Learning on the Job: Improving Robot Perception Through Experience

Corina Gurău*, Jeffrey Hawke*, Chi Hay Tong and Ingmar Posner
Mobile Robotics Group, Oxford University, United Kingdom
{corina, jhawke, chi, ingmar}@robots.ox.ac.uk

Abstract

This paper is about robots that autonomously learn how to interpret their environment through use. Specifically, robot perception is improved with every outing through effortless human assistance. This is made possible by the fact that robots operate repeatedly in specific application domains. Our approach, which we call Experience-Based Classification (EBC), is similar in spirit to the concept of hard negative mining (HNM), but it is entirely *self-supervised*. In the context of autonomous driving we use 17km of data to show that EBC is a practical alternative to HNM, and demonstrate the advantages of experience-specific classifiers. We believe that our approach presents a fundamental shift in how robot perception is viewed, and advocate for lifelong learning systems that excel in a specific application domain instead of providing mediocre performance everywhere.

1 Introduction

Accurate situational awareness is a pivotal requirement for safe autonomous operation in complex environments. Object detection lies at the heart of this, with the conventional desire of fast, reliable performance *across* a number of workspaces. This is explicitly encouraged in the machine vision community by competitions such as the ImageNet Large Scale Visual Recognition Challenge [1].

While much progress is being made, the error rates of state-of-the-art approaches are still prohibitive, particularly for safety critical applications. We believe that the desire for generality leads to mediocre performance everywhere. Instead, we advocate for a radically different detection deployment model from the status quo. If we admit that robots will only be used for specific tasks, then we can alternatively strive for excellence in a particular application domain. This leads to three major points. First, learning a single model for object detection is insufficient for robust performance. Second, operation in dynamic environments implies the need for a robot to adapt over time. Third, a robot should make effective use of its feedback from cycles of interaction with the world. One way detectors can improve is by identifying additional relevant negative samples from the environment, whose variation is not captured in the original training data.

The standard method for obtaining relevant negative samples is known as hard negative mining (HNM) [2, 3]. In HNM, the classifier is first trained on the original training data and then used for object detection on a *labelled* dataset. False positives are identified using the ground truth labels provided and included for classifier retraining. While this environment-specific tuning provides considerable improvement over the original classifier, the labelling effort required for HNM is labour-intensive and limited to a selected set of data.

However, a key characteristic of robotics not available in general computer vision problems is that we work in specific application domains, and often traverse the same workspace over and over again. As a result, we can exploit scene context. This context – often obtained through online sensing or contained in (semantic) map priors – is commonly leveraged as a filter (e.g. [4, 5]) to

[*] C. Gurău and J. Hawke contributed equally to this work.

