Boosted Pedestrian Detector Adaptation in Specific Scenes

Puhao Ma, Lei Sun, Haizhou Ai
Computer Sci. & Tech. Dept., Tsinghua University
Beijing, 100084, China
ahz@mail.tsinghua.edu.cn

Shun Sakai
OMRON Corporation
Kusatsu-city, Shiga, 525-0025, Japan
shun_sakai@omron.co.jp

Abstract

In detector adaptation, the quality and quantity of collected online samples are of fundamental importance, yet have not been thoroughly investigated. In this paper, we present an efficient detector adaptation approach with a novel unsupervised online sample collection scheme, which can obtain sufficient aligned samples in a specific video. Unlike other methods that collect samples by only leveraging the detection confidence or track, we select aligned samples by evaluating the alignment scores using a pixel-wise Gaussian Model. Since this selection would lead to an inadequate number of positive samples, we synthesize positive samples by composing the pedestrian foreground in each aligned positive samples with the scene background at different locations. In this way, we can obtain a large number of qualified aligned positive samples encoding new scene information. With sufficient samples, we adopt a simple yet effective method to obtain an adaptive detector, which not only preserves the effective part of the offline boosted detector but also well adapts to the new scene by adding some new trained classifiers. Experiments demonstrate the efficacy of our sample collection scheme and that our approach significantly improves the performance.

1 Instruction

Object detection is an important task in computer vision and typically an object detector is trained from a large set of labeled samples. Unfortunately, the trained detector would inevitably suffer from a large performance degradation when it is applied to a new scene, especially when the scene is very different from the original training scene. Therefore, how to adapt such an offline detector to a specific novel scene, i.e. detector adaptation, becomes very important and has attracted much attention recently.

In detector adaptation, online samples from the new scene are required. Some previous methods collect online samples in a supervised [1, 2] or semi-supervised [3, 4] way, which require manual labelling. There are some approaches [5, 6, 7] attempting to obtain online samples with no supervision. For example, [5] uses the offline detector to collect samples with high detection confidences as online positive samples. In [7], they combine the offline detector and a tracker to collect online samples. However, these online samples collected using unsupervised methods may not have been well spatially aligned. In addition, only discarding unaligned samples is also infeasible, since it leads to a large decrease of the number of positive samples, thereby making them insufficient for adaptation.

With collected online samples, some approaches such as incremental learning [1, 6] which optimize the baseline detector using gradient descent methods, while others attempt to design an additional mechanism such as SVM classifier [4] and random fern classifier [7]. We believe that, it is preferable to directly obtain an adaptive detector rather than to add some post-processing schemes. Therefore, our adaptive boosted detector replace the inadequate layers with new trained classifiers while preserving the effective part of the offline detector. Compared with incremental learning [1, 6], our method is simple but effective.

In this paper, we propose an efficient boosted pedestrian detector adaptation approach with a novel unsupervised online sample collection scheme. Given a training video from a new scene, we first collect online positive samples by leveraging both detection and tracking results. Then we select aligned samples using a pixel-wise Gaussian Model. For each confident aligned sample, we segment the pedestrian foreground out by reconstructing the foreground mask from background subtraction using a dictionary of pedestrian binary masks. Then we compose the pedestrian foreground with the background centered at a different position to generate a synthesized sample. In this way, we can obtain sufficient aligned samples and encode the background information in the new scene. As for online negative samples, we collect them both around the pedestrian and uniformly in the background. With these online samples collected, we adapt the offline detector by cutting some rest strong classifiers of the offline boosted detector. The flowchart of our approach is illustrated in Fig. 1.

2 Online Sample Collection

2.1 Collecting Aligned Positive Samples

Collecting positive samples with the offline detector is very likely to lose scene specific positive samples. Meanwhile, tracking based sample collection is able to collect some lost detection samples, but it is prone to noise and may not spatially align with the ground truth. Taking all these factors into consideration, we first use a tracker[8] to collect online positive samples and then use the offline detector to further select the tracks with high confidence scores. At last, we select aligned positive samples with estimating spatially align errors within each track.

