Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2009, Article ID 164019, 12 pages
doi:10.1155/2009/164019

Research Article

RVM-Based Human Action Classification in Crowd through Projection and Star Skeletonization

B. Yogameena, S. Veeralakshmi, E. Komagal, S. Raju, and V. Abhaikumar

Department of Electronics and Communication Engineering, Thiagarajar College of Engineering, Madurai 625015, Tamil Nadu, India

Correspondence should be addressed to B. Yogameena, ymece@tce.edu

Received 1 February 2009; Revised 17 May 2009; Accepted 26 August 2009

Recommended by Amit Roy-Chowdhury

Detection of abnormal human actions in the crowd has become a critical problem in video surveillance applications like terrorist attacks. This paper proposes a real-time video surveillance system which is capable of classifying normal and abnormal actions of individuals in a crowd. The abnormal actions of human such as running, jumping, waving hand, bending, walking and fighting with each other in a crowded environment are considered. In this paper, Relevance Vector Machine (RVM) is used to classify the abnormal actions of an individual in the crowd based on the results obtained from projection and skeletonization methods. Experimental results on benchmark datasets demonstrate that the proposed system is robust and efficient. A comparative study of classification accuracy between Relevance Vector Machine and Support Vector Machine (SVM) classification is also presented.

Copyright © 2009 B. Yogameena et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Security of citizens in public places such as Hotels, Markets, Airports, and Train stations is increasingly becoming a crucial issue. A number of video surveillance systems for multiple people detection and tracking in a crowded environment have been reported in literature [1–3]. Describing an unusual activity or behaviour from a video is a challenging problem. It involves representation and interpretation of the visual information for behaviour learning and recognition [4–8]. Turaga et al. have summarized the approaches that have been pursued over the last 20 years to address the problem of activity recognition. They have discussed the problem at two levels of complexity: "actions" and "activities." "Actions" are characterized by simple motion patterns typically executed by an individual whereas "activities" involve coordinated actions among a group [9].

Though alternate methods are available, background subtraction continues to be a method of importance in video surveillance. Piccardi has reviewed a number of background subtraction approaches [10]. Wren et al. [11] have proposed a statistical method, in which a single Gaussian function was used to model the distribution of background. Later Mittal and Paragios have proposed a novel kernel-based multivariate density estimation technique that adapts the bandwidth according to the uncertainties [12]. Yet there are issues like the robustness to illumination changes, the effectiveness in suppressing shadows, and the smoothness of foreground's boundary which need to be addressed in indoor and outdoor environments [13].

The real-time visual surveillance system employs combination of detection and tracking group of people as well as monitors their activities even in the presence of occlusion. However, labeling an individual becomes less feasible in the case of crowds, where people are typically severely occluded. Wren et al. applied a statistical model for color and shape to segment a person, tracked heads and hands, and identified gestures, mainly dealt with individuals [11]. W^4 was another system for detecting and tracking individuals based on shape models [14]. Zhao and Nevatia proposed a Bayesian model-based segmentation algorithm using shape models which segmented each individual from a scene. This method was based on Markov Chain Monte Carlo sampling and was prohibitively slow for large crowds [15]. The major drawback of most of the methods mentioned above is that they assume that there is a distinct visual separation

between individuals, so that the motion-segmented image contains enough visual information to separate individuals moving as a group. However, this is not always true in dense groups, when people are visually inseparable. A novel approach to segment individual in the crowd is required which reduces the effects of certain problems like occlusions and overlap which are faced by conventional techniques.

Many action recognition methods have been applied to classify an individual's action as normal or abnormal. Most frequent activities like walking and sitting are considered as normal and the activities like jumping, running, waving hand, bending, and fighting are considered as abnormal. A method to tackle activity recognition using descriptive local features of actions performed by humans at multiple scales and temporal speeds has been proposed [16]. Song et al. [17] used a triangular lattice of grouped point features to encode layout. Lee and Xu proposed a method using the velocity of body parts for learning human actions [18]. Stauffer and Grimson [19] proposed a system which accumulates joint cooccurrence statistics to create a hierarchical binary-tree classification of the representations for classifying sequences, as well as individual instances of activities in a site. The feature selection also plays an important role for any classification system. In [20] the authors proposed Hidden Markov Models for action classification using the features based on the position and velocity of body parts for learning. Wu et al. proposed a method which employs optical flow for detecting abnormal human blobs and Principal Component Analysis for feature selection and Support Vector Machine for classification of human actions. The number of primary features, the selection of primary features, and loss in the original information content formulate the PCA feature selection to be more complex [21]. Therefore features which significantly improve the classification accuracy with reduced complexity are essential.

There are many types of neural networks that can be used for a binary classification problem, such as Support Vector Machines, Radial Basis Function Networks, Nearest Neighbor Algorithm, and Fisher Linear Discriminant. Machine learning techniques can be considered as linear methods in a high-dimensional feature space nonlinearly related to the input space. Using appropriate kernel functions, it is possible to compute the hyperplane which separates the two classes. Consequently a machine learning technique which minimizes the number of active kernel functions to reduce computation time is required [22–24]. The objective of this paper is to classify normal and abnormal actions of an individual in a crowd using appropriate feature selection and classification technique to improve the classification accuracy. Hence a novel system is proposed to classify the action of an individual. First the foreground blobs are detected using a background subtraction technique which is to be robust to illumination changes and shadows. Then projection is used to segment an individual in the crowd even in the presence of occlusion. Finally skeleton features and RVM are used to improve the classification accuracy and thereby reducing the computational complexity by selecting an appropriate kernel. Experimental results demonstrated that the proposed approach is robust in classifying human actions.

