Scene-Adaptive Human Detection with Incremental Active Learning

Ajay J. Joshi*
University of Minnesota, Twin Cities
ajay@cs.umn.edu

Fatih Porikli
Mitsubishi Electric Research Laboratories
fatih@merl.com

Abstract

In many computer vision tasks, scene changes hinder the generalization ability of trained classifiers. For instance, a human detector trained with one set of images is unlikely to perform well in different scene conditions. In this paper, we propose an incremental learning method for human detection that can take generic training data and build a new classifier adapted to the new deployment scene. Two operation modes are proposed: i) a completely autonomous mode wherein first few empty frames of video are used for adaptation, and ii) an active learning approach with user in the loop, for more challenging scenarios including situations where empty initialization frames may not exist. Results show the strength of the proposed methods for quick adaptation.

1 Introduction

Most learning methods for detecting or classifying objects in images are trained by providing annotated samples. Such methods perform well when training and testing is done in similar conditions, such as on the same scene. However, conditions often change since training and deployment can be in different locations with widely varying illumination, camera position, apparent object sizes, pose of the subject/object. The generalization ability of trained classifiers is compromised in the presence of such changes. For example, Figure 1[1] shows the output of a human detector on a test frame from the CAVIAR dataset[2], when trained using Histogram of Oriented Gradients (HOG) features [2] on a subset of the INRIA pedestrian dataset.

For each frame of test video, we employ a sliding window of 75 pixels by 50 pixels wide, with horizontal and vertical overlap of 50 and 30 pixels respectively. HOG features are extracted for each window, and the obtained vector is passed through the trained Support Vector Machine (SVM) classifier. Red boxes indicate a positive classifier output, i.e., the particular bounding box contains a human according to the trained classifier. The figure shows an extremely large number of false positive detections, primarily due to misleading texture in the upper part of the frame.

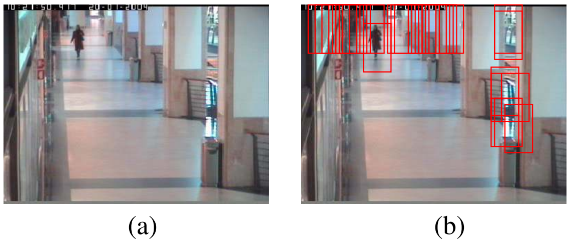


Figure 1. (a) Original frame. (b) Output of a human detector trained on the INRIA dataset.

From the previous example, it is clear that the learned human models *do not generalize well*, and heavily rely on the specifics of the training data. The background texture is never seen in training, and is consequently classified as a human in the new frame. On the other hand, we can also see that the human is detected correctly in the frame. The model therefore correctly captures some aspects of the detection problem, specifically, the appearance of the human.

Motivated by the partial correctness of the learned model, our objective is to *adapt* it to the new scene efficiently and quickly, i.e., with little or no user input. The goal is to retain the informative aspects of the previous training data, while also gathering more information about the new classification task, thereby constructing a scene-specific classifier from a generic one. In this paper, we focus on the application of human detection, which has been an area of significant recent research [2–4, 10, 12, 14]. However, note that the approach can be applied to other detection tasks as well.

Broadly, our method works via performing incremental updates by *actively selecting* new instances for training and removing old uninformative instances. The removal of training examples allows us to maintain fixed training sizes, so training is efficient, and can work on a fixed memory budget.

Consider the following setting. We have access to a

[*] Work done while at MERL.
[1] All figures best viewed in color.
[2] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.676

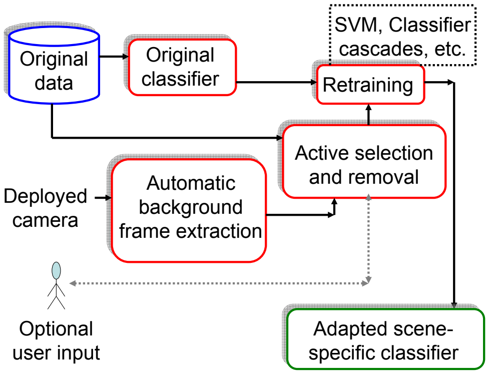

SVM, Classifier cascades, etc.
Original data
Original classifier
Retraining
Active selection and removal
Deployed camera
Automatic background frame extraction
Optional user input
Adapted scene-specific classifier