First, the offline detector is applied at a high precision setting for each frame in the video and we obtain all the detection responses $\{\boldsymbol{S_k}\}$, then the tracker is used to get all tracking sequences $\{\boldsymbol{T_i}|i = 1, 2, ..., m\}$. $\boldsymbol{T_i}$ is a tracking sequence, which is represented as $\boldsymbol{T_i} = \{t_{i,j}|j = 1, 2, ..., n\}$, where $t_{i,j}$ is the j_{th} tracking response in the tracking sequence $\boldsymbol{T_i}$. Each $t_{i,j}$ is a combination of tracking window $w_{i,j}$, frame index $f_{i,j}$ and confidence score $c_{i,j}$, denoted as $t_{i,j} =$

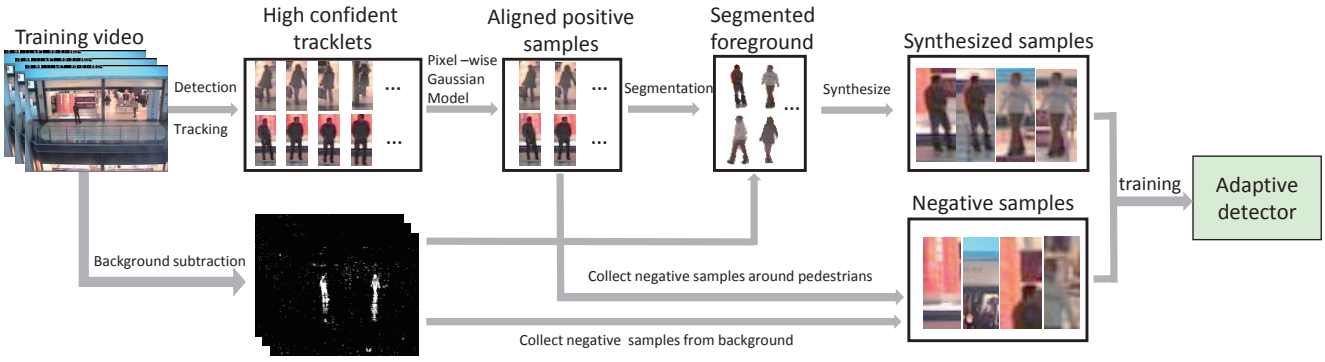


Figure 1. Overview of our approach.

$\{\omega_{i,j}, f_{i,j}, c_{i,j}\}$. We compute the overlap between the detection responses and tracking responses for each frame. The confidence score for a tracking response $t_{i,j}$ to be a positive online sample is calulated by

$$c_{i,j} = \begin{cases} 1, & O(S_k, t_{i,j}) > \theta_1 \\ \mu c_{i,j-1}, & O(S_k, t_{i,j}) \leq \theta_1 \end{cases} \quad (1)$$

where O is the overlap of the bounding boxes of S_k and $t_{i,j}$, θ_1 is a threshold value, μ is the confidence lost coefficient and $\mu < 1$. In this case, if the previous tracking response has a high confidence, it is very likely that the current sample is a confident sample.

We select some tracking responses $t_{i,j}$ with a high confidence score within each track $\boldsymbol{T}_i$, and thus form a new subsequence $\boldsymbol{T}'_i$. For each tracking subsequence $\boldsymbol{T}'_i$, we resize all the responses in the subsequence $\boldsymbol{T}'_i$ to the same size and assume that the pixels' value at the same position obeys Gaussian distribution, which results in a pixel-wise Gaussian Model for each track. It is reasonable because all the pedestrians in the same tracking sequence are the same pedestrian. Then we can estimate the probability of each sample to be an aligned sample in the tracking sequence as follow:

$$\Phi(t'_{i,j}) = \sum_{x,y} N(\omega_{i,j}(x,y) | \mu'_i(x,y), \Sigma'_i(x,y)) \quad (2)$$

where $N(\omega|\mu, \Sigma)$ is the Gaussian function, (x, y) is the position. $\mu'_i(x,y)$ and $\Sigma'_i(x,y)$ and the average value and the variance of pixels at the position (x, y).

We select the positive samples with $\Phi(t'_{i,j}) > \varepsilon$ as aligned positive samples, where ε is a threshold and

$$\varepsilon = \frac{\eta}{|\boldsymbol{T}'_i|} \sum_{t'_{i,j} \in \boldsymbol{T}'_i} \Phi(t'_{i,j}) \quad (3)$$

where η is a coefficient. In this way, we can get aligned positive samples $\boldsymbol{P_o}$.

2.2 Synthesizing Positive Samples

In a new scene, the amount of individuals in the training video is limited. And the background information is very important. In order to collect sufficient positive samples, we synthesize positive samples by composing the segmented pedestrian area of aligned positive samples with the scene background. In order to segment the pedestrian, we first apply background subtraction to get a coarse foreground mask, then refine the mask using a dictionary of pedestrian binary masks. We also obtain an edge-preserving mask by Guided Image Filtering [9]. At last, we merge two binary masks together to get a mask that has both good shape and edge. A large number of synthesized samples can be generated by merging the segmented pedestrian area with scene background at different positions.