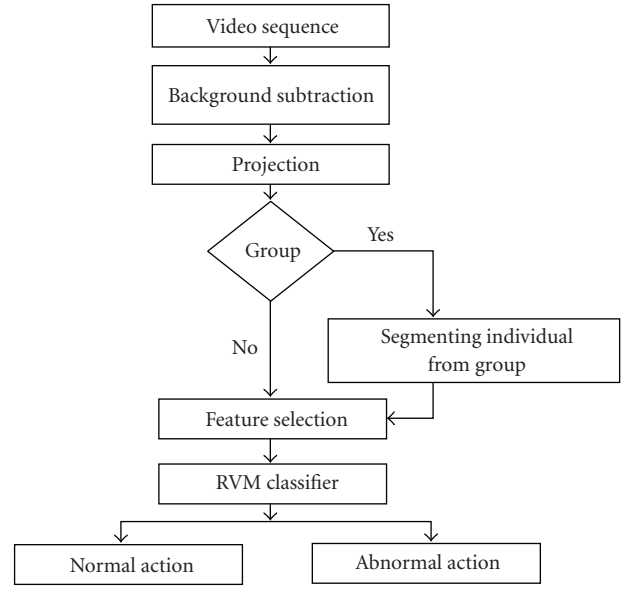


Figure 1: The block diagram of the Video Surveillance System.

The remainder of this paper is organized as follows. Section 2 describes the methodology. Section 3 describes background subtraction and projection techniques to identify individuals in groups and explains star skeleton feature extraction method. Section 4 depicts RVM learning system for classification of human actions. Section 5 discusses the experimental results. Finally, the conclusion is presented.

2. Methodology

An overview of the system is shown in Figure 1. The first stage of the surveillance system is background subtraction. This blob detection subsystem detects the foreground pixels by subtracting a statistical background model. Then, the foreground pixels are grouped into blobs. Each foreground region is labelled as an individual or a group. These classifications are based on the projected sizes and velocities of the regions. In the second stage, the foreground blob containing multiple people is divided using a projection method such that the individuals are identified. The tracker is then automatically initialized for each foreground blob that is identified as an individual. Then the individual's action is classified as normal or abnormal action using Relevance Vector Machine (RVM). Any generic machine learning system needs features to detect abnormal actions. The skeleton points and the motion cues for each blob are selected as features. Moreover the classification accuracy is improved by choosing appropriate kernel function and relevant input vectors.

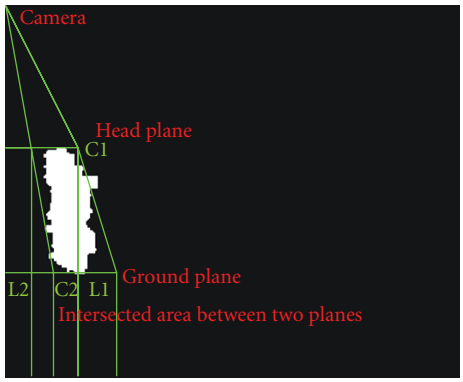


Figure 2: Intersected area between the ground plane and the head plane projection of an individual.

3. Action Analysis

3.1. Background Subtraction and Projection. In this work, background substraction is accomplished in real-time using the adaptive mixture of Gaussians method proposed by Atev et al. [25]. There are some of the practical issues concerning the use of the existing algorithm based on mixtures of Gaussians for background segmentation in outdoor scenes, including the choice of parameters [26]. The proposed system analyzes the choice of different parameter values and their performance impact is obtained to get robust background model. In addition, the motivation for adopting this method stems from its simplicity and efficiency in meeting with sudden global illumination changes based on the contrast changes over time.

Subsequently, the individual is to be identified for further analysis leading to action classification. An extracted blob in a frame, representing an individual is subjected to action analysis described in subsequent sections. If there exists more than one blob, but with connectivity, there is a likelihood to be considered as single entity. This results in the identification of a group as "individual." This makes recognition of individual's action in a crowd more difficult. Therefore, a geometric projection on the blob is proposed to separate an individual from the group for analysing his or her actions. The blob is projected to head and ground plane from the camera view point leading to intersected area in world coordinates. Such projection shown in Figure 2 eliminates the variation of area with the distance from the camera so that it identifies only humans [27]. The success of human identification lies on segmentation of individual human in a given frame as a single blob. However, there is a chance of multiple blobs representing an individual human due to oversegmentation. But, since the projection of a blob is accomplished from head plane to ground plane, any discontinuity in a blob representing an individual is achieved by linking discontinuous blobs covered by bounding rectangle.

3.2. Individual and Group Identification. The head plane is fixed such that all the individuals in the scene are detected.

Figure 3: Examples for Labelled as "Group."

Having a head plane height too large will result in zero-intersected area for shorter people. On the other hand, setting very small head plane heights will result in detecting shorter objects. Therefore, a head plane height of 160 cm is assumed as a balancing height for the intersected area computation [27]. This makes the method robust to false detections from other objects commonly found in urban environments and also reduces the effects of shadows on the ground plane in cases where the shadow of a person or group appears shorter in the image than the height of the person or group. For all blobs a rectangular area is formed by connecting the opposite points C1 and C2. If this area is less than the area threshold, then the region is classified as an individual. Subsequently, the classified individual foreground blobs are tracked using the centroid. The blobs whose projected area exceeds a threshold are classified as a group and are shown in Figure 3. The threshold is selected to be just under the area corresponding to two individuals in the real world. The individuals in a group are recognized and labelled by (1), (2) and (3).

$$\text{Projected Area in the Head Plane} = \text{C1} - \text{L2}, \quad (1)$$

$$\text{Projected Area in the Ground Plane} = \text{L1} - \text{C2}, \quad (2)$$

$$\text{Resultant (overlapped) Intersected portion} = \text{C1} - \text{C2}, \quad (3)$$

where C1 and L2 are the projected points of the head and Leg on the Head Plane, respectively; L1 and C2 are the projected points of the head and Leg on the Ground Plane, respectively.

Given the intersected area of the group being labelled, the count estimation is applied for this current frame using (4) and is shown in Figure 4:

$$\text{Count} = \frac{\text{Area}}{K}, \quad (4)$$

where K is the individual's intersected area.

According to the estimated intersected area for an individual, the group is separated as individuals and is shown in Figure 5. The individuals are labelled as I1, I2, and so on. The tracker is then automatically initialized for each foreground blob that is identified as an individual and the count estimation is initiated. Each of these instantaneous estimates is stored, along with an initial estimate age of zero. Every time a new instantaneous estimate is stored, the ages of all the previous estimates are updated. This history of count estimates is attached to the tracker, and all of this together is referred to as a group tracker [27].