Figure 2. Block schematic of the proposed system.

large set of training examples from a standard dataset, such as INRIA pedestrian data (generic data). The objective is to deploy a classifier (human detector) on a new scene wherein we can access frames from the video sequence captured by the camera. We propose two modes of system operation. The first mode is that of semi-supervised adaptation, with user in the learning loop. The system adapts to the new scene based on a few queries made to the user (such as showing an image window and querying whether it consists of a human or not). In the second autonomous mode, the system uses generic data along with the first few frames from the new video (which does not contain any motion) to learn a scene-specific classifier. The first mode is for more challenging environments where human appearance may differ significantly or where empty frames are not available for autonomous adaptation.

2 Adaptation with user in the loop

2.1 Active learning

In this section, we give a short overview of active learning, followed by our proposed active selection method. The basic idea in active learning is to query the user for "informative examples", so as to learn faster than passive methods, i.e., with fewer training samples. Active learning has been employed in applications such as text classification [9], and more recently computer vision [5–7], including pedestrian detection [13].

The active selection process is usually iterative, wherein the algorithm queries the user for annotation on selected unlabeled examples, obtains user feedback, and appends the data to the training set. The classifiers are retrained at each round, and the process continues until the desired accuracy level is reached, or until no more training data can be obtained. Through intelligent query selection, active learning can learn good classifiers with very few training examples.

The most crucial aspect is active learning is the query selection mechanism. Measuring the potential informativeness (in terms of future classification rate) of an unlabeled data point is difficult, as is the case for query selection. Most methods use proxies such as uncertainty sampling – i.e., selecting data points on which the current model is most uncertain, or in other words the most confusing samples. For example, for a Support Vector Machine (SVM) classifier, data points closest to the classification boundary are confusing and can be potentially informative if labeled. We focus on uncertainty sampling in this work.

2.2 Incremental learning and forgetting

In this section, we employ active learning and forgetting for incremental learning. The main idea is that given a set of *generic* training images, new informative images *from the location of deployment* can be queried to the user for adding to the training set, while old uninformative images can be removed. The selection and deletion (forgetting) processes both work through active selection. For deletion, the active selection measure is inverted – i.e., examples which are least informative are selected. To our knowledge, this is the first work that employs active forgetting, and combines it with active learning for incremental classifier training.

Figure 2 shows our learning setup. Given a new scene for deployment along with generic training data, the method queries the user and adds a few training images from the new frame. This little training data allows the classifier to quickly adapt to the new scene. At the same time, old uninformative data is removed from the training set, thus limiting the total memory and training time. As the examples to be removed are selected actively, they are relatively uninformative and the removal does not significantly hurt accuracy. This process is performed iteratively, and it results in a new classifier that is scene-specific, achieved by adapting the generic training data with little human input. In general, at a new deployment location, the first few frames of video can be used for performing the update, and the resulting classifier can then be deployed on the location.

2.2.1 Uncertainty-based selection measure

The selection measure we employ is based on distance to the hyperplane of the SVM classifier. Since our application is expected to run in real time via incremental updates, fast training is crucial. As such, we use linear SVM, but the formulation allows other kernel functions as well. After an SVM is trained, it is used to estimate class membership probability values for the unlabeled images. We give a brief overview of the probability estimation technique in the following.

Probability estimation based on margins
In order to obtain estimates of class membership probability from margins, we follow the approach proposed by [8], which is a modified version of Platt's method to extract probabilistic outputs from SVM [11].

The basic idea is to approximate the class probability using a sigmoid function. Suppose that $x_i \in \mathbb{R}^n$ are the feature vectors, $y_i \in \{-1, 1\}$ are their corresponding labels, and $f(x)$ is the decision function of the SVM. The conditional probability of class membership $P(y = 1|x)$ can be approximated using

$$p(y = 1|x) = \frac{1}{1 + \exp(Af(x) + B)}, \tag{1}$$

where A and B are parameters estimated using maximum likelihood [8, 11]. We employ the LIBSVM toolbox [1] for implementation.