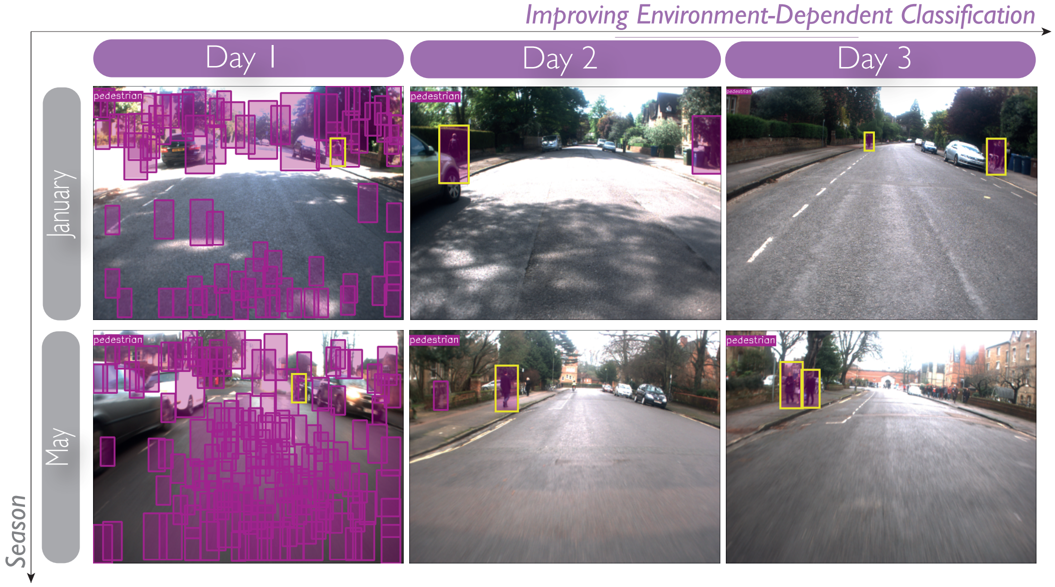


Figure 1: Images from a route in Oxford at two different times (January and May) on which we performed pedestrian detection. We demonstrate a great improvement in performance over a few EBC iterations by automatically adapting the detector to different seasons. False positives are shown in purple, while true positives have a yellow bounding box. Figure best viewed in colour.

discard detections if certain validation criteria are not met. Instead of just filtering, we also use the scene context to replace the need for human labels. Concretely, we conduct hard negative mining in a *self-supervised* manner by continuously feeding back any false positives identified by the scene context into the detector training process throughout the lifetime of the system. We call this process *Experience-Based Classification* (EBC).

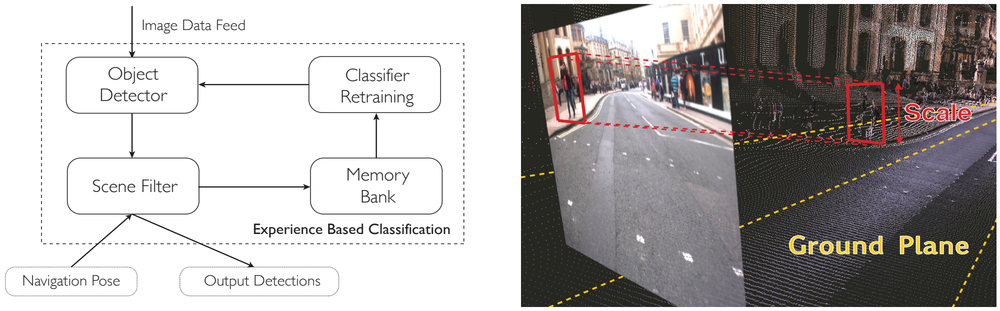
In effect, EBC automatically adapts detectors to specific environments, which permits us to easily train new detectors as desired. While this may lead to overfitting to the background encountered, we argue that this is exactly what is required in mobile robotics. EBC is a self-supervised and environment-dependent approach that is able to incorporate considerably more data without the need for human fine-tuning. We evaluate this approach for a pedestrian detection problem (see Figure 1) using 17km of urban driving data gathered in Oxford over two seasons.

2 Related Work

3D scene information has been primarily used in object detection to generate Regions of Interest (ROIs). For example, a ground plane computed from stereo imagery can provide a search space for positive detections (e.g. [4, 5]), or enforce scale [6]. Instead of generating ROIs to present to our classifier, we invert the order and apply scene information after we compute detections. While both approaches provide us with a set of valid positive classifications, this ordering also allows us to obtain a set of hard negatives that can be used for classifier improvement.

As mentioned in the introduction, the conventional approach for obtaining these hard negatives is HNM. Initially introduced by Sung and Poggio [2] as a bootstrap method for expanding the training set, Felzenszwalb et al. [3] tailored it for structural SVMs by defining 'hard' negatives as examples that are incorrectly classified or within the margin of the classifier. Instead of HNM, Henriques et al. [7] used block-circulant decomposition to train SVMs with an approximation to the set of all negative samples from a series of images. In effect, training with a vast set of negatives reduces the need to specifically mine for hard negatives. While efficient, the training remains limited by computational resources, and does not escape the core requirement of labelled data.

We share some similarities with the concept of group induction [8], where self-supervised training is performed by alternating between classifying unlabelled tracks and incorporating the most confident



(a) Our implementation of EBC which makes use of navigation information.

