A Framework of Spatio-Temporal Analysis for Video Surveillance

Duan-Yu Chen
Institute of Information Science, Academia Sinica, Taiwan
dychen@iis.sinica.edu.tw

Kevin Cannons
Department of Computer Science and Engineering, York University, Canada
kcannons@cse.yorku.ca

Hsiao-Rong Tyan
Department of Information and Computer Engineering, Chung Yuan Christian University, Taiwan
tyan@ice.cycu.edu.tw

Sheng-Wen Shih
Department of Computer Science and Information Engineering,
National Chi Nan University
stone@csie.ncnu.edu.tw

Hong-Yuan Mark Liao
Institute of Information Science, Academia Sinica, Taiwan
liao@iis.sinica.edu.tw

Abstract

This paper presents a video surveillance system that is capable of detecting and classifying moving targets in real-time. The system extracts moving targets from a video stream and classifies them into predefined categories according to their spatiotemporal properties. Classification of the moving targets is completed via a combination of a temporal boosted classifier and spatiotemporal "motion energy" analysis. We illustrate that a temporal boosted classifier can be designed that successfully recognizes five object categories: person(s), bicycle, motorcycle, vehicle, and person with umbrella. The proposed temporal boosted classifier has the unique ability to improve weak classifiers by allowing them to make use of previous information when evaluating the current frame. In addition, we demonstrate a method to further process targets in the "person(s)" category to determine if they are single moving individuals or crowds. It is shown that this challenging task of moving crowd recognition can be effectively performed using spatiotemporal motion energies. These motion energies provide a rich description of a target's dynamic characteristics, from which classification can be performed. Our empirical evaluations demonstrate that the proposed system is extremely effective at recognizing all predefined object classes.

Keywords : video surveillance, object classification, spatiotemporal analysis

1. Introduction

Automatic moving target classification has attracted much attention in recent years [1], especially in the field of surveillance. Though much progress has been made [2-5][14], recognizing moving targets with high accuracy remains a challenging,

unsolved problem. Significant difficulties of moving target recognition in the surveillance setting include the facts that targets often have complex shapes due to their non-rigidness and can be of low-resolution because of the nature of a video camcorder. Matters are complicated further because of the sheer range of moving objects that can be observed within an environment.

In the context of surveillance systems, a variety of machine learning classification techniques have been investigated, including support vector machines [5], naïve Bayes classifier [6], and AdaBoost [3]. AdaBoost is especially suitable for surveillance scenarios since it has achieved high detection rates using simple Haar-like features in real-time [3]. Though boosting paradigms have attracted significant attention recently, most previous moving target classification work has focused on boosting within a single frame. In fact, the use of temporal features as inputs to both weak and strong classifier levels has not been carefully studied in the past. However, in video frames, if a moving target is seen in one frame, it is very likely that it will be present in the next frame. Therefore, one critical contribution of this work is to present a novel method of introducing dynamic information into the AdaBoost framework.

Within the surveillance domain, the targets that are typically of primary interest are people. In this paper, our second goal is to further analyze moving targets that our temporal boosted AdaBoost system classifies as “person(s)”. The purpose of post-processing this class of targets is so that we can further identify two sub-classes: moving crowds and moving single persons. Automatically identifying the number of people that are involved in an activity under surveillance is a critical component to attaining higher level scene understanding.

Here, we propose to incorporate spatiotemporal information into our sub-classification module through the use of oriented energies. The general applicability of spatiotemporal orientations and their relationship to motion perception was first realized in [8]. One of the applications of this field initially considered was the recovery of optical flow using filters in space-time [11]. It was illustrated [13] that qualitative descriptors can be assigned to a local spatiotemporal region using oriented energy signatures. Spatiotemporal orientation-selective filters are already starting to be adopted in the tracking and surveillance domains. In [10], orientation selective filtering of the spatiotemporal domain was performed to obtain a pixel-wise measure of coherent motion.

In light of previous research, the main contributions of our proposed spatio-temporal analysis for crowd detection system are as follows. First, efficient self-similarities are computed in spatial domain to detect candidates of moving crowds. Second, we apply powerful, oriented energy descriptors in temporal domain to recognize the real moving crowds within the candidates.

The remainder of this paper is organized as follows. In Section 2 details the features used for describing spatial temporal patterns of moving targets and the proposed temporal boosted learning algorithm. Section 3 introduces the moving crowd detection module based on spatiotemporal motion energies. Experimental results are shown in Section 4 and the conclusion is drawn in Section 5.

