Sharif et al. EURASIP Journal on Image and Video Processing (2017) 2017:89
DOI 10.1186/s13640-017-0236-8

EURASIP Journal on Image and Video Processing

RESEARCH Open Access

A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection

Muhammad Sharif[1], Muhammad Attique Khan[1*], Tallha Akram[1], Muhammad Younus Javed[2], Tanzila Saba[3] and Amjad Rehman[4]

Abstract

Human activity monitoring in the video sequences is an intriguing computer vision domain which incorporates colossal applications, e.g., surveillance systems, human-computer interaction, and traffic control systems. In this research, our primary focus is in proposing a hybrid strategy for efficient classification of human activities from a given video sequence. The proposed method integrates four major steps: (a) segment the moving objects by fusing novel uniform segmentation and expectation maximization, (b) extract a new set of fused features using local binary patterns with histogram oriented gradient and Harlick features, (c) feature selection by novel Euclidean distance and joint entropy-PCA-based method, and (d) feature classification using multi-class support vector machine. The three benchmark datasets (MIT, CAVIAR, and BMW-10) are used for training the classifier for human classification; and for testing, we utilized multi-camera pedestrian videos along with MSR Action dataset, INRIA, and CASIA dataset. Additionally, the results are also validated using dataset recorded by our research group. For action recognition, four publicly available datasets are selected such as Weizmann, KTH, UIUC, and Muhavi to achieve recognition rates of 95.80, 99.30, 99, and 99.40%, respectively, which confirm the authenticity of our proposed work. Promising results are achieved in terms of greater precision compared to existing techniques.

Keywords: Human detection, Preprocessing, Segmentation, Feature extraction, Fusion, Feature selection, Action recognition

1 Introduction

A system which intelligently detects a human from an image or a video is a challenging task of the modern era. From the last decade, computer vision and pattern recognition community concentrated on the human detection largely due to the variety of industrial applications, which include video surveillance [1], traffic surveillance [2], human-computer interaction [3], automotive safety [4], real-time tracking [5], human-robot interaction [6], search and rescue missions [7], humans' collective behavior analysis [8], anti-terrorist applications [9], pedestrian detection [10], etc.

This research addresses human detection in the recorded videos, which is a challenging task in terms of variations in color, movement, appearance, etc. [11]. Furthermore, some other complex problems are also considered such as light variations, poor background, etc.

In literature, several techniques are implemented for human detection which includes segmentation techniques, feature extraction techniques, classification-based detections, etc. Segmentation techniques for human detection include template matching [12], foreground detection [2], and background subtraction [13], but these methods failed with multiple humans in a scene. Additionally, several

* Correspondence: attique.khan440@gmail.com
[1]COMSATS Institute of Information technology, Wah Cant 40470, Pakistan
Full list of author information is available at the end of the article

feature extraction techniques are implemented for human detection such as histogram of oriented gradient (HOG) [14], HWF [15], Haar-like features [16], motion features [17], edge features [18], ACF [19], ISM [20], etc. These feature extraction techniques do not perform well when humans are not clearly visible or having extreme variations in their poses. Also, we noticed that the selection of relevant features significantly improves the classification results of human activities.

2 Method/experimental

To resolve the above-mentioned problems, we proposed a hybrid methodology which initially enhances the frames to extract the moving objects and later classifies the regions based on feature vector. The preprocessing step is very important to resolve the problems related to contrast and noise; therefore, we are giving a good weight to this step. Overall, the proposed method is divided into four primary steps: (a) frame acquisition and enhancement, (b) segmentation of moving region, (c) feature extraction and fusion, and (d) feature selection and action recognition. Also, in the proposed method, the classification of human is done with other objects such as vehicles. The sample proposed classification results are shown in Fig. 1. Our major contributions are enumerated below:

a) Implementation of a contrast stretching technique to make foreground object (human) maximally differentiable compared to the background.
b) Implementing velocity estimation to identify the motion regions which are later segmented using fusion of uniform distribution-based method and expectation-maximization (EM) segmentation.
c) Utilizing serial-based fusion technique which integrates HOG and texture features with LBP features [21].
d) Implementation of a joint entropy-PCA-based feature selection, based on maximal score. The selected features are later classified using a multi-class SVM for action recognition.
e) A detailed comparison of proposed action classification method with existing algorithms.

The selected datasets include MSR Action [22], CASIA [23], INRIA [24], Weizmann [25], KTH [25], UIUC [26], and Muhavi [27]. The proposed method is verified with five classification method, while multi-class SVM acts as a base classifier. The performance of our proposed algorithm is based on multiple measures which include recall rate, false positive rate, false negative rate, accuracy, and precision.

The schemes of this article are as follows: Method/experiments are presented in "Section 2." Related work is presented in "Section 3." "Section 4" elucidates the proposed framework. Experimental results are provided and discussed in "Section 5." Finally, "Section 6" concludes the article. The declaration is provided after "Section 6."