Refining the pedestrian mask: Given a pedestrian binary mask $\boldsymbol{F}$ from the background subtraction, we refine it using a dictionary of pedestrian masks. Specifically, we follow the method in [10] for pedestrian mask dictionary learning. Given a training set of pedestrian masks $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}]$, we learn a dictionary $\boldsymbol{D} = [\boldsymbol{d_1}, ..., \boldsymbol{d_m}]$ by minimizing the reconstruction error

$$\min_{\boldsymbol{D},\boldsymbol{\alpha}} \frac{1}{n} \sum_{i=1}^{n} (\frac{1}{2} \|\boldsymbol{x_i} - \boldsymbol{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \|\boldsymbol{\alpha}_i\|_1) \quad (4)$$

where λ is a regularization parameter, and $\boldsymbol{\alpha}_i$ is a sparse coefficient. Then with the learned dictionary $\boldsymbol{D} = [\boldsymbol{d_1}, ..., \boldsymbol{d_m}]$, for each pedestrian binary mask $\boldsymbol{F}$, we can obtain a reconstruction parameter $\boldsymbol{\beta}$ by Lasso, expressed as

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{F} - \boldsymbol{D}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq s \quad (5)$$

where $\boldsymbol{\beta}$ contains sparse coefficients. s is a parameter to control sparsity. Then the reconstructed foreground mask $\boldsymbol{R}$ can be presented as $\boldsymbol{R} = \boldsymbol{D}\boldsymbol{\beta}$, which usually has a good pedestrian shape. But it often has minor flaws at the pedestrian edge. To get a more exact and smooth pedestrian edge, we use Guided Image Filtering [9] to do edge-preserving smoothing.

Obtaining edge-preserving mask: The model of Guided Image Filtering [9] assumes a local linear relationship between the guidance $\boldsymbol{I}$ and the filter output $\boldsymbol{Q}$

$$\boldsymbol{Q}_i = a_h \boldsymbol{I}_i + b_h, \forall i \in W_h \quad (6)$$

where W_h is a window centered at the pixel h. a_h, b_h are some linear coefficients depending on W_h and an input $\boldsymbol{R}$. It ensures that the output image $\boldsymbol{Q}$ can have an edge only when the guided image $\boldsymbol{I}$ has an edge.

We use the original positive sample map $\boldsymbol{I}$ as the guided image and use the reconstructed mask $\boldsymbol{R}$ as the filtering input. Even when the reconstructed mask

has an inaccurate edge, we can get the exact edge by doing edge-preserving smoothing.

Merging two masks: We merge the reconstructed mask $\boldsymbol{R}$ and the filtering mask $\boldsymbol{Q}$ to get the final aligned positive segmentation mask $\boldsymbol{A}$ by

$$\boldsymbol{A}(x,y) = min(\boldsymbol{Q}(x,y), \boldsymbol{R}(x,y)), \tag{7}$$

where (x,y) is the position, $\boldsymbol{A}(x,y), \boldsymbol{Q}(x,y), \boldsymbol{R}(x,y)$ are the pixels at the positions (x,y) for $\boldsymbol{A}, \boldsymbol{Q}, \boldsymbol{R}$ respectively. In this way, the segment mask has both good edge and good shape.

Synthesizing samples: we use the matting method to synthesize new positive samples, which regards the segment mask $\boldsymbol{A}$ as the pixel's opacity component. Then a synthesized sample $\boldsymbol{P}$ can be formulated as

$$\boldsymbol{P}_z = \boldsymbol{A}_z \boldsymbol{I}_z + (1 - \boldsymbol{A}_z)\boldsymbol{B}_z, \tag{8}$$

where $z = (x,y)$ is the position, $\boldsymbol{I}$ and $\boldsymbol{B}$ are the pixel's foreground and background colors.

2.3 Negative Samples

The negative samples are collected from the background and around pedestrians. In particular, we collect negative samples on the background generated by background subtraction, denoted as $\boldsymbol{N}_b$. And we also collect negative samples around the pedestrian area, which are referred as $\boldsymbol{N}_r$. Each $s \in \boldsymbol{N}_r$ is collected using the following equation:

$$O(t,s) < \theta_2, \forall t \in \boldsymbol{P_o} \quad and \quad O(t,s) > \theta_3, \exists t \in \boldsymbol{P_o} \tag{9}$$

where θ_2 and θ_3 are threshold values, $\boldsymbol{P_o}$ is the online aligned positive sample dataset. Then the negative sample dataset $\boldsymbol{N} = \boldsymbol{N}_b \bigcup \boldsymbol{N}_r$.