(b) False detections are filtered using heuristics such as scale and ground connection.

Figure 2: The EBC architecture and our scene filter implementation which employs a 3D prior. The detector improves through successive outings as it automatically adjusts to its experiences.

positive classifications in retraining. Our approach differs by the fact that we use an external signal in the form of an environmental prior to provide labels for the whole scene. This allows us to focus only on hard samples and provides a means to automatically tune to specific environments.

3 Framework Description

EBC augments the standard perception pipeline by introducing a scene filtering step after object detection, a memory bank of negative samples and classifier retraining. Our implementation of this system is depicted in Figure 2a.

3.1 Object Detector

In general terms, an object detector processes a data stream and produces detections. In this work we employ a linear SVM classifier trained on Histogram of Oriented Gradients (HOG) features [9] for pedestrian detection. Given an input image, we first compute HOG features for the entire image, and then employ a sliding window approach to obtain classification scores. Multiscale detection is performed by resizing the image and repeating the process. Finally, non-maximal suppression is used to filter out overlapping detections. The output is a set of bounding boxes which correspond to subwindows that score above a threshold, which are deemed to be positive detections.

3.2 Scene Filter

The scene filter is the core component of the EBC framework. Given a set of detections, the scene filter employs local context to filter out false positives according to strong heuristics. Accepted detections are passed on to the remainder of the perception pipeline, while rejected detections are stored in the memory bank. Since the rejected samples are detections that scored highly in the previous step, these are by definition hard negatives.

Given localisation information and a 3D scene prior, we first project the local ground plane into the image. This is used by a first filter, which rejects detections that lie outside of a ground polygon. Our second filter then projects each remaining bounding box into the 3D scene to ensure detections are of a viable scale. The application of these heuristics is illustrated in Figure 2b.

The scene filtering step should be conservative and avoid false negative errors. This is because incorrectly classified positives will lead to semantic drift [10]. For false positives, we appeal to the fact that we can use multiple filters, and that classifier performance will improve overall in retraining.

3.3 Memory Bank and Retraining

The final step of the EBC cycle augments the original training set with the rejected samples and retrains the classifier model. Since these additional negatives are obtained during operation, each subsequent training cycle further adapts the classifier to the specific environment. It should be

	North Oxford January			North Oxford May		
	Average per EBC training cycle	Total training	Total testing	Average per EBC training cycle	Total training	Total testing
Kilometers	2.11	8.45	1.99	1.35	5.40	1.01
Images	9651	38605	9155	5680	22720	4656
Minutes	8.05	32.2	7.63	4.73	18.9	3.88

Table 1: A summary of the datasets used for evaluation.

noted that data streams gathered from mobile robotic platforms tend to be spatially and temporally correlated. This can cause problems in retraining as most classifiers assume independent, identically distributed data. Subsampling may be required to avoid these issues.

4 Experimental Evaluation

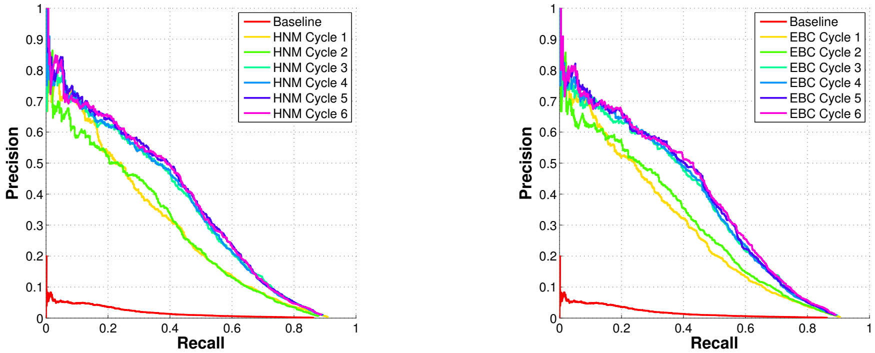
Our baseline classifier was trained using LIBLINEAR [11] and OpenCV [12] on the Daimler Pedestrian Detection Benchmark Dataset [13]. We performed ten-fold cross validation on the training set consisting of 52112 positive and 32465 negative samples.

