Adaptive Object Classification in Surveillance System by Exploiting Scene Context

Jitao Sang Zhen Lei Shengcai Liao Stan Z. Li *

Center for Biometrics and Security Research &
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun Donglu, Beijing 100190, China.
{jtsang,zlei,scliao,szli}@cbsr.ia.ac.cn

Abstract

Surveillance system involving hundreds of cameras becomes very popular. Due to various positions and orientations of camera, object appearance changes dramatically in different scenes. Traditional appearance based object classification methods tend to fail under these situations. We approach the problem by designing an adaptive object classification framework which automatically adjust to different scenes. Firstly, a baseline object classifier is applied to specific scene, generating training samples with extracted scene-specific features (such as object position). Based on that, bilateral weighted LDA is trained under the guide of sample confidence. Moreover, we propose a bayesian classifier based method to detect and remove outliers to cope with contingent generalization disaster resulted from utilizing high confidence but incorrectly classified training samples. To validate these ideas, we realize the framework into an intelligent surveillance system. Experimental results demonstrate the effectiveness of this adaptive object classification framework.

1. Introduction and Related Work

With the rapid development of video capture technology and great demanding for Intelligent Video Surveillance(IVS) systems, there has been significant interest in classification of moving objects in video sequences. Most methods[6, 9, 8] focus on classifying foregrounds extracted from fixed background. While moving objects being detected by background modeling successfully, object recognition is reduced to correctly classifying the moving foregrounds. Specifically, our goal is to classify the moving foregrounds into pedestrian and vehicle. Typical applications include intelligent traffic surveillance. Such far-field surveillance systems deployed throughout cities usually involve hundreds of cameras which has different positions and zooms. Thus objects extracted from these cameras may have diverse visual appearance and vary significantly, which straightly leads to interclass diversity. Obviously, traditional supervised learning on gathering enough labeled training samples for each scene is impractical due to the incredible workload on manually labeling training samples and fixing the parameters of each camera. Further, cameras used in surveillance systems usually set at relatively high position and have a depressing angle to the moving objects, making the classification task even more difficult: size of moving objects is low and change greatly with the distance to the camera. Last but not least, time complexity is utmost because applications embedded in video surveillance systems should always be real-time.

Considering the above-mentioned aspects, constructing such a real-time robust object classification system is desired and challenging. We proceed with further discussion on these three problems and simply introduce our corresponding solution in the rest of this section.

Exploiting unlabeled data to help classifier training has become a hot topic during the past few years. Currently there are three main paradigms for learning with unlabeled data[12], i.e., semi-supervised learning, transductive learning and active learning. One common principle is exploiting large numbers of unlabeled samples to help improve learning performance. As one paradigm of semi-supervised learning, co-training[2] initially trains two separate classifiers with few labeled samples on two respective sub-feature sets and then teach each other using the most confident predicted labels generated by classifying the unlabeled samples. Different from co-training, which chooses and de-

*Stan Z. Li is the corresponding author.

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

ploys unlabeled samples relying on the confidence of the other classifier, tri-training[13] inducts the third classifier and carries out updating based on voting rule. In this paper, we design a full-automatically updating process to construct a real on-line adaptive framework. The assumption is that only one baseline classifier (which is trained off-line and independent of scenes) is available and no labeled samples is offered. Under this situations where no labeled samples be used to help updating multi-classifiers synchronously, we directly apply the baseline classifier to classify the unlabeled samples in each scene and utilize the predicted label(not reliable, we call it soft label[5]) to achieve a scene-specific classifier.

In far-field video sequences, object detected often has very few pixels. So it's hard to extract appearance-based features reliably. Besides, scene-independent features (e.g. relative size and shape) to discriminate between vehicles and pedestrians will change greatly because of the significant projective distortion. However, scene-dependent context features (e.g. object-position, orientation and direction of moving objects) will provide useful knowledge about scene constraints. Bose et al.[4] demonstrate the effectiveness of incorporating scene context features into adaptive classifier learning. Enlightened by this, we exploit underlying regularities within scenes by combining scene context features with scene independent features to help improve the performance of scene-specific classifier.

