Learning Robust Object Recognition Strategies

Ilya Levner, Vadim Bulitko, Lihong Li, Greg Lee, Russell Greiner
University of Alberta
Department of Computing Science
Edmonton, Alberta, T6G 2E8, CANADA
ilya|bulitko|lihong|greglee|greiner@cs.ualberta.ca

Abstract

Automated image interpretation is an important task in numerous applications ranging from security systems to natural resource inventorization based on remote-sensing. Recently, a second generation of adaptive machine-learned image interpretation systems have shown expert-level performance in several challenging domains. While demonstrating an unprecedented improvement over hand-engineered and first generation machine-learned systems in terms of cross-domain portability, and design-cycle time, such systems have yet to be rigorously tested. This paper inspects the anatomy of the state-of-the-art Multi Resolution Adaptive Object Recognition framework (MR ADORE) and presents experimental results aimed at establishing the robustness of the system to real-world image perturbations. Tested in a challenging domain of forestry, MR ADORE is shown to be robust to changes in sun angle, camera angle and training signal accuracy.

Keywords: *Adaptive and Machine Learning, Intelligent Image Processing and Computer Vision.*

1 Introduction & Related Research

Image interpretation is an important and highly challenging problem with numerous practical applications. Hand engineering an image interpretation system requires a long and expensive design cycle as well as subject matter and computer vision expertise. Furthermore, hand-engineered systems are difficult to maintain, port to other domains, and tend to perform adequately only within a narrow range of operating conditions atypical of real world scenarios. In response to the aforementioned problems, various *automated* ways of constructing image interpretation systems have been explored in the last three decades [8].

Based on the notion of "goal-directed vision" [7], a promising approach for autonomous system creation lies with treating computer vision as a control problem over a space of image processing operators. Initial systems, such as the Schema system [7], had control policies consisting of *ad-hoc*, hand-engineered rules. While presenting a systemic way of designing image interpretation systems, the approach still required a large degree of human intervention. In the 1990's 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 image processing operators. The two well-known pioneering examples are a Bayes net system [15] and a Markov decision process (MDP) based system [6].

Our research efforts have focused on automating the latter system, called ADaptive Object REcognition system (ADORE), which learned dynamic image interpretation strategies for finding buildings in aerial images [6]. 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, or curves. ADORE modelled image interpretation as a Markov decision process, where the intermediate representations were continuous state spaces, and the vision procedures were actions. The goal was to learn a dynamic control policy that selects the next action (i.e., image processing operator) at each step so as to maximize the quality of the final image interpretation.

As a pioneering system, ADORE proved that a machine learned control policy was much more adaptive that its hand-engineered counterparts by outperforming any hand-crafted sequence of operators within its library. In addition, the system was easily ported to recognize stationary (staplers, white-out, etc.) in office scenes and again was shown to outperform operator sequences designed by human domain experts [5]. In [13], such a system was used to identify individual trees from aerial images of forest plantation scenes, and was again shown to outperform *the best* static

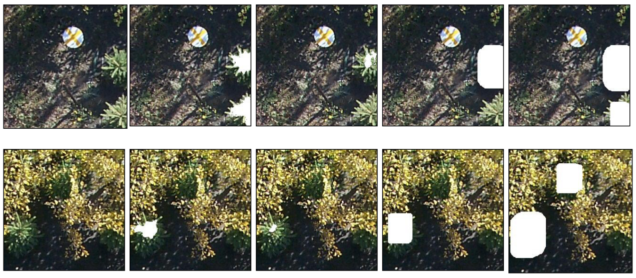


Figure 1. Each row from left to right: the original image, desired user-provided labeling (ground-truth), optimal off-line labeling, best static policy of length 4 labeling, best-first policy labeling. The adaptive policy used by MR ADORE has been observed to outperform best static policy (top row) and sometimes the human experts as well (bottom row).