3 Related work

Hou [28] introduced a fast human detection and classification method using HOG features and Support Vector Machine (SVM) classifier. The proposed method comprised of three primary steps: (a) detection of moving regions from the video sequence, (b) extraction of HOG features of moving regions, and (c) utilizing SVM to classify the moving regions. Similarly, Qixiang [29] presented an Error-Correcting Output Code (ECOC) based on manifold clustering strategy for human detection. The proposed technique worked efficiently with multi-view and multi-posture problems. Dewei [30] introduced an online expectation-maximization (EM) algorithm in order to estimate foreground and background. Later, the human samples are cropped from the estimated foreground for HOG feature extraction. The SVM classification is employed. Javier et al. [31] introduced random subspace (RSM) algorithm for partially occluded human detection based on HOG and HOH-LBP feature. The extracted features were later classified by SVM method. Chung [32] proposed a hybrid technique to classify moving objects in the video sequences. Segmentation was performed using a histogram-based prominence approach; in

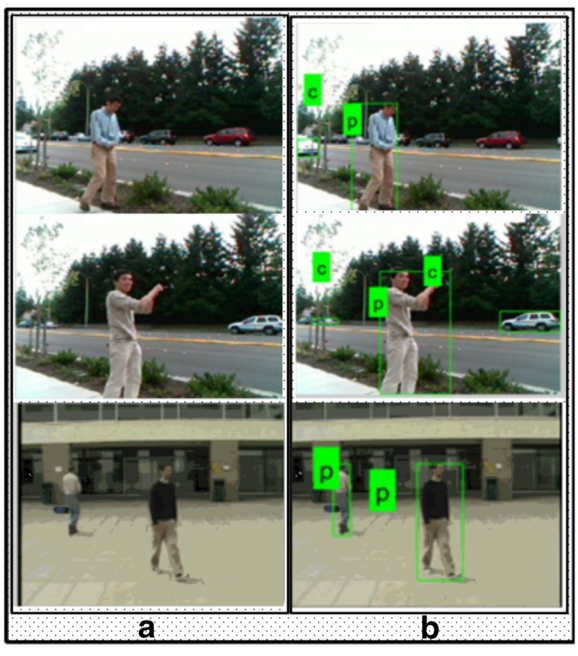

Fig. 1 Proposed human detection and classification results using EPFL and MSRA Action dataset

addition, shadow removal technique was also implemented to improve the performance of human classification. HOG and wavelets were fused with local shape features to acquire the vector of principle values, and finally, the classification was carried out with SVM classifier. Tudor [33] introduced a human detection approach using temporal difference-based procedure and morphological operations such as closing and filling. Also, HOG-based matching features were introduced for detection. Van [34] introduced a variant scale block-based HOG features for human detection. The extracted features were constructed with multiple blocks of a variable size which were easily distinguishing the positive and negative human samples. In addition, they also merged SVM with boosted tree algorithm to construct a state-of-the-art classifier which improved the efficiency of the classification problem. Conde [35] presented a human detection method for video surveillance that worked in an uncontrolled environment. The strategy was based on the integration of HOG and Gabor features [36]. The performance was compared with existing methods like HOG-based human detection [14] and boosted tree classifier for object detection [37]. This proposed method also showed improved performance under complex situations like blocking area, partly covered and with baggage. Kim. D [38] introduced two novel feature descriptors: (a) binary HOG and local gradient patterns and (2) fusion of local transform features with a combination of several local features namely LBP, LGP, and HOG by Adaboost feature selection technique. The described feature selection method greatly improved the performance of human detection with the fact that the selection of extracted features plays a vital role. Qiming [39] introduced a robust discriminant-weighted sparse partial least squares feature selection algorithm for human detection. This method reduced the dimensions of extracted features like partial least square to efficiently recognize positive and negative human samples based on the latent matrix. Finally, the selected features were classified with a boosted tree algorithm. Lee [40] introduced a new technique to recognize human head and shoulders by extracting their edge information and geometric features.

4 Proposed work

In this research, a new algorithm is presented for human detection by employing multiple video frames. The proposed algorithm performed better than several conventional techniques in terms of processing overhead which is achieved by minimizing the number of scanned regions. For this purpose, we implement sliding window concept by considering regions with variation in each successive frame while the static/unnecessary regions, mostly background, are ignored. The proposed work is mainly a conjunction of four primary sub-blocks: first, acquisition and refinement block; second, regions of interest detection block; third feature extraction and fusion block; and fourth feature selection and recognition block. The detailed flow of proposed algorithm is shown in Fig. 2.

4.1 Preprocessing

Preprocessing is a common name of operations with input video frame at the lowest level of observations. The input frame is captured by the given video sequence, which is originally RGB format. The major aim of preprocessing is an improvement of the frame data that enhances some foreground features for further processing. The steps of preprocessing are described below.