Usually low-confidence samples are close to the decision boundary, thus more informative for classifier learning. However, these samples are more likely to be incorrectly labeled. Similar work like[3] impose the problem of a trade-off between the risk of using low-confidence samples and their value of discriminative information. They propose a weighted SVM which varies the Lagrange multipliers for soft labeled samples in proportion to their confidences. As a large Lagrange multiplier heavily penalizes an incorrect classification, it allows samples near the decision boundary to modify the adapted solution slightly while reduces the risk of disrupting the training process by incorrect samples with low-confidence. Different from weighted SVM used in[3], for each sample we assign different weights to both classes. This can be regarded as constructing two instances from the original sample and assigning its memberships of positive and negative classes respectively. The intuition is that we can make more efficient use of the training sample. In addition, to achieve real-time performance, we deploy LDA instead of SVM as the classifier methods. One problem in consequence of weighted classifier is also emphasized: incorrectly classified samples with high-confidence will greatly affect the training process and thus disrupt it much more. In this paper, we introduce outlier detection to remove those incorrectly labeled samples and possible samples of unknown classes.

Our method performs object classification on detected moving objects. Simple background subtraction based on Gaussian Mixture Model (GMM) [11] is used to detect the moving objects. The baseline classifier is trained on large number of samples off line. We apply AdaBoost learning algorithm with appearance-based features[14]. The framework of online updating comprises three steps:

1. applying the baseline classifier on objects to obtain soft labeled training samples for each scene;
2. performing outlier detection in each scene to remove incorrectly classified samples with high confidence;
3. training scene-specific classifier using BW-LDA based on extracted scene context features.

We realize this framework into an intelligent surveillance system and demonstrate the effectiveness on a large data from different scenes.

The rest of this paper is organized as follows. Section 2 describes the scene context features we used and introduces original classifier updating process. In section 3 we deduce the BW-LDA and explain the biased naive bayesian based outlier detection. Experimental results are shown in Section 5. The final section concludes our contributions.

2. Scene Context Features and Classifier Updating

With the assumption that moving objects being detected and tracked successfully, we focus on classifying the separated foregrounds. One important step in all object classification methods is to extract effective features for data representation. Features that help discriminate between objects of interest like vehicles and pedestrians involve two types: scene-independent features(e.g. appearance-based features like descriptor of shape[1] and common features like relative size) and scene-dependent context features(e.g. object-position, orientation and direction of moving objects). Because size of objects is relative low and may change greatly with the distance to the camera, scene-independent features may fail to be reliably extracted. Bose et al.[4] demonstrate the effectiveness of incorporating scene context features into adaptive classifier learning.

Within the standard supervised learning paradigm, large number of hand-labeled training samples are required in each scene. It is too cumbersome to implement this scheme in practical surveillance systems with hundreds of cameras. We first apply the baseline object classifier to unlabeled samples from each scene, and then use the obtained soft label to update a LDA based scene-specific classifier.

2.1. Scene Context Features

Scene context features are defined as those are useful for classification in any single scene, but fail when training in one scene and testing in another scene. In other words,

context features are scene-dependent, which cannot transfer across scenes because their different distributions in different scenes.

For classification task of separating pedestrians from vehicles, features like object position and direction of motion are demonstrated as scene-context features[3]. In traffic scenes where car lane, bicycle lane and crosswalk are fixed, by simply using spatial information of objects, we may by and large discriminate pedestrians from vehicles. This is clearly demonstrated in Fig. 1

We define bounding-box as the minimum area of encircling the moving object, occupancy-percentage as the ratio of moving objects to the bounding-box. Obviously, these features are scene-independent features and can be used to separate pedestrian from vehicle. Discriminative features used in this paper include scene-context features like x- and y- object coordinates and aspect ratio. Simple scene-independent features like area of the bounding-box, occupancy-percentage and speed of motion are combined with the above scene context features in the classifier updating process.

2.2. LDA Based Scene Specific Classifier

Linear Discriminant Analysis (LDA)[10] has been successfully used as a dimensionality reduction technique to many classification problems, such as speech recognition, face recognition and multimedia information retrieval. The objective is to find a projection W that maximizes the ratio of between-class scatter S_B against within-class scatter S_W (Fisher's criterion):

$$\arg\max_W \frac{W^T S_B W}{W^T S_W W}$$

First the baseline classifier is applied to unlabeled samples from each new scene and soft-labeled samples with extracted features mentioned in Section 2.1 are obtained. With the assumption that incorrectly classified objects will have lower confidence, obviously it is risky to utilize all soft-labeled samples (this will be further discussed in Section 4.2.1). Thus we choose the top 50% highest confident samples in each scene to train LDA based scene-specific classifier. In this way, we manage to afford some robustness to gross outliers. The updated scene-specific classifiers will then act on corresponding scenes.