Consider that $\mathcal{L}$ is the set of labeled training data at any instant. Let x be the unlabeled example for which the active selection measure (confusion score) is to be computed. Let y be the true label of x, which is unknown during selection. We define the selection measure as the difference between the estimated probabilities for the two classes $|P(y = 1|\mathcal{L}) - P(y = 0|\mathcal{L})|$. Thus, active example selection from a large pool $\mathcal{A}$ can be formulated as

$$x^* = \underset{x_i \in \mathcal{A}}{\operatorname{argmin}} \quad |P(y_i = 1|\mathcal{L}) - P(y_i = 0|\mathcal{L})| \tag{2}$$

The above score represents how confused the classifier is about an unlabeled example. The lower the score, higher is the confusion (smaller margin), and the example is more likely to update the current classifier.

We can use the same confusion score above, and remove examples having the highest score, indicating that they are farthest away from the classifier boundary. For SVM classifiers, these examples *are not support vectors* and hence removing them does not change the classifier. Note that adding new examples might make the removed examples potential support vectors, however, in practice we observed that this happens extremely rarely. Consequently, example removal using the proposed measure does not hurt accuracy.

For binary classification, distance to the margin suffices. However, using estimated probability values, we can extend the above method for multi-class settings as well. The selection measure for a k-class problem is

$$x^* = \underset{x_i \in \mathcal{A}}{\operatorname{argmin}} \quad |P(y_{k_1}|\mathcal{L}) - P(y_{k_2}|\mathcal{L})|, \text{ where} \tag{3}$$
$$k_1 = \underset{i=1:k}{\operatorname{argmax}}\, P(y_i), \; k_2 = \underset{i=1:k, i \neq k_1}{\operatorname{argmax}} \; P(y_i).$$

2.3 Results with semi-supervised adaptation

The experiments are performed on an two video sequences, one from a local parking lot obtained with a surveillance camera (referred to as VID), and another from the CAVIAR dataset. The extracted image frames have a lot of confusing reflections and texture.

Since the mistakes in predictions are false positives, the primary evaluation measure used is False Positives per Window (FPPW) for each frame. All the classifiers detect the human in the frame correctly.

In this section, we compare our method of incremental active learning (IL+A) with two baselines: i) using the generic classifier on VID (called GC), and ii) incremental learning, but with Random selection of examples to add and remove, instead of active selection (IL+R). Figure 3(d) shows the achieved FPPW rate over multiple frames on VID, alongside the number of training examples used from VID. Figures 3(a)–(c) shows sample detection results on one frame of CAVIAR. The improvement of incremental active learning over the generic classifier demonstrates the importance of scene-specific training, whereas the improvement over random selection demonstrates the importance of active selection.

Note that our proposed method is not intended to replace other detection techniques, but rather *complement* them by adding incremental active learning. As such the proposed approach can be used with other existing techniques that perform well in particular domains, such as classifier cascades which have been demonstrated to give good performance in human detection applications [4].

The above method of semi-supervised adaptation can be applied to many incremental learning tasks, even when training and test conditions differ substantially and no other information is available. In many real human detection applications, more information is available. For example, at the deployment location, we might have access to a few frames of video without any human in the scene. Alternatively, motion sensors are often available in surveillance environments – these motion sensors can be used as a primary sensor to indicate the presence of a frame without a human. In these scenarios, we can adapt the generic classifier to the new scene completely autonomously as follows.

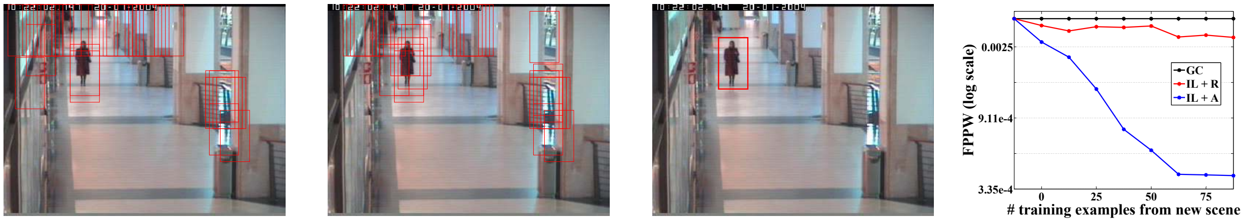
3 Autonomous adaptation

In the example of Figure 1, there are a large number of false positives we aim to eradicate, while keeping the correct detection as is. If we have access to the video frames when there is no human in the scene, we can use image windows from that frame to gather more *negative* training images.

