Boosting Object Detection Performance in Crowded Surveillance Videos

Rogerio Feris, Ankur Datta, Sharath Pankanti
IBM T. J. Watson Research Center, New York

Ming-Ting Sun
University of Washington, Seattle

Contact: Rogerio Feris (rsferis@us.ibm.com)

Abstract

We present a novel approach to automatically create efficient and accurate object detectors tailored to work well on specific video surveillance cameras (specific-domain detectors), using samples acquired with the help of a more expensive, general-domain detector (trained using images from multiple cameras). Our method requires no manual labels from the target domain. We automatically collect training data using tracking over short periods of time from high-confidence samples selected by the general-domain detector. In this context, a novel confidence measure is proposed for detectors based on a cascade of classifiers, which are frequently adopted for computer vision applications that require real-time processing. We demonstrate our proposed approach on the problem of vehicle detection in crowded surveillance videos, showing that an automatically generated detector significantly outperforms the original general-domain detector with much less feature computations.

1. Introduction

Object detection plays a fundamental role in intelligent video surveillance systems. The ability to automatically search for objects of interest in large video databases or in real-time video streams often involves, as a pre-requisite, the detection and localization of objects in the video frames.

Traditional surveillance systems usually apply background modeling techniques [20, 21] for detecting moving objects in the scene, which are efficient and work reasonably well in low-activity scenarios. However, they are limited in their ability to handle typical urban conditions such as crowded scenes and environmental changes like rain, snow, reflections, and shadows. In crowded scenarios, multiple objects are frequently merged into a single motion blob, thereby compromising higher-level tasks such as object classification and extraction of attributes.

Appearance-based object detectors [2, 4] arise as a promising direction to deal with these challenging conditions. Specifically for applications that require real-time processing, cascade detectors based on Haar-like features have been widely used for detection of faces [22], pedestrians [23] and vehicles [6]. Although significant progress has been made in this area, state-of-the-art object detectors are still not able to generalize well to different camera angles and lighting conditions.

As real deployments commonly involve a large number of surveillance cameras, training per-camera detectors is not feasible due to the annotation cost. Online adaptation methods [11, 16] have been proposed to adapt a general detector to specific domains, but they usually require a small number of manual labels from the target domain. Most methods rely on adaptation of weights only, while keeping the same features and the same computational complexity of the original detector.

In this paper we propose a novel method for creating specific-domain object detectors without requiring manual labels from the target domain. Our first contribution is the derivation of a *confidence measure for cascade detectors*. To our knowledge, this is a largely unaddressed problem in computer vision, as current work only treats cascade detectors as binary output classifiers without associated confidence. Our second contribution is a *method to automatically collect training samples from the target domain*. We use our proposed confidence measure to select high-confidence detected samples from a general detector trained using images from multiple cameras, and then collect positive samples from tracking over short periods of time (tracklets). These positive samples contain variations such as occlusions which may be complementary to the general detector. Negative samples are also generated by using regions around high-confidence samples as well as samples with different aspect ratio of the object of interest. As a third contribution, we show that by training a specific-domain object detector from the automatically collected data, *we obtain significant accuracy improvement over the general detector with much less feature computations*. Our experimental analysis shows the usefulness of the proposed approach on the problem of vehicle detection in crowded surveillance videos.

2. Related Work

Various methods have been proposed for object detection in images and videos. Deformable part-based models [4], classifiers based on histograms of oriented gradient features [2], and convolutional neural networks [14] are examples of successful approaches that have achieved state-of-the-art results in several standard datasets. In general, however, these methods run at less than 15 frames per second on conventional machines and therefore may not be applicable to surveillance applications that require processing many video channels per server.

Cascade detectors [3, 22] have been commonly adopted for efficient processing. Viola and Jones [22] introduced a robust and efficient detector based on a cascade of Adaboost classifiers, using fast-to-compute Haar-like features. Many variants of this algorithm, including different boosting models and different features have been proposed in the past few years. Confidence measures for cascade detectors have not been well studied.