3. Bilateral Weighted LDA and Outlier Removing

For unlabeled samples, true labels are not available. It's a waste throwing away low-confidence samples, as they are relatively close to the decision boundary and thus more informative. A balance can be obtained by assigning different weights to samples with weight in proportion to confidence.

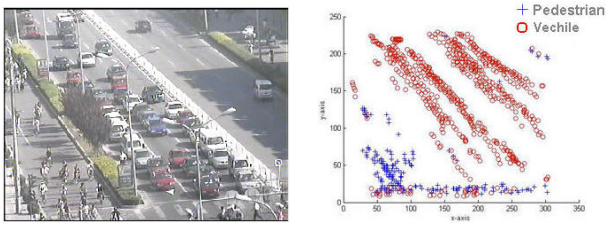

Figure 1. (a)Video frame showing scene3. (b)Scatter plot illustrating spatial distribution of pedestrians and vehicles in scene3.

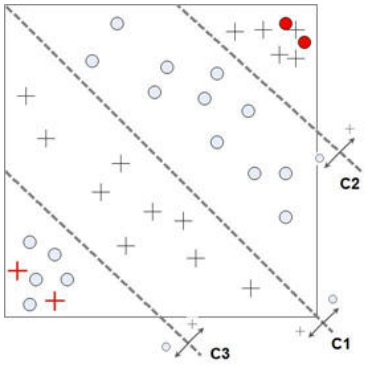

Figure 2. Decision surface for outlier detection. Red-marked samples which incorrectly classified by mutually biased classifier C2 and C3 are detected as outliers

One problem resulted from assigning different weights to samples is that high-confidence samples with high weight tend to affect the classifier-training process even more. Thus high-confidence but incorrectly classified samples will disrupt the training process severely. Outlier detection and removing is proposed to help concern and handle with this trouble.

3.1. Bilateral Weighted LDA

Assigning weights according to confidence is also presented in[3]. We have two improvements. First, in previous proposed weighted SVM one sample uniquely belongs to single class, while in our new method we treat each sample as both of positive and negative classes. The intuition is we can make more use of training sample and manage to achieve better generalization ability. As in the new method each sample contributes two errors to the total error term in the final objective function, we call it as Bilateral Weighted LDA(BW-LDA), while the previous method as unilateral-weighted SVM(UW-SVM). Second, the issue of time complexity is considered. With different distribution of training data, the number of support vector in the SVM classifier becomes unpredictable, thus making time cost beyond control. To achieve real-time performance, we deploy LDA instead of SVM as the classifier methods.

One can easily gets the matrix format of LDA[10]. Original between-class scatter matrix S_B and within-class scat-

ter matrix S_W are defined as follows:

$$\begin{aligned} S_B &= \sum_{l=1}^{K} N_l(\mu_l - \mu)(\mu_l - \mu)^T \\ &= \sum_{l=1}^{K} N_l \mu_l \mu_l{}^T - N\mu\mu^T \\ &= Y\Lambda_c Y^T - N\mu\mu^T \end{aligned} \tag{1}$$

$$\begin{aligned} S_W &= \sum_{l=1}^{K} \sum_{x_i \in C_l} (x_i - \mu_l)(x_i - \mu_l)^T \\ &= \sum_{l=1}^{K} (X_l X_l{}^T - N_l \mu_l \mu_l{}^T) \\ &= XX^T - Y\Lambda_K Y^T \end{aligned} \tag{2}$$

where K is the number of class, N_l is the number of samples in the l-th class, N is the number of all samples, $\mu_l = \frac{1}{N_l}\sum_{x_i \in C_l} x_i$ is the mean vector of the l-th class, $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the total mean vector. $X = [X_1, X_2, \cdots, X_K] = (x_1, x_2, \cdots, x_N)$ be the data matrix of training sets with N labeled samples belonging to K classes, $Y = [\mu_l, \mu_2, \cdots, \mu_K]$ is matrix composed by mean vector of K class, and $\Lambda_K = \begin{pmatrix} N_1 & & & \\ & N_2 & & \\ & & \ddots & \\ & & & N_K \end{pmatrix}$ is diagonal matrix whose diagonal element is sample number of each class.