To show the impact of environmental variation and to evaluate our self-supervised approach, we used ten different datasets gathered with a Bumblebee2 stereo camera mounted on our vehicle driving in an urban environment. These datasets consisted of two routes (over the same location) split into four runs for training and one for testing. They are referred to in this section as *North Oxford January* and *North Oxford May*. In both datasets, we used only the left image with a capture rate of 20Hz. This provided a total of 14km of unlabelled training data and 3km of labelled test data. The datasets are summarised in Table 1.

4.1 Comparison with Hard Negative Mining

We initially compared EBC to traditional HNM to see if we could achieve comparable improvement without the need for labelled data. Since HNM requires ground truth labels, we split the labelled North Oxford January test set into two parts, with 70% allocated for training and 30% for testing. Figure 3 shows the six training iterations performed for each method and the raw detector performance without scene filtering.

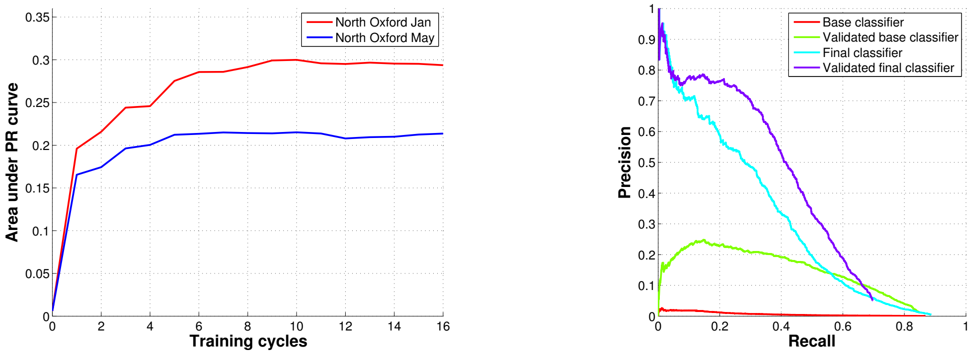
While the classifier performed well in training, we see that the baseline detector performance was poor. This was because detection employed a sliding window, which resulted in tens of thousands of windows being presented to the classifier per image with the majority being negatives. Other reasons



(a) Hard Negative Mining

(b) Experience-Based Classification

Figure 3: Precision-Recall curves comparing HNM to EBC iterations for a portion of the North Oxford January dataset. There is a dramatic improvement after the first iteration, and similar performance from both approaches. This demonstrates the value of environment-specific tuning and that EBC is a practical alternative to HNM. Figures best viewed in colour.



(a) The area under the precision-recall curves for EBC classifiers trained on one set of routes and tested on the other. The plot was split by test set, with each line in the graph indicating a different set of training data. As can be seen, a classifier performed better when trained and tested on data from the same season.

(b) Performance increase provided by the scene filter (referred to as 'validating' a classifier) when applied to both the base classifier and the final EBC classifier on the North Oxford January test dataset. Though the EBC classifier was trained using samples rejected by the scene filter, a sizeable performance boost remains.

Figure 4: Plots depicting the experience-specific performance and the continued value of the scene filter. Figures best viewed in colour.

for the performance drop were sampling resolution and change in test environment. In effect, the training set was not representative of the test data. However, we see a dramatic improvement in performance when retraining, and that EBC has similar performance to HNM. This demonstrates that using scene context is a practical alternative to using manually labelled data.

4.2 Environment-Specific Tuning

After establishing EBC as a viable alternative to HNM, we investigated the effects of tuning to specific environments. We do this by training separate classifiers for each route, and presented the datasets in the order in which they were collected. This represented different outings on successive days. Four EBC iterations were performed for each dataset. Figure 4a suggests that there are perceptual differences between the two seasons and a detector trained in January has a better performance in January rather than May. This corroborates our argument for experience-specific classifiers. Additionally, there is an improvement over time, which demonstrates the capability of the EBC framework to facilitate lifelong learning.

