Real-time Object Classification and Novelty Detection for Collaborative Video Surveillance

Christopher P. Diehl[‡]
Applied Physics Laboratory
Johns Hopkins University
Laurel, MD 20723

John B. Hampshire II
exScientia, LLC
85 Speen Street, Lower Level
Framingham, MA 01701

Abstract

To conduct real-time video surveillance using low-cost commercial off-the-shelf hardware, system designers typically define the classifiers prior to the deployment of the system so that the performance of the system can be optimized for a particular mission. This implies the system is restricted to interpreting activity in the environment in terms of the original context specified. Ideally the system should allow the user to provide additional context in an incremental fashion as conditions change. Given the volumes of data produced by the system, it is impractical for the user to periodically review and label a significant fraction of the available data. We explore a strategy for designing a real-time object classification process that aids the user in identifying novel, informative examples for efficient incremental learning.

1 Introduction

A distributed video surveillance system provides a low-cost means of monitoring activity within a given environment. Yet without automation, the user will be unable to effectively utilize the system to detect relevant activity in a timely fashion. Ideally, a video surveillance system should provide the user with an integrated view of the environment that allows the user to detect and respond to ongoing events. For many scenarios within the civilian and military sectors, real-time interpretation is required for the information produced by the system to be valuable.

Given the challenge of achieving real-time performance on low-cost commercial off-the-shelf hardware, system designers typically define the classifiers prior to the deployment of the system so that the performance of the system can be optimized for a particular mission. This implies the system is restricted to interpreting activity in the environment in terms of the original context specified. For uncertain or changing environments, the necessary context to interpret the environment cannot be completely defined beforehand. Therefore the system should allow the user to provide additional context in an incremental fashion so that the classifiers continue to provide the relevant information meeting the user's needs.

When designing a classifier to interpret the video data, we require a set of labeled training examples to estimate the classifier parameters. While the surveillance system is in operation, we obtain volumes of data. Yet it will often be impractical to review and label a significant fraction of the data. What is needed is a collaborative process that focuses the user's attention on the novel, informative examples for review and labeling. By selecting the most informative examples, our goal is to minimize the burden on the user while maximizing the rate at which the system learns the underlying concepts.

In this paper, we discuss the issues associated with the design of a real-time process for object classification and novelty detection. We present a strategy for addressing the problem and investigate the performance of the classifier designed for Carnegie Mellon's *CyberScout* semi-autonomous, mobile video surveillance platform [15].

2 The Design Philosophy

For many of the real-time video surveillance systems discussed in the literature, the common objectives are to detect, classify and track objects of interest in the environment [1, 2, 6, 7, 8, 10]. Typically the video stream is analyzed by a series of processes that perform each of these tasks. The motion detector nominates candidate motion regions in each video frame. The tracker associates motion regions across video frames, producing image sequences for each candidate moving object. The classifier analyzes the image sequences incrementally and associates an image sequence with the most likely object class.

We will focus on designing the classifier responsible for efficiently labeling the image sequences and assessing classification confidence. Our objective will be to simultaneously optimize the following design parameters: *computational complexity*, *classification performance* and *rejection performance*. The issue of computational complexity will be the dominant concern in the design. Given that our goal is to deliver information about ongoing activity to the user in real-time, simplicity will be stressed whenever possible.

In order to specify the image sequence classification process, the following components must be defined: the representation of the image sequence, the classifier's decision process and the procedure for learning the classifier from the data. We consider each of these components next.

3 The Image Sequence Representation

The role of the representation is to provide a description of the image sequence that captures the necessary information for object classification and novelty detection in a form that supports real-time processing. The real-time constraint dictates

[‡]This work was conducted while CPD was at Carnegie Mellon.