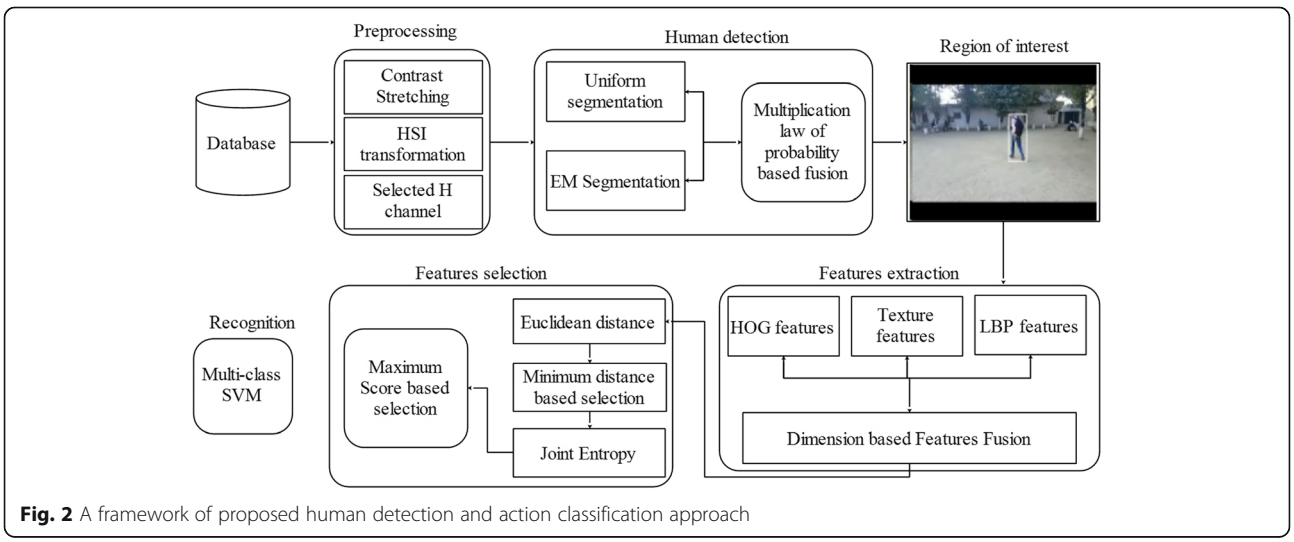

Fig. 2 A framework of proposed human detection and action classification approach

4.1.1 Frame enhancement

For the videos, the processing is done frame-by-frame because each processed frame provides us with different results having single or multiple moving objects. For this specific case, each frame can have a different number of moving objects. In the designed algorithm, firstly, image frames are enhanced and then transformed into hue-saturation-intensity (HSI) color space.

In the first step, contrast enhancement [41] is implemented for each RGB color channel, utilizing histogram equalization detailed with the Algorithm 1.

Algorithm 1. Histogram equalization of gray channels

Step 1: For each color channel with L gray levels and with pixels' intensity value k^1,

$$hist[k^1] = hist[k^1] + 1, \text{ when } i=0 \text{ to } L-1$$

Step 2: The cumulative frequency of histogram H_{cf} is given as:

$$H_{cf}[k^1] = h_{cf}[k^1 - 1] + hist[k^1]$$

Step 3: The equalized histogram is generated by H_{cf} and total number of pixels 'N' in the frame .

$$H_{eq}[k^1] = \lfloor \frac{L * h_{cf}[k^1] - N^2}{N^2} \rfloor$$

Step 4: For each k^1, replace previous values with the new mapping gray value $H_{eq}[k^1]$.

HSI transformation [42] is applied after contrast stretching of each color channel in RGB color space. The enhanced channels for red, blue, and green are calculated as:

$$\phi^{\sim R} = \frac{\phi^R}{\sum_{j=1}^{3}\phi^j} \tag{1}$$

$$\phi^{\sim G} = \frac{\phi^G}{\sum_{j=1}^{3}\phi^j} \tag{2}$$

$$\phi^{\sim B} = \frac{\phi^B}{\sum_{j=1}^{3}\phi^j} \tag{3}$$

Where $j = \{1, 2, 3\}$ for ϕ^R, ϕ^G, and ϕ^B which are original red, green, and blue channels, and $\phi^{\sim R}$, $\phi^{\sim G}$, $\phi^{\sim B}$ are modified red, green, and blue channels, respectively. The revised RGB color space is now transformed to HSI color space. To calculate saturation channel, given relations are:

$$\phi^{\sim S} = 1 - \frac{3}{\sum_{j=1}^{3}\phi^j} \times \alpha \tag{4}$$

where, i is the index for red, green, and blue channels, respectively, $\alpha = \min(\varsigma)$, where $\varsigma = \min\left(\phi^{\sim R}, , \phi^{\sim G}, , \phi^{\sim B}\right)$. The intensity channel is calculated using relation: $n(\varsigma)$

$$\phi^{\sim I} = \frac{\sum_{i=1}^{3}\phi^i}{3} \tag{5}$$

$$\phi^{\sim H} = cos^{-1}\frac{\left(0.5 \times \left(\phi^{\sim R} - \phi^{\sim G}\right) + \left(\phi^{\sim R} - \phi^{\sim B}\right)\right)}{\left(\sqrt{\left(\phi^{\sim R} - \phi^{\sim G}\right)^2 + \left(\phi^{\sim R} - \phi^{\sim B}\right)\left(\phi^{\sim G} - \phi^{\sim B}\right)}\right)} \tag{6}$$

where $\phi^{\sim H}$ is the hue channel with the conditions, if $\left(\frac{\phi^{\sim B}}{\phi^{\sim I}}\right) > \left(\frac{\phi^{\sim G}}{\phi^{\sim I}}\right)$ then $\phi^{\sim H} = 360 - \phi^{\sim H}$ and normalized to the range of $0 \rightarrow 1$. Finally, Hue $\phi^{\sim H}$ is utilized for further segmentation process. Figure 3 shows the enhanced sample L channel and histogram equalized color channel.

