
Open Challenges in Learning Vision Systems

Vadim Bulitko

Greg Lee

Ilya Levner

Lihong Li

Department of Computing Science

University of Alberta

Edmonton, Alberta T6G 2E8, CANADA

BULITKO@UALBERTA.CA

GREGLEE@CS.UALBERTA.CA

ILYA@CS.UALBERTA.CA

LIHONG@CS.UALBERTA.CA

Abstract

Automated image interpretation and object recognition is an important task in numerous applications ranging from video surveillance to brain tumor identification. While demands for vision systems steadily grow, manual engineering is still among the core development techniques. In this paper we discuss adaptive learning vision systems. The challenges related to task-oriented image interpretation and object categorization and recognition are illustrated with recent research efforts.

Keywords: adaptive learning vision systems, task-oriented image interpretation and object recognition, automated construction of vision systems, reinforcement learning for vision control.

1. Introduction

The field of task oriented image interpretation and object recognition is a well developed discipline with numerous practical applications. Demands for automated vision systems steadily grow and yet manual engineering is still a core development technique. Frequently the designers specify and code the flow of information through the system and then fine-tune numerous parameters (either manually or in an automated fashion). The limitations common to this approach include the following. The design cycle is lengthy and expensive: projects of two to fifteen years in duration are not uncommon [8]. Considerable amounts of human expertise are necessary both on the computer vision and the subject matter sides. The system produced is usually tuned for specific types of images. Correspondingly, it may work well in the lab settings but not necessarily in the field conditions. Adjusting the system for field conditions requires on-going involvement of computer vision and subject matter experts as well as extensive experimentation, which leads to high maintenance costs. Finally, a system developed using a certain approach for a particular task (e.g., the “valley following” technique for tree canopy

segmentation [8]) may not be effective or even applicable to a sufficiently different domain (e.g., recognition of human faces in surveillance images). Indeed, when domain constraints are incorporated in the design, porting the system can be cost-prohibitive.

In response to these challenges, various *automated* ways of constructing image interpretation systems have been explored in the last three decades [4]. Early systems, such as the Schema System [6], had control policies consisting of *ad hoc* hand-engineered rules. In the nineties the second generation of control policy-based-image interpretation systems came into existence. More than a systematic design methodology, such systems used theoretically well-founded machine learning frameworks for automatic acquisition of control strategies over a space of vision operators. Pioneering examples include a Bayesian net system [15] and a Markov decision process (MDP) based system [5].

An implementation of the latter approach, called ADaptive Object REcognition (ADORE), learned dynamic image interpretation strategies for finding buildings in aerial images. As with many vision systems, it identified objects (in this case buildings) in a multi-step process. Raw images were the initial input data, while image regions containing identified buildings constituted the final output data. In between the data could be represented as intensity images, probability images, edges, lines, curves, etc. ADORE modeled image interpretation as an MDP, where the intermediate representations were states and the vision procedures were actions. It succeeded in learning a dynamic control policy that selects the next action at each step so as to maximize the quality of the final image interpretation in two domains [5].

2. Open Challenges

We propose an open research program with the **long-term objective** of creating a domain-independent image interpretation kit that can automatically learn a particular image interpretation task provided with some training data. The learning should occur much like it happens with human image interpreters – by observing training examples rather

than via step-by-step supervisory instructions. Given this formulation of the problem, reinforcement learning methods appear suitable [17]. However, the scale of learning in vision challenges traditional reinforcement learning methods and leads to the following open questions.

[Memory] ADORE [5] used collections of artificial neural networks to generalize the action values backed-up from the explored part of the state-action space onto the entire space. In order to make this approach feasible, a set of state feature functions were *manually* designed for the specific application domain. Removing the need for human expertise, [12] explored several automatically generated as well as off-the-shelf feature sets. In particular, Principal Component Analysis (PCA) methods were used to reduce the dimensionality of the state space for action-value function approximation. These efforts have met with a mixed success leaving automated feature interpretation as an open question. Note, that feature selection for sequential decision making may be different from feature selection for extensively studied image classification tasks.

