

## Human Behavior Classification Using Geometrical Features of Skeleton and Support Vector Machines

Syed Muhammad Saqlain Shah<sup>1,\*</sup>, Tahir Afzal Malik<sup>2</sup>, Robina khatoon<sup>1</sup>, Syed Saqlain Hassan<sup>3</sup> and Faiz Ali Shah<sup>4</sup>

**Abstract:** Classification of human actions under video surveillance is gaining a lot of attention from computer vision researchers. In this paper, we have presented methodology to recognize human behavior in thin crowd which may be very helpful in surveillance. Research have mostly focused the problem of human detection in thin crowd, overall behavior of the crowd and actions of individuals in video sequences. Vision based Human behavior modeling is a complex task as it involves human detection, tracking, classifying normal and abnormal behavior. The proposed methodology takes input video and applies Gaussian based segmentation technique followed by post processing through presenting hole filling algorithm i.e., fill hole inside objects algorithm. Human detection is performed by presenting human detection algorithm and then geometrical features from human skeleton are extracted using feature extraction algorithm. The classification task is achieved using binary and multi class support vector machines. The proposed technique is validated through accuracy, precision, recall and F-measure metrics.

**Keywords:** Human behavior classification, segmentation, human detection, support vector machine.

### Abbreviations

SVM: Support Vector Machines	GMM: Gaussian Mixture Model
RBF: Radial Basis Function	HB: Human Blob
HBH: Human Blob Height	HBW: Human Blob Width
FHO: Fill Holes inside Objects	HDA: Human Detection Algorithm
FEA: Feature Extraction Algorithm	UL: Upper Left
UR: Upper Right	BL: Bottom Left
BR: Bottom Right	

---

<sup>1</sup> Department of CS&SE, International Islamic University, H-10 Sector, Islamabad, 44000, Pakistan.

<sup>2</sup> Department of Management Information Systems, Ibn Rushd College of Management Sciences, Abha, Saudi Arabia.

<sup>3</sup> Department of Computer Science, Bahria University, Islamabad, 44000, Pakistan.

<sup>4</sup> Institute of Computer Science, University of Tartu, Tartu, Estonia.

\* Corresponding Author: Syed Muhammad Saqlain Shah. Email: syed.saqlain@iiu.edu.pk.

## **1 Introduction**

Surveillance is the monitoring of behaviors, activities or other changing information in a secret manner. The surveillance can be carried out in many different ways i.e., biometric surveillance, camera surveillance, aerial surveillance and computer surveillance, etc. [Lyon (2010)]. Prior to the automatic video surveillance, human supervision was needed to monitor the camera footages that not only needed more effort but remained lesser efficient. During the last decade, automatic surveillance has been focused by researchers involving development of computer vision based algorithms for analyzing the abnormal events from videos. Automatic surveillance may be on different crowded areas like airports, shopping malls, stations and private residences for recognizing and monitoring threats, anticipating and preventing the criminal activities. The main tasks involved in visual surveillance include object detection, tracking and behavior classification, widely studied in computer vision. During the last few years, researchers have achieved progress in this, however yet there exist open issues that are needed to be addressed prior to implementation of a robust video surveillance system. A number of visual surveillance systems have been developed for recognition of abnormality in the crowded scenes, by analyzing crowd flow. A little work is available in recognizing individual's behavior from crowd. Some of the abnormal events may not change the overall behavior of the crowd but may affect only some of the individuals (two or three persons) in the crowd. The aim of this study is to recognize the individual's normal and abnormal behavior in the crowd and present a real-time surveillance system.

The proposed technique presents solution of classifying normal and abnormal behaviors of the human in context of recognizing abnormal activities in the crowded scenes. The proposed solution is addressed for constrained environment single and fixed monitoring camera. Rest of the paper is organized as follows: Section 2 presents work related to the proposed research, Section 3 contains proposed methodology, while Section 4 is dedicated to experimental results. Finally, Section 5 concludes the presented research.

## **2 Related work**

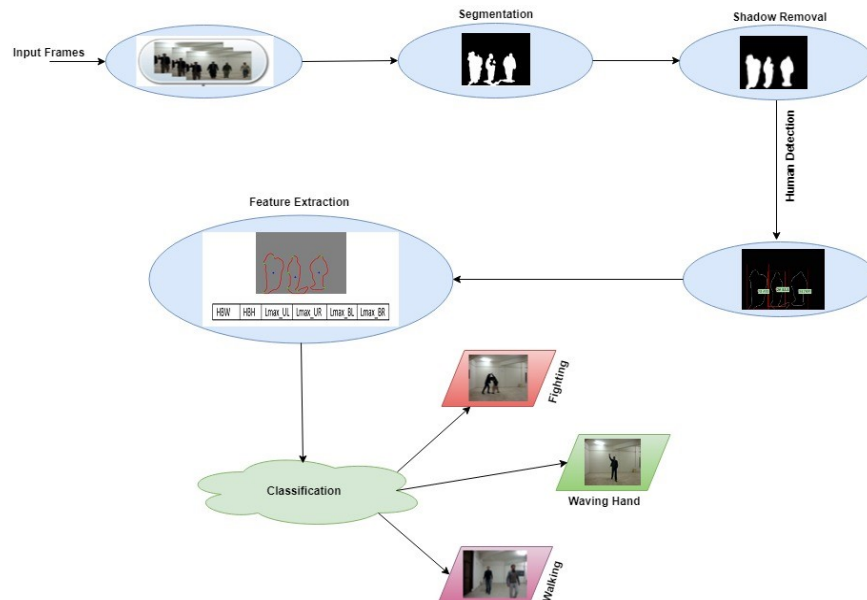
Human behavior classification can be applied in many real-time situations. Andrade et al. [Andrade, Ernesto, Fisher et al. (2006)] presented an approach to detect normal and abnormal events from crowd in context of emergency situations. The experimental results showed that presented models were quite efficient in detecting simulated emergency situation in a dense crowd. In Hsieh et al. [Hsieh and Hsu (2007)], Hsieh and Hsu proposed a simple and rapid surveillance system that achieved human tracking along with classification between normal and abnormal behaviors. Abnormal behaviors included climbing, falling, stopping, and disappearing. Experimental results showed that the system dealt with occlusion and moving objects tracking in an efficient manner. To detect and monitor the human aggressive behaviors, Chen et al. [Chen, Wactlar and Chen (2008)] presented an approach using local binary descriptor for human detection. The proposed approach modelled the actions of arm, body, and the object together. The top 10 retrieval results include about 80% aggressive behaviors, which is much better than the random accuracy of 36.2%. Kiryati et al. [Kiryati, Raviv, Ivanchenko et al. (2008)] presented a novel approach for real time detection of abnormal event. This approach is well suited for applications where limited computing power is available near the camera for compression