Co-training techniques [12, 17] have been applied to boost the performance of object detection in specific domains, by training separate classifiers on different views of the data. The confidently labeled samples from the first classifier are used to augment the training set of the second classifier and vice-versa. The underlying assumption of co-training is that the two views of the data are statistically independent, which may be violated especially when the features are extracted from a single modality.

Several on-line adaptation methods [11, 16] have been proposed to adapt general detectors to specific domains. Usually these techniques either require few manual labels from the target domain or suffer from inaccuracies in capturing online data to correctly update the classifier. With few exceptions [9], only feature weights are adapted and not the features themselves. As a result, the adapted classifier is generally at least as expensive as the original detector. Online learning has also been applied to improve tracking [10, 1], with the assumption that an object appears in one location only.

Feris et al [5] proposed a technique to automatically collect training data from the target domain and learn a classifier. However, they require user input to specify regions-of-interest and attributes such as motion direction and acceptable aspect ratios of the object of interest. More recently, Siddiquie et al [19] proposed a method that takes into account scene geometry constrains to transfer knowledge from source domains to target domains. This approach can even achieve better performance than a detector trained with samples from the target domain, but requires a large battery of source domain detectors covering different poses and lighting conditions.

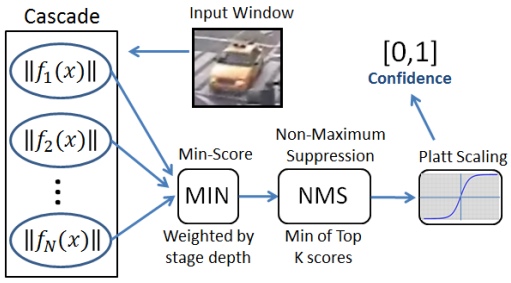

Figure 1. Our confidence measure for cascade detectors.

3. Confidence Measure for Cascade Detectors

Cascade detectors consist of a set of stage classifiers which are applied sequentially to classify a particular image sample. During this process, if any stage detector classifies the sample as negative, the process ends and the sample is immediately considered as negative. The result is positive only if all stage detectors in the cascade classy the sample as positive. In this section, we propose a confidence measure associated with the output of cascade detectors, allowing the results to be ranked according to confidence. Our goal is to make sure the high-confidence positive samples are true positives which can be used for automatic data collection, as we will describe in the following section.

Consider a cascade detector $F(x)$ composed of N stage classifiers $f_i(x), i = 1..N$. In our implementation, we assume the stage classifiers are based on boosting and similar to [22] defined as a linear combination of weak classifiers h_t with a bias θ:

$$f_i(x) = \sum_{t=1}^{T} w_t^i h_t^i(x) - \theta^i. \quad (1)$$

Note that, given an input image sample x, the stage classifier $f_i(x)$ generates a scalar output whose polarity - sign of $f_i(x)$ – determines class membership. The magnitude $||f_i(x)||$ can usually be interpreted as a measure of belief or certainty in the decision made. Nearly all binary classifiers can be viewed in these terms; for density-based classifiers (Linear, Quadratic and Fisher) the output function $f_i(x)$ is a log likelihood ratio, whereas for kernel-based classifiers (Nearest-Neighbor, RBFs and SVMs) the output is a "potential field" related to the distance from the separating boundary.

A key important aspect to be noted is that, according to the cascade principle, if a single stage f_i has low confidence $||f_i(x)||$, the cascade output confidence cannot be high, even if all other stages have high confidence scores. In other words, a high-confidence sample must have high confidences in *all* stage classifiers. This property prevents us from using measurements like max-confidence or even the sum of confidences if they are not properly weighted.

We adopt a simple strategy by taking the *minimum* of all stage classifier scores. This ensures that high-confidence samples will do consistently well in all stage classifiers. Assuming that a sample x progresses through all the stages of the cascade, we define an intermediate cascade confidence score $\delta(x)$ as:

$$\delta(x) = w_d(\min_{1 \leq i \leq N} ||f_i(x)||) \quad (2)$$

where w_d is a weight proportional to the the depth of the minimum stage classifier score, so that samples that have low confidence in early stages are penalized.