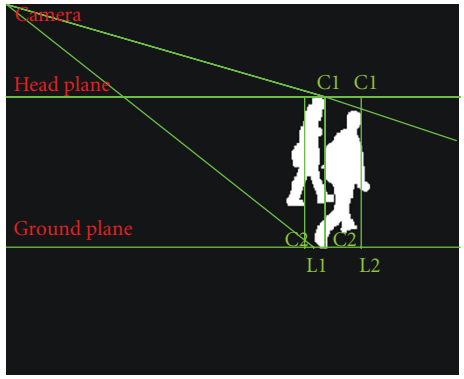


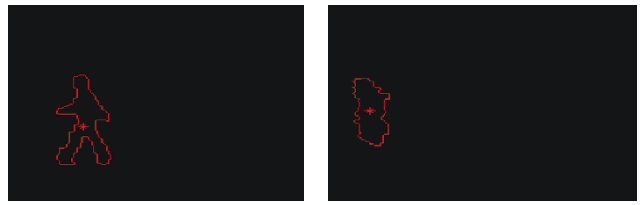
Figure 4: Intersected area between the ground plane and the head plane projection.

Figure 5: Segmented Individual.

3.3. Star Skeletonization. Consequently the individual is labelled then the system tracks consecutive blobs for that individual. Fujiyoshi et al. propose the use of star skeletonization procedure for analysing the motion of human targets. The standard star skeleton techniques for skeletonization such as distance transformation and thinning are computationally expensive and moreover are highly susceptible to noise in the target boundary. The method adapted in this paper provides a simple way of detecting only the gross extremities of the target to produce star skeleton. It also reduces the noise for the splotchy motion blobs such as Figure 2 by smoothing the distance from the centroid to contour points, d_i plot by moving average. The main idea is, the simple form of skeletonization extracts the broad internal motion features of a target and is employed to analyze the target's motion [28]. Then the contour for a human blob is extracted as shown in Figure 6. The centroid (x_c, y_c) of the human blob is determined by using the following (5) and (6) and is also shown in Figure 6:

$$x_c = \frac{1}{N}\sum_{j=1}^{N} x_i, \tag{5}$$

$$y_c = \frac{1}{N}\sum_{j=1}^{N} y_i, \tag{6}$$



(a) Apparent motion blob (b) Splotchy motion blob

Figure 6: Border points and Centroid for motion blobs.

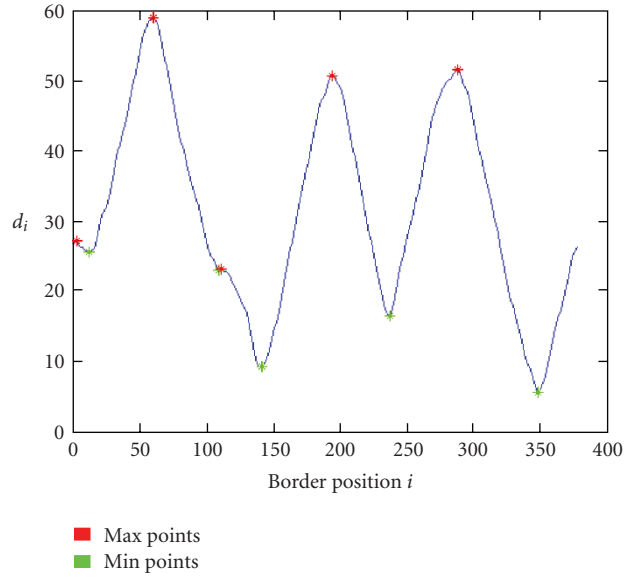


Figure 7: Plot of skeleton extreme points.

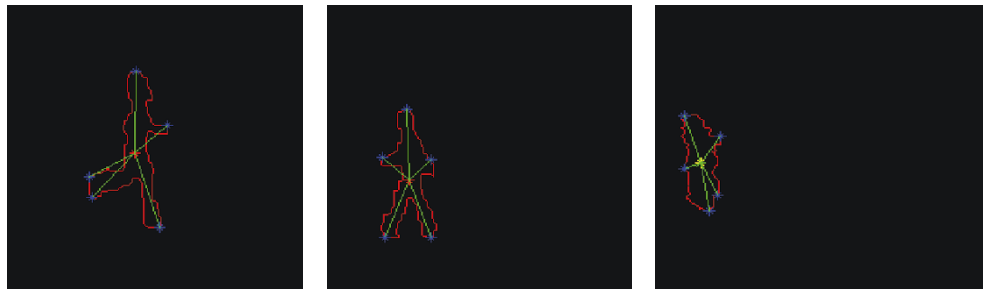
where (x_c, y_c) represent the average contour pixel position, (x_i, y_i) represent the points on the human blob contour and there are a total of N number of points on the contour. The distance d_i from the centroid to contour points is given by (7):

$$d_i = \sqrt{\left((x_i - x_c)^2 + (y_i - y_c)^2\right)}. \tag{7}$$

From the d_i plot, the local maximum points are collected and their corresponding plot is shown in Figure 7.