and communication. The experimental results showed that the system is reliable for the real-time operation. The abnormal action videos on which system was tested include running, jumping, and grass crossing actions. In Wang et al. [Wang and Mori (2010)], authors proposed human action recognition based on topic models. Yogameena et al. [Yogameena, Veeralakshmi, Komagal et al. (2009)] worked on a real-time video surveillance system, classifying normal and abnormal actions of persons in crowd. Abnormal actions include running, jumping, bending, walking, waving hand and fighting. RVM (Relevance Vector Machine) is used to control huge number of vectors problem. G´arate et al. [Garate, Bilinsky and Bremond (2009)] presented an approach for crowd event recognition that used HOG (Histogram of Gradients). The events included crowd splitting, formation, walking, running, evacuation etc. This approach dealt with overall behavior of the crowd for recognizing crowd events. There are still some errors in the recognized events. This technique needs to improve the threshold computation at the level of scenario models. In Zweng et al. [Zweng and Kampel (2010)], research is related to the unexpected behavior recognition in highly dense density crowded scenes. The actions recognized by the system include running and fall detection. Lin et al. [Lin, Hsu and Lin (2010)] recognized human actions using NWF (Nonparametric Weighted Feature Extraction) based histogram. Research classified ten actions including running, jumping, walking, bending etc. To achieve the lower time complexity for a huge sized dataset, dimensionality was reduced using PCA. Popoola et al. [Popoola and Wang (2012)] presented a critical survey to identify the limitation in existing techniques of abnormal human activity analysis. They have discussed the techniques from the methodologies to applications. Authors Chaaaraoui et al. [Chaaaraoui, Climent-Pérez and Florez-Revuelta (2012)] presented a review over human behavior analysis. Li et al. [Li, Han, Ye et al. (2013)] used sparse reconstruction analysis (SRA) for detection of abnormal behavior. They obtained normal dictionary set for normal behaviors through control point features of cubic B-spline curves and used minimal residue to classify the normal and abnormal behaviors. Cristani et al. [Cristani, Raghavendra, Del Bue et al. (2013)] analyzed human activities on the basis of social signal processing that deals with social, affective, and psychological literature notions. Jiang et al. [Jiang, Bhattacharya, Chang et al. (2013)] provided a review over the techniques dealing with high-level event recognition in unconstrained videos. The idea behind their review was to tackle the problem of analyzing the videos developed by the non-professionals and the videos widely available over the web. Weiyao et al. [Lin, Chen, Wu et al. (2014)] proposed an algorithm which takes a scene from video and represents the video in the form of the network architecture. They termed it as network-transmission-based (NTB) algorithm. The presented algorithm represented scenes as nodes and edges represented the correlation between the scenes. They used their model for classifying the abnormal events. Tran et al. [Tran, Gala, Kakadiaris et al. (2014)] proposed a framework for group activity analysis. They took the liberty of graph by representing human as nodes of the graph and edges as interaction between those humans. Bag-of Words were used as features and SVM was used for classification. Elloumi et al. [Elloumi, Cosar, Pusiol et al. (2015)] in their paper presented different features for recognizing the human activities in unstructured environment. They have used unsupervised learning and tested their technique over the video dataset of medical field for patients monitoring suffering from Alzheimer’s and dementia. Vishwakarma et al. [Vishwakarma and Kapoor (2015)] proposed a hybrid classifier for

analyzing the human activities from videos. The presented classifier used K-NN and SVM at its baseline and authors named it SVM-NN. They tested it over Weizmann, KTH, and Ballet Movement datasets. Eweiwi et al. [Eweiwi, Cheema and Bauckhage (2015)] proposed a new research for classifying human actions in still images. They used local descriptive from images which are supported by their presence in particular areas evident through different videos. Experimental results showed promising outcomes. Vignesh et al. [Ramanathan, Huang, Abu-El-Haija et al. (2016)] proposed their technique for recognizing multi-person event recognition. They used recurrent neural network for tracking individual humans and then model supported the extraction of individuals responsible for the activity. On the next level they again used recurrent neural network for classification of activity. Yogameena et al. [Yogameena, Komagal, Archana et al. (2010)] classified actions like person carrying a long bar, walking, bending and waving hand in the crowd. Human features were extracted using star skeletonization [Fujiyoshi, Lipton and Kanade (2004)] which gives five extreme points of the human skeleton and motion cues. These features are then classified using SVM (Support Vector Machine) classifier. Human body is treated as the interconnections of the five rigid bodies [Guo, Li and Shao (2018)]. The motion of each part was taken as discriminating factor to distinguish between the actions. Bag of features of each part was calculated and the classification was performed through SVMs. Human action classification, involving upper and lower part of the human body, is presented [Lai and Lau (2018)]. The main features were based on the detection of wrist, shoulder and elbow points. K-means clustering was used to imply the classification task.