3.1 Which negative examples to select?

The number of sliding windows per frame can be very large, because of the small window size and substantial overlap. As such, it is impractical to use all of the windows as negative training instances, from both perspectives of training set size, and retraining time. In this section, we discuss our method of example selection and removal.

The generic classifier is deployed on the empty frame, and all the windows on which it gives a *positive* response are selected for training. As the frame is known to be empty, the positive detections are essentially **misclassifications** by the generic classifier. Therefore, adding them to the training data is likely to



(a) (b) (c) (d)

Figure 3. Sample results with 75 training examples from CAVIAR, (a) Generic classifier, (b) Incremental learning with random selection, (c) Incremental learning with active selection. (d) FPPW values of different methods with varying number of training examples on VID.

change the classifier, and reduce the number of false positive detections.

3.2 Maintaining training set sizes

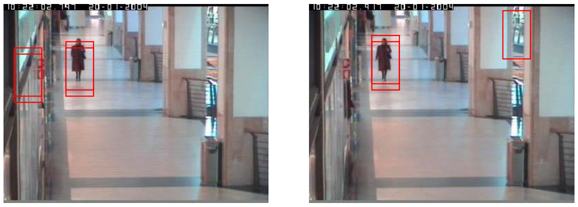
On the other hand, adding new training instances increases the size of the training set. This is undesirable in memory-constrained situations and where processing rate is critical. Therefore, we also propose to *remove* an equal number of old negative examples from the generic data. This is accomplished by using the method of the previous section, i.e., removing examples that are farthest away from the boundary.

3.3 Results with autonomous adaptation

Figure 4 shows the results of using initial background frames to extract false negatives along with the generic training data. As the number of background frames used increases, the number of false positives goes down. The method is thus a viable candidate to adapt a classifier to a new location *without the need for any human supervision.* Furthermore, in many cases scene conditions such as illumination levels change over time. One can use autonomous updates using empty frames to adapt the classifier when such changes occur, so that detection quality is consistently maintained.

4 Conclusion

We propose two approaches, one completely autonomous, and one with little user supervision to adapt generic training data to provide scene-specific detectors. The discussed methods address the important issue of quick deployment in various locations, without involving expensive operations of data collection at the location. Using incremental learning, the classifiers can combine the advantages of available generic data as well as scene-specific data.



(a) (b)

Figure 4. (a),(b) show results with using 1 and 2 background frames respectively for autonomous updates.

References

[1] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[3] T. Haga, K. Sumi, and Y. Yagi. Human detection in outdoor scene using spatio-temporal motion analysis. In *ICPR*, 2004.

[4] M. Hussein, F. Porikli, and L. Davis. A comprehensive evaluation framework and a comparative study for human detectors. *IEEE Trans. Intelligent Transportation Systems*, 10(3):417–427, 2009.

[5] P. Jain and A. Kapoor. Active learning for large multi-class problems. In *CVPR*, 2009.

[6] A. J. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009.

[7] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with Gaussian Processes for object categorization. In *ICCV*, 2007.

[8] H.-T. Lin, C.-J. Lin, and R. C. Weng. A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*, 68:267–276, 2007.

[9] A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. In *ICML*, pages 350–358, 1998.

[10] C. Orrite-Urunuela, J. M. del Rincón, J. E. Herrero-Jaraba, and G. Rogez. 2D Silhouette and 3D skeletal models for human detection and tracking. In *ICPR*, 2004.

[11] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*. MIT Press, 2000.

[12] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on Riemannian manifolds. In *CVPR*, 2007.

[13] T. Yang, J. Li, Q. Pan, C. Zhao, and Y. Zhu. Active learning based pedestrian detection in real scenes. In *ICPR*, 2006.

[14] Q. Zhu, S. Avidan, M. C. Yeh, and K. T. Cheng. Fast human detection using a cascade of histograms of oriented gradients. In *CVPR*, 2006.