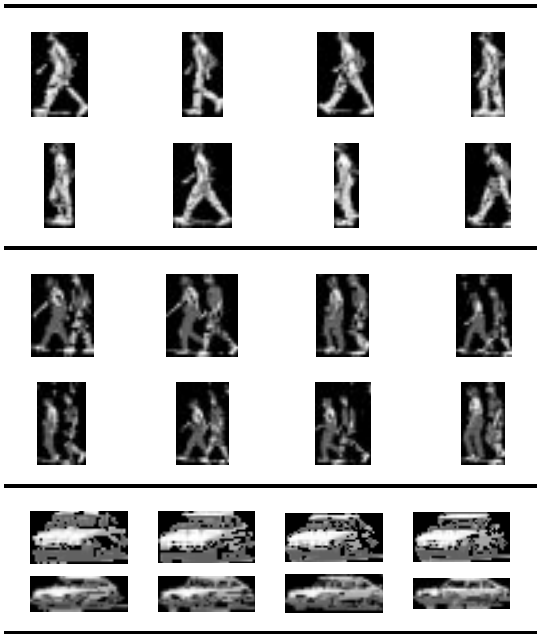

Figure 1: Example image sequences

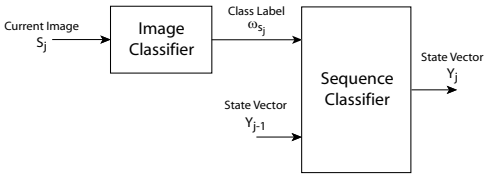

Figure 2: Image sequence classification process

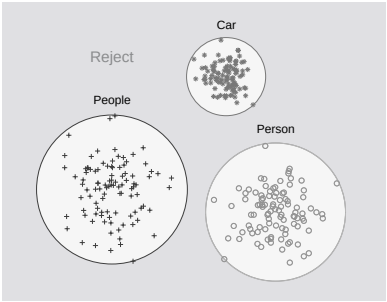

Figure 3: A sample partition

computational simplicity; whereas novelty detection requires a high-dimensional feature space to aid in discriminating between the known and unknown object classes. Therefore the challenge is to identify a representation with the appropriate level of complexity that achieves a balance between these conflicting constraints.

The image sequences assembled by the motion detector and tracker provide spatiotemporal data about the moving objects in the scene. Several examples are shown in figure 1. In these examples, one can observe several sources of variability that we must contend with. Images within a given sequence vary in size due to changes in object aspect and shape. Images also vary in resolution as the range from the object to the sensor changes. Object positions within the image are offset when partial detections or other image variations occur. Other challenges are caused by occlusion, lighting variation and nonuniform sampling of the environment.

For real-time object classification, we argue that the cost of exploiting the variation in appearance as a discriminator does not justify the potential benefits in classification performance. To leverage this information, significant effort will be required to compensate for the malicious effects caused by changes in the environment, occlusion and nonuniform sampling. At the same time, the data requirements for acceptable generalization performance will be substantial since the system must learn to classify appearance variations at various resolutions and aspects.

By limiting our focus to spatial features, the representation design task becomes one of defining an image representation. In recent work on appearance-based object detection and classification in static imagery, excellent performance has been achieved by classifying the images directly or features derived from the images using (over)complete wavelet dictionaries [11, 13, 14, 16, 18]. Ideally we would like to learn a representation that is well matched to the specific object classification task. By specifying a parameterized or overcomplete representation, the learning procedure can search over multiple feature spaces for a particular representation that optimizes the specified objective function. In the work cited above, the objective is to discriminate between the various object classes which are all represented in the training data. We will now consider the problem of learning to discriminate between the known object classes in a manner that supports novelty detection.

4 Image Sequence Classification

Since the classification process will examine only the spatial features of each image, the classification of an image sequence becomes a two-step process as shown in figure 2. In the image classification phase, the current image S_j from the image sequence S is analyzed to determine the most likely class label ω_{S_j}. Then in the sequence classification phase, the evidence from the classification of the current image is integrated with past evidence to produce an overall decision with a confidence level.

4.1 Partitioning the Image Feature Space

Learning to map a given image feature vector S_j to the most likely class label ω_{S_j} involves learning a partition of the image feature space. Figure 3 illustrates our objective graphically. Given a set of training images and their corresponding class labels, we want to learn a partition that maps regions of the image feature space to one of the specified object classes where significant data exists to support the decision. In other regions of feature space where little to no data exists, feature vectors will be rejected.