### 3 Proposed methodology

The proposed methodology classifies the normal and ab-normal human actions in a thin crowd. The graphical representation of the proposed research is shown in Fig. 1.



**Figure 1:** Flow diagram of the proposed technique

### 3.1 Segmentation and post-processing

Segmentation of foreground from the complex background is achieved by using Gaussian Mixture [Atev, Masoud and Papanikolopoulos (2004)]. It segmented the foreground moving pixels by finding the probability of each pixel in the image. Every new pixel value is checked against the existing K Gaussian distributions, until value is less than the standard deviation.

$$\sqrt{(c - \mu_j)^j \Sigma_i^{-1}(c - \mu_j)} < \delta \quad (1)$$

Likelihood of all the unmatched components is updated and if match is found for any of the K distribution then mean, covariance and likelihood of k matched components are updated. If c matches a mixture component,  $M_k$ , it must be determined that it is a part of the background or foreground. Given a threshold  $B \in [0 \dots 1]$ ,  $M_k$  is considered a part of the background if

$$\sum_{i=1}^{k-1} (M_k) < B \quad (2)$$

Morphological dilation using square as a structuring element is applied over the segmented regions. The size of the square is kept as 7 pixels resulting 49 neighbourhood for dilating the segmented regions. Over the dilated regions, we have applied Gaussian filter for noise reduction. Unwanted pixels are removed and marked as part of background. Holes within the blobs are filled using FHO algorithm followed by retaining the larger blobs than the specified size ones.

---

#### Algorithm 1: Fill Holes inside Objects (FHO)

---

**INPUT:** Segmented and Processed Image having Blobs

**OUTPUT:** Image with filled holes within Blobs

- 1: Find the boundaries of processed objects.
  - 2: Build boundry set  $BOS_1, BOS_2, \dots, BOS_p$  for each object  $OB_1, OB_2, \dots, OB_p$ .
  - 3: Find minimum and maximum along horizontal and vertical for each set. i.e,  
 $BOS_j x_{min}, BOS_j x_{max}, BOS_j y_{min}$  and  $BOS_j y_{max}$ .  $1 \leq j \leq p$
  - 4: For each  $BOS_t : 1 \leq t \leq p$
  - 5:  $\exists OB_s$  such that  $OB_s \subseteq OB_t$  i.e,  
 if ( (  $BOS_s x_{min} > BOS_t x_{min}$  )  $\wedge$  (  $BOS_s x_{max} < BOS_t x_{max}$  )  $\wedge$  (  $BOS_s y_{min} > BOS_t y_{min}$  )  $\wedge$  (  $BOS_s y_{max} < BOS_t y_{max}$  ) )
  - 6:  $OB_s$  is a hole within  $OB_t$
  - 7: For each point  $p_m \in OB_s$
  - 8: If  $p_m$  is background
  - 9: Mark  $p_m$  as foreground
  - 10: end for
  - 11: end for
-

### 3.2 Shadow removal

The foreground extracted from the segmentation is subject to shadow detection [Khatoun, Saqlain and Bibi (2012); Kelly, Agapito, Conaire et al. (2010)] where first frame of the video sequence is considered as the background image. A color brightness difference value  $D$  is calculated as follows:

$$D = 18 \times D_{gb} + \left| \log \left[ \frac{V_{BG}}{V_{curr}} \right] \right| \quad (3)$$

where  $D_{gb}$  is the distance between the current pixel and the background pixel which is normalized by **gb** space; background brightness is represented by  $V_{BG}$  and  $V_{curr}$  is the brightness of current pixel. If color/brightness level has difference value  $D < 0.5$ , then the foreground is marked as a shadow pixel and gets discarded. Post-processing is then applied to the resultant image for suppressing the noise and fill the holes in silhouette. The segmented image is reconstructed after marking the shadow pixels.

### 3.3 Human detection

Human detection is achieved by using local HOGs and SVM. The local HOGs are calculated over the four partitioned areas of the window having segmented object. The details of the proposed method are presented in human detection algorithm (HDA).

---

#### Algorithm 1: Human Detection Algorithm (HDA)

---

**INPUT:** Segmented Image SI having  $n$  Blobs

**OUTPUT:**  $k$  Blobs detected as Human

**INPUT:** Segmented Image SI having  $n$  Blobs

**OUTPUT:**  $k$  Blobs detected as Human

1: For each blob  $B_j$ :  $1 \leq j \leq n$ .

2: Find enclosed rectangle around  $B_j$  by finding points  $(x_{min}, y_{min}), (x_{max}, y_{min}), (x_{min}, y_{max}), (x_{max}, y_{max})$  lying over  $B_j$ .

3: Calculate center of rectangle around  $B_j$  by using  $x_{rc} = \frac{x_{max} - x_{min}}{2}$  and  $y_{rc} = \frac{y_{max} - y_{min}}{2}$

4: With reference to  $(x_{rc}, y_{rc})$  divide  $B_j$  into four Regions  $R_i$ :  $1 \leq i \leq 4$  i.e.,

Upper Left defined by  $\{(x_{min}, y_{min}), (x_{rc}, y_{min}), (x_{min}, y_{rc}), (x_{rc}, y_{rc})\}$ ,

Upper Right defined by  $\{(x_{rc}, y_{min}), (x_{max}, y_{min}), (x_{rc}, y_{rc}), (x_{max}, y_{rc})\}$ ,

Bottom Right defined by  $\{(x_{rc}, y_{min}), (x_{rc}, y_{rc}), (x_{min}, y_{max}), (x_{rc}, y_{max})\}$  and

Bottom Left through  $\{(x_{rc}, y_{rc}), (x_{max}, y_{rc}), (x_{rc}, y_{max}), (x_{max}, y_{max})\}$

5: For each  $R_i$ :  $i \leq 1 \leq 4$

6: Find Oriented Descriptors by using HOG method. Parameters used in HOG are:

Size of region:  $W \times H$  where  $W = x_{rc} - x_{min}$  and  $H = y_{rc} - y_{min}$ .