For bilateral weighted LDA (BW-LDA), we first introduce the weight matrix $F \in R^{N \times K}$, and F_{il} is the probability that the i-th sample belongs to the l-th class. It is obvious that $\sum_{l=1}^{K} F_{il} = 1$. We further define

$$n_l = \sum_{i=1}^{N} F_{il}$$

$$n = \sum_{l=1}^{K} n_l$$

$$\hat{\mu_l} = \frac{\sum_{i=1}^{N} F_{il} x_i}{n_l}$$

$$\hat{\mu} = \frac{\sum_{l=1}^{K}\sum_{i=1}^{N} F_{il} x_i}{n}$$

The new between-class scatter matrix and within-class scatter matrix can be written as following:

$$n\hat{S_B} = \sum_{l=1}^{K} n_l(\hat{\mu_l} - \hat{\mu})(\hat{\mu_l} - \hat{\mu})^T \tag{3}$$

$$n\hat{S_W} = \sum_{l=1}^{K}\sum_{i=1}^{N} F_{il}(x_i - \hat{\mu_l})(x_i - \hat{\mu_l})^T \tag{4}$$

Equ.(1) and Equ.(2) are particular cases of Equ.(3) and Equ.(4) when $F_{il} = \begin{cases} 1 & \text{if } x_i \in C_l \\ 0 & \text{otherwise} \end{cases}$.

With

$$\begin{aligned} n &= \sum_{l=1}^{K} n_l \\ &= \sum_{l=1}^{K}\sum_{i=1}^{N} F_{il} = \sum_{i=1}^{N}(\sum_{l=1}^{K} F_{il}) = N \end{aligned}$$

By analogy, Equ.(3) and Equ.(4) becomes

$$N\hat{S_B} = \hat{Y}\hat{\Lambda_K}\hat{Y}T - N\hat{\mu}\hat{\mu}^T \tag{5}$$

$$N\hat{S_W} = XX^T - \hat{Y}\hat{\Lambda_K}\hat{Y}^T \tag{6}$$

where

$$\hat{Y} = [\hat{\mu_1}, \hat{\mu_2}, \cdots, \hat{\mu_K}]$$

$$\hat{\Lambda_K} = \begin{pmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_K \end{pmatrix}$$

With weight matrix F defined, the proposed BW-LDA is reduced to solving the following objective function:

$$\arg\max_W \frac{W^T\hat{S_B}W}{W^T\hat{S_W}W}$$

3.2. Biased Naive Bayesian Based Outlier Detection and Removing

Delegation learning algorithm[7] focuses on separating difficult samples and delegates them to train another classifier. It decomposes the classification task into two steps: a first classifier chooses which samples to classify and then delegates the 'difficult' samples to train a second classifier. As in Fig.2, $C2$ and $C3$ are the classifier in the first step, which classifies the 'easy' samples in the top right and left bottom corner. 'Difficult' samples near the centerline are left for classifier $C1$ to handle.

Delegation is utilized to detect outliers. To identify samples that are confidently classified, biased classifiers can be utilized. As classifier biased towards predicting positives usually has a high precision on negative samples and vice versa, we treat those who are incorrectly classified by both the bias-towards-positive and bias-towards-negative classifiers as incorrectly labeled samples or samples of unknown classes, i.e. outliers. In Fig.2, classifier $C2$ biases towards the cross class and classifier $C3$ biases towards the circle

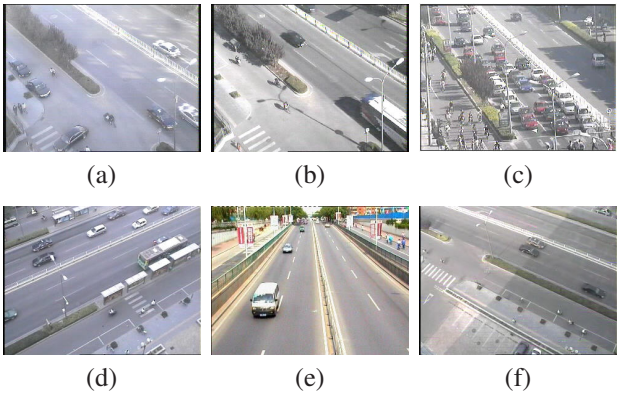

Figure 3. Clips from different testing scenes. (a)scene 1 (b)scene 2 (c) scene 3 (d) scene 4 (e) scene 5 (f) scene 6.