2. Temporal Boosted Learning

In this section, we first describe the features used for recognizing the five predefined target classes with our system. Once the features utilized in this work have been detailed, the actual temporal boosted classifier is described.

2.1 Moving Target Features

In real-time scenarios, it is important to use features that are computationally inexpensive and invariant to lighting condition. Therefore, in this work, the features employed include: (i) the

eccentricity of the bounding ellipse of a moving target; (ii) the orientation of the major axis of the bounding ellipse; (iii) the peak position of the normalized horizontal and vertical projection of a moving target; (iv) the pixel percentage of the peak in the normalized horizontal and vertical projection; (v) the difference in pixel density within the moving target's bounding box for two consecutive frames; (vi) the difference between the first four eigenvalues (computed by PCA) for the moving target in two consecutive frames. Therefore, the total dimensionality of the feature vector is 11.

2.2 Temporal Boosted Learning Algorithm

For our primary classification module, we use a modified version of AdaBoost [7] as the underlying learning algorithm. The primary and significant alteration we make to the AdaBoost paradigm is the inclusion of dynamic target information so that a so-called temporal boosted binary classifier is constructed for each object category.

In this work, we propose a new learning strategy. Usually, temporal coherence/ incoherence is ignored in the training process. However, the temporal information is particularly important in moving target classification since a single frame will often contain insufficient or unreliable data for proper recognition. Therefore, we propose a temporal boosted learning algorithm. Moving objects are tracked using our proposed tracking method [16]. For the training algorithm, the error function is defined to analyze $2s+1$ consecutive frames so that temporal consistency of the object class can be evaluated. In this work, we train five classifiers, one for each object category (**V**ehicle, **B**icycle, **P**erson(s), **M**otorcycle, and **P**erson with **U**mbrella). The details of the algorithm are as follows:

Temporal Boosting Learning Algorithm

Input: (1) sequence of N labeled examples $\{(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)\}$

(2) distribution D over the N examples

(3) weak learning algorithm **WeakLearn**

(4) integer T specifying number of iterations

Initialization: the weight vector:

Do for $t=1, 2, \ldots, T$

1. Set $p^t = w^t / \sum_{i=1}^{N} \omega_i^t$
2. Call **WeakLearn**, providing it with the distribution $\boldsymbol{p}^t$; get back a hypothesis $h_t: X \rightarrow [0,1]$.
3. Calculate the error of h_t:
$$\varepsilon_t = \sum_{i=1}^{N} p_i^t \, round\left(\left| \frac{1}{2s} \sum_{f=-s}^{s} h_t(x_{i+f}) - y_i \right| \right).$$
4. Set $\beta_t = \varepsilon_t / (1-\varepsilon_t)$.
5. Set the new weights vector to be
$$\omega_i^{t+1} = \omega_i^t \beta_t^{1-|h_t(x_i)-y_i|}$$

Output: the hypothesis
$$h_f(x) = \begin{cases} 1 & \text{if } \sum_{t=1}^{T} \left(\log \frac{1}{\beta_t}\right) h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \log \frac{1}{\beta_t} \\ 0 & \text{otherwise} \end{cases}.$$

If a moving target is hypothesized to belong to more than one category, the final decision, the classified category, C, is the class that yields the maximum output value according to

$$C = \underset{c=\{V,B,P,M,PU\}}{\arg\max} \sum_{t=1}^{T} \frac{\alpha_t}{\sum_{t=1}^{T} \alpha_t} h_t^c(x), \qquad (1)$$

where the weighting factor α_t is normalized by the summation of all α for each iteration.

3. Moving Crowd Detection Using Spatiotemporal Energies

The second stage of our system is to further process targets which are identified as the "Person(s)" class in the first stage. The purpose of this additional processing is to determine if a target under analysis is a single moving person or multiple individuals. We shall first describe the self-similarities based approach in Section 4.1. Section 4.2 describes the computation of motion energies and how they can be employed for detecting moving crowds.

3.1 Self-Similarity based Moving Crowd

Detection in the Spatial Domain

In Fig. 1, we can clearly notice that under a somewhat elevated, top-down viewpoint, significant portions of each individuals' torso is visible. The local intensity patterns exhibited by torso regions are often repeated in nearby image locations of moving crowds. Therefore, the problem of moving crowd detection can be transformed to the problem of computing local self-similarities in a region of interest.