Finally, we note that the results so far only show the raw detector performance. Since the scene filter is already incorporated into the EBC framework, we can also validate our detections while running online if a 3D scene prior and localisation information is available. However, one may wonder if the scene filter has any effect if samples rejected by the scene prior are already used for retraining. This is answered in Figure 4b, which shows that the scene filter still provides substantial improvement for both the baseline and EBC classifiers. This is due to the fact that there remains ambiguities that cannot be resolved through appearance alone.

5 Conclusion

Though general object detection remains a noble goal, applications in robotics tend to be constrained to particular operating environments. We can exploit this fact to obtain practical systems which excel in a specific application domain. This is a major step towards reliable performance for real-world safety-critical systems. In particular, we make use of scene context to validate detections, and feed the rejected samples back to retrain the detector. This augmentation to the standard perception pipeline provides self-supervised, environment-dependent improvement over the lifetime of the system. We call this process Experience-Based Classification.

We believe that this approach offers robots the ability to iteratively improve their perceptual ability over time with effortless human assistance. In the autonomous driving context, a human simply drives the car around its operating environment while the detector learns in the background. Using urban driving data gathered over two seasons, we demonstrated the advantages of experience-specific models, constant adaptation and exploitation of environmental feedback. This was accomplished by showing that EBC provides comparable performance to HNM without the impractical requirement of manually-labelled data, and that EBC continually improves with experience.

Our experimental results show that environment-specific tuning provides benefits in performance at the cost of generality. While we manually divided the datasets in this paper, we require an automated method to determine when to train new classifiers. This may be achieved by reinitialising the training framework according to localisation estimates, but we may also find benefit in transferring classifiers to different locations with similar environmental conditions. Probabilistic topic modelling [14] offers a possible alternative. Finally, as we desire lifelong learning, we must address the issues of positive mining and semantic drift [10].

Acknowledgements

The authors gratefully acknowledge the support of this work by the European Community's Seventh Framework Programme under grant agreement no FP7-610603 (EUROPA2), by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/J012017/1 as well as by the UK Space Agency (grant number ST/L002981/1) as part of the CREST-2 Initiative.

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Miami, Florida, USA, 20-25 June 2009, pp. 248–255.

[2] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39–51, Jan. 1998.

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[4] D. M. Gavrila and S. Munder, "Multi-cue pedestrian detection and tracking from a moving vehicle," *International Journal of Computer Vision*, vol. 73, no. 1, pp. 41–59, Jun. 2007.

[5] P. Sudowe and B. Leibe, "Efficient use of geometric constraints for sliding-window object detection in video," in *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, Sophia Antipolis, France, Sep. 2011.

[6] D. Gerónimo, A. D. Sappa, D. Ponsa, and A. M. López, "2d–3d-based on-board pedestrian detection system," *Computer Vision and Image Understanding (CVIU)*, vol. 114, no. 5, pp. 583–595, May 2010.

[7] J. Henriques, J. Carreira, R. Caseiro, and J. Batista, "Beyond hard negative mining: Efficient detector learning via block-circulant decomposition," in *2013 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2013, pp. 2760–2767.

[8] A. Teichman and S. Thrun, "Group induction," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, Nov. 2013, pp. 2757–2763.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, USA, 25 June 2005, pp. 886–893.

[10] J. R. Curran, T. Murphy, and B. Scholz, "Minimising semantic drift with mutual exclusion bootstrapping," in *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, Sep. 2007, pp. 172–180.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research (JMLR)*, vol. 9, pp. 1871–1874, Aug. 2008.

[12] The OpenCV library. [Online]. Available: http://opencv.org/

[13] M. Enzweiler, A. Eigenstetter, B. Schiele, and D. Gavrila, "Multi-cue pedestrian classification with partial occlusion handling," in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2010, pp. 990–997.

[14] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.