We also take into account the non-maximum suppression step to improve our confidence measure. In object detection, multiple detection windows in nearby image locations and scales are grouped together to produce a single output detection window. Indeed, most implementations require a minimum number of neighboring detection windows for firing the detector. For a particular image sample x, let us consider the set $\Delta = \{\delta(x)\} \cup \{\delta_{m_1}, \delta_{m_2}, ..., \delta_{m_M}\}$ where $\delta(x)$ is the confidence score for the sample x as defined in equation 2 and $\{\delta_{m_i}\}, i = 1..M$, corresponds to the set of scores of the M neighboring detection windows. Let $topK(.)$ be a function that returns the top K largest elements of a set and fills out with zeros in case the cardinality of the set is less than K. The refined confidence score for an image sample x is defined as:

$$\alpha(x) = \min\{topK(\Delta)\} \quad (3)$$

The intuition is that a high-confidence example should have at least K high-confidence neighboring window scores. In our implementation, we set $K = 3$.

The final confidence measure $C(x)$ for the cascade detector is obtained by normalizing the score $\alpha(x)$ to the range $[0, 1]$. This is done by using Platt Scaling [13], i.e., we pass the output through a sigmoid to get calibrated probabilities:

$$C(x) = \frac{1}{1 + \exp(A\alpha(x) + B)} \quad (4)$$

where the parameters A and B are fitted using maximum likelihood estimation from the training set. Figure 1 summarizes the main steps of our confidence measure for cascade detectors.

4. Learning Specific-Domain Detectors

Given a general detector (e.g., a car detector trained with images from multiple cameras), and a video from a specific domain (e.g., video from a particular surveillance camera), our goal is to create a more efficient and more accurate detector for the target domain. Our approach is to automatically collect positive and negative samples from the target domain using the general detector, and then create a new specific-domain detector using the collected samples. In the following we describe our method for automatic data collection and detector re-training. In our implementation, we utilize boosted cascade detectors [22] for both general and specialized detectors.

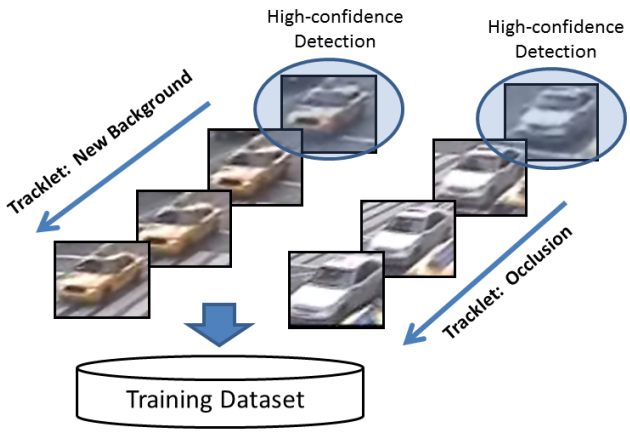

Figure 2. Examples of high-confidence detections and corresponding tracklets. Notice that the collected samples contain additional information such as occlusions and background texture specific to the target domain.

4.1. Positive Samples from Tracklets

Our method for automatic extraction of positive samples in the target domain consists of two steps: 1) detection of high-confidence samples using the general detector and 2) for each high-confidence detection, we perform tracking over a short number of frames (tracklets) and add those tracklet samples to the set of positive training data. This process is illustrated in figure 2. The key intuition is that the tracklet samples will have variations that are specific to the target domain and potentially different from the training data distribution of the general detector. As can be seen in figure 2, the tracklet samples include different information from the original high-confidence detection window, such as a cluttered background and occlusions. This is extremely important to train a detector with more robustness to crowded scenes and specific background texture of the target domain.

High-confidence detection windows are selected by thresholding the confidence measure described in section 3. We tuned the threshold based on a validation set, enforcing zero or very few false alarms while tolerating more false negatives.