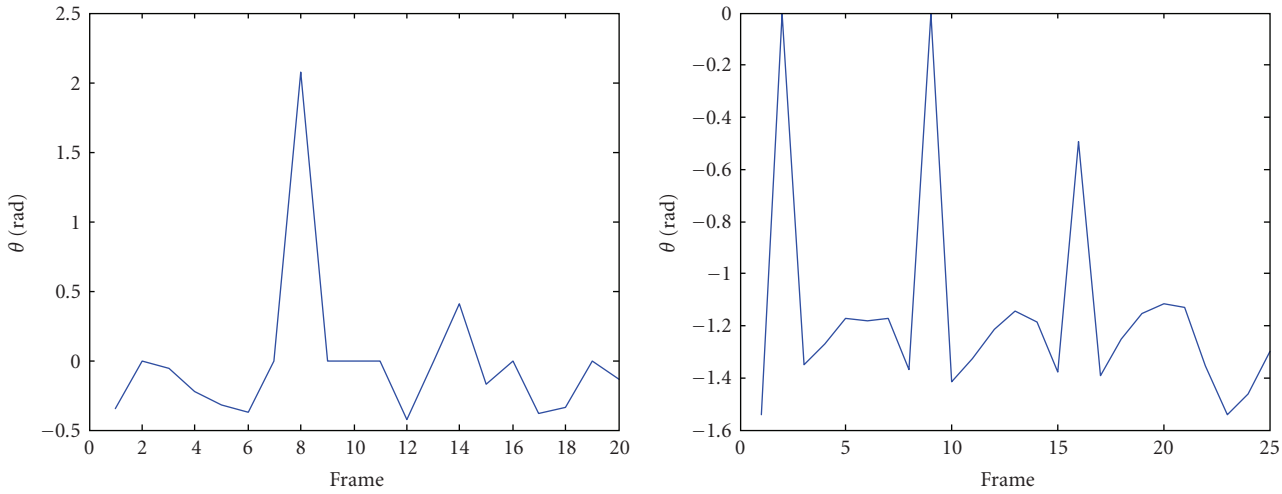
The star skeletonization is formed as shown in Figure 8.

Another prompt to analyze the motion of the target is its posture. Using motion cues based on the star skeleton, it is possible to determine the posture of a moving human. For the cases in which a human is moving in an upright position, it can be assumed that the lower extreme points are legs, and so choosing these points to analyze cyclic motion seems to be a reasonable approach [28]. In particular, the left-most lower extreme points (l_x, l_y) are used as the cyclic points. However, it is not necessary that



(a) Skeleton features for Running (b) Skeleton features for an apparent motion blob (c) Skeleton features for a splotchy motion blob

Figure 8: Skeleton features for different motion blobs.



(a) Motion analysis plot using leg angle for walking (b) Motion analysis plot using leg angle for running

Figure 9: Star skeletonization motion cues.

the same leg is detected at all times, because the cyclic structure of the motion will still be evident from this point's motion. If $\{(x_i^s, y_i^s)\}$ is the set of extreme points, (l_x, l_y) is chosen according to the following condition given in (8) and (9):

$$\left(l_x, l_y\right) = \left(x_i^s, y_i^s\right) : x_i^s = \min_{y_i^s < y_c} x_i^s. \tag{8}$$

Then, the leg angle θ is calculated by making use of the values of (l_x, l_y), such as

$$\theta = \tan^{-1} \frac{l_x - x_c}{l_y - y_c}. \tag{9}$$

A further cue to determine the posture of moving human is the inclination of the torso. This can be approximated by the angle of upper-most extreme point of the target. This torso angle Φ can be determined in exactly the same manner as θ and the leg angle for walking and running is shown in Figure 9. Another feature which can be clearly observed is that the frequency of the cyclic motion point is clearly higher in the case of running person; so this can be used as a good metric for the classification. The cutoff frequency was set as 0.1 to get appropriate extreme points in this proposed work. At last the leg angle θ, torso angle Φ, and the skeleton motion in a sequence are given as input vectors for the Relevance Vector Machine.

4. Classification Using RVM Learning

Posture classification is a key process for analyzing the human action. Computer vision techniques are helpful in automating this process, but cluttered environments and consequent occlusions often make this task difficult [29]. There are numerous methods for incremental model-based pose estimation where a model of an articulated structure (person) is specified [30–32]. Many types of neural networks are used for a binary classification problem like individual's activity classification as normal or abnormal. By training the systems, the difference between normal and abnormal

Table 1: Datasets from different outdoor sequences.

		Dataset	Indoor/outdoor	Sequence length	Frame size
College campus		An individual bends down while most of walking-DS-I(DS represents Dataset)	Indoor/outdoor	1628	240 × 320
		A person waving hand in a group-DS-II	Indoor/outdoor	864	240 × 320
		Running human in a group-DS-III	outdoor	1320	240 × 320
		A person carrying a bar in a group DS-IV	outdoor	920	240 × 320
Benchmark dataset	IBM Dataset	Two persons walking	Indoor	781	240 × 320
	Weizmann Dataset	Eli-Walk	Outdoor	645	240 × 320
		Eli-Run	Outdoor	712	240 × 320
		Moshe-Bend	Outdoor	786	240 × 320
		Eli-Jump	Outdoor	855	240 × 320
	CMU Dataset	Person A fights with person B	Indoor	1234	240 × 320
		Person A pulls the person B	Indoor	1065	240 × 320
	CAVIAR	An individual with his hand lifted up	Indoor	1187	240 × 320

human actions, the computational action models built inside the trained machines can automatically identify whether the action is normal or abnormal. The action classification system proposed in this paper is trained for both normal and abnormal actions so that testing becomes a two class hypothesis problem. SVM is classical training algorithm because it has stronger theory-interpretation and better generalization than the other neural networks mentioned earlier. The decision function of the SVM classification system cannot be much sparser; that is, the number of support vectors can be much larger. This problem can be partially overcome by the state-of-the-art model RVM.

The proposed Relevance Vector Machine (RVM) classification technique has been applied in many different areas of pattern recognition, including functional neuro images analysis [33], facial expressions recognition [34], and pose estimation [35]. The RVM is a Bayesian regression framework, in which the weights of each input vector are governed by a set of hyper parameters. These hyperparameters describe the posterior distribution of the weights and are estimated iteratively during training. Most hyper parameters approach infinity, causing the posterior distributions of the corresponding weights to zero. The remaining vectors with nonzero weights are called relevance vectors. RVM does not need the tuning of a regularization parameter and also the inversion of a large matrix is not required during the training phase. This makes this methodology appropriate for large datasets. In this paper, Relevance Vector Machine (RVM) technique is used for the classification of human action such as normal or abnormal.