4.2 Frame segmentation

In this article, optical flow is used to identify motion of pixels in each frame sequence. After velocity estimation, a segmentation technique is implemented, named as uniform segmentation, which is improved with EM segmentation. The purpose of segmentation is to collect common features of an image such as color and texture. The fundamental problem faced was "how to exact foreground information with prominent variations in the contrast?" To deal with this problem, the proposed segmentation method worked significantly well. The detailed description of each section presented below.

4.2.1 Velocity estimation

To calculate the velocity estimation of motion regions in the video sequences is still an existing research topic in the area of pattern recognition. To estimate the flow of moving objects in the video sequences, we implemented an optical flow algorithm [43, 44]. The optical flow algorithm identifies the active pixels in the video sequences

Fig. 3 Preprocessing results. **a** Original frame. **b** Contrast stretching. **c** Hue channel. **d** Mesh plot

from time t to time $t+1$. It provides active pixel information in all three directions as horizontal, vertical, and time. The detailed description of optical flow algorithm is presented in the Algorithm 3, where h and v are motion vectors, h_{av} and v_{av} are the average of four neighbors, ζ_x, ζ_y are the displacement functions in x and y direction, ζ_t is the function of time, γ is the smoothing parameter, P is parallel flow, and D is normal flow. The motion information is segmented by uniform segmentation and EM segmentation [45] and then fuse both segmented frames by implementing multiplication law of probability.

Algorithm 3. Velocity estimation using optical flow

Step 1: $i = 0$

Step 2: Initialize $h^i v^i$ randomly

Step 3: Repeat until Convergence

{

$$h = h_{av} - \zeta_x \frac{P}{D}$$

and v is calculated using equation:

$$v = v_{av} - \zeta_y \frac{P}{D}$$

Where;

$$P = \zeta_x h_{av} + \zeta_y v_{av} + \zeta_t$$

and D is calculated as:

$$D = \gamma + \zeta_x^2 + \zeta_y^2$$

}

Step 4: Return

4.2.2 EM segmentation

Human detection under different conditions of visual surveillance is a key problem which requires prudent decisions. The proposed technique deals with moving object detection and classification by utilizing consecutive frame subtraction. In the real-time, video frames may contain multiple moving objects, e.g., humans and vehicles, and the proposed hybrid strategy classifies the moving regions with maximum accuracy. The central concept revolves around the detection of the motion vector from the optical flow which is embedded into the video sequence using segmentation of moving regions. For the detection of motion regions, we implement a hybrid technique, which is a combination of uniform distribution and EM segmentation. The implementation of EM segmentation is given as follows:

The EM segmentation [46] is an unsupervised clustering method and utilized for density estimation of the data points. The EM consists of two steps: (1) expectation and (2) maximization.

Supposedly, we have a set of observations; in our research, $\phi^H\sim$ frame is utilized as a input with $\xi_i = \varsigma_i^H$, for $i = 1$ with the ith pixel's value in $\phi^H\sim$ channel. The data are represented as $(1 \times \mathrm{D})$ matrix where dimension D represent hue pixels in the frame. To calculate the K-number of mixture densities, the following equation is used:

$$p\left(\varsigma_i, |\phi_j\right) = \sum_{j=1}^{k} \alpha_j p_j\left(\varsigma_i, ; m_j, , \sigma_j\right) \tag{7}$$

where α_j is a mixing parameter $\sum_{j=1}^{k} \alpha_j = 1$ for each Gaussian mixture model $\phi_j = (\partial_j, m_j)$, where ∂_j, m_j are the mean and standard deviations of mixtures. The variance is fixed to 1. A K-dimensional binary random variable z is introduced with all zero entries except the K^{th} entry $z_j = z_{j1}, z_{j2}, \ldots, z_{jk}$. The value of z_j satisfies the condition $z_j \in [0, 1]$. The joint distribution $p(\xi, z)$ is defined in terms of marginal distribution $p(z)$ and conditional distribution $p(\xi|\, z)$ given by$p(z)p(\xi|\, z)$.

$$\sum_z p(z)p(\xi|z) = \sum_{j=1}^{k} \alpha_j N\left(\xi|m_j, , \partial_j\right) \tag{8}$$

Let $g(\alpha_1, \alpha_2, \ldots, \alpha_{(k-1)}; m_1, m_2, \ldots, m_k; \sigma_1, \sigma_2, \ldots, \sigma_k)$ be a vector of estimated parameters.