sequence of operators. However, to date there have been no studies aimed at exploring the robustness of such systems. In response, this paper explores ability of the MR ADORE system to adapt to several perturbations typical of real-world scenarios. To evaluate the performance of MR ADORE, the test domain of forestry is used, which presents the following problems typical of other real-world domains. Special purpose algorithms designed to identify individual trees and their corresponding species class, have been known to be highly sensitive to the position of the camera and the angle of the sun with respect to the orientation of target objects (in this case trees) [4]. In addition, both forestry specific and general purpose vision systems, such as MR ADORE, require a set of training images (Figure 1). However, manual interpretation of aerial forest scenes is an error prone procedure [10]. Since training images are likely to contain annotation errors, the robustness of the system to labeling errors is of utmost interest. Hence, in addition to testing the robustness of the system to changes in sun angle and camera angle, this paper reports empirical evidence on robustness of the MR ADORE system to labeling errors.

The rest of the paper is organized as follows. First, we review the requirements and design of MR ADORE, in order to demonstrate the critical assumptions made and the resulting difficulties. Second, we briefly present the domain of forestry and outline the three challenges it presents. The paper then goes on to outline the experiments and their corresponding results and concludes with future research directions and closing comments.

2 MR ADORE Operation

MR ADORE starts with a Markov decision process (MDP) [16] as the basic mathematical model by casting the IPL operators as MDP **actions** and the results of their applications (i.e., data tokens) as MDP **states** (Figure 2).

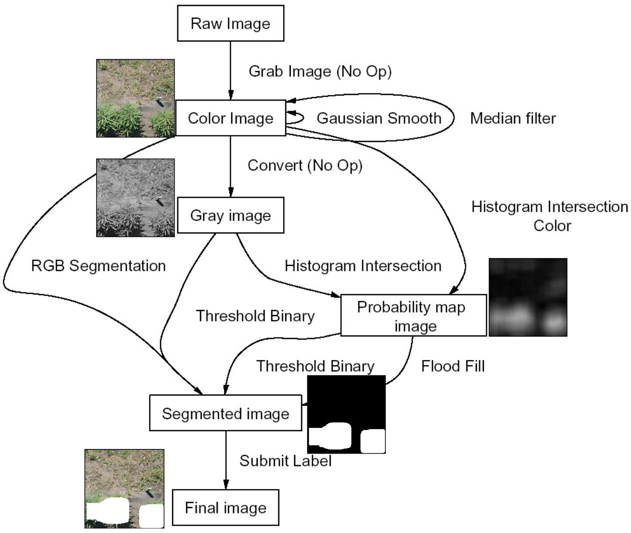


Figure 2. Partial operator graph for the domain of forest image interpretation. The nodes and the corresponding example images depict that data processing layers, which in turn describe the *type* of MDP states present with MR ADORE. The edges represent vision routines, typically ported from the Intel OpenCV and IPL libraries, that transform one state to another (i.e., the MDP actions).

First, the domain expertise is encoded in the form of training data. Each training datum consists of two images, the input image, and its user-annotated counterpart allowing the output of the system to be compared to the desired image labeling. Figure 1 (columns 1 - 2) demonstrates two training pairs (image and correct labeling) for the forestry image interpretation domain. Second, during the off-line stage the state space is explored via limited depth expansions of all training image pairs. Within a single expansion all sequences of IPL operators up to a certain user-controlled length are applied to a training image. Since training images are user-annotated with the desired output, terminal rewards can be computed based on the difference between the produced labeling and the desired labeling. System **rewards** are thus defined by creating a scoring metric that evaluates the quality of the final image interpretation with respect to the desired (used-provided) interpretation[*]. Then, dynamic programming methods [2] are used to compute the value function for the explored parts of the state space. We represent the value function as $Q : S \times A \rightarrow R$ where S is the set of states and A is the set of actions (operators). The true $Q(s, a)$ computes the maximum cumulative reward the

[*] For all experiments presented, the intersection over union scoring metric, $\frac{A \cap B}{A \cup B}$ is used. This pixel-based scoring metric computes the overlap between the set of hypothesis pixels produced by the system (A) and the set of pixels within the ground-truth image (B). If set A and B are identical then their intersection is equal to their union and the score/reward is 1. As the two sets become more and more disjoint the reward decreases, indicating that the produced hypothesis corresponds poorly to the ground-truth.

policy can expect to collect by taking action a in state s and acting optimally thereafter.