Given a high-confidence detection window, we perform KLT feature tracking [18] to follow the object across the video. We make three improvements to the standard KLT feature tracking: 1) Not all pixels within the detection window belong to the object since some of the pixels belong to the background. Tracking all the features within the win-

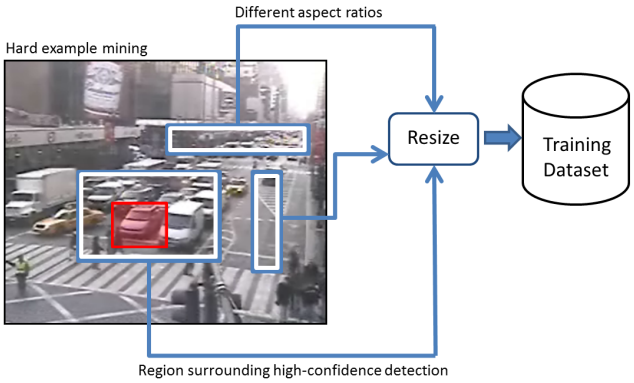

Figure 3. Automatic extraction of negative samples from a crowded traffic scene.

dow may mislead the tracklet, therefore, we utilize background subtraction to only track features that lie on the foreground. 2) In a busy scene, there is potential for occlusions from neighboring tracklets, therefore, we perform a robust estimation of motion using Random Sample and Consensus (RANSAC) [7]. 3) Finally, since certain parts of a vehicle may be textureless or under blur, certain feature may be less reliable than others. We detect less reliable features by accumulating the error attributed to each feature over tracking duration and assigning less weight to more error-prone features during RANSAC robust motion estimation.

Tracking an object over a long period of time is a very challenging problem. As we only consider short tracks over 10 frames in our implementation, the results are very reliable. Indeed, as we will show in our experimental analysis, by combining high-confidence detections with tracklets we are able to obtain a large number of positive samples from the target domain without false alarms.

4.2. Negative Samples

We extract patches from a collection of around 1000 web images that do not contain the object of interest to create a large set of negative samples. In addition, we automatically capture negative samples from the target domain using two strategies: extracting samples related to groups or parts of objects and samples that have a different aspect ratio of the considered object.

Our first strategy consists of extracting windows that are located close to a high-confidence detection window, but with different sizes. As an example, in a crowded traffic scene as depicted in figure 3, a negative window containing a high-confidence detection could comprise a group of vehicles. Capturing such kind of negative samples is relevant as false alarms consisting of groups or parts of vehicles are typical in traffic scenes.

Our second strategy is to sample negative patches from the video frames using windows with different aspect ratios of the considered object and re-sizing them to the object size. The intuition is that the re-sized windows could be potentially similar to false alarms in the scene that have the object size. Since there are many possible windows to be considered as negative samples over the frames, we just select those that are considered "hard examples", i.e., the negative windows with high detection confidence. Figure 3 shows examples of negative patches automatically extracted from a crowded scene.

4.3. Detector Learning

Both the general and the specialized detectors are trained using a framework similar to the work of Viola and Jones[22]. It consists of a cascade of Adaboost classifiers, where the weak learners are simple thresholds over Haar-like features. Each stage of the cascade is tuned to minimize false negatives at the expense of a larger number of false positives - this allows fast inference by quickly discarding background images. Bootstrapping is also employed by selecting negatives examples where the previous stages have failed. We used Gentle Adaboost learning instead of the traditional discrete Adaboost classifiers as it has proven to achieve superior results with decision stumps [15, 8]. At test time, the detectors are applied using a standard sliding window scheme.

5. Experiments

We demonstrate our approach on the problem of vehicle detection in surveillance videos. Our general-domain detector consists of a cascade Adaboost detector trained with 4000 vehicle images obtained from 20+ surveillance cameras. In this study we consider a single vehicle pose only, with slight variation (around 30 degrees maximum pose variation). The negative set was composed of around 1000 images obtained from the web and also from surveillance videos at selected times where no vehicles were present in the scene. We performed several bootstrap rounds during training to improve accuracy, obtaining a detector with 40 stages.

Figure 4 shows some examples of high-confidence examples selected by the general detector using our confidence measure. Our proposed measure allowed us to automatically collect useful data for training without false alarms as we will describe next. The same level of accuracy was not reached with other confidence measures that we have tested, such as relying only on the confidence of the last stage classifier, which focus on discrimination from vehicle-like patterns.

In order to evaluate our approach, we collected a challenging test set from a specific surveillance camera (target domain) containing 229 images and 374 vehicles of a single pose. The images were captured in different months, covering different weather conditions including sunny and rainy days, different lighting effects, such as shadows and

Figure 4. Examples of high-confidence samples selected using our proposed confidence measure for cascade object detectors.