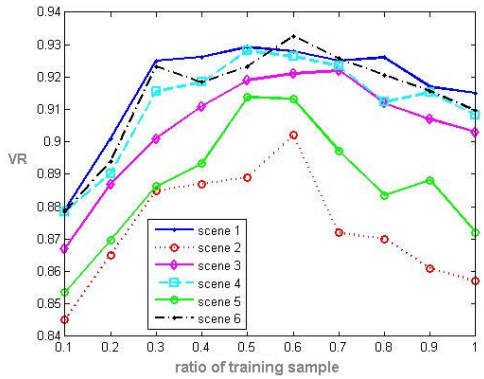

Figure 4. Classification correct ratio versus the ratio of training samples in each scene.

class. Red-marked samples are detected as outliers and removed from the training set.

Considering m classes, $C = C_1, C_2, \cdots, C_m$. One sample X is presented as $x_1, x_2, \cdots, x_n$. Naive Bayesian classifier (NBC) can be constructed as follows:

$$c(X) = \arg\max_{C_i \in C} \prod_{k=1}^{n} P(C_i) P(x_k|C_i)$$

where $P(C_i)$ is the prior probability for i-th class, $P(x_k|C_i)$ is the conditional probability for feature x_k in i-th class.

We utilize NBC to construct biased classifiers. It is easy to bias a Bayesian classifier by either modifying the prior probabilities or to impose biased thresholds on the posterior probabilities. In this paper, negative-biased classifier is formed by setting $P(C_{positive}) = 2 \times P(C_{negative}))$ and positive-biased classifier by setting $P(C_{negative}) = 2 \times P(C_{positive}))$

Further remarks are discussed. While outliers are those deviate from distribution of the same class, there are good chances that these samples are more informative to training. How can we distinguish informative samples from outliers (noise)? In Fig.2, we assume that the centerline is the ground truth decision boundary. Thus samples close with the centerline are regarded as informative samples, those far away from centerline and incorrectly classifier by both biased classifiers are recognized as outliers. Experimental results in following section validate this assumption.

4. Experiment

4.1. Data Set

Video sequences are collected from outdoor far-field cameras. Clips from 6 scenes for test are shown in Fig.3.

We construct a surveillance system which involves background subtraction, moving objects extraction and tracking. For every scene, we first collect more than $4,000$ samples for each class to obtain a scene-specific classifier and then apply the updated classifier to the corresponding scene for classification task. There are totally 24,218 pedestrians and 36,092 vehicles in the collected training samples. Testing samples are extracted from tracked object sequences involving $4,010$ pedestrians and $4,822$ vehicles.

4.2. Experimental Results

4.2.1 Using All Soft-Labeled Samples

In order to measure the influence with the ratio of soft-labeled samples utilized, we vary the ratios of training samples in each scene for constructing the new scene-specific classifier. Soft-labeled samples are ranked according to confidence in each scene and only the certain top ratio of them are utilized to train the new classifier.

As shown in Fig.4, for each scene when ratio exceeds 50%, correct rates even tend to decline. Lower ratio leads to less training samples and the achieved decision boundary is more likely far from the real one, thus classifier generalization ability is limited. However, higher ratio increase the risk of using low-confidence samples which have higher probability to be incorrectly labeled. It also narrow the performance of the new classifier.

Enlightened by this, it is unreasonable to directly increase the ratio or even exploit all training samples. However, cautiously employing the low-confidence with low weight to construct the new classifier achieve a balance.

4.2.2 Performance Evaluation and Analysis

For BW-LDA, the construction of weight matrix F is very important. In this paper, we make use of the confidence obtained by applying the baseline classifier to unlabel samples to construct it. For each scene, confidence of samples from every class is first normalized to $0 - 1$. It is noted as S_{il},

Table 1. Correct classification rates compared between different methods

		Scene1	Scene2	Scene3	Scene4	Scene5	Scene6
Sample number (p/v)	Training Samples	4000/6567	4000/6117	4000/5459	4000/4175	4218/4000	4000/5774
	Testing Samples	539/769	510/707	818/1021	677/712	877/812	589/801
Classification correct ratio	Baseline Classifier	90.6%	81.5%	90.4%	90.6%	86.7%	89.5%
	LDA	92.9%	88.9%	91.5%	93.7%	91.4%	92.3%
	UW-SVM	93.8%	89.2%	92.9%	95.5%	93.1%	94.1%
	BW-LDA	93.7%	89.6%	94.1%	94.9%	93.3%	94.4%
	BW-LDA + Outlier Removing	95.4%	91.1%	95.3%	96.1%	94.2%	95.1%