The partition is generally represented in terms of a set of $\mathcal{C}$

parameterized *discriminant functions* $g_k(X|\theta)$ where

$$g_k(X|\theta) - \max_{i,i\neq k} g_i(X|\theta) > 0 \quad (1)$$

when the feature vector X maps to the class label ω_k [5]. The set of feature vectors satisfying equation 1 define the *decision region* R_k. Given our objective, it may seem natural to derive estimates of the *a posteriori* class probabilities $\hat{\mathrm{P}}(\omega_k|S_j)$ and define the discriminant functions as

$$g_k(S_j) = \hat{\mathrm{P}}(\omega_k|S_j). \quad (2)$$

Coupling this set of discriminant functions with the rejection rule

$$g_*(S_j) = \max_{k\in[1,\mathcal{C}]} g_k(S_j) < 1-\delta \longrightarrow \text{reject decision} \quad (3)$$

leads to an approximation of the Bayes-optimal classification and rejection strategies [3].

Unfortunately there are problems with this approach. Assuming for the moment that we can estimate the posterior probabilities for a given feature vector S_j, the estimate of the probability of correct classification $g_*(S_j)$ will provide misleading results when the training data is not consistent with the underlying distribution [12]. To understand this problem, consider a simple two class problem. The posterior probabilities $\mathrm{P}(\omega_k|X), k \in \{1,2\}$ are by definition equal to

$$\mathrm{P}(\omega_k|X) = \frac{\mathrm{p}(X|\omega_k)\mathrm{P}(\omega_k)}{\mathrm{p}(X|\omega_k)\mathrm{P}(\omega_k) + \mathrm{p}(X|\overline{\omega_k})\mathrm{P}(\overline{\omega_k})}. \quad (4)$$

If we evaluate the posterior probabilities for samples from an unknown outlier process, we will find that the probability of correct classification is nearly equal to 1 for some samples with very small likelihoods $\mathrm{p}(X|\omega_k), k \in \{1,2\}$. This problem stems from the fact that the unconditional density $\mathrm{p}(X)$ in the denominator of equation 4 deviates significantly from the true unconditional density in certain regions of the feature space. This indicates that in domains such as surveillance where the training sample is not representative of the underlying distribution, estimating the probability of correct classification will not provide a valid estimate of classification confidence.

Another approach for defining the partition involves estimating closed decision regions with minimal volume that encompass the majority of training examples. The decision regions can be obtained indirectly by thresholding density estimates or directly by estimating indicator functions that define the regions [17]. We will directly estimate closed decision regions with minimal volume that encompass a specified fraction of the training examples.

Given we will not be able to employ a bootstrapping process to obtain a large number of representative false alarm examples, minimizing the volume of the decision regions appears to be the only means by which the image classifier's false alarm rate can be controlled. By definition the false alarm rate P_{FA} of the classifier is

$$\mathrm{P}_{\mathrm{FA}} = \sum_{k=1}^{\mathcal{C}} \int_{R_k} \mathrm{p}(X|\omega_{other})dX. \quad (5)$$

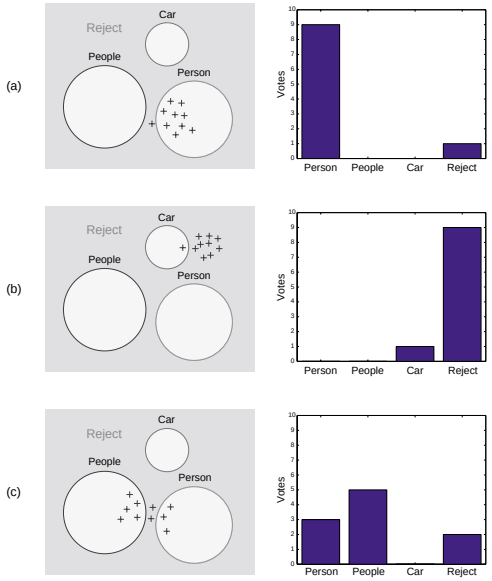

Figure 4: Examples of image sequence class label distributions: (a) confident classification: vast majority of the images lie in one decision region (b) consistent rejection: significant fraction of the images lie in the rejection region (c) classifier confusion: images distributed over two or more regions