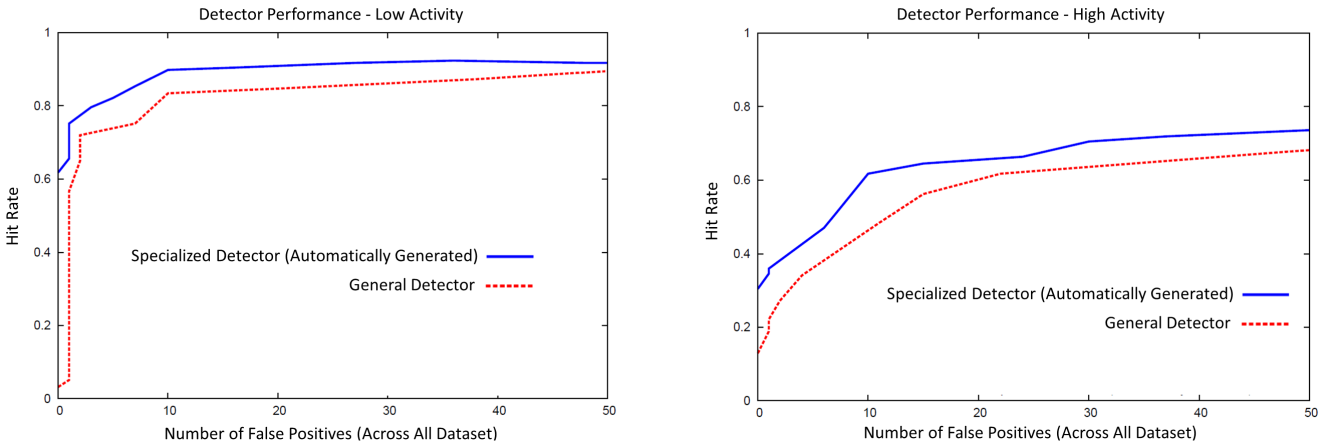

Figure 5. Comparison of a general-domain detector with an automatically generated specific-domain detector in low activity (left) and crowded scenes (right).

specularities, and different periods of time such as morning and evening. In addition, we split our test set into two groups: high activity, i.e., crowded scenes with many occlusions (104 images and 217 vehicles) and low activity (125 images and 157 vehicles).

We applied our automatic data collection technique described in section 4 to a 5 hours (from 2pm to 7pm) video sequence of the same camera but in a different day/month of the period used to capture the test images. This way we could collect 4000 positive training samples automatically without any false alarms. For the negative data, we used the same set of non-vehicle images used to train the general detector (around 1000 images) plus thousands of negative samples automatically collected from the target domain. Using these training samples collected from the target domain, a 20-stage cascade Adaboost classifier was learnt. We refer to this detector as specialized or specific-domain detector.

Figure 5 shows a comparison of the general-domain detector with the automatically generated detector in the target domain. Note that our approach outperforms the general detector in both low activity and crowded scenes. In fact, our data collection technique is capable of capturing data in highly crowded scenarios.

A key advantage of our approach is that we obtain improved accuracy in the target domain with significant gains in terms of efficiency. Our specialized detector achieves superior performance with only 20 stages, half of the number of stages of the general detector. Figure 6 shows a plot comparing the number of features of the initial 20 stages of the general detector with the specialized detector. Note that the specific-domain detector has much fewer features in each of the stages and therefore is significantly more efficient. The general detector has additional 20 stages that are not shown in the plot. The reason for this computational gain is that the target domain data has substantially less appearance variations than the general domain, therefore requiring much less features for discrimination. This is an important advantage over online adaptation methods, which tend to keep the complexity of the original classifier without improving the efficiency.