5. Results and Discussion

The efficiency of the proposed system has been evaluated by carrying out extensive works on the simulation of the algorithm on benchmark datasets. In this paper the video files used are taken in both indoor and outdoor. It is assumed that according to the camera view point, people move on the ground plane in the real world. The method processes about 24 frames per second for colour images. To demonstrate the performance of the proposed method, different abnormal action sequences are taken and shown in Table 1. They are taken from Weizmann (http://www.wisdom.weizmann.ac.il) and CAVIAR (http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/) datasets for the action of an individual such as walking, running, jumping, bending, and a person lifting his hand. In order to reveal complexity, IBM dataset having multiple people walking and (http://domino.research.ibm.com/comm/research_projects.nsf/pages/s3.performanceevaluation.html) CMU (http://mocap.cs.cmu.edu) database which contains two persons are fighting with each other in one sequence and in another sequence a person pulling up the hands of other person is used.

The video sequences were converted into frames and the background subtraction was obtained using GMM. To evaluate the proposed approach, the foreground detection results were compared with different illumination conditions as shown in Figure 16. In the model, several parameters have been set. In our experiment, the model's sensitivity has been analyzed to each parameter by observing the variation of the false negative (foreground pixel that were missed, FN) and the false positive (background pixels that were marked as

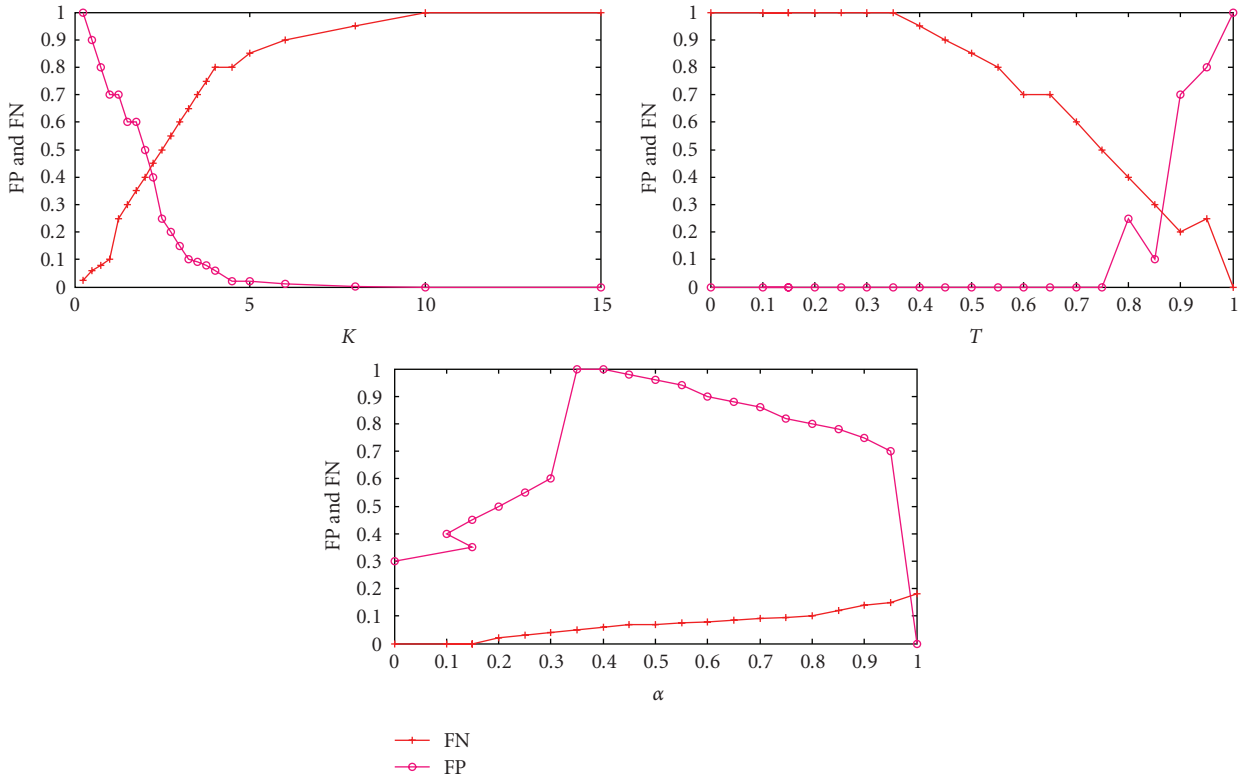
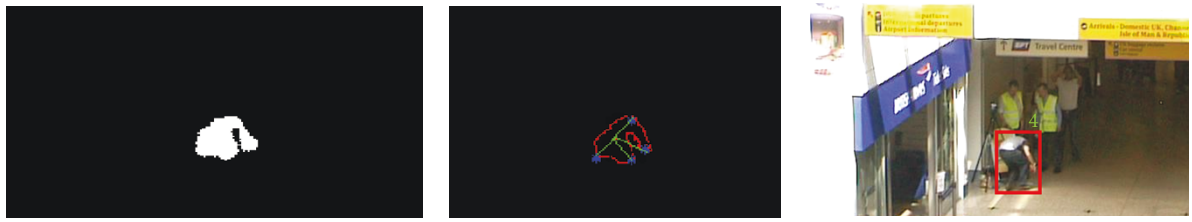


Figure 10: Parameter settings for GMM.



(a) Foreground image (b) Skeleton image (c) Abnormal classified image

Figure 11: A bending down movement of the person classified as abnormal action.