When measuring the self-similarity within a region of interest (ROI), we have a template image A and a target image G. The similarity between template A with size $m \times n$ and a patch B with size $m \times n$ within image G can be revealed using straightforward similarity measures, such as a simple correlation measurement, defined as[16]

$$r = \sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B}) \Big/ \sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)} \quad (2)$$

where $\overline{A}$ and $\overline{B}$ are the average of template A and a patch B, respectively.

However, the value of the correlation measurement defined in (2) depends on the color consistency between the template and the patches within an ROI. In order to enhance the color contrast between neighboring uniform color regions, we use a histogram equalization approach. This technique enhances the ROI by mapping the gray levels to a small range of n levels. An example of ROI enhancement is shown in Fig. 1. Figure 1(b) shows an ROI and Fig.1(c) shows the corresponding enhanced image.

To compute the self-similarities within an ROI, a template that represents the rough approximation of a person's head and torso is used. Figure 1(a) shows the specific template. The correlation surface computed using (2) is shown in Fig.1(d). Since people are of many different sizes and can be found at various distances from the camera, we compute convolution surfaces using the template of Fig. 1(a) at multiple scales. To combine information from multiple scales, we calculate the entropy E_{Ai}.

$$E_{Ai} = -\sum_{(x,y)\in(X\times Y)} \hat{G}_r(x,y) \log \hat{G}_r(x,y) \cdot \quad (3)$$

We compute the entropy of the normalized correlation surface, $\hat{G}_{r,Ai}$, to determine the importance of a template scale, A_i. Thus, each correlation surface $\hat{G}_{r,Ai}$ is weighted according to its entropy and combined to form a normalized correlation surface $\overline{G}_r$, as follows:

$$\overline{G}_r(x,y) = \sum_{i=1}^{s} E_{Ai} \times \hat{G}_{r,Ai}(x,y) \cdot \quad (4)$$

Hence, a region is labeled as a candidates crowd if there is a unique peak detected in the neighboring region.

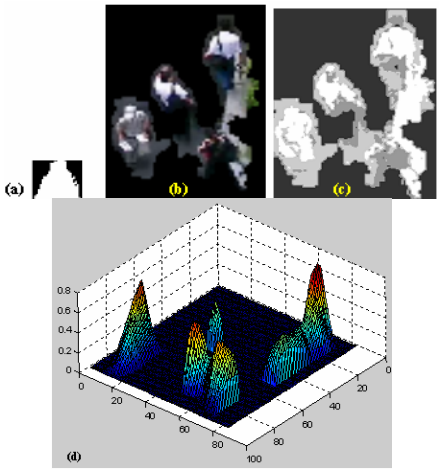

Fig. 1. Self-similarities in a moving crowd. (a) A single template in the shape of head and torso (30×30) (b) A foreground region with a moving crowd. (c) Enhanced contrast of the foreground region in (b) obtained using histogram equalization and $n=5$ intensity levels. (d) Self-similarities are computed by the correlation between the template and image in (c) (red corresponds to the highest values).

3.2 Moving Crowd Detection in the Spatiotemporal Domain using Motion Energies

Once the candidate crowd regions have been identified using our self-similarity based approach, spatiotemporal analysis via motion energies can be used to eliminate false positives and detect true crowds. In this section, we shall first describe the technique of computing oriented energies. Subsequently, we will propose a method of processing these motion energies for the

application of detecting crowds.

3.2.1 Oriented Energy Computation

When performing spatiotemporal analysis, significant information can be obtained by filtering the spatiotemporal volume representation of a video sequence with orientation selective filters. For this work, the filtering of space-time volumes was performed using broadly tuned, steerable, separable filters based on the second derivative of a Gaussian, G_2, and their corresponding Hilbert transforms, H_2 [5], with responses pointwise rectified (squared) and summed. Filtering was completed across a total of four 3D orientations $\theta = (\eta, \xi)$ where η and ξ specify polar angles. The four orientations that were selected correspond to upward, downward, leftward, and rightward motion. Thus, a measure of local motion energy, *e*, can be computed using

$$e(x;\theta) = [G_2(\theta) * I(x)]^2 + [H_2(\theta) * I(x)]^2, \quad (5)$$

where $x = (x, y, t)$ are spatiotemporal image coordinates, I is the image sequence, and * denotes convolution. This initial measure of local energy, (5) is dependent on image contrast. However, a purer measure of oriented energies that is less affected by contrast can be obtained through normalization,

$$\hat{e}(x;\theta) = e(x;\theta) / \left(\sum_{\tilde{\theta}} e(x;\tilde{\theta}) + \varepsilon\right), \quad (6)$$

where ε is a small bias term to prevent instabilities when the overall energy content is small and the summation in the denominator covers all orientations. To ensure that the energies obtained are valid for an entire candidate crop box, additional padding pixels (both in space and in time) must be used during filtering. Specifically, in all dimensions, padding of size $f/2$ must be used, where f is the length of the separable 3D filters [9]. In this work, a padding size of 6 was needed, which means that 6 frames on either side of the current frame (yielding a total of 13 frames) were used when computing the motion energies. Figure 2 provides an illustrative example of the motion energies that are utilized throughout the remainder of the crowd detection system.