6. Conclusions

We have proposed a novel approach for learning specific-domain detectors in surveillance videos. Our method relies on a general-domain detector and assumes no labels from the target domain. A novel confidence measure is proposed for cascade object detectors, which is utilized to select high-confidence examples in the target domain, using the general detector. We then perform tracking over short periods of time to collect new samples that may include new information such as occlusions, background texture, and slight variations in object pose, all specific to the target domain. Negative examples are also automatically collected from the target domain. We demonstrate our approach on the problem

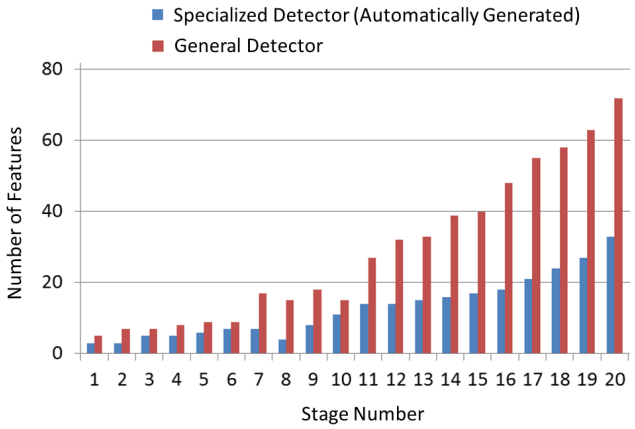

Figure 6. Plot comparing the number of features of the initial 20 stages of the general detector with the specialized detector. The general detector is significantly more expensive and has also additional 20 stages with increasing number of features not shown in the plot.

of vehicle detection in complex surveillance videos, showing that an automatically generated specific-domain detector significantly outperforms the original general detector, not only in accuracy, but also in efficiency, as it requires much less feature computations.

As future work, we plan to investigate the use of high-confidence tracks obtained by background subtraction to augment data collection with more diverse data. Especially in low-activity scenarios, background modeling techniques work very reliably. Extracting samples from both tracklets and motion blobs obtained by background subtraction could produce a richer data collection mechanism in the target domain and potentially improve accuracy.

In this paper we have trained a specific domain-detector using automatically collected data from a single day. We believe that collecting more training data over extended periods of time would improve the robustness of the generated classifier.

References

[1] S. Avidan. Ensemble tracking. *IEEE Transactions on PAMI*, 2007.

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

[3] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *CVPR*, 2010.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on PAMI*, 2010.

[5] R. S. Feris, J. Petterson, B. Siddiquie, L. Brown, and S. Pankanti. Large-scale vehicle detection in challenging urban surveillance environments. In *WACV*, 2011.

[6] R. S. Feris, B. Siddiquie, Y. Zhai, J. Petterson, L. Brown, and S. Pankanti. Attribute-based vehicle search in crowded surveillance videos. In *ICMR*, 2011.

[7] M. Fischler and R. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 1981.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 38(2):337–374, 2000.

[9] H. Grabner and H. Bischof. Online boosting and vision. In *CVPR*, 2006.

[10] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *ECCV*, 2008.

[11] V. Jain and E. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *CVPR*, 2011.

[12] O. Javed, S. Ali, and M. Shah. Online detection and classification of moving objects using progressively improving detectors. In *CVPR*, 2005.

[13] J.Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

[14] Y. LeCun, K. Kavukvuoglu, and C. Farabet. Convolutional networks and applications in vision. In *ISCAS*, 2010.

[15] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *DAGM 25th Pattern Recognition Symposium*, 2003.

[16] S. Pan, I. Tsang, J. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011.

[17] P. Roth, H. Grabner, D. Skocaj, H.Bischof, and Leonardis. On-line conservative learning for person detection. In *PETS Workshop*, 2005.

[18] J. Shi and C. Tomasi. Good features to track. In *CVPR*, 1994.

[19] B. Siddiquie, R. Feris, A. Datta, and L. Davis. Unsupervised model selection for view-invariant object detection in surveillance environments. In *ICPR*, 2012.

[20] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1998.

[21] Y. Tian, M. Lu, and A. Hampapur. Robust and efficient foreground analysis for real-time video surveillance. In *CVPR*, 2005.

[22] P. Viola and M. Jones. Robust Real-time Object Detection. In *International Journal of Computer Vision*, 2004.

[23] P. Viola, M. Jones, and D. Snowi. Detecting pedestrians using patterns of motion and appearance. In *ICCV*, 2003.