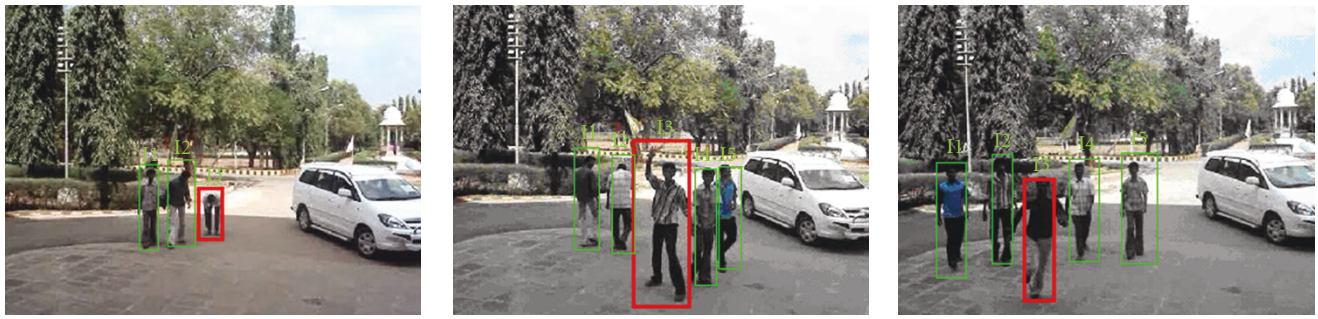
foreground, FP) caused by the change of one parameter while keeping the others unchanged. The measurements were performed on different video sequences. Figure 10 illustrates the FN and FP for different parameter values and clearly shows that 4 is suitable value for K (number of Gaussian mixtures) and 0.004 is an appropriate value for T. The proper values for the learning rate β and the match threshold are 0.001 and 0.4, respectively.

As mentioned in Table 1 dataset DS I, along with a person bending down blob two other foreground blobs were detected and in DS III five fore ground blobs were obtained including the running person's blob. Here, in DS I and in DS III, the car was in the static position. Hence it has been considered as background. Even when the datasets were taken in different illumination conditions, the foreground blobs obtained as shown in Figure 16 are robust. After foreground blobs were detected, using projection they were separated as individuals and groups. In DS II the persons away from the camera view point were easily identified as individuals and the remaining as group. Using the intersected area the individuals were separated from group and labeled as individuals. As illustrated in Figure 16 all individuals were clearly projected and also in the CMU Fighting dataset the individual who holds the stool and the other plunged down were separated.

Then the star skeletonization was used to obtain the motion cues as shown in Figures 9(a) and 9(b). For an individual blob like walking, running, and jumping, skeleton features have been obtained clearly. But for bending action in DSI, the skeleton features were varied from the Weizmann dataset. In the weizmann dataset the skeleton points were depicted as a short human being walking

Table 2: The results of RVM classification.

Datasets	Actions		Vectors		
			SVM		RVM
			Multiclass SVM with PCA features	Multiclass SVM with skeleton Features	Multiclass RVM with skeleton Points
College Campus	Normal	Walking	186	145	11
		Running	146	123	18
	Abnormal	Carrying Bar	212	133	23
		Bending	137	120	12
		Waving hand	174	118	16
Benchmark Datasets	Normal	IBM	183	138	12
		Eli-Walk	179	124	15
		Eli-Run	154	116	18
	Abnormal	Moshe-Bend	132	98	20
		Eli-Jump	145	112	16
		CMU1	157	87	19
		CMU2	165	76	14
		CAVIAR	176	94	12



(a) One person bending down (b) One person waving hand (c) One person running in the crowd

Figure 12: Results of classified abnormal actions for College Campus dataset.

where as Figure 11(b) shows the skeleton points of a man "bending." But in the CMU meeting dataset both were clearly skeletonized. However in the CMU Fighting dataset the person who holds the stool is skeletonized including the stool because of the connectivity. Finally the skeleton features and the motion cues were given as input for the relevance vector machine algorithm to classify the abnormal action from the normal one which was indicated in red color as shown in Figure 11 and Figures 12(a), 12(b), and 12(c) and also for benchmark datasets as shown in Figure 16.

Relevance Vector Machine uses a suitable kernel for the classification task. In this paper Gaussian kernel was used. To determine the fine-tuning parameters of the RVM classifier model for optimal performance, a tenfold cross validation (CV) has been applied in the training dataset. In each dataset, 80% of the sequence has been used for training and the remaining 20% for testing. The best error level 6.89% was obtained by using the Gaussian kernel. Gaussian kernel was the best with around 100.0% for training, 94.0% for CV, and 96.0% for testing the given feature vectors. The error rate of the Gaussian kernel was lower than that of other kernels in terms of classification rate. For comparison, an SVM classifier is trained using the same dataset. Indeed, the RVM classifier is much sparser than the SVM. The number of relevance vectors (produced during training) was found to be minimal for the RVM classifier, while comparing with SVM and is shown in Table 2. On an average, six datasets have been tested for classifying each activity like running, jumping, waving hand, and bending from weizmann dataset containing a total of 4 actions performed by 9 people totalling 36 videos.

The error levels for the above mentioned datasets are compared between the SVM and RVM classifiers as shown in Figure 13. The percentage of error level is considered for both

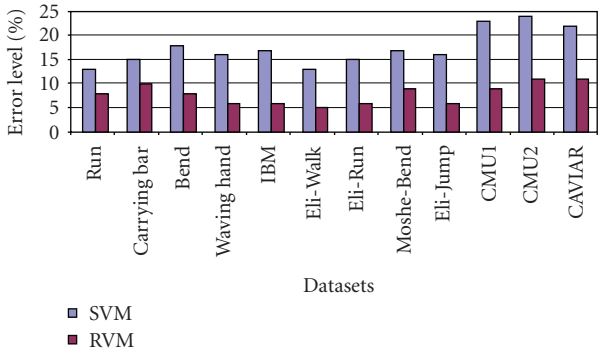
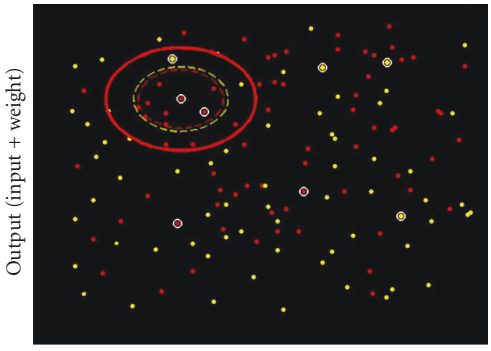
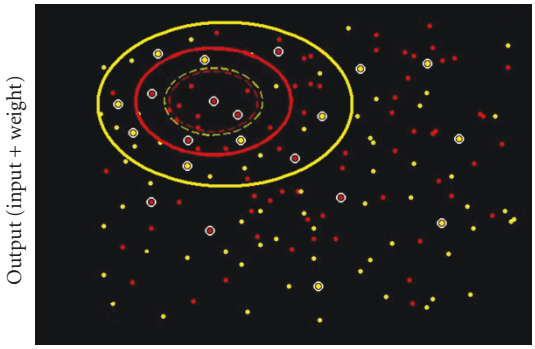


Figure 13: Comparison of RVM and SVM errors.