E-Step: Calculate the post probability with heuristic initialized means, fixed variances, and randomly selected alpha. Evaluating the responsibilities:

$$\beta_{ij}^{(u)} = \frac{\alpha_i p(\xi; m_j^{(u)}, \partial_j^{(u)}}{\sum_{j=1}^{k} \alpha_j p\left(\varsigma_i; m_j^{(u)}, \sigma_j^{(u)}\right)}$$

M-Step: Re-estimating the parameters.

$$m_j^{(u+1)} = \frac{\sum_{i=1}^{N} \beta_{ij}^{(u)} \xi_i}{\sum_{i=1}^{N} \beta_{ij}^{(u)}} \tag{9}$$

$$\partial_j^{(u+1)} = \frac{\sum_{i=1}^{N} \beta_{ij}^{(u)} \left(\xi_i - m_j^{(u)}\right) T\left(\xi_i - m_j^{(u)}\right)}{\sum_{i=1}^{N} \beta_{ij}^{(u)}} \tag{10}$$

$$\alpha_j^{(u+1)} = \frac{1}{N} \sum_{i=1}^{N} \beta_{ij}^{(u)} \tag{11}$$

for convergence:

$$\widetilde{\Phi_{EM}} = \max_j \frac{e^{(-1/2)\left(\xi_i - m_j^{EM}\right)\left(\partial_j^{EM}\right)-1\left(\xi_i - m_j^{EM}\right)T}}{\left|\partial_j^{EM}\right|^{-1/2}} \tag{12}$$

where ξ_i represents data, $\widetilde{\Phi_{EM}}$ is final EM frame, and m_j^{EM}, ∂_j^{EM} are the means and the standard deviation respectively.

4.2.3 Uniform segmentation

The uniform distribution based segmentation technique utilized for accurate detection of multiple humans in a given scenario. This technique is also well performed in low-resolution sequences and high variation. The uniform segmentation work based on mean and variances of motion regions. The idea behind uniform segmentation is that equally utilized each motion pixel and create a border based on their mean and change in variances. The mean and variances of uniform distribution are calculated as follows:

$$\mu = \int_q^r \phi f(\phi) d\phi, \frac{1}{(r-q)} \int_q^r \phi d\phi \tag{13}$$

$$= \frac{1}{r-q}\left[\frac{\phi^2}{2}\right]_q^r \tag{14}$$

$$\mu = \frac{r+q}{2} \tag{15}$$

where r and q denote the maximum and minimum motion pixels of the processed frame. Then calculate the variances of motion pixels as follows:

$$E(\phi^2) = \int_q^r \phi^2 f(\phi) d\phi \tag{16}$$

$$= \frac{1}{(r-q)}\left[\frac{\phi^3}{3}\right]_q^r \tag{17}$$

$$E(\phi^2) = \frac{[r^2 + q^2 - rq]}{3} \tag{18}$$

Hence, the variance is defined as:

$$\sigma^2 = E(\phi^2) - [E(\phi)]^2 \tag{19}$$

$$\frac{[r^2 + q^2 - rq]}{3} - \frac{(r+q)^2}{4} \tag{20}$$

$$\sigma^2 = \left[\frac{r^2 + q^2 - 2rq}{12}\right] \tag{21}$$

Now inserting μ and σ^2 in below equation and extract the human from video sequence.

$$\xi(\phi) = \phi^n \left(2\mu - \sigma^2\right)^{\frac{3}{2}} \tag{22}$$

$$\xi(\phi) = \phi^n \left(2\mu - \sigma^2\right) \sqrt{(2\mu - \sigma^2)} \tag{23}$$

The graphical representation of EM and uniform segmentation is shown in Fig. 4. After this, both segmented frames are fused by implementing of multiplication law of probability. The fused frame has more information as compared to the individual frame. The basic goal of frame fusion is to improve the detection results in terms of graphical and tabular.

4.2.4 Frames fusion

After foreground segmentation, both segmented frames are fused to get a new foreground which embeds more information compared to a single segmented frame. The main goal of image fusion is to integrate the common information of two images into one image, which contains more information and is easier for human and machine perception compared to individual image [47]. In this article, the fusion of the two segmented frames is done based on the multiplicative law of probability. The fusion using multiplication law is described in the following.

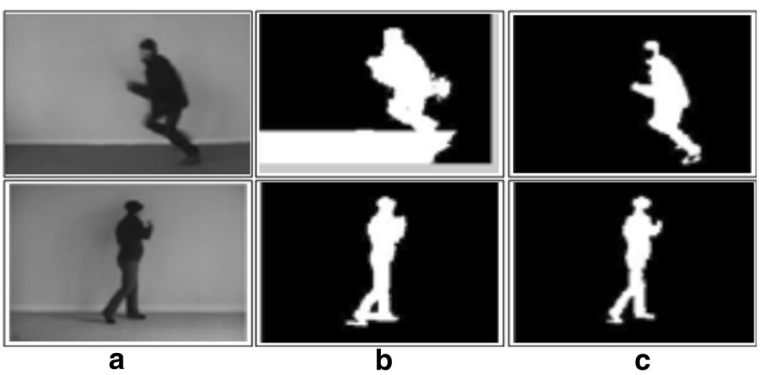

Fig. 4 Segmentation results. **a** Original frame. **b** EM segmentation. **c** Uniform segmentation

Let m_1 denotes the number of $\xi(\phi)$ pixels, and m_2 denotes the number of $\phi_{EM}\sim$ pixels, where $\xi(\phi)$ is the uniform segmentation frame and $\phi_{EM}\sim$ is the EM segmented frame. Let m_3 denotes the matching pixels between $\xi(\phi)$ and $\phi_{EM}\sim$. Then the fusion of both segmented frames is defined as:

$$\widetilde{\phi_{fusion}} = P\left(\xi(\phi), \widetilde{\phi_{EM}}\right) = \frac{m_3}{n}, \frac{m_3}{n} \times \frac{m_1}{m_1}, = \frac{m_1}{n} \times \frac{m_3}{m_1}, \tag{24}$$

$$\phi_{fusion}\sim = \xi(\phi) \times \mathrm{P}(\xi(\phi)\phi_{EM}\sim) \tag{25}$$

The fusion results are shown in Figs. 5, 6, and 7. Also, the multiple human detection results are shown in Figs. 8 and 9.