where $i(i = 1, 2, \cdots, N)$ means the i-th sample and l denotes the l-th class ($l = 0$ denotes positive, $l = 1$ denotes negative). Then we define the weight

$$F_{il} = \frac{1}{1 + e^{-S_{il}}}$$

and

$$F_{i,1-l} = 1 - F_{il}$$

In original LDA based method, we choose top 50% high-confidence training samples for constructing scene-specific classifier. We also implement UW-SVM [3] and compare its performance with the proposed BW-LDA method. As low-confidence samples may disturb the performance of the NBC and outliers in low-confidence samples affect slightly to the final updated scene-specific classifier, NBC based outlier detection is only applied to the top 50% high-confidence soft-labeled samples in each scene.

Table.1 illustrates number of training and testing samples (p/v indicates pedestrians vs. vehicles) for each scene and the correct classification rates between different methods. Scene1 and scene2 are video sequences from the same camera, extracted from different period of time. Fig.3 shows that foreground objects in scene2 have obvious shadows, having an heavy influence on the appearance-based baseline classifier. It is shown that exploiting scene context knowledge greatly improve the performance. Comparing row6 with 7, BW-LDA achieves better results than UW-SVM. More prior improvement lies in time complexity of the new algorithm. We achieve a real-time performance on a 4 channel surveillance system in a PC with 3.0GHz dual CPUs and 1GB memory. Experimental results from the bottom row shows that implement of outlier removing reduces the risky that high-confidence but incorrectly classified training samples breach fatally to the training process.

5. Conclusions

In this paper we exploit the scene context information and propose a full-automatically updating framework for object classification in practical surveillance systems. Our main contributions can be summarized as follows: First, we construct an intelligent surveillance system based on proposed adaptive object classification framework. Then Bilateral Weighted LDA (BW-LDA), which exemplifies the reliability of the predictions is introduced. Last, Naive Bayesian Classifier (NBC) based outlier detection and removing is employed to reduce the risk of high-confidence but incorrectly classified samples and manage to exploit the unlabel samples to the utmost. Experimental results validate the effectiveness of our method on a large data from different scenes. Future research will focus on cooperating outlier detection into construction of weight matrix.

Acknowledgement

This work was supported by the following fundings: National Natural Science Foundation Project #60518002, National Science and Technology Support Program Project #2006BAK08B06, National Hi-Tech (863) Program Projects #2006AA01Z192, #2006AA01Z193, and #2008AA01Z124, Chinese Academy of Sciences 100 People Project, and AuthenMetric R&D Funds.

References

[1] B. Belongie, J. Malik, and J. Puzicha. "Shape matching and object rcognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.

[2] A. Blum and T. Mitchell. "Combining labeled and unlabeled data with co-training". *COLT 1998*.

[3] B. Bose and E. Grimson. "Improving Object Classification in Far-Field Video". *CVPR 2004*.

[4] B. Bose and E. Grimson. "Learning to Use Scene Context for Object Classification in Surveillance". *In Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2003.

[5] O. Chapelle, B. Scholkopf, and A. Zien. *"Semi-Supervised Learning"*. MIT Press, Cambridge, MA, 2006.

[6] R. Cutler and L. Davis. "Robust real-time periodic motion detection, analysis, and applications". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[7] C. Ferri, P. Flach, and J. Hernandez-Orallo. "Delegating Classifiers". *ICML*, 2004.

[8] O. Javed and M. Shah. "Tracking and object classification for automated surveillance". *ECCV*, 2002.

[9] A. Lipton, H. Fujiyoshi, and R. Patil. "Moving target classification and tracking from real-time video". *IEEE Workshop on Applications of Computer Vision*, 1998.

[10] N. Peter, P. Joao, and J. Devid. "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997.

[11] C. Stauffer and W. Grimson. "Adaptive background mixture models for real-time tracking". *CVPR 1999*.

[12] Z. Z-H. "Learning with unlabeled data and its application to image retrieval". *PRICAI 2006*.

[13] Z. Z-H and M. Li. "Tri-training: Exploiting unlabeled data using three classifiers". *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[14] L. Zhang, S. Z. Li, X. Yuan, and S. Xiang. "Real-time Object Classification in Video Surveillance Based on Appearance Learning". *CVPR 2007*.