Features (f), used as **observations** by the on-line system component, represent relevant attributes extracted from the unmanageably large states (i.e., data tokens). Features make supervised machine learning methods practically feasible, which in turn are needed to extrapolate the sampled Q-values (computed by dynamic programming on the explored fraction of the state space) onto the entire space.

Finally, when presented with a novel input image, MR ADORE exploits the machine-learned heuristic value function $Q(f(s),a)$ over the abstracted state space, $f(S)$, in order to intelligently select operators from the IPL. The process terminates when the policy executes the action `Submit(`$\langle labeling \rangle$`)`, which becomes the final output of the system. Both the off-line and on-line processes are illustrated in Figure 3.

2.1 Adaptive Control Policies

The purpose of the off-line learning phase within MR ADORE is to construct an on-line control policy. While best-first policies are theoretically capable of much more flexibility than static policies, they depend crucially on (i) data token features for *all* levels and (ii) adequate amounts of training data to train the Q-functions for *all* levels. Feature selection/creation can be substantially harder for earlier data processing levels, where the data tokens exhibit less structure [8, 12]. Compounding the problem, a single user-labeled training image delivers exponentially larger numbers of training tuples, $\langle$ state, action, reward $\rangle$, at later processing levels. However, the first processing level gets the mere $|A_1|$ tuples per training image since there is only one data token (the input image itself) and $|A_1|$ actions. As a net result, best-first control policies have been shown to backtrack frequently [6] as well as produce highly suboptimal interpretations [3], due to poor decision making at the top processing layers.

Rather than making control decisions at every level based on the frequently incomplete information provided by imperfect features, the **least-commitment policies** postpone their decisions until more structured and refined data tokens are derived. That is, all operator sequences up to a predefined depth are applied and only then the machine-learned control policy is engaged to select the appropriate action. Doing so allows the control system to make decisions based on high-quality informative features, resulting in an overall increase in interpretation quality. As a side benefit, the machine learning process is greatly simplified since feature selection and value function approximation are performed for considerably fewer processing levels while benefiting from the largest amount of training data. In [13] such a policy was shown to outperform the *best* static policy.