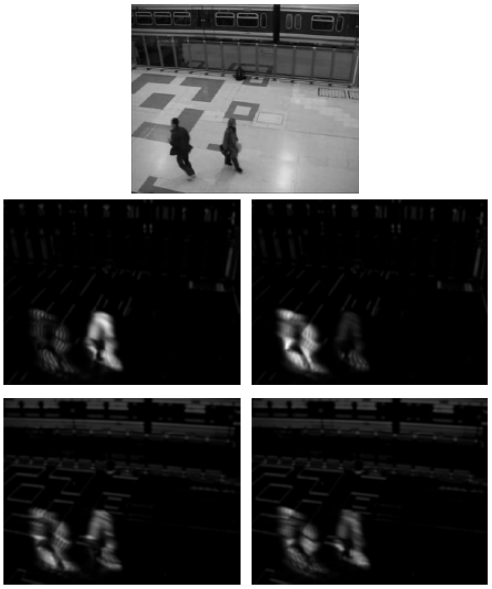

Fig.2. A sample frame of the PETS 2006 [12] data set with its corresponding motion energies. The final two rows show the rightward, leftward, downward, and upward motion energies, from left to right, top to bottom.

3.2.2 Moving Crowd Detection Using Motion Energies

Once the motion energies for a candidate region have been computed, they can be used as a vehicle for classifying between crowds and non-crowds. Our proposed moving crowd detector evaluates three criteria to determine if a candidate is, in fact, a crowd. First, we check whether there is a significant amount of motion energy in more than one orientation within the candidate region. Second, we determine if there are any large separations between regions that contain strong energy responses. Third, we check whether there are motion energy signatures that are too complex to represent a single person. If any of the three criteria are satisfied, the candidate is classified as a moving crowd. In what follows, we shall describe in detail how these criteria are realized.

- *Criterion #1: Multiple Dominant Motion Orientations in Target Region*

The first criterion is to analyze the number of orientations that contain significant levels

of energy. Specifically, for each of the four orientations, we compute the sum of the energies across the target support

$$S(\theta)=\sum_{i=1}^{n}\hat{e}(X_i^*;\theta), \quad (7)$$

where $X_i^*=(x^*,y^*)$ is a single target candidate pixel at some temporal instant and i ranges such that X_i^* covers the target's support.

The summed energies of (7) can be represented as a normalized histogram and can subsequently be used to determine if a crowd is present within a candidate crop box. Typically, if a candidate region contains a single moving object, one motion channel will contain the vast majority of the energy. Thus, if there are two or more orientations with substantial motion energy, we conclude that there are multiple people in the target region. Correspondingly, we use the following rule to decide if a crowd is present:

$$D_1=\begin{cases}1, \text{if } \left[\max_{i\in[1\ldots4],\theta_i\neq\theta_{\max}} S(\theta_i)\right] > \alpha S(\theta_{\max}) \\ 0, \quad \text{otherwise}\end{cases}, \quad (8)$$

where α is an empirically selected threshold and θ_{max} is defined as the orientation with the maximal summed energy response across the target support. Furthermore, D_1 = 0 and D_1 = 1 indicate that a crowd is not present and present by Criterion #1, respectively.

- *Criterion #2: Detecting Separations between High Motion Energy Regions*

The goal of the second criterion is to recognize crowds when the individuals involved are far apart from one another. For this work, we assume that the camera angle is similar to that of the PETS 2006 image shown in Fig. 2. Under this assumption, we project the single dominant motion energy of the candidate crowd region onto the X-axis. Projection is performed onto the X-axis because most crowds will predominately contain motion in the horizontal directions.