(a) Initial stage



(b) Sixth iteration

Figure 14: Relevance vector obtained after training.

training and testing sequences of the dataset in terms of True Positive and False Negative. Higher error level leads to poor action classification. In all the datasets the percentage of error level of SVM is higher than proposed RVM classification. Moreover it is inferred from Figures 14(a) and 14(b) that as the number of iteration increase the better convergence of relevance vectors is achieved for the experimental datasets.

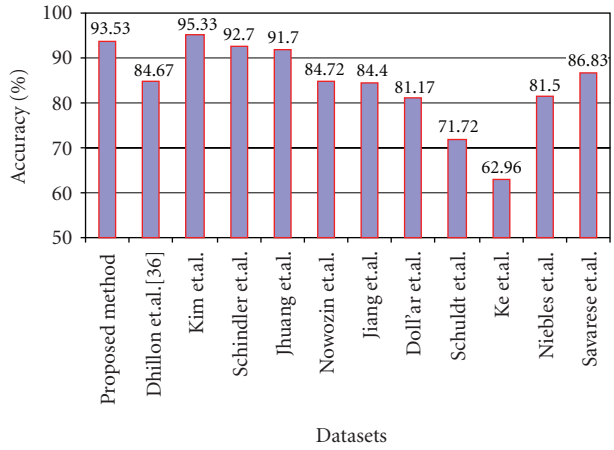


Figure 15: Comparative results of action classification.

To evaluate the proposed approach the classification accuracy has been computed and compared with existing state-of-the-art methods. On an average the performance of the linear SVM classifier [36] was 84.67%, for all activities, with a standard deviation of 0.56%, and the performance of the proposed method was 93.53% as shown in Figure 15. The majority of the abnormal actions are classified except the Weizmann bending and CMU pulling as illustrated in Figure 16. Weizmann bending is classified as normal action (may be walking) and also in CMU pulling dataset both individuals were considered as walking.

6. Conclusion

In this paper, a novel, real-time video surveillance system for classifying human normal and abnormal actions is described. First the foreground blobs are detected using adaptive mixtures of Gaussians which is to be robust to illumination changes and shadows. Then projection is applied to segment an individual in the crowd. This has helped to formulate the method to be robust from occlusion and false detections like other objects and shadows. Subsequently, skeleton features are extracted for each individual. These features reduced the training time and also improved the classification accuracy. The features are learnt through a relevance vector machine to classify the individual's actions into two classes. The number of relevance vectors obtained is smaller than SVM and it did not require the tuning of a regularization parameter during the training phase. The error rate is also reduced by selecting the appropriate kernel Gaussian which also reduces the computational complexity. The distinct contribution of the proposed work is in classifying the action of an individual in a crowded scene even in the case of partial occlusion. This facilitates the proposed system that is able to detect abnormal actions of an individual such as running, bending down, waving hand while others walk, and persons fighting with each other with high accuracy.