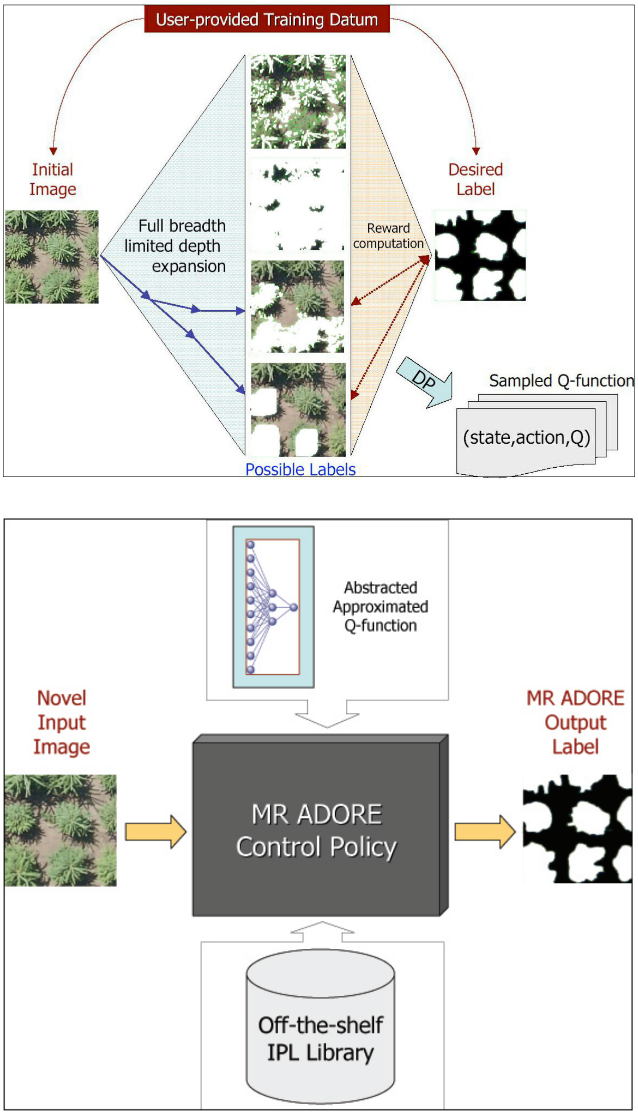


Figure 3. Top: Off-line training phase. Exploration of the state space is done by applying all possible operator sequences to a number of training images for which ground truth is provided. By comparing the interpretation resulting from an application of a sequence of operators to the ground truth, each hypothesis is assigned a quality measure (i.e., reward). The rewards are then propagated up the expansion tree in order to calculate q-values to the intermediate data tokens. Function approximators are trained on the features extracted from the data tokens produced during the exploration phase. **Bottom:** On-line operation. Using the machine leaned q-function approximators the on-line policy greedily selects the state-action pair expected to yield maximum reward. the process terminates then a interpretation hypothesis is submitted to the user.

3 Forestry Domain

Forest maps and inventories have become a critical tool for wood resource management (planting and cutting), ecosystem management and wild-life research. Canada alone has approximately 10^{11} harvestable trees, making manual forest inventorization completely infeasible. In order to automatically create forest inventories, an image interpretation system needs to measure the type (species), position, height, crown diameter, wood volume and age class for every tree in the survey area. This paper focuses on the tree labeling problem as a sub-process within a larger, complete system aimed at extracting the aforementioned parameters from individual trees. Namely, we consider the pixel-level labeling of aerial tree images and apply the adaptive object recognition approach that competes with previous state-of-the-art research methodologies and sometimes outperforms human interpreters as demonstrated in Figure 1.

A number of approaches have been proposed for extracting the aforementioned information about individual trees from aerial images. Model-free (image-based) approaches attempt to delineate individual trees, and subsequently classify each tree instance to a species class[10]. On the other hand, model-based approaches employ template matching methods to extract regions of interest and then delineate individual trees [11]. Regardless of the approach used, modern systems are highly sensitive to image variations, especially those resulting from sun angle and camera angle changes. For instance, the performance of image-based algorithms has been reported to degrade in proportion to the off-nadir view angle of a given forest scene. Likewise, off-midday sun angles negatively effect the performance of both the image-based and model based approaches [4]. Lastly, expert-annotated images, needed to train image-based, model-based, and MR ADORE-type systems, typically contain 10-40% miss-labeling errors when compared to ground-based forest surveys [9]. Thus, the use of MR ADORE is motivated by the inability of any single classical approach to perform adequately under the multitude of conditions typical of real-world scenarios within present-day remote sensing applications.

4 Empirical Evaluation

In order to test the feasibility of the MR ADORE in remote sensing applications three sets of experiments were carried out to test the robustness of the system to labeling errors, changes in overall illumination and changes in camera angle.

4.1 Labeling Errors Experiment

In Figure 1, a small pine tree was left unlabeled by the expert in row two, column two. However, the adaptive policy of MR ADORE was able to find it. To study the robustness of the off-line phase of the system, a synthetic forest scene and the corresponding ground truth were created. By spatially dilating/erroding the set of ground-truth pixels, labeling errors (false positives or false negatives respectively) were introduced into the expert labeled interpretation of the image as illustrated in Figure 4.