1-D Derivative masks used are:  $[-1 \ 0 \ +1]$  and  $\begin{bmatrix} -1 \\ 0 \\ +1 \end{bmatrix}$

Cell size:  $8 \times 8$

Block Size:  $2 \times 2$

Block overlap:  $1 \times 1$

Weighted Gradients are used for each pixel orientation.

- 7: Create 1-D vector  $v_i$  having descriptors obtained through HOG, corresponding to  $R_i$
  - 8: end for
  - 9:  $v_i: 1 \leq i \leq 4$  are provided to 4-SVM classifiers each classifying one of four UL,UR,LL and LR. If three of the regions are classified as human parts. We classify blob as human.
  - 10: end for
  - 11: Blobs identified as humans are kept enrectangled. Non Humans are turned as background.
- 

Each of the segmented blob,  $B_j: j=1 \leq n$ , is provided as an input to the HAD that finds the  $\min_x, \min_y$  and  $\max_x, \max_y$  points of the rectangle enclosing  $B_j$ . By finding the centroid of each of the enclosing rectangle, they are divided into four sub rectangles with reference to the center points. HOG is calculated for each of the four sub rectangles and their histograms are fed to the corresponding SVMs trained over upper left(UL), bottom left(BL), upper right(UR) and lower right(LR) regions. All the four trained SVM models produce an output whether the region represents a human part or not. If three out of the four or all the models produce positive response, the blob in the rectangle is classified as human. Otherwise, it is a non-human is made part of the background.

### 3.4 Feature extraction

The task of feature extraction is accomplished by proposing feature extraction algorithm(FEO). It takes each of the potential human blob as input and divides it into four regions which are computed with reference to the centroid of the blob. The extreme points in each of the four regions are computed which are then combined with the width and height of the blob to form a 6-dimentional feature vector. The FEO is presented as:

---

#### Algorithm 2: Feature Extraction Algorithm (FEA)

---

**INPUT:** Image Having k-Human Blobs

**OUTPUT:** k-vectors having feature set for each Identified Human

- 1: For each human blob  $HB_k: 1 \leq k \leq m$ , m are total number of blobs identified as human.
- 2: Find Width ( $HBW_k$ ) and Height ( $HBH_k$ ) using  $xb_{max} - xb_{min}$  and  $yb_{max} - yb_{min}$  respectively, where  $xb_p, yb_p$  are point lying in  $HB_k$ .
- 3: Centroid of Human blob is calculated using  $X_{cb} = \frac{1}{n} \sum_{i=1}^n x_i$  and  $Y_{cb} = \frac{1}{n} \sum_{i=1}^n y_i$
- 4: Divide the human contour in four sub-regions i.e., Upper Left(UL),Upper Right(UR), Bottom Left(BL) and Bottom Right(BR).
- 5: For each sub-region  $SR_q \in \{UL, UR, BL, BR\} \subseteq \{HB_k\}$

- 6: Calculate distance of the centroid from each boundary pixel using:  

$$d = \sqrt{(x_i - X_{cb})^2 + (y_i - Y_{cb})^2}$$
 resulting a 1-D discrete signal.
  - 7: 1-D discrete signal is smoothed using low pass filter for noise reduction
  - 8: Local maxima  $LM_q$  in 1-D discrete signal is taken as extreme points which is detected by finding zero-crossing difference function  $\delta(i) = d^{(i)} - d^{(i-1)}$
  - 9: end for
  - 10: Create a 1-D vector  $v_k$  corresponding to  $HB_k$  having  $HBW_k$ ,  $HBH_k$  and  $LM_q$   
 where  $q \in \{UL, UR, BL, BR\}$ .
  - 11: end for
- 

### 3.5 Classification

In the proposed technique, Support Vector Machines (SVM) are used for the action classification. The solution to the problem of multi-class is achieved through the combination of multiple binary class SVMs. The decision functions for the binary-class linear SVMs are represented by  $(u \cdot fs_j + m) \leq -1$  if  $c_j = -1$  and  $(u \cdot fs_j + m) \geq 1$  if  $c_j = 1$ . The two relations can be combined to cover both the cases i.e.,  $c_j(u \cdot fs_j + m) \geq 1$  where  $c$  represents the class,  $fs$  is the input feature set,  $u$  represents weight and  $m$  is the margin. Decision function for the non-linear classification is represented as  $f(fs) = u \cdot \Phi(fs) + m$ . This is presented as we are dealing with the research problem having input feature subset instead of a single dimensional problem and we have performed experiments with SVM using other modalities than linear functions. In order to have convergence for the non-linearity, SVM kernel functions are widely used. In a kernel-based SVM,  $u$  may be represented as  $u = \sum_{j=1}^n \beta_j \phi(fs_j)$ ,  $n$  is number of samples. Decision function for a non-linear classification problem having multidimensional feature set can be represented as:

$$f(fs) = \sum_{j=1}^n \beta_j \phi(fs_j) \cdot \phi(fs) + m \quad (4)$$

where  $\Phi(s_j)$ .  $\Phi(s)$  is defined as the kernel function and denoted as  $KF(s, s_j)$ . By replacing  $\Phi(s_j)$ .  $\Phi(s)$  by  $KF(s, s_j)$  the decision function is given as:

$$f(fs) = \sum_{j=1}^n \beta_j KF(fs, fs_j) + m \quad (5)$$

In order to implement non-linear classification task, radial basis function(RBF) based kernel is identified as the most suitable. This is due to the RBF kernel being known to be used in solving infinite dimension problems. The RBF kernel can be represented as

$$KF(fs, fs_j) = e^{(-\gamma \|fs - fs_j\|^2)} \quad (6)$$

By substituting (6) into (5), the required decision function becomes:

$$f(fs) = \sum_{j=1}^m \beta_j e^{(-\gamma \|fs - fs_j\|^2)} + m \quad (7)$$

In order to adopt the binary class SVM for the multi-class problem of  $K$  categories and using the one-vs-one approach, we need to build  $\frac{K(K-1)}{2}$  binary SVM classifiers [Lee, Lin and Wahba (2004)]. In the presented research  $K=3$ , so  $\frac{3(3-1)}{2}$  binary SVM classifiers are modeled i.e.,  $C=3$ . On the other hand, the one-vs-all technique has three binary classifiers are modeled as well. The classifiers for one-vs-one are {Walking-vs-Waving, Walking-vs-