4.3 Frames representation

In computer vision, feature extraction is a major step for a description of the input query. For this purpose, many feature extraction techniques are implemented as discussed in literature review. In this article, we extract three types of features as HOG and texture features with LBP features. The texture features are also represented as GLCM (gray-level covariance matrix) features. The HOG features are originally introduced by Dalal et al. [14] which produce shape-based features. The HOG features are implemented in four steps: (1) gradient computation using magnitude and orientation, (2) cell and blocks creation, (3) votes, and (4) normalize the block size. The block size is 16×16, and the size of the input query is fixed at 128×64. Hence, the size of HOG feature vector is 1×3780.

Secondly, extracted eight GLCM features which include contrast, cluster shade, homogeneity, the sum of variances, autocorrelation, energy, inverse difference moment normalized, and difference entropy. Then, calculate the mean, standard deviation, and skewness against each feature and obtain a new feature vector of size 1×24. Thirdly, extract local binary patterns (LBP) of detected regions having feature vector size (1×59). Whereas, LBP features which are originally introduced in [21] represent human silhouette more evident and also resolve the problem of contrast of bright object against a dark background.

The LBP features are calculated as follows:

$$\phi_{u,v} = \sum_{\Omega=1}^{\varrho-1} s(\Phi_\Gamma \Phi_\lambda), \text{where } s(\Gamma) = \begin{cases} 1 & \Gamma \geq 0 \\ 0 & \Gamma < 0 \end{cases} \tag{26}$$

$\varrho = 8$, which are the total number of neighboring pixels, Φ_λ is the value of the pixels at (u, v), and Φ_Ω is the value of pixels in the Ωth location on the circle of the radius R around Φ_λ. The size of LBP feature vector is 1×59 that are further fused with HOG and GLCM features based on their vector size.

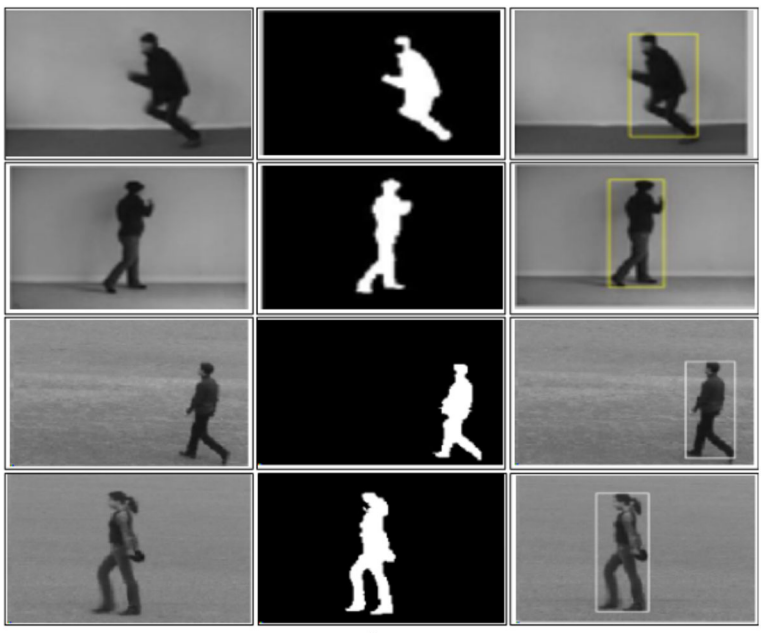

Fig. 5 Frames fusion results. **a** Original. **b** Fused frame. **c** ROI detection