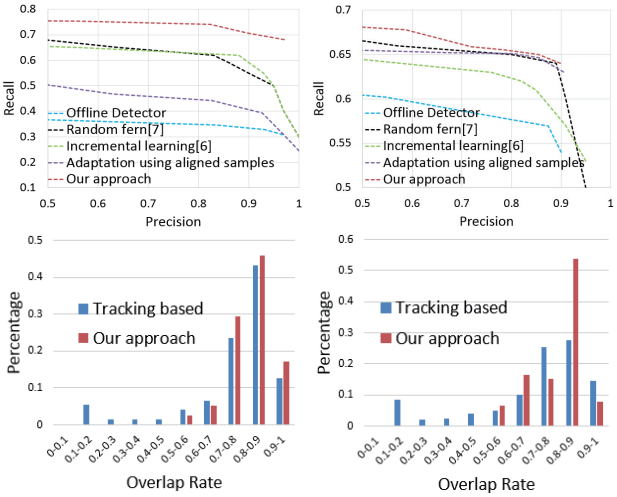
3 Training for Adaptation

A boosted cascade detector can be formulated as

$$F(x) = \prod_{i=1}^{K} f_i(x) \tag{10}$$

where $f_i(x)$ is the i-th strong classifier in the cascade. For a sub-window x, if it is rejected by the i-th strong classifier then $f_i(x) = 0$, otherwise $f_i(x) = 1$.

For an offline boosted detector, the main reason of the performance degradation in a new scene is that some of the strong classifiers in the cascade have lost their efficacy. And it is infeasible to retrain a new detector using the collected online samples, considering the over-fitting problem caused by insufficient samples. Besides, we observe that, there are some strong classifiers in the offline detector that remains effective. Therefore, we remove several strong classifiers and retrain several new strong classifiers using collected online positive and negative samples. In this way, we preserve the efficient part of the offline boosted detectors, while also add some new strong classifiers specifically designed for the new scene.



(a) CAVIAR1 (b) CAVIAR2

Figure 2. Precision-Recall curves and the percentages of samples of different overlap on CAVIAR dataset.

4 Experiment

We validate our approach on two publicly available datasets: CAVIAR dataset and Munich airport dataset. In this section, we will provide implementation details and show the adaptation performance.

Implementation Details: We train an AdaBoosted cascade detector of 20 layers [11] as the offline detector. For aligned sample selection, we first label 3000 binary pedestrian masks, and learn a dictionary with 30 pedestrian mask words using dictionary learning [10]. The size of each window W_k is 4×4 in Guided Image Filtering. The parameter μ is set to 0.8, thresholds θ_1, θ_2 and θ_3 are 0.5, 0.3 and 0.1 respectively, η is set to 1. We conduct experiments by varying the number of synthesized samples between 2000 to 10000, which show that 5000 is sufficient. So we fix the number of 5000 in the following experiments.

Computation Time Cost: We perform all experiments on a 2.8GHz, Xeon computer. It takes about tens of milliseconds to synthesize a sample, thus leading to several minutes to collect thousands of samples. For adaptation, the overall time is about several hours, which is much less than the time cost of retraining an object detector that typically is about 2-4 days [12].

4.1 CAVIAR Dataset

CAVIAR Dataset contains multiple night scenes, which is rather challenging for a generic detector due to illumination variations. We use two scenes, CAVIAR1 and CAVIAR2, and correspondingly select four videos (*OneShopOneWait1front, OneShopOneWait2front, OneShopLeave2Enter and WalkByShop2front*) which are used for experiments. In CAVIAR1, we test our approach on *OneShopLeave2Enter*, and compare with two other approaches. *OneShopLeave2Enter* contains 1200 frames of size 384×288 and 290 ground-truth (GT) instances. In CAVIAR2, we test on *WalkByShop2front*, which has 2360 frames and 1012 ground-truth instances.