Fighting, Waving-vs-Fighting} and for the one-vs-all methodology the same are {Walking-vs-All, Waving-vs-All, Fighting-vs-All}. The term ‘All’ is different in all the tree cases i.e., it is {Waving, Fighting} for Walking-vs-All, {Walking, Fighting} for Waving-vs-All and {Walking, Waving} for the classifier Fighting-vs-All. The final classification of the input feature set,  $fs$ , through the one-vs-one multi-class classifiers is achieved as:

$$f(fs) = \underset{p}{\operatorname{argmax}}(\sum_q f_{p,q}(fs)) \quad (8)$$

where  $p, q \in \{\text{Walking, Waving, Fighting}\}$  and  $f_{p,q} \in \{\text{Walking-vs-Waving, Walking-vs-Fighting, Waving-vs-Fighting}\}$ . The classification of a feature set,  $fs$ , using one-vs-all modality is computed as:

$$f(fs) = \underset{i=1..N}{\operatorname{argmax}}(cp(i)) \quad (9)$$

where  $cp(i)$  is the class probability of the  $i^{\text{th}}$  class.

## 4 Experimental setup, results & discussion

### 4.1 Experimental setup

In this subsection, the detail of dataset, specification of hardware system used for implementation, description of the tool used and the evaluation metrics will be presented.

#### 4.1.1 Dataset & system platform

The proposed method is experimented on a dataset of self-created video sequences and publically available datasets for classifying walking, waving hand and fighting in the crowd. The proposed technique is evaluated over two types of experiments i.e., individual action classification (walking, waving hands and fighting) and normal-abnormal action detection. In case of second type experiment, walking and waving hands is taken as normal action while fighting is classified as abnormal action. Total number of tested frames is 1000. The proposed technique is tested by implementing in MATLAB R2015 on a machine with multicore 1.8 GHZ processor. Following datasets are used other than the self-created dataset:

- Weizmann (<http://www.wisdom.weizmann.ac.il>)
- CAVIAR <http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1>

Video sequences taken from the above-mentioned datasets represent different scenarios and it will be considered while performing the experiments. Following are the seven ones:

*Scenario 1:* One person walking in corridor

*Scenario 2:* Two persons walking in corridor

*Scenario 3:* Three persons walking and waving hand in corridor

*Scenario 4:* One person walking in outdoor scene and shadow appears large

*Scenario 5:* One person walking with dog in outdoor scene

*Scenario 6:* Four person walking and waving hand in corridor

*Scenario 7:* Two persons fighting in corridor

#### 4.1.2 Quantitative measuring parameters

Precision, recall, f-measure and accuracy are used to compute quantitative performance of the proposed technique i.e.

$$\text{Precision} = \frac{\text{Number of true positives detected}}{\text{Total number of positives detected}} \quad (10)$$

$$\text{Recall} = \frac{\text{Number of true positives detected}}{\text{Total number of true positives}} \quad (11)$$




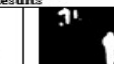
























$$\text{F - measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

$$\text{Accuracy} = \frac{\text{No. of true positives}}{\text{Total number of positives}} \quad (13)$$

#### 4.2 Results & discussion

Results of the presented methods for each of the scenarios are presented in Figs. 2-7. The results of segmentation of two consecutive frames for each scenario are shown in Fig. 2. In the scenario 1, although there are different lightning conditions in the scenes, yet the results of segmentation are quite promising as mixture of Gaussian deals with such conditions. In the scenario 4 the shadow of person appears large in outdoor scene and which is little bit handled by segmentation and will be totally removed after applying shadow removal technique. In scenario 5 there is flickering of leaves in the background that is well handled by mixture of Gaussian as it eliminates misclassification due to cyclic motion in the background.

After applying the shadow removal, results get much simplified and clear as the shadow is removed and false human detection would be minimized. The fact may be evidently visualized for the scenarios 1 and 4 in Fig. 2 and Fig. 3, where larger sized shadows are removed and chance of misclassification of those shadows as humans is tackled.

Sr #	Original Images		Segmentation Results	
1				
2				
3				
4				
5				
6				
7				

**Figure 2:** Consecutive frames and corresponding segmentation results for each scenario

Sr #	Original Images		Shadow Removal Results	
1				
2				
3				
4				
5				
6				
7				

Figure 3: Results of shadow removal for consecutive frames of each scenario

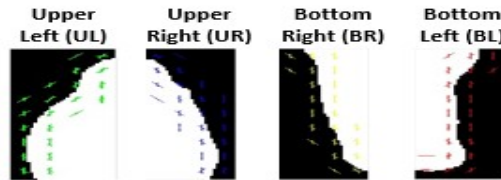


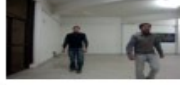













Figure 4: Partwise feature extraction for human detection

Fig. 4 is showing the division of the extracted blobs, which were obtained through the segmentation and shadow removal steps. Each of the blob in the segmented image is divided into the four parts and their corresponding HOG features are extracted. If any of the blob has a greater size than a threshold, there is a chance of occurring multiple humans. In this case, with reference to the centroid the blob is divided into multiple set of regions where each of the set looks for occurrence of a unique human. The HOG feature is fed to the four SVM classifiers using RBF kernels resulting either a part of human or not. In case, if the majority of SVM models result in positive the blob is labelled as human otherwise it is considered as non-human.