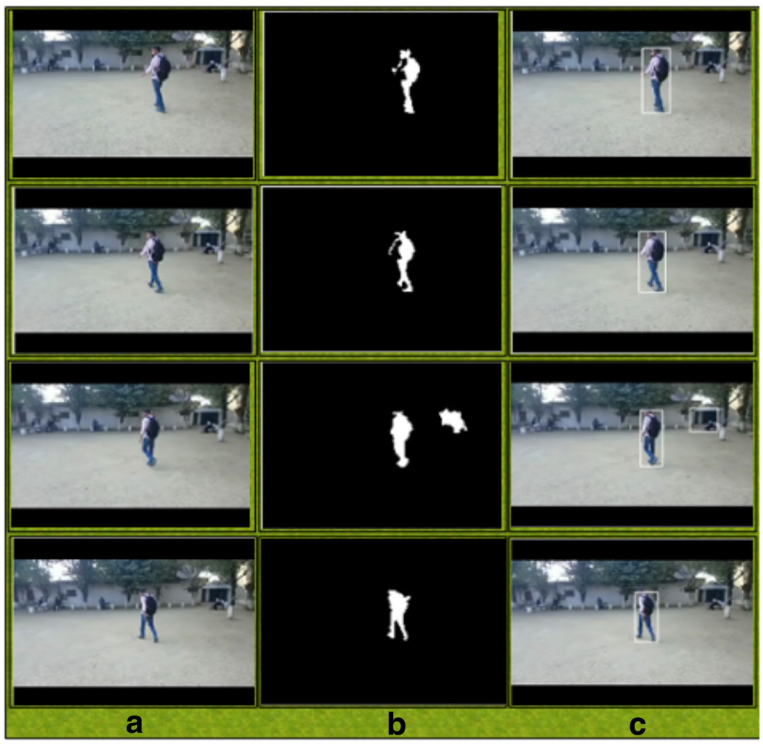

Fig. 6 Segmentation results of own recorded videos

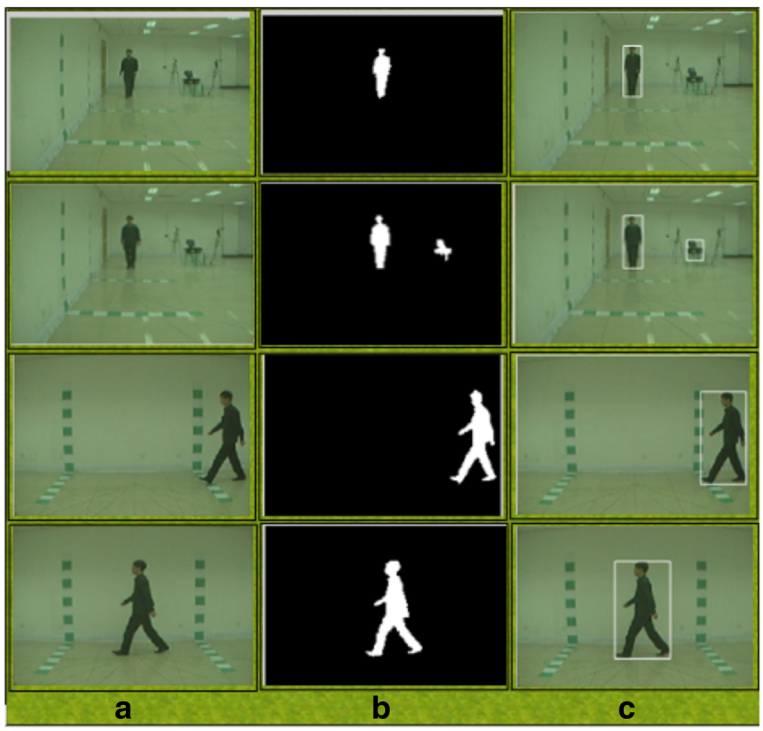

Fig. 7 Segmentation results using CASIA dataset

Fig. 8 Detection and classification results of proposed algorithm using different video sequences

The main reason of fusing these three types of feature sets is to increase the performance of human action recognition and also improve the classification rate of human and other objects (i.e., vehicles) in the complex scenarios as high brightness environment and in the dark background. The proposed feature fusion method not just works better in the area of high brightness and dark background, but also produces a significant improvement of detection performance with respect to the original HOG and LBP features.

For feature fusion, a new technique is implemented which concentrate on the size of vector dimension. As explained above, the size of the extracted feature vector is 1×3780, 1×24, and 1×59 for HOG, texture, and LBP, respectively. Then the fusion is defined as follows:

Suppose C_1, C_2, ... C_n are known human action classes, which have to be classified. Let $\Delta = \{\phi \vee \phi \in R^N\}$ denotes the number of training samples. The three extracted feature vectors are $\{\Upsilon_{HOG}, , \Upsilon_{txt}, , \Upsilon_{LBP}\} \in R^{N_{HOG+txt+LBP}}$. The size of each extracted feature vector is defined as:

$$W_1 = \{h_1, \ldots h_k\}$$
$$W_2 = \{t_1, \ldots t_k\}$$
$$W_3 = \{l_1, \ldots l_k\}$$

where W_1, W_2, W_3 denote the size of HOG, texture, and LBP features. The number of features is represented by $k \in \{3780,24,59\}$, respectively. As we know the size of the extracted set of feature vector (i.e., $\Upsilon_{HOG} \rightarrow 1 \times 3780$, $\Upsilon_{txt} \rightarrow (1 \times 24)$, $\Upsilon_{LBP} \rightarrow (1 \times 59)$, the extracted vector is integrated as:

$$F(\phi) = \sum_h^{W_1} \sum_t^{W_2} \sum_l^{W_3} \{\Upsilon_{W_1+W_2+W_3}\} \quad (27)$$

$$F(\phi) = \sum_h^{W_1} \Upsilon_{S_1} + \sum_t^{W_2} \Upsilon_{S_2} + \sum_t^{W_3} \Upsilon_{S_3}$$

$$F(\phi) = \{(1 \times 3780) + (1 \times 24) + (1 \times 59)\} \quad (28)$$

$$Final(\phi) = \{1 \times 3863\}$$

4.4 Feature selection

In literature, several feature reduction techniques are implemented for human detection and action recognition; but up to our limited knowledge, no one has implemented a feature selection technique. The purpose of feature selection technique is to select a subgroup of features from the high dimensional feature set for a compact and accurate data classification. The main reason is to select the most prominent feature to build simpler and faster model. Another reason of feature selection is to find the smallest group of features that maximally increase the performance of proposed method.