This allows us to state unequivocally that if a decision region R_a is a proper subset of R_b ($R_a \subset R_b$), R_a will yield a lower false alarm rate. Assuming R_a and R_b produce the same probability of detection, R_a clearly offers superior performance. In the more general case where $R_a \cap R_b \neq \emptyset$ and $R_a \not\subset R_b$, it is impossible to make any definitive claims about the relative false alarm performance; yet we believe on average that the smaller decision region will provide a lower false alarm rate.

If the representation allows one to parameterize the underlying feature space in multiple ways, the learning procedure can search for the parameterization that allows one to construct the smallest decision regions with a specified probability of detection. The challenge is to define a proper measure of decision region volume that allows one to compare decision regions across parameterizations. In this paper, we do not address this issue in rigorous detail. In our initial experiments, we employ a simpler measure that is related to the decision region volume. We will discuss our specific approach in the next section.

4.2 Classifying Class Label Sequences

As the image classifier processes a series of images of a given object, a sequence of class labels is produced. Based on this class label sequence, we wish to assign a class label to the image sequence along with a level of classification confidence. The class label sequence provides two types of information. First, a set of possible class labels is obtained along with the relative frequencies of occurrence of each class label. In addition, the sequence captures the transitions between the class labels as the appearance of the object changes over time. The question we wish to address at this point is whether the class label distribution is sufficient to reliably classify image

Figure 5: CyberScout all-terrain vehicle

sequences of known objects and detect image sequences of unknown objects and novel views of known objects.

Whether or not the class label distribution is sufficient is actually determined by the image representation and the partition. In order to reliably classify known objects and detect unknown objects and novel views of known objects based on the class label distribution, we must be able to successfully discriminate between such examples in the image feature space. If the combination of image representation and partition provides the necessary discrimination power, the class label distributions induced by known objects, unknown objects and novel views of known objects will be sufficiently separable in class label distribution space. Ideally one would hope that the image classifier reliably and consistently classifies or rejects the images in a given sequence as illustrated in figures 4(a) and (b), thereby simplifying the image sequence classification task. Yet the reality is that image sequences will often induce a mixture of classifier outputs as illustrated in figure 4(c) indicating classifier confusion. When classifier confusion does occur, our ability to discriminate between known and unknown objects does not necessarily decrease significantly. As we shall see later, certain unknown object classes may actually display patterns of classifier confusion that are different from those associated with the known object classes. Therefore the task of identifying the unknown object image sequences remains tractable.

Since it is not clear what additional patterns could be efficiently exploited in the history of the class label transitions, our approach to sequence classification will entail mapping class label distributions to one of the known object classes. We will learn a partition of the class label distribution space from the training data and classification confidence will be assessed in a principled manner. We will withhold our discussion of classification confidence until later, since our approach is connected with the learning procedure.

5 Experimental Results

5.1 The Classification Task

To evaluate this general strategy, we designed an object classifier to classify image sequences observed by the *CyberScout* mobile video surveillance platform as either individuals (*person*), groups of people (*people*) or cars (*car*). The image and sequence classifiers were trained on image sequences obtained from several data collections on the Carnegie Mellon campus. Since the motion detector generally provides a reasonable segmentation of the moving objects, we chose to classify size-normalized binary images of the moving objects. Size normalization is achieved by first resizing each binary image so that the largest dimension is a fixed dimension N. The resized image is then zero-padded to produce a square $N \times N$ pixel image with the original image in the center.