We collect 1500 aligned positive samples in CAVIAR1 and synthesize 5000 positive samples. In

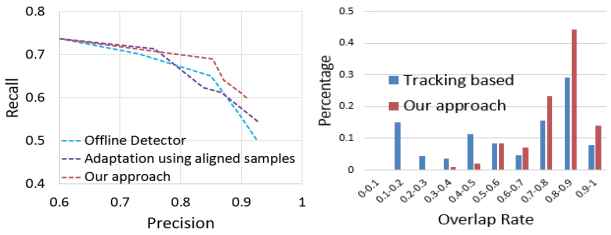


Figure 3. Precision-Recall curves and the percentages of samples of different overlap on Munich airport dataset.

Figure 4. Some adaptation results for different scenes. The first row shows the results by offline detector, the second row shows the results of our approach.

CAVIAR2, 310 aligned positive samples are collected. To show the effectiveness of our aligned sample selection, we calculate the overlap of samples with the ground truth, and compare to a baseline method that uses a tracker and background subtraction. Fig. 2 shows the percentages of samples of different overlap on CAVIAR1 and CAVIAR2. We can see that samples collected using our method mostly have 80% overlap with the ground truth, while the baseline contains many undesirable samples with overlap less than 50%.

We compare the performance with two state-of-the-art approaches. One is an incremental learning approach [6], the other is a random fern approach [7]. We also compare with two baselines: The first uses the offline detector with no adaptation, and the second uses only aligned samples for adaptation. The ROC curves results are shown in Fig. 2. We can see that, our approach outperforms both methods in [6] and [7], and has a significant improvement over the offline detector method. We can also see that the performance of the baseline using only aligned samples is only better than the offline detector baseline. It is mainly due to the inadequate number of online positive samples, which proves the efficacy of our synthesized samples.

4.2 Munich airport Dataset

This dataset consists of 200 images with a resolution of 720×480 and 1829 annotated pedestrians. It is a challenging dataset due to large viewpoint variations, low contrast, small obstacles and occlusions.

We use the first 100 frames to collect samples, and the rest 100 frames for test. We collect about 200 aligned online positive samples and synthesize 5000 positive samples. We also evaluate the distribution of samples of different overlap with ground truth, which is shown in Fig. 3. We compare our approach with the offline detector baseline and aligned sample baseline. The precision-recall curves of different methods are illustrated in Fig. 3. From the results, we can see that our approach achieves the best performance. Note that the improvement is not as significant as in CAVIAR dataset, this is mainly because that the training video is too short (only 100 frames), which lead to a large resemblance in collected online samples, and decreases the adaptation performance. Even though, there is still a 5 percent improvement of recall rate over the offline detector baseline when the precision is 0.9.

5 Conclusion

We present an efficient adaptation approach to transfer an AdaBoost detector to a new scene. The key step in our approach is the online sample collection. We can collect sufficient and qualified online samples without supervision, including aligned online samples and synthesized samples. Experiments on two public datasets demonstrate the power of our approach and further demonstrate the efficacy of our aligned samples and synthesized samples.

References

[1] C. Huang, H. Ai, T. Yamashita, S. Lao, and M. Kawade, "Incremental learning of boosted face detector.," in *ICCV*, 2007.

[2] C. Zhang, R. Hamid, and Z. Zhang, "Taylor expansion based classifier adaptation: Application to person detection.," in *CVPR*, 2008.

[3] B. Zeisl, C. Leistner, A. Saffari, and H. Bischof, "Online semi-supervised multiple-instance boosting.," in *CVPR*, 2010.

[4] G. Shu, A. Dehghan, and M. Shah, "Improving an object detector and extracting regions using superpixels.," in *CVPR*, 2013.

[5] B. Wu and R. Nevatia, "Improving part based object detection by unsupervised, online boosting.," in *CVPR*, 2007.

[6] P. Sharma, C. Huang, and R. Nevatia, "Unsupervised incremental learning for improved object detection in a video.," in *CVPR*, 2012.

[7] P. Sharma and R. Nevatia, "Efficient detector adaptation for object detection in a video.," in *CVPR*, 2013.

[8] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses.," in *CVPR*, 2009.

[9] K. He, J. Sun, and X. Tang, "Guided image filtering.," *PAMI*, vol. 35, no. 6, pp. 1397–1409, 2013.

[10] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding.," in *ICML*, 2009.

[11] G. Duan, C. Huang, H. Ai, and S. Lao, "Boosting associated pairing comparison features for pedestrian detection," in *Ninth IEEE International Workshop on Visual Surveillance*, 2009.

[12] Z. Liu, G. Duan, H. Ai, and T. Yamashita, "Adaptation of boosted pedestrian detectors by feature reselection," in *ICIP*, 2012.