Fig. 5 is showing the results of the human detection through HDA. It may be observed that HDA detected all the human blobs present in the frames and for all the seven scenarios.

Sr #	Original Images	Human Detection Results
1		
2		
3		
4		
5		
6		
7		

**Figure 5:** Results of human detection for each scenario

Once the human is detected, his action needs to be classified. In order to classify the human action, features are extracted through FEA. A feature vector is of dimension  $1 \times 6$  and shown in the Fig. 6.

HBW	HBH	Lmax_UL	Lmax_UR	Lmax_BL	Lmax_BR
-----	-----	---------	---------	---------	---------

**Figure 6:** Feature vector for action classification

where HBW, HBH represent human blob width, human blob height, while Lmax\_UL, Lmax\_UR, Lmax\_BL and Lmax\_BR represent farthest point from centroid in upper left, upper right, bottom left and bottom right regions. For the actions like walking sidewise, waving hand and fighting the width of blob is more than those of walking towards camera but all the three actions have discriminant region wise features i.e., sidewise walking have higher values for Lmax\_BL or Lmax\_BR or both while the waving hands has higher values for Lmax\_UL or Lmax\_UR. The fighting action may have all the region values different from the other two actions.

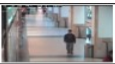


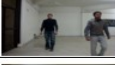


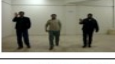
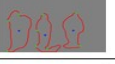








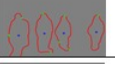

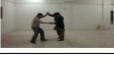


The results of feature extraction and action classifications is shown in Fig. 7. The green bounding box represents normal and red bounding box represents abnormal action. The multiclass problem is solved through the adoption of binary class SVMs. In the current research, two types of adoptions are used i.e., one-vs-all and pairwise classification. Here, the adoption of pairwise classification is called binary class SVM while one-vs-all is termed as multi-class SVM. In order to implement pairwise classification modality, following set of binary classifiers i.e., {Walking-vs-Waving, Walking-vs-Fighting, Waving-vs-Fighting}. The task of one-vs-all is adopted through following set of classifiers

i.e., {Walking-vs-All, Waving-vs-All, Fighting-vs-All}. The pairwise classification is performed using radial basis function kernel and linear kernel, while one-vs-all modality is performed using RBF kernel bases SVM classifiers only. In case of pairwise classification (Binary class SVM) majority voting criteria is adopted. In order to predict using one-vs-all (Multiclass SVM), the test instance belongs to class whose class probability is highest amongst the three classes i.e., walking, waving and fighting.

For an example, an image frame which needs to be classified after the feature extraction has an action of waving hands. In order to predict the action using binary class SVM, the extracted features are fed into all the three pairwise classifiers. The binary class classifier, walking-vs-waving, predicts the action as waving, walking-vs-fighting classifier predicts and action as walking and waving-vs-fighting gives prediction output as waving. The majority of the votes are for the waving class i.e., 2, so the action is classified as waving.

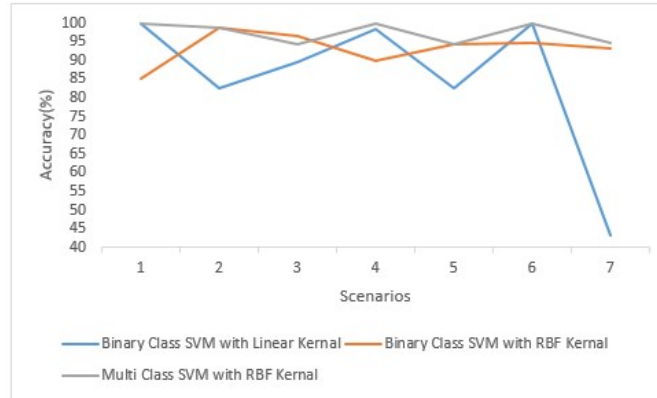
While testing through multiclass SVM, the features are fed to the walking-vs-all, waving-vs-all, and fighting-vs-all classifiers. The output class probabilities by all the three classifiers were 0.45, 0.81 and 0.13 for walking, waving and fighting class. The maximum class probability is for the waving class, so the action is classified as waving.

Fig. 8 is showing the comparative results of the proposed methodology in terms of accuracy using three different modalities i.e., binary class SVM (pairwise classification) using linear kernel, binary class SVM using radial basis function (RBF) kernel and multiclass SVM (one-vs-all) using RBF kernel. The graph shows the comparative accuracies for all the seven scenarios as defined in dataset description. The highest of the accuracies for scenario 1 are 100% for multiclass SVM using RBF and binary class SVM using linear kernel while it remained 85% for the and binary class SVM with linear kernel. The highest accuracies for the reaming six scenarios are 98.82%, 96.53%, 100%, 95.0%, 100%, 94.83% for radial basis binary class SVM, RBF based binary class SVM, multiclass SVM with RBF kernel, RBF based binary class SVM, multiclass SVM with RBF kernel(combined with linear kernel SVM) and multiclass SVM with RBF kernel respectively.