5.2 Image Classification and Rejection

To estimate a set of closed decision regions encompassing the training examples from the various object classes, we first learn a *large margin partition* of the image feature space by minimizing the objective function

$$\frac{1}{2}\|\theta\|^2 - \beta \sum_{i=1}^{M} \sigma(\delta(X_i|\theta), \psi) \tag{6}$$

over the discriminant function parameter space θ [9]. The *discriminant differential* (*multi-class margin*) $\delta(X|\theta)$ is defined as the difference between the discriminant function associated with the correct class and the largest other discriminant function. If a given feature vector X has a class label ω_c, then the discriminant differential for this example is denoted as

$$\delta(X|\theta) = g_c(X|\theta) - \max_{k,k\neq c} g_k(X|\theta). \tag{7}$$

$\sigma(\delta, \psi)$ is a parameterized, monotonically increasing function of the discriminant differential that allows one to vary the tradeoff between *margin maximization* and *training error minimization*. Note that this objective function is a generalization of a standard formulation of *support vector learning* which is ideally suited for learning partitions of high-dimensional spaces from sparse datasets [4]. Once the initial partition is learned, we define the *rejection region* R_{reject} by estimating $\mathcal{C}$ differential thresholds δ_{R_k} that yield a given class-conditional probability of detection on a validation set. All images that lie within the portion of the decision region R_k defined by

$$0 \leq g_k(X|\theta) - \max_{i,i\neq k} g_i(X|\theta) \leq \delta_{R_k} \tag{8}$$

will be rejected.

When selecting a class of discriminant functions to use, we need to ensure that the discriminant function class induces closed decision boundaries in the space where the data lies. Let us consider the logistic linear form. The logistic linear discriminant function $g_k(X|\theta_k, \theta_{b_k})$ for class ω_k is defined as

$$g_k(X|\theta_k, \theta_{b_k}) = f(\theta_k^T X + \theta_{b_k}) \tag{9}$$

where

$$f(x) = \frac{1}{1+e^{-x}}, \tag{10}$$

For classification problems where $\mathcal{C} > 2$, thresholding the discriminant differential leads to decision boundaries that form semi-infinite wedges in feature space [4]. Since the set of $N \times N$

Resolution	30x30	30x30	20x20
Normalization	None	Center of Mass	Differential
Median Rejection Rate	0.38	0.42	0.72

Table 1: Median foliage rejection rates obtained over a series of 50 trials where the classifier is trained on different random partitions of the training data

binary images define the vertices of an N^2-dimensional hypercube, the logistic linear form induces closed decision regions on the surface of the hypercube as desired. By increasing the threshold, the volume of the closed decision region decreases. Therefore we will control the size of the decision regions indirectly by maximizing the separability of the data.

We conducted a series of experiments to evaluate the performance of the classifier when trained on normalized and unnormalized images of various resolutions [4]. Two different normalization techniques were investigated. The first involved centering the images horizontally based on the center of mass. The second involved horizontally translating the images to maximize the resulting differential during training and testing.

After training on normalized or unnormalized images, the classifier generalizes successfully on the test set. Yet the classifiers trained on the unnormalized and centered images fail to reject a majority of a set of false alarm images generated by moving foliage as indicated in table 1. Examining the false alarm results, one notices that the majority of false alarms are classified as people. This is not surprising. Groups of people walking together can be observed in a variety of configurations relative to one another. This leads to a highly unstructured weight layer for the people discriminant function and a long-tailed distribution for the differential δ_{people}. To achieve a high probability of detection for the people class, we are forced to set the differential threshold low, which leads to the high false alarm rate.

To overcome this deficiency, we investigated the effectiveness of the procedure whereby the image is horizontally translated to determine the translation that maximizes the difference between the largest and next largest discriminant function output. When employing this strategy, we are simultaneously learning a transformation of the image space and a partition of the resulting space that are complimentary. The classifier normalizing *20x20 pixel images* in this manner yields a 30% median improvement in rejection performance when compared with the 30x30 pixel image classifier processing centered images. This comes at the cost of a minor reduction in classification performance and added computational complexity due to the need to evaluate the classifier for multiple translations of the size-normalized binary image.