One of the reasons PCA-derived features are so successful in classifier systems is the fact that images are often registered over some critical regions (e.g., eyes in cats and dogs [1]). On the other hand, a typical aerial forest image contains numerous randomly positioned canopies resulting in a much greater inter-image variance [11]. Consequently, an interesting line of research lies with developing adaptive focus-of-attention techniques. Within the MDP-based framework such techniques can be conveniently implemented as a collection of operators that split an image into several sub-images for further processing. Then a control policy will learn to use these operators adaptively to focus on salient image regions.

[Learning & Adaptation] ADORE used a best-first control policy guided by the learned action-value function. The errors in such a function may lead the control policy along a seemingly best but actually suboptimal path all the way to a suboptimal image interpretation. Additionally, such a best-first control requires state features at *all* processing levels. Not only additional human expertise is required but also it is more difficult to extract informative features at early levels of image processing [4]. In [13] we suggested an alternative control policy that performs a full-width tree expansion up to a certain depth and only then uses the learned action-value heuristic in the best-first fashion. Consequently, the demand for features is reduced several folds and the expected image interpretation quality rises. Presently, the open questions include: identifying the optimal depth of the full-width expansion followed by best-first action-value guided policy; developing algorithms for selecting the depth adaptively on an image per image basis and comparing them to the base-line greedy and fixed depth policies in terms of image interpretation quality and running time.

In [5] a full-width fixed-depth exploration policy was used during off-line training. For each input image, all limited-length operator sequences were executed. The produced alternative image interpretations were scored based on their proximity to the expert-supplied interpretation. As longer sequences were shown to be beneficial, [11] introduced two state pruning techniques to combat the exponential explosion of the state space. The experiments show nearly an order of magnitude reduction in the computational time with no or minor losses in quality.

Despite the initial success, large operator libraries render the control policy learning task intractable. Therefore, the success of learning MDP-based vision systems depends on having a compact and domain-targeted operator library. The original ADORE [5] used a manually selected library of vision routines. The required human engineering can be reduced by replacing the custom-tailored vision operator library with an off-the-shelf library. Being domain-independent and thus fairly universal, such a library is expected to have multiple *redundant* operators for any particular domain. In [2] a machine learning solution to this problem was presented. Namely, stochastic heuristic search (simulated annealing and evolutionary algorithms) automatically produced small domain-specific operator *sub-libraries*. The heuristic fitness of an operator sub-library traded off the computational resources needed to run it and the expected image interpretation quality. Combining the strengths of filter and wrapper models, secondary machine learning acquired the heuristic fitness function *automatically*. This novel technique accelerated the training phase 16-fold while maintaining the interpretation quality of 97% of that possible with the full operator library.

In these initial experiments the parameterized operators in the library were “bundled” together. For instance, invoking `BinaryThreshold` operator actually executed the bundle of operators `BinaryThreshold(10)`, `BinaryThreshold(20)`, ..., `BinaryThreshold(200)`. This technique reduced the branching factor within the control policy from 20 to 1 but increased the running time substantially. Correspondingly, the open challenge in this line of research lies with scaling the automated sub-library selection methods to much larger parameterized libraries.

In [5, 13] training data were in the form of input images and the corresponding desired interpretations. In numerous domains manually interpreting images is an expensive and error-prone process. In particular, the error in human expert interpretation of aerial forest photographs was found to be 18-40% [7]. On the other hand, unlabeled images have zero interpretation cost and are abundant in many domains. Using unlabeled data for learning vision systems is an open line of research. Initial progress is reported in [3].

[14] is one of the first applications of ensemble methods to reinforcement learning. In using the novel RESLEV lever-

aging algorithm, we observed that *improving* the regression mean squared error (MSE) can actually *worsen* the expected policy value $E[V^\pi]$ and vice-versa. Accordingly, an open line of research is to analyze this phenomena with different ensemble learning approaches and extend RESLEV to optimize the expected policy value as opposed to the MSE.

