are still limited. They have not yet created systems capable of recognizing a wide variety of objects in unconstrained or loosely constrained domains. Instead, they have created systems that robustly interpret scenes from limited domains, such as images of table settings [23] or aerial views of industrial buildings [9] or tightly controlled office scenes [8]. We have to do better.

I believe that part of the problem is the dimensionality of the control decision space. Bayes nets and Markov models do a good job of selecting actions based on the information they are given. The problem is, we don't give them enough information.

[2] ADORE: *Ad*aptive *O*bject *Re*cognition

Let me motivate this claim with an example. ADORE is an object recognition system with a Markov-based control system. As mentioned earlier, it has been used to identify industrial buildings in aerial images and objects in office scenes. At run-time, its control decisions are based on what it knows about the current state of the interpretation, as well as what it has previously learned about the visual routines and the domain. The state of the current interpretation is represented by one of several types of visual data; for example, intensity images, probability images, image regions, active contours, or sets of line segments [9]. The run-time control decisions, however, are not actually based on the current state of the interpretation. They are based instead on a small set of features that describe the current data, rather than the data itself.

In theory, this makes ADORE a partially observable Markov decision process (POMDP). In practice, it means that ADORE depends heavily on the quality of its features. For example, we have run informal tests where ADORE is given no features at all. Unable to distinguish one image from another, it learns the best static control policy, i.e. the policy with the highest average reward over the training set, which it applies to all images. This is theoretically correct, but not very useful. Alternatively, when we run ADORE with the best features we know of for a domain, it outperforms all static control policies.

Unfortunately, this suggests that we are once again applying informal control policies under the guise of a theoretically sound system. Feature quality determines the quality of the image interpretation, but the features are selected heuristically. The same would be true of a Bayes net, where both the network structure and the observable features are heuristic.

Despite being a Markov-based system, ADORE's control policy allows backtracking. We can estimate how often it is correct in a control decision by seeing how often and where it backtracks. What we observe tells us something interesting about our features. When ADORE backtracks it is almost always the first control decision that is undone. This is because it is harder to represent an image by a small set of heuristic features than a higher level construct such as an active contour. It is easier to construct meaningful features for more abstract and focused hypotheses.

In order to build robust yet general systems, we will need to build systems that can make control decisions based on better information. Logically there are two approaches. We can either (1) automatically learn a small number of highly descriptive features from training data, or (2) learn to make control decisions in higher dimensional spaces. The logical extreme of option 2 is to abolish the role of features altogether, and make control decisions based directly on the data. Unfortunately, in the case of images this is a very high dimensional space, and training instances are usually limited. An intermediate position is to build systems that use projection techniques such as principle component analysis (PCA [15]) or independent component analysis (ICA [14]) to reduce the dimensionality of the

data, while also learning to make control decisions in higher dimensional spaces.

Bulitko et. al are in the early stages of trying to build a system (called MR ADORE[3]) that does exactly that [6]. It uses no user-defined features. Instead, PCA compresses the data, and then nearest neighbor techniques and neural networks are used to learn expected future reward functions from the training set over this compressed space.

MR ADORE is still in the early stages of development, but already it can learn control strategies that outperform all possible static control policies, given a domain and a set of visual routines. We are initially applying it to the problem of recognizing trees in aerial images of forests, but nothing about MR ADORE is domain specific. The goal is to build a system that makes better control decisions because it has better information, while still operating within a Markov model framework.

Conclusion

Vision is not one problem, but many. According to cognitive theories, what should bind the many special-purpose vision systems together are a common set of visual routines, and a common mechanism for inferring task-specific control policies. The first generation of control systems for computer vision exploited common libraries of visual routines, but used heuristic methods to control them. The second generation of control systems combined visual routines into task-specific systems using general purpose and well understood mechanisms (Bayes nets and Markov models).

Second generation control technology has been successful at building effective and efficient recognition systems in very limited domains. I believe that part of this success, however, has come from heuristically selected features. In order to build more broadly applicable systems, we need to avoid relying on heuristically selected, domain specific features.

Our approach is to borrow a page from appearance-based matching methods, and use projection methods to compress and characterize the data, and high-dimensional function approximation techniques to learn expected reward functions. Other approaches would be to avoid data compression altogether, or to automatically learn domain-specific image features.

Time limitations have forced me to focus on one aspect of visual control in this talk, namely the progression from model-free approaches to low-dimensional control models to higher dimensional control models. Two other aspects of visual control deserve to be mentioned, however.