The ideas of Criterion #2 can be presented more formally as

$$p(x^*)=\sum_{\tilde{y}}\hat{e}(x^*,\tilde{y};\theta_{\max}), \quad (9)$$

where $\tilde{y}$ varies over all rows in the candidate crowd region and x^* corresponds to a column of the region. The projected energies can be visualized as a normalized histogram. Notice how the projected energies are close to zero for the image columns between the two individuals. The goal of our second criterion is to identify when these low energy "gaps" occur. If such a gap exists, it is concluded that the candidate region is a crowd. Mathematically, our methodology for finding gaps can be written as

$$D_2=\begin{cases}1, \text{if } p(x_i^* < \beta P_{\max,} \text{ where } i\in 1,2,\ldots C) \\ 0, \quad \text{otherwise}\end{cases}, \quad (10)$$

where x_i^* corresponds to a column of the candidate crop box, β is a user-defined threshold, i varies over all C columns in the target support, and $P_{\max}$ is the largest projected sum, (9), for any of the candidate's columns. Furthermore, D_2 = 0 and D_2 = 1 indicate that a crowd is not present and present by Criterion #2, respectively.

- *Criterion #3: Detecting Complex Energy Patterns*

In a similar manner to Criterion #2, the third criterion is employed in cases where the crowd candidate contains motion in only one dominant orientation. The third criterion in our system detects crowds based on the fact that the projected energy signatures of a crowd are typically much more complex than those created by a single person consisting of multitude peaks and valleys.

4. Experimental Results

In this section, we evaluate the performance of the temporal boosted classifier and the spatiotemporal-based moving crowd detector.

4.1 Performance of Temporal Boosted Classifier

In the training process, 200 samples were used for each object category. Since our system design uses multiple one-against-all classifiers, 200 samples for one category

were included as positive examples while the other 800 samples from the other four categories were used as negatives. In the testing process, we made the system run for several days from 9 am to 5 pm under distinct weather conditions, including sunny, cloudy and rainy days.

In order to compare with the proposed temporal boosted classifier, the original AdaBoost algorithm [7] was implemented. Table 1 shows the confusion matrix obtained using the original AdaBoost algorithm shown in light gray and that obtained when using our proposed method shown in deep gray, respectively. It is clear that the classification rate was substantially improved when temporal coherence was put into consideration.

TABLE 1. CONFUSION MATRIX OF MOVING TARGET CLASSIFICATION OBTAINED BY USING ADABOOST.M1 [7] AND BY USING THE PROPOSED TEMPORAL BOOSTING ALGORITHM

	V	B	P	M	PU
V	100% 100%	0	0	0	0
B	0	89.3% 94.6%	0	10.7% 5.40%	0
P	0	0.20% 0	99.8% 100%	0	0
M	0	8.60% 8.0%	0	91.4% 92.0%	0
PU	0	0	6.90% 3.30%	0	93.1% 96.7%

Among the categories that gained improvement, we analyzed the situation between the bicycle and motorcycle classes in more detail. For the bicycle, it has a sparse wheel structure. In contrast to a bicycle, the wheel structure of a motorcycle is relatively dense. However, a sparse wheel cannot always be obtained for a bicycle due to lighting conditions and the cluttered background. Under these circumstances, temporal coherence associated with a tracked moving target is particularly important for improving the classification accuracy.

4.2 Detection of Moving Crowds

The performance of the moving crowd detection system was tested on a variety of video sequences within the PETS 2006 data set. The following parameters were empirically selected and were held constant for all experiments: $\alpha = 1/3$ and $\beta = 0.05$.

The moving crowd detection system was evaluated qualitatively using the videos from the PETS 2006 database. Specifically, we wanted to observe that the system was capable of identifying crowds that satisfied each of the three separate criteria. Furthermore, it had to be ensured that the system did not produce a large number of false positives. Consequently, the system was presented with many examples of single moving persons to ensure that they were not misclassified as crowds. Fig.3 shows a variety candidate crowd and non-crowd regions obtained from the PETS 2006 videos.

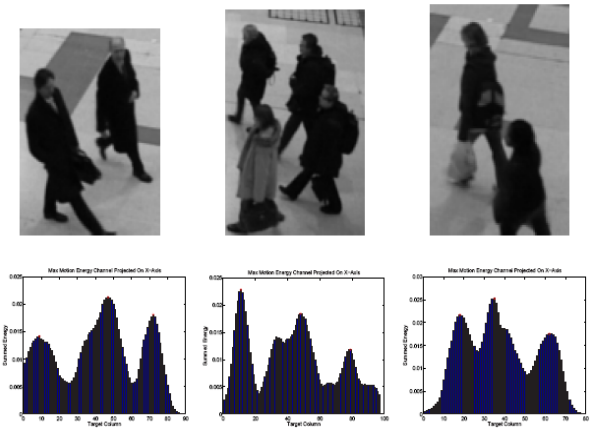

Fig.3. Examples of candidate crowd regions and their projected dominant energy. These examples were correctly classified using Criterion #3.