Sr #	Original Images	Feature Extraction Results	Classified Results
1			
2			
3			
4			
5			
6			
7			

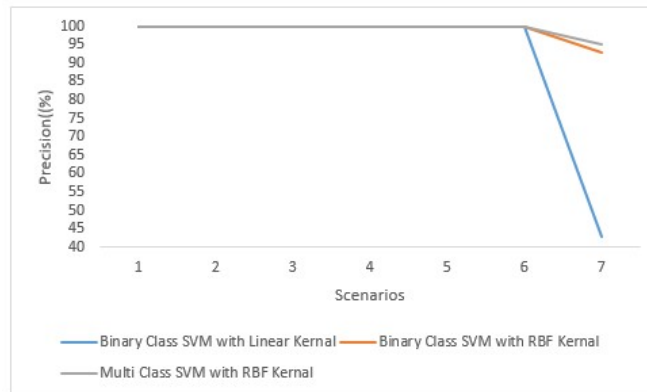
**Figure 7:** Results of feature extraction and classification for each scenario

On the other hand, RBF based binary SVM has least accuracy for scenario one, four and seven while binary SVM with linear kernel retained lower accuracy for the second, third, fifth and seventh scenarios.



**Figure 8:** Comparative analysis of classification accuracies using variant of SVM modalities

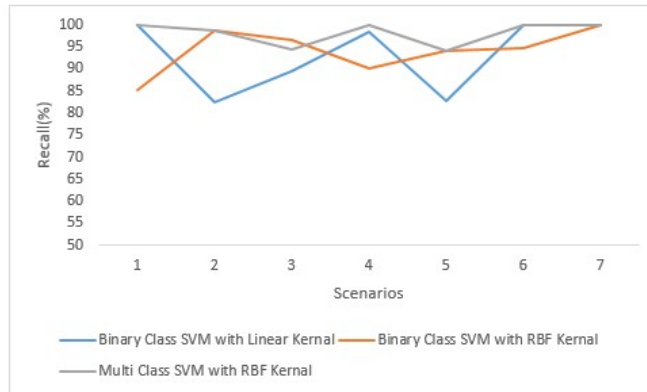
In Fig. 9, the comparison of the above stated three modalities for the SVM classifier is presented in the form of precision. It is quite clear from the graph that all the three modalities attained 100% precision for all the first six scenarios. In case of seventh scenarios, the precision values are 93.10%, 43.10% and 94.83% for RBF kernel based binary class SVM, linear kernel based binary class SVM and multiclass SVM with RBF kernel respectively.



**Figure 9:** Comparative analysis of classification precisions using variant of SVM modalities

Fig. 10 is showing the comparative analysis of the proposed technique evaluated through recall measurement. The comparison is presented for all the three variants of SVMs. The binary SVM classifier with RBF kernel achieved recall values 85.19%, 98.80%, 96.53%, 90%, 94.23%, 94.57% and 100% for the scenarios 1-7 respectively. In case of linear kernel based binary SVM, the same evaluation parameter and for the same one to seven scenarios the achieved results are 100%, 85.25%, 89.58%, 98.33%, 82.69%, 100% and 100% respectively. While using RBF based multiclass SVM, the proposed technique has achieve

following recall values i.e., 100%, 98.80%, 94.44%, 100%, 94.23%, 100% and 100% for all the seven scenarios in a sequence respectively.



**Figure 10:** Comparative analysis of classification recall using variant of SVM modalities

The statistics of the third evaluation metric, F-measure, is graphically presented in Fig. 11. It shows the comparative results of the proposed technique using all the three SVM modalities. It may be observed that while using RBF based SVM, the f-measures of the proposed technique on all the seven scenarios are: 92%, 99.9%, 98.23%, 94.74, 97.03%, 97.21% and 96.43%. The same metric and over the same scenarios but using linear kernel based SVM the results are 100%, 99.043%, 94.51%, 99.16%, 90.53%, 100% and 60.24%. The results are in sequence corresponding to the sequential scenarios in an order of one to seven. At last, the f-measure results through the proposed technique and using RBF based multiclass SVMs are presented as: 100%, 99.39%, 97.14%, 100%, 97.03%, 100% and 97.35% for the scenarios 1-7 respectively.



**Figure 11:** Comparative analysis of classification f-measure score using variant of SVM modalities

While performing the second set of experiments i.e., abnormal action detection over the available dataset, the normal class is defined as Normal={walking Images}U{Waving Hands Images} and the abnormal class is Abnormal={Fighting Images}. Tab. 1 presents

the comparison of the proposed technique with the existing one [Yogameena, Komagal, Archana et al. (2010)]. The results show that the proposed system out-performs than the existing technique [Yogameena, Komagal, Archana et al. (2010)] for normal behavior classification using both the SVM classifier with RBF kernel and multi-class SVMs. The same results were achieved for the abnormal behaviors through the binary class SVMs with RBF kernels but the proposed technique was out-performed by the existing technique [Yogameena, Komagal, Archana et al. (2010)] through the multi-class SVMs for abnormal behavior classification.

**Table 1:** Accuracy based comparative analysis of the proposed technique

	Existing Technique [Yogameena, Komagal, Archana et al. (2010)] using SVM	Proposed Technique using SVM	Existing Technique [Yogameena, Komagal, Archana et al. (2010)] using Multi-class SVM	Proposed Technique using Multi-class SVM
Normal Actions	89.85%	93.22%	96.45%	97.91%
Abnormal Actions	82.50%	93.10%	96.70%	94.83%

## 5 Conclusion

In this research, a methodology for human behavior classification is presented. The proposed methodology comprised on five modules including segmentation, shadow removal, human detection, feature extraction and for achieving the required goals fill hole inside objects (FHO), human detection algorithm and feature extraction algorithm (FEA) are presented. The presented human detection technique is robust as applied in both indoor and outdoor scenes and give good results. The proposed methodology is implemented to classify the different human actions along with normal and abnormal human behaviors in thin crowded videos. It was tested over different public datasets and the performance of the proposed technique is evaluated through accuracy, F-measure, precision and recall metrics.

## Declarations

### *Availability of supporting data*

Our dataset may be provided on request from the corresponding author.

### *Competing interests*

The authors declare that they have no competing interests.