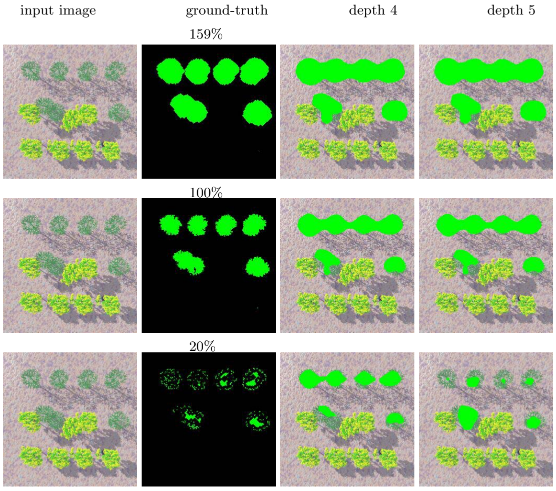


Figure 4. Column 1: Synthetic Input Image. **Column 2:** Ground-truth corresponding to the input image. Actual ground truth (100%, in row 2) is perturbed in rows 1 and 3 to simulate the error made by domain experts. **Columns 3 and 4:** Best interpretation found by MR ADORE during off-line expansion as a function of search depth.

In order to assess performance degradation resulting from labeling errors, each interpretation (columns 3 and 4 in Figure 4) was re-evaluated against the true hypothesis (row 2 of Figure 4). The results are presented in Table 1.

The initial results indicate that adding pixels (i.e, false positives) to the ground-truth interpretation is less detrimental to performance than missing target class pixels (i.e. false negatives). Table 1 shows optimal results produced with respect to the erroneous interpretation but evaluated against the error-free ground-truth image.

4.2 Sun Angle Experiments

Changes in overall scene illumination were created by simulating various time of the day as shown in Figure 5. Table 2 shows the off-line performance on the sun angle experiment. Using the SNoW algorithm [1] in conjunction with local binary patterns serving as texture features [14], the on-line policy achieved an **89%** accuracy with respect to off-line optimal during the leave-one-out cross-validation experiment. In comparison, the best static sequence could only achieve a segmentation accuracy of **74%**. In essence, using a single sequence of operators cannot produce adequate results. Rather for each position of the sun there exists a specific sequence of operators that works best. Therefore in order to achieve robust performance an adaptive policy,

Table 1. Off-line robustness to noise. When maximal sequence length is set to four, the experimental results indicate a negligible performance degradation when false positives or false negatives are added to the ground-truth image. When maximal sequence length is set to five (and six), false positives appear less detrimental to the performance of the system when compared to false negatives. Middle row (highlighted) represents the base line performance given the correct ground-truth. Erroneously adding pixels to the set of target concept pixels appears to be less detrimental than omitting pixels belonging to the target concept for longer sequence lengths.

Pixels	%	depth 4 False Reward	depth 4 Real Reward	depth 5 False Reward	depth 5 Real Reward
12000	159%	0.77	0.554	0.77	0.554
10557	140%	0.74	0.592	0.74	0.592
8915	118%	0.703	0.622	0.703	0.622
7538	100%	0.639	0.639	0.64	0.64
4635	61%	0.455	0.606	0.458	0.592
2626	35%	0.267	0.553	0.308	0.389
1510	20%	0.147	0.553	0.2	0.284

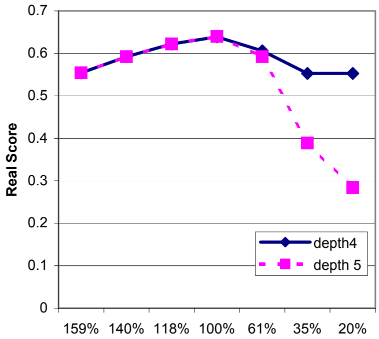
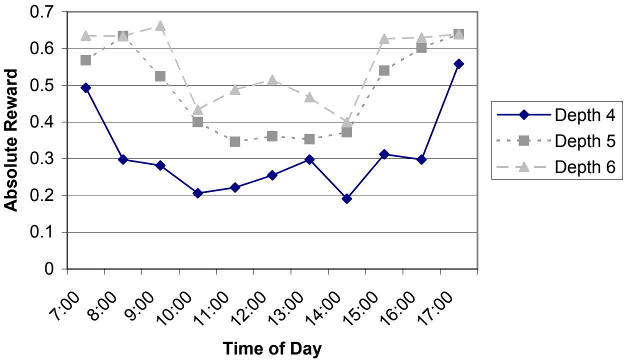


Table 2. Off-line performance on the Sun Angle experiment. The current library of operators appears to perform poorly in mid-day sun as the results indicate.