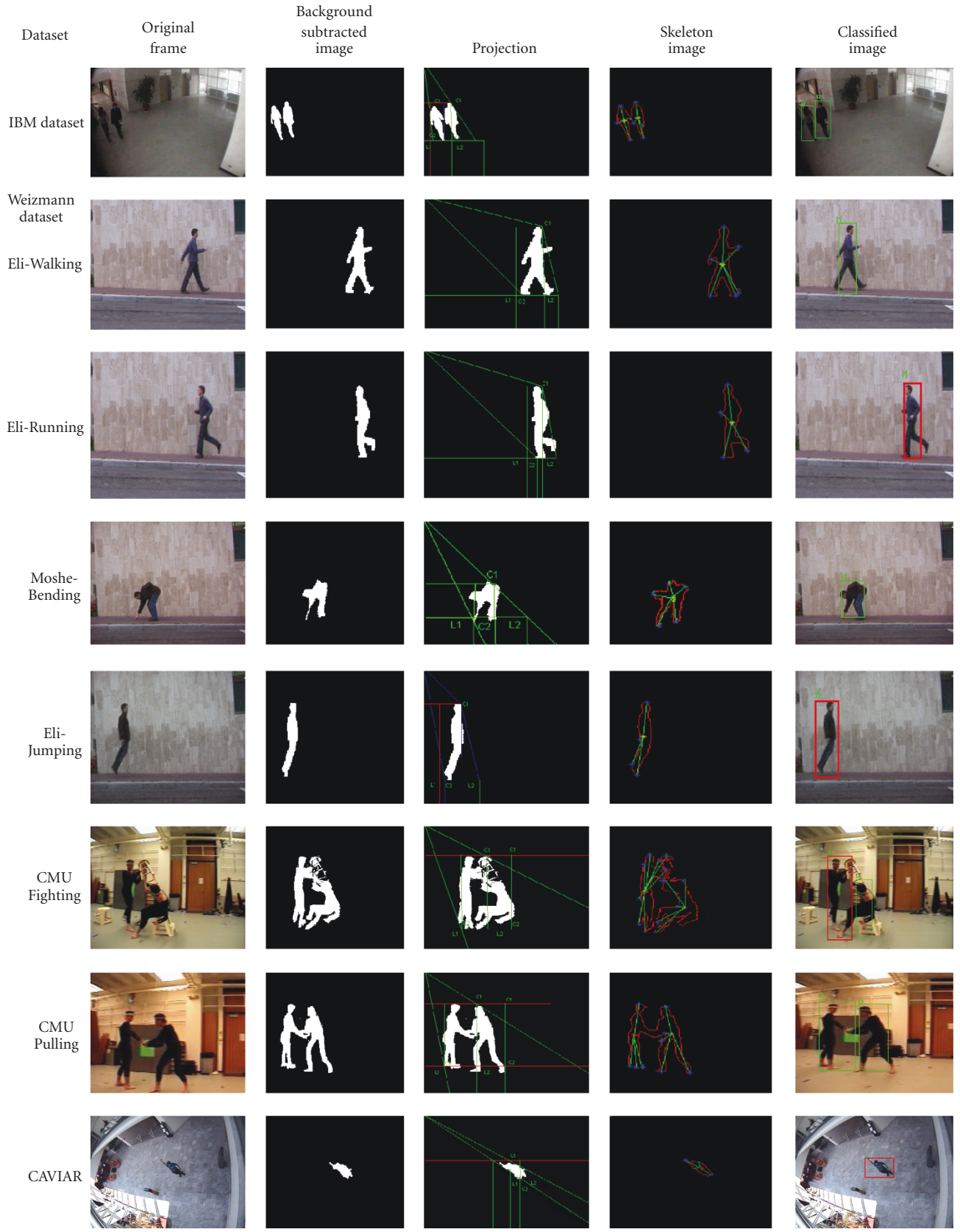


Figure 16: Results for benchmark datasets.

References

[1] M. A. Ali, S. Indupalli, and B. Boufama, "Tracking multiple people in the context of video surveillance," in *Proceedings of the 1st International Workshop on Video Processing for Security (VP4S '06)*, Quebec City, Canada, June 2006.

[2] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07)*, 2007.

[3] T. Zhao and R. Nevatia, "Tracking multiple humans in crowded environment," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. 406–413, 2004.

[4] J. K. Aggarwal and Q. Cai, "Human motion analysis: a review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428–440, 1999.

[5] C. Cédras and M. Shah, "Motion-based recognition a survey," *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, 1995.

[6] D. M. Gavrila, "The visual analysis of human movement: a survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[7] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man and Cybernetics Part C*, vol. 34, no. 3, pp. 334–352, 2004.

[8] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1114–1127, 2008.

[9] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[10] M. Piccardi, "Background subtraction techniques: a review," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3099–3104, October 2004.

[11] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[12] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*,, vol. 2, pp. 302–309, 2004.

[13] J.-S. Hu and T.-M. Su, "Robust background subtraction with shadow and highlight removal for indoor surveillance," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 14 pages, 2007.

[14] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: a real time system for detecting and tracking people," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 962–968, 1998.

[15] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, pp. 459–466, 2003.

[16] A. Gilbert, J. Illingworth, and R. Bowden, "Scale invariant action recognition using compound features mined from dense spatio-temporal corners," in *Proceedings on the 10th European Conference on Computer Vision (ECCV '08)*, D. Forsyth, P. Torr, and A. Zisserman, Eds., vol. 5302 of *Lecture Notes in Computer Science*, pp. 222–233, Springer, Berlin, Germany, 2008.

[17] Y. Song, L. Goncalves, and P. Perona, "Unsupervised learning of human motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 814–827, 2003.

[18] K. K. C. Lee and Y. Xu, "Modeling human actions from learning," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '04)*, vol. 3, pp. 2787–2792, October 2004.

[19] C. Stauffer and W. E.L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[20] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1091–1104, 2002.

[21] X. Wu, Y. Ou, H. Qian, and Y. Xu, "A detection system for human abnormal behaviour," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '05)*, vol. 2, pp. 1204–1208, Edmonton, Canada, August 2005.

[22] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, no. 3, pp. 211–244, 2001.

[23] M. E. Tipping and A. Faul, "Fast marginal likelihood maximization for sparse Bayesian models," in *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, Key West, Fla, USA, January 2003.

[24] O. Williams, A. Blake, and R. Cipolla, "A sparse probabilistic learning algorithm for real-time tracking," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, vol. 1, pp. 353–360, Nice, France, October 2003.

[25] S. Atev, O. Masoud, and N. Papanikolopoulos, "Practical mixtures of gaussians with brightness monitoring," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems (ITSC '04)*, pp. 423–428, October 2004.

[26] C. Stauffer and W. E.L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, pp. 246–252, 1999.

[27] P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos, "Estimating pedestrian counts in groups," *Computer Vision and Image Understanding*, vol. 110, no. 1, pp. 43–59, 2008.

[28] H. Fujiyoshi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization," *IEICE Transactions on Information and Systems*, vol. E87-D, no. 1, pp. 113–120, 2004.

[29] R. Chellappa, A. K. Roy-Chowdhury, and S. K. Zhou, *Recognition of Humans and Their Activities Using Video*, Morgan and Claypool, San Francisco, Calif, USA, 1st edition, 2005.

[30] H. Ren and G. Xu, "Human action recognition in smart classroom," in *Proceedingsof the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 399–404, Washington, DC, USA, May 2002.

[31] A. Mittal, L. Zhao, and L.S. Davis, "Human body pose estimation using silhouette shape analysis," in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '03)*, 2003.

[32] R. Cucchiara, A. Prati, and R. Vezzani, "Posture classification in a multi-camera indoor environment," in *Proceedings of the International Conference on Image Processing (ICIP '05)*, vol. 1, pp. 725–728, 2005.

[33] D. G. Tzikas, A. Likas, N. P. Galatsanos, A. S. Lukic, and M. N. Wernick, "Relevance vector machine analysis of functional neuroimages," in *Proceedings of the 2nd IEEE International Symposium on Biomedical Imaging*, vol. 1, pp. 1004–1007, 2004.

[34] H. C. Lian and B. L. Lu, "Multi-view gender classification using local binary patterns and support vector machines," in *Proceedings of the 5th International Conference on Artificial Neural Networks (ISNN '06)*, vol. 2, pp. 202–209, Chengdu, China, 2006.

[35] A. Thayananthan, R. Navaratnam, B. Stenger, P. H. S. Torr, and R. Cipolla, "Pose estimation and tracking using multivariate regression," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1302–1310, 2008.

[36] P. S. Dhillon, S. Nowozin, and C. Lampert, "Combining appearance and motion for human action classification in videos," in *Proceedings of the 1st International Workshop on Visual Scene Understanding (ViSU '09)*, Miami, Fla, USA, 2009.