### *Funding*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.



**Acknowledgements:** Authors are thankful to Ms. Shafina Bibi at International Islamic University, Islamabad, Pakistan for her continuous support during the conduct of whole research.

## References

- Andrade, L.; Ernesto, B.; Fisher, R.; Blunsden, S.** (2006): Detection of emergency events in crowded scenes. *Proceedings of IET Conference on Crime and Security*, pp. 528-533.
- Atev, S.; Masoud, O.; Papanikolopoulos, N.** (2004): Practical mixtures of Gaussians with brightness monitoring. *Proceedings of 7th International IEEE Conference on Intelligent Transportation Systems*, pp. 423-428.
- Chen, D.; Wactlar, H.; Chen, M. Y.; Gao, C.; Bharucha, A. et al.** (2008): Hauptmann, recognition of aggressive human behavior using binary local motion descriptors. *Proceedings of 30th Annual International Conference of the IEEE on Engineering in Medicine and Biology Society*, pp. 5238-5241.
- Chaaroui, A. A.; Climent-Pérez, P.; Florez-Revuelta, F.** (2012): A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, vol. 39, no. 12, pp. 10873-10888.
- Lyon, D.** (2010): Surveillance, power and everyday life. *Proceedings of Emerging Digital Spaces in Contemporary society*, pp. 107-120.
- Cristani, M.; Raghavendra, R.; Del Bue, A.; Murino, V.** (2013): Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, vol. 100, pp. 86-97.
- Elloumi, S.; Cosar, S.; Pusiol, G.; Bremond, F.; Thonnat, M.** (2015): Unsupervised discovery of human activities from long-time videos. *IET Computer Vision*, vol. 9, no. 4, pp. 522-530.
- Eweiwi, A.; Cheema, M. S.; Bauckhage, C.** (2015): Action recognition in still images by learning spatial interest regions from videos. *Pattern Recognition Letters*, vol. 51, pp. 8-15.
- Fujiyoshi, H.; Lipton, A. J.; Kanade, T.** (2004): Real-time human motion analysis by image skeletonization. *IEICE TRANSACTIONS on Information and Systems*, vol. 87, no. 1, pp. 113-120.
- Garate, C.; Bilinsky, P.; Bremond, F.** (2009): Crowd event recognition using hog tracker. *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pp. 1-6.
- Guo, Y.; Li, Y.; Shao, Z.** (2018): DSRF: a flexible trajectory descriptor for articulated human action recognition. *Pattern Recognition*, vol. 76, pp. 137-148.
- Hsieh, C. C.; Hsu, S. S.** (2007): A simple and fast surveillance system for human tracking and behavior analysis. *Proceedings of Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, pp. 812-818.
- Jiang, Y. G.; Bhattacharya, S.; Chang, S. F.; Shah, M.** (2013): High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, vol. 2, no. 2, pp. 73-101.

- Kelly, P.; Agapito, J. D. P. M.; Conaire, C.; Monaghan, D.; Kuklyte, J. et al.** (2010): A low-cost performance analysis and coaching system for tennis. *Proceedings of ACM Multimedia*.
- Khatoon, R.; Saqlain, S. M.; Bibi, S.** (2012): A robust and enhanced approach for human detection in crowd. *Proceedings of 15th International Multitopic Conference*, pp. 215-221.
- Kiryati, N.; Raviv, T. R.; Ivanchenko, Y.; Rochel, S.** (2008): Real-time abnormal motion detection in surveillance video. *Proceedings of 19th International Conference on Pattern Recognition*, pp. 1-4.
- Lai, S. C.; Lau, P. Y.** (2018): Upper body action classification for multiview images using K-means. *International Workshop on Advanced Image Technology*, pp. 1-4.
- Li, C.; Han, Z.; Ye, Q.; Jiao, J.** (2013): Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. *Neurocomputing*, vol. 119, pp. 94-100.
- Lin, C. H.; Hsu, F. S.; Lin, W. Y.** (2010): Recognizing human actions using NWFE-based histogram vectors. *EURASIP Journal on Advances in Signal Processing*, vol. 9.
- Lin, W.; Chen, Y.; Wu, J.; Wang, H.; Sheng, B. et al.** (2014): A new network-based algorithm for human activity recognition in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 826-841.
- Lee, Y.; Lin, Y.; Wahba, G.** (2004): Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67-81.
- Popoola, O. P.; Wang, K.** (2012): Video-based abnormal human behavior recognition-A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865-878.
- Ramanathan, V.; Huang, J.; Abu-El-Haija, S.; Gorban, A.; Murphy, K. et al.** (2016): Detecting events and key actors in multi-person videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3043-3053.
- Tran, K. N.; Gala, A.; Kakadiaris, I. A.; Shah, S. K.** (2014): Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, vol. 44, pp. 49-57.
- Vishwakarma, D. K.; Kapoor, R.** (2015): Hybrid classifier based human activity recognition using the silhouette and cells. *Expert Systems with Applications*, vol. 42, no. 20, pp. 6957-6965.
- Wang, Y.; Mori, G.** (2010): Human action recognition by semilattent topic models. *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 10, pp. 1762-1774.
- Yogameena, B.; Komagal, E.; Archana, M.; Abhaikumar, S. R.** (2010): Support vector machine-based human behavior classification in crowd through projection and star skeletonization. *Journal of Computational Science*, vol. 6, pp. 1008-1013.
- Yogameena, B.; Veeralakshmi, S.; Komagal, E.; Raju, S.; Abhaikumar, V.** (2009): RVM-based human action classification in crowd through projection and star skeletonization. *Journal on Image and Video Processing*, vol. 4.

**Zweng, A.; Kampel, M.** (2010): Unexpected human behavior recognition in image sequences using multiple features. *Proceedings of 20th International Conference on Pattern Recognition*, pp. 368-371.