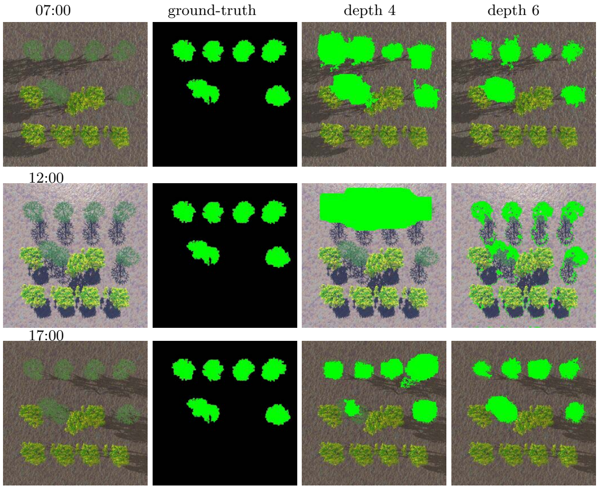


Figure 5. Various angles of the sun can produce vastly different image interpretations for a given scene. While sun positions at 07:00, 12:00, and 17:00 are shown in column one, a total of eleven images was used in the experiment containing sun angles between 07:00 and 17:00 hours at one hour increments. A static, expert labelled ground-truth image (column 2) is used to compare the hypotheses produced by MR ADORE at various search depths (columns 3 and 4).

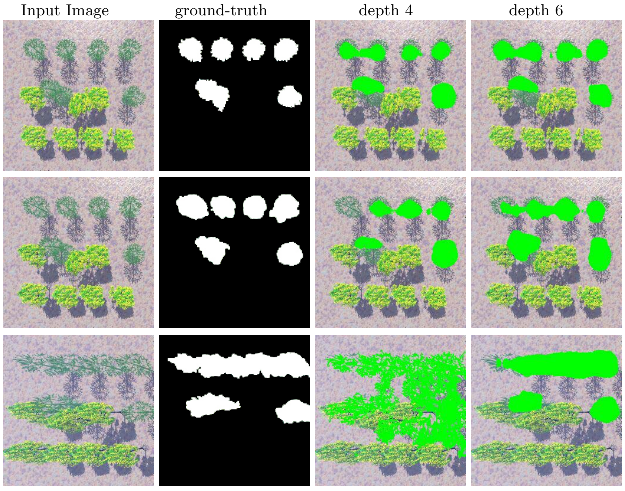


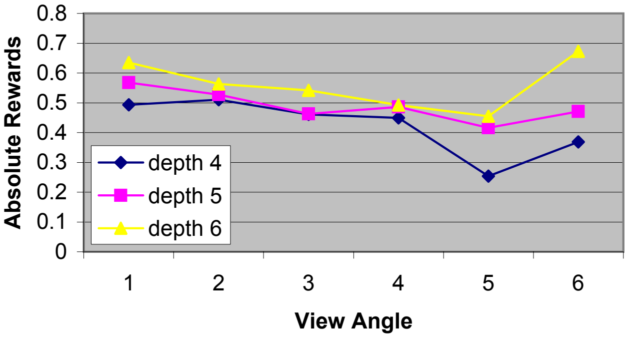
Figure 6. Various view angles produce different image interpretations for a given scene. The hypotheses produced by MR ADORE at various search depths (columns 3 and 4) are compared to ground-truth (column 2).

such as the least-commitment policy used by MR ADORE, is needed.

4.3 Camera Angle Experiments

To simulate changes in camera angle, the view of the synthetic image was altered by translating the camera

Table 3. Off-line performance on the View Angle experiment. Since the operator library uses a single template taken at nadir view angle, the interpretation quality gracefully degrades as the scene is viewed from larger off-nadir angles. At extreme view angles the best interpretation is increased by employing a significantly different sequence of operators from the previous 5 view angles.