Examining the class-conditional differential distributions, it is clear that this technique eliminates the long-tailed distributions that plagued the previous classifiers. This allows one to set the differential thresholds significantly higher to minimize the decision regions. Yet in retrospect, it is not clear that we

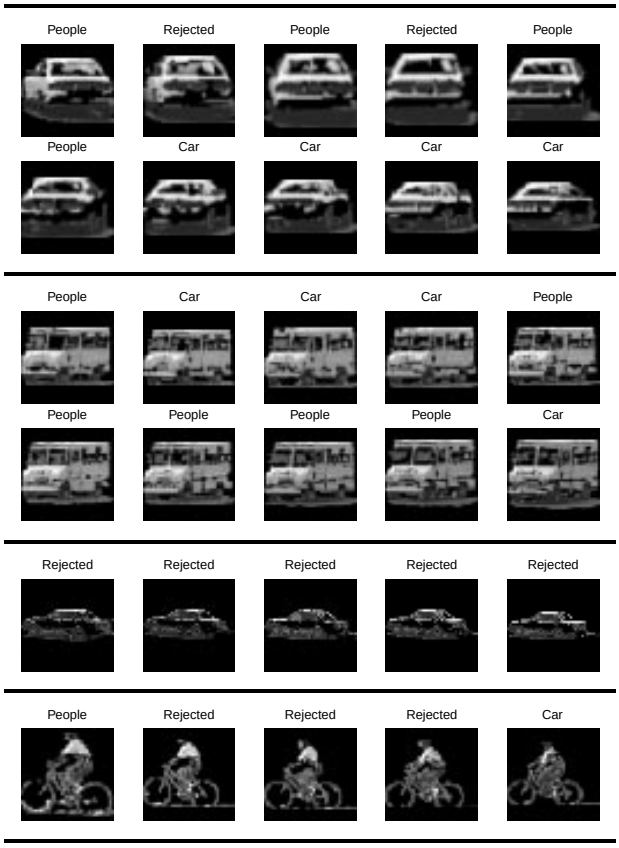

Figure 6: Novel image sequences identified by the image sequence classifier

can attribute the dramatic improvement in performance to reductions in the size of the decision regions. The transformation of the image space that we have learned is a many-to-one mapping. Therefore it is difficult to argue geometrically that we have effectively controlled the false alarm rate. More work is needed here to complete this story. At this point, our major question is whether such a simple, real-time classifier processing low resolution images of moving objects can provide a useful novelty detection capability. We pursue this next.

5.3 Novelty Detection

Once the image classifier is trained, we process the training image sequences to derive the associated class labels for training the sequence classifier. Each training image sequence has a corresponding class label distribution which is simply the relative frequencies of each class label in the class label sequence. Using these class label distributions, we trained a logistic linear classifier to partition the class label distribution space.

As the sequence classifier processes the observed image sequences, the image sequences assigned to a given class ω_k are rank ordered based on the likelihood $\mathrm{p}(\delta|\omega_k)$ [4]. Given that the likelihood is generally monotonically increasing with increasing differential, we sort the image sequences based on the differential produced by the sequence classifier. This allows the user to quickly focus on the examples, such as those in fig-

ure 6, that cause the greatest degree of confusion for the current classifier.

In order to evaluate the utility of the 20x20 pixel image sequence classifier, we classified and sorted a test set of image sequences consisting of examples from the three known classes along with image sequences of bicycles, trucks and vans. Our main objective was to determine whether this process would allow the user to easily detect many of the examples from the unknown classes by simply scanning through a small subset of the sorted observations. In the case of the bicycle class, 70% of the observed examples were in the top 20% of the data assigned to the people class. Bicycles typically caused unique patterns of classifier confusion which simplified their identification. If the classifier was able to view the bicycle from multiple perspectives, it was more likely to produce a mixture of rejections and class labels. Trucks and vans, on the other hand, could not be successfully discriminated from cars. Only vehicles like FedEx trucks produced classifier confusion because their appearance varies significantly from cars. At such low resolution, it is nearly impossible to reliably distinguish between cars, trucks and vans.