1. Biological models suggest that visual systems should be divided more by task than by domain or object. Most of us divide our system by domains for convenience, however, because we have access to images from multiple domains. More research needs to be done on a truly task-oriented division of labor.

2. The ideal size (complexity) of visual routines is still unclear. Smaller

[3] MR ADORE: Multi-Resolution ADORE

visual routines provide better opportunities for control, but may waste resources.

References

1. Aliomonos, J. *Purposive and Qualitative Active Vision*. in *Image Understanding Workshop*. 1990. Pittsburgh, PA: Morgan Kaufman.
2. Andress, K.M. and A.C. Kak, *Evidence Accumulation & Flow of Control in a Hierarchical Spatial Reasoning System.* AI Magazine, 1988. **9**(2): p. 75-94.
3. Arbib, M.A., *The Metaphorical Brain: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory*. 1972, New York: Wiley-Interscience.
4. Arbib, M.A., *Segmentation, schemas, and cooperative computation*, in *MAA Studies in Mathematics*, S. Levin, Editor. 1978. p. 118-155.
5. Ballard, D.H., *Animate Vision.* Artificial Intelligence, 1991. **48**: p. 57-86.
6. Bulitko, V., I. Levner, and B.A. Draper. *Machine Learning for Adaptive Object Recognition*. in *Unpublished*. 2003.
7. Clement, V. and M. Thonnat, *A Knowledge-Based Approach to Integration of Image Processing.* Computer Vision, Graphics and Image Processing, 1993. **57**(2): p. 166-184.
8. Draper, B.A., U. Ahlrichs, and D. Paulus. *Adapting Object Recognition Across Domains: A Demonstration*. in *International Workshop on Vision Systems*. 2001. Vancouver, CA.
9. Draper, B.A., J. Bins, and K. Baek, *ADORE: Adaptive Object Recognition.* Videre, 2000. **1**(4): p. 86-99.
10. Draper, B.A., et al., *The Schema System.* International Journal of Computer Vision, 1989. **2**(2): p. 209-250.
11. Draper, B.A., A.R. Hanson, and E.M. Riseman, *Knowledge-Directed Vision: Control, Learning and Integration.* Proceedings of the IEEE, 1996. **84**(11): p. 1625-1637.
12. Freuder, E.C. *A Computer System for Visual Recognition using Active Knowledge Sources*. in *International Joint Conference on Artificial Intelligence*. 1977. Cambridge, MA.
13. Hwang, V.S.-S., L.S. Davis, and T. Matsuyama, *Hypothesis Integration in Image Understanding Systems.* Computer Vision, Graphics and Image Processing, 1986. **36**(2): p. 321-371.
14. Hyvärinen, A., J. Karhunen, and E. Oja, *Independent Component Analysis*. 2001, New York: John Wiley & Sons. 481.
15. Kirby, M., *Geometric Data Analysis*. 2001, New York: John Wiley & Sons. 363.
16. McKeown, D.M., W.A. Harvey, and J. McDermott, *Rule-Based Interpretation of Aerial imagery.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1985. **7**(5): p. 570-585.
17. Milner, A.D. and M.A. Goodale, *The Visual Brain in Action*. Oxford Psychology Series. 1995,

Oxford: Oxford University Press. 248.
18. Nagao, M. and T. Matsuyama, *A Structural Analysis of Complex Aerial Photographs*. 1980, New York: Plenum Press.
19. Nakamura, K., et al., *Functional delineation of the human occipito-temporal areas related to face and scene processing: a PET study.* Brain, 2000. **123**: p. 1903-1912.
20. O'Craven, K.M. and N. Kanwisher, *Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions.* Journal of Cognitive Neuroscience, 2000. **12**(6): p. 1013-1023.
21. Ohta, Y.-i., *A Region-oriented Image-analysis System by Computer*. 1980, Kyoto University: Kyoto, Japan.
22. Peng, J. and B. Bhanu, *Closed-Loop Object Recognition Using Reinforcement Learning.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998. **20**(2): p. 139-154.
23. Rimey, R.D. and C.M. Brown, *Control of Selective Perception using Bayes Nets and Decision Theory.* International Journal of Computer Vision, 1994. **12**: p. 173-207.
24. Ullman, S., *Visual Routines.* Cognition, 1984. **18**: p. 97-106.
25. Ungeleider, L.G. and M. Mishkin, *Two cortical visual systems*, in *Analysis of visual behavior*, D.J. Ingle, M.A. Goodale, and R.J.W. Mansfield, Editors. 1982, MIT Press: Cambridge, MA. p. 549-586.