[Categorization] Known successful efforts on learning a domain-specific optimal policy over a generic library of vision operators focused on pixel-level interpretation [5, 13]. An open line of research is to extend this work to higher-level object recognition and categorization. A particularly intriguing application would be to model biological cognitive vision (e.g., [16]).

Acknowledgements

Bruce Draper participated in the initial design stage. Russ Greiner, Omid Madani, Guanwen Zhang, Dorothy Lau, Li Cheng, Joan Fang, Terry Caelli, David H. McNabb, Rongzhou Man, Jane Hilderman, and Ken Greenway have contributed in various ways. We are grateful for the funding from the University of Alberta, NSERC, and the Alberta Ingenuity Centre for Machine Learning.

References

- [1] J. Bins and B. Draper. Feature selection from huge feature sets. In *Proceedings of International Conference on Computer Vision*, volume 2, pages 159–165, 2001.
- [2] V. Bulitko, G. Lee, and I. Levner. Evolutionary algorithms for operator selection in vision. In *Proceedings of the 7th Joint Conference on Information Sciences : the 5th International Workshop on Frontiers in Evolutionary Algorithms*, Cary, NC, 2003.
- [3] V. Bulitko, G. Lee, I. Levner, L. Li, and R. Greiner. Adaptive image interpretation : A spectrum of machine learning problems. In *Proceedings of the ICML Workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*, Washington, DC, 2003.
- [4] B. Draper. From knowledge bases to Markov models to PCA. In *Proceedings of Workshop on Computer Vision System Control Architectures*, page 8, Graz, Austria, 2003.
- [5] B. Draper, J. Bins, and K. Baek. ADORE: adaptive object recognition. *Videre*, 1(4):86–99, 2000.
- [6] B. Draper, A. Hanson, and E. Riseman. Knowledge-directed vision: Control, learning and integration. *Proceedings of the IEEE*, 84(11):1625–1637, 1996.
- [7] F.A. Gougeon. A system of individual tree crown classification on conifer stands at high spatial resolutions. In *Proceedings of the 17th Canadian Symposium on Remote Sensing*, pages 635–642, 1995.
- [8] F.A. Gougeon and D.G. Leckie. Forest information extraction from high spatial resolution images using an individual tree crown approach. Technical report, Pacific Forestry Centre, 2003.
- [9] F.H. Hsu, M.S. Campbell, and A.J.J. Hoane. Deep Blue system overview. In *Proceedings of the 9th ACM Int. Conf. on Supercomputing*, pages 240–244, 1995.
- [10] R.E. Korf. Real-time heuristic search. *Artificial Intelligence*, 42(2-3):189–211, 1990.
- [11] I. Levner. Multi resolution adaptive object recognition system: A step towards autonomous vision systems. Master’s thesis, University of Alberta, Edmonton, Alberta, September 2003.
- [12] I. Levner, V. Bulitko, L. Li, G. Lee, and R. Greiner. Automated feature extraction for object recognition. In *Proceedings of the Image and Vision Computing New Zealand conference*, Palmerston North, NZ, 2003.
- [13] I. Levner, V. Bulitko, L. Li, G. Lee, and R. Greiner. Towards automated creation of image interpretation systems. In *Proceedings of the 16th Australian Joint Conference on Artificial Intelligence*, Perth, Australia, 2003.
- [14] L. Li, V. Bulitko, R. Greiner, and I. Levner. Improving an adaptive image interpretation system by leveraging. In *Proceedings of the 8th Australian and New Zealand Conference on Intelligent Information Systems*, Sydney, Australia, 2003.
- [15] R. Rimey and C. Brown. Control of selective perception using Bayes nets and decision theory. *International Journal of Computer Vision*, 12:173–207, 1994.
- [16] M. Spetch and A. Friedman. Recognizing rotated views of objects: Interpolation versus generalization by humans and pigeons. *Psychonomic Bulletin & Review*, 10:135–140, 2003.
- [17] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2000.