Other image sequences from the known object classes producing small differentials illustrate concepts the current classifier has not captured. For example, since many of the car training examples show side views, vehicle observations exhibiting major aspect change caused classifier confusion. In several preliminary experiments, we have seen that actively selecting such examples in an incremental learning framework can accelerate concept acquisition [4].

6 Conclusions

The initial results we have obtained are encouraging; but it is clear much work remains. In order to detect more subtle class distinctions, we must operate in higher dimensional spaces. The challenge is to preserve real-time performance as we increase the dimensionality. We must also explore methods for jointly learning a feature space and partition that are complimentary. Efficiency in representation is critical to meet our performance objectives. We will continue to explore these issues in our future work.

Acknowledgments

We gratefully acknowledge funding provided for this research under DARPA grant F04701-97-C-0022. Thanks to Mahesh Saptharishi for his helpful suggestions that improved the manuscript.

References

[1] Marc Bogaert, Nicolas Chleq, Philippe Cornez, Carlo Regazzoni, Andrea Teschioni, and Monique Thonnat. The PASSWORDS project. In *Proceedings, International Conference on Image Processing*, pages 675–678, 1996.

[2] Hilary Buxton and Shaogang Gong. Advanced visual surveillance using Bayesian networks. In *Proceedings, IEEE Workshop on Context-Based Vision*, 1995.

[3] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, IT-16(1):41–46, 1970.

[4] Christopher P. Diehl. *Toward Efficient Collaborative Classification for Distributed Video Surveillance*. PhD thesis, Carnegie Mellon University, December 2000.

[5] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Inc., 1973.

[6] Bruce E. Flinchbaugh and Thomas J. Olson. Autonomous video surveillance. *Proceedings of the SPIE*, 2962:144–151, 1997.

[7] G. L. Foresti. Object recognition and tracking for remote video surveillance. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(7):1045–1062, October 1999.

[8] W. E. L. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site. In *Proceedings, IEEE Conference on Computer Vision and Pattern Recognition*, pages 22–29, 1998.

[9] John B. Hampshire II. *A Differential Theory of Learning for Efficient Statistical Pattern Recognition*. PhD thesis, Carnegie Mellon University, September 1993.

[10] Takeo Kanade, Robert T. Collins, Alan J. Lipton, Peter Burt, and Lambert Wixson. Advances in cooperative multi-sensor video surveillance. In *Proceedings, DARPA Image Understanding Workshop*, 1998.

[11] Constantine P. Papageorgiou and Tomaso Poggio. Trainable pedestrian detection. In *Proceedings, International Conference on Image Processing*, volume 4, pages 35–39, 1999.

[12] Stephen J. Roberts and William Penny. Novelty, confidence and errors in connectionist systems. Technical report, Neural Systems Research Group, Imperial College of Science, Technology and Medicine, April 1997.

[13] Danny Roobaert. Improving the generalization of linear support vector machines: an application to 3D object recognition with cluttered background. In *Proceedings, Support Vector Machine Workshop at the 16th International Joint Conference on Artificial Intelligence*, pages 857–862, August 1999.

[14] Danny Roobaert. View-based 3D object recognition with support vector machines. In *Proceedings, IEEE Workshop on Neural Networks for Signal Processing*, pages 77–84, August 1999.

[15] M. Saptharishi, K. S. Bhat, C. P. Diehl, J. M. Dolan, and P. K. Khosla. CyberScout: Distributed agents for autonomous reconnaissance and surveillance. In *Proceedings of the Conference on Mechatronics and Machine Vision in Practice*, pages 93–100, September 2000.

[16] Henry Schneiderman and Takeo Kanade. A statistical model for 3D object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[17] Bernhard Schölkopf, Robert C. Williamson, Alex Smola, and John Shawe-Taylor. SV estimation of a distribution's support. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*. Morgan Kaufmann, 2000.

[18] Paul Viola and Michael J. Jones. Robust real-time object detection. Technical Report CRL 2001/01, Compaq Cambridge Research Laboratory, February 2001.