The performance of the crowd detection module was also measured quantitatively on the PETS 2006 dataset. In total, our test set was comprised of 111 crowd data samples and 179 non-crowd samples. Of the non-crowd examples, 6 contained only background portions of the scene, while 21 displayed less than half of a single moving person. The remaining non-crowd examples showed a single moving individual.

The overall classification rate on all data samples (crowd and non-crowd) was computed to be 85.2%. The confusion matrix of Table 2 provides additional details

regarding the system's performance for the two classes. As can be seen, the false negative rate was slightly higher than the false positive rate. The general limitations of the system described above were indeed the cause of the majority of the misclassifications during our quantitative evaluation. Nonetheless, the overall performance of our crowd detection system was quite competitive when compared to other such systems in the literature.

TABLE 2. CONFUSION MATRIX FOR OUR PROPOSED MOVING CROWD DETECTION SYSTEM ON THE PETS 2006 DATASET

	Crowd	Non-Crowd
Crowd	82%	18%
Non-Crowd	12.8%	87.2%

5. Conclusion

In this paper, a novel approach has been proposed for detecting and classifying targets of interest into predefined categories according to their spatial-temporal properties. A novel temporal boosted classifier was designed and used to classify moving targets into five categories: person(s), bicycle, motorcycle, vehicle, and person with umbrella. Significantly, the proposed method improves weak classifiers by allowing them to use information from previous and future frames when processing the current frame. Additionally, a hierarchical system is proposed to further analyze regions which are known to exhibit human motion. This system uses a combination of template matching and powerful spatiotemporal oriented energies to differentiate between single persons and crowds. Experiments demonstrate that our system is extremely effective both at identifying the class of objects as well as detecting human crowds.

ACKNOWLEDGMENT

This work is supported by the Ministry of Economic Affairs under Contract No. 96-EC-17-A-02-S1-032.

References

[1] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for Cooperative Multisensor Surveillance," *Proc. of IEEE*, Vol. 89, No.10, pp. 1456-1477, Oct. 2001.

[2] O. Javed, and M. Shah, "Tracking and Object Classification for Automated Surveillance," Proc. of *ECCV*, pp. 439-443, May 2002.

[3] P. Viola, M. J. Jones, and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance," Proc. of *IEEE ICCV*, 2003.

[4] M. Shah, O. Javed, and K. Shafique, "Automated Visual Surveillance in Realistic Scenarios," *IEEE Multimedia*, Vol.14, No.1, pp. 30-39, Jan.-Mar. 2007.

[5] C. Papageorgiou and T. Poggio, "Trainable Pedestrian Detections," Proc. of *ICIP*, 1999.

[6] H. Schneiderman and T. Kanade, "A Statistical Method for 3D Object Detection Applied to Faces and Cars," Proc. of *CVPR*, 2000.

[7] Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and An Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].

[8] E. Adelson and J. Bergen, "Spatiotemporal energy models for the perception of motion," *JOSA A*, 2(2):284–299, 1985.

[9] K. Derpanis and J. Gryn, "Three-dimensional nth derivative of Gaussian separable steerable filters," *ICIP*, 3:553–556, 2005.

[10] M. Enzweiler, R. Wildes, and R. Herpers, "Unified target detection and tracking using motion coherence," *Wrkshp. Motion & Video Comp.*, 2:66–71, 2005.

[11] D. Heeger, "Optical flow from spatiotemporal filters," *IJCV*, 1(4):297–302, 1988.

[12] PETS.http://www.cvg.rdg.ac.uk/PETS2006/data.html, 2006.

[13] R.Wildes and J. Bergen, "Qualitative spatiotemporal analysis using an oriented energy representation," *ECCV*, 2:784–796, 2000.

[14] Y. T. Hsu, J. W. Hsieh, and H. Y. Mark Liao, "Video-based human movement analysis and its application to surveillance systems," Accepted and to appear in *IEEE Trans. on Multimedia*.

[15] J. S. Bendat, and A. G. Piersol. Engineering applications of correlation and spectral analysis. New York, Wiley-Interscience, 1980.

[16] H. Y. Mark Liao, D. Y. Chen, C. W. Su, and H. R. Tyan, "Real-time Event Detection and Its Application to Surveillance Systems, P*roc. IEEE International Symposium on Circuits and Systems*, Kos, Greece, May 21-24, 2006.