(0,10,20,40,80 or 160 units) left from the center of the scene and then rotating (pan) it back to view the center of the scene. Illustrated in Table 3, the system demonstrates a graceful degradation in performance as camera angle becomes more and more pronounced. The on-line performance unfortunately is not as successful as in the sun angle experiment. The best static sequence was able to achieve a 77% accuracy, while the least-commitment policy was only able to attain a 70% accuracy on the leave-one-out experiment. The exact causes for the rapid degradation of the online policy are currently under investigation.

5 Conclusion

This paper presented three sets of experiments aimed at determining the robustness of the MR ADORE system. First, effects of labeling errors were explored and the off-line system component was shown to be robust with respect to false positives but not false negatives. Next, the sun angle was varied for a static scene and corresponding ground-truth. The on-line component of the system learned to select near optimal interpretations demonstrating the system's robustness to changes in overall scene illumination. Respectively, the experiments varied the image or its ground-truth. By changing the camera angle in the final experiment, both the scene and the correct labeling were varied from one image to another. While the system gracefully degraded in its off-line performance, the on-line component was unable to cope with the changes in the test images. However, the poor performance of the on-line least-commitment policy may be the result of using a limiting set of training samples (5 images). Hence two main avenues of research still remain: (i) determining the cause of poor on-line performance on the view angle experiment and (ii) determining the on-line performance for the labeling errors experiment.

Acknowledgements

Bruce Draper participated in the initial MR ADORE design stage. Funding provided by University of Alberta, NSERC, and AICML.

References

[1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *7th European Conference on Computer Vision*, volume 4, pages 113–130, Copenhagen, Denmark, 2002.

[2] A. G. Barto, S. J. Bradtke, and S. P. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72(1):81–138, 1995.

[3] V. Bulitko and I. Levner. Improving learnability of adaptive image interpretation systems. Technical report, University of Alberta, 2003.

[4] D. Culvenor. Tida: an algorithm for the delineation of tree crowns in high spatial resolution remotely sensed imagery. *Computers & Geosciences*, 28(1):33–44, 2002.

[5] B. Draper, U. Ahlrichs, and D. Paulus. Adapting object recognition across domains: A demonstration. In *Proceedings of International Conference on Vision Systems*, pages 256–267, Vancouver, B.C., 2001.

[6] B. Draper, J. Bins, and K. Baek. ADORE: adaptive object recognition. *Videre*, 1(4):86–99, 2000.

[7] B. Draper, A. Hanson, and E. Riseman. Knowledge-directed vision: Control, learning and integration. *Proceedings of the IEEE*, 84(11):1625–1637, 1996.

[8] B. A. Draper. From knowledge bases to Markov models to PCA. In *Proceedings of Workshop on Computer Vision System Control Architectures*, Graz, Austria, 2003.

[9] F. Gougeon. A crown-following approach to the automatic delineation of individual tree crowns in high spatial resolution meis images. *Canadian Journal of Remote Sensing*, 21(1):274–284, 1995.

[10] F. Gougeon and D. Leckie. Forest information extraction from high spatial resolution images using an individual tree crown approach. Technical report, Pacific Forestry Centre, 2003.

[11] M. Larsen. Individual tree top position estimation by template voting. In *Proceedings of the Fourth International Airborne Remote Sensing Conference and Exhibition*, 1999.

[12] I. Levner. Multi resolution adaptive object recognition system: A step towards autonomous vision systems. Master's thesis, Department of Computer Science, University of Alberta, 2003.

[13] I. Levner, V. Bulitko, G. Lee, L. Li, and R. Greiner. Towards automated creation of image interpretation systems. In *Australian Joint Conference on Artificial Intelligence (To appear)*, 2003.

[14] T. Ojala and M. Pietikinen. Unsupervised texture segmentation using feature distributions. *Pattern Recognition*, 32(3):477–486, 1999.

[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] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2000.