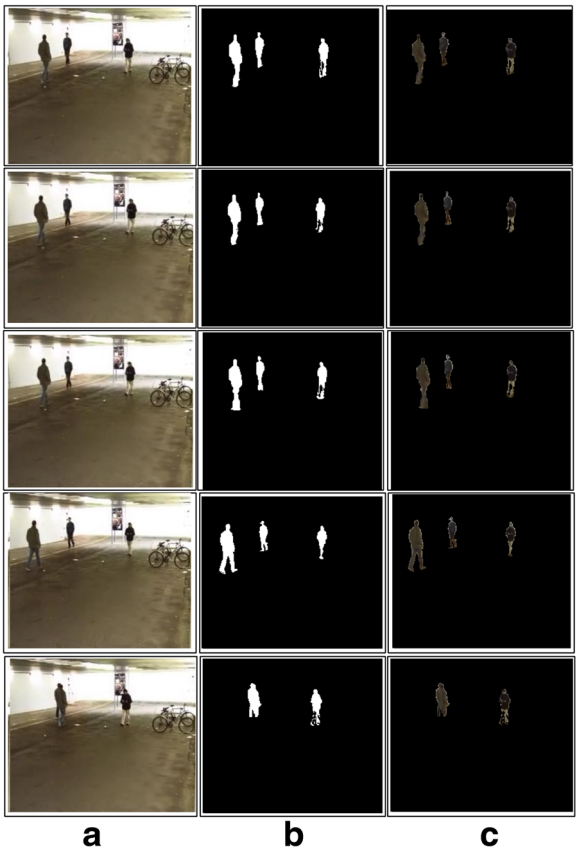

Fig. 9 Sample results of multiples human detection from EPFL and own recorded videos

In this article, we implement a new feature selection technique based on Euclidean distance and joint entropy with PCA. The proposed selection method consists of two major steps. First, calculate the Euclidean distance between fused features and select the best 500 features based on minimum distance.

$$\overrightarrow{D} = \sqrt{\sum_{f=1}^{d} \left(F(\phi)_{i+1} - F(\phi)\right)^2} \tag{29}$$

where $F(\phi)_{i+1}$ denotes the $i+1$ frame, and $F(\phi)$ is the current frame. This process continues up to d features, where d denotes the dimension of fused feature vector, which is 1×3863.

$$M\left(\overrightarrow{D}\right) = \text{Minmum}\left(\overrightarrow{D}, \delta\right) \tag{30}$$

where the selected parameter $\delta = 500$. Then, perform PCA on selected vector $M\left(\overrightarrow{D}\right)$ and find out the score of each feature. The joint entropy is performed on score features, and the best 356 score features are selected for classification and recognition.

$$\text{Entropy} = \sum_{f_1}^{d} \sum_{f_2}^{d} P(f_1, , f_2)\ log\ P(f_1, , f_2) \tag{31}$$

where f_1, f_2 are the current and previous minimum distance based on the selected features. Hence, the final feature vector has a dimension of 1×356, which is later fed to multi-class SVM [48] for classification of human and recognition of their actions. The human classification results are shown in Figs. 10 and 8.

5 Results and discussion

5.1 Evaluation protocol

The proposed algorithm is validated in two major steps: (a) human classification and (b) action recognition. In the first step, human classification is done as shown in Figs. 2 and 11. For human classification, multiple humans and background moving objects are detected and classified using the extracted features. Three publicly available datasets (i.e., MIT pedestrian [49], CAVIAR [47], and BMW [21]) are utilized for training the classifier for human classification. For testing our proposed algorithm with other human detection, approaches are tested on three publicly available datasets, multicamera pedestrian video (EPFL) [50], CASIA [23], and MSR Action [22] dataset, and also tested on our own collected videos. Each dataset with their classification results on three classifiers, M-class SVM, EBT, KNN, and linear discriminant analysis (LDA), are described in detail. Three performance measures, area under the curve (AUC), false negative rate (FNR), and accuracy, are considered.

In the second phase, four publicly available datasets are utilized for human action recognition. The selected datasets are Weizmann [25], KTH [25], UIUC [26], and Muhavi [27], which are utilized for both training and testing of ration 50:50. Four classification methods (i.e., weighted KNN (W-KNN), subspace discriminant analysis (SDA), logistic regression (LR), and multi-class SVM) are used for validation. The simulations are done in Matlab R2016a utilizing a personal computer, Intel Core I7, a 3.40 GHz processor with 8GB of RAM. The description of training dataset is listed below.

5.2 Results

The results section consists of two subsections: (a) human classification results and (b) action recognition results. The detail of each section is described below.

5.2.1 Human classification results

In this subsection, the algorithm's performance is analyzed and validated through experiments. The testing datasets are multicamera pedestrian video (EPFL), CASIA, and MSR Action. Also, to validate the performance of the proposed algorithm, we recorded our own videos.

A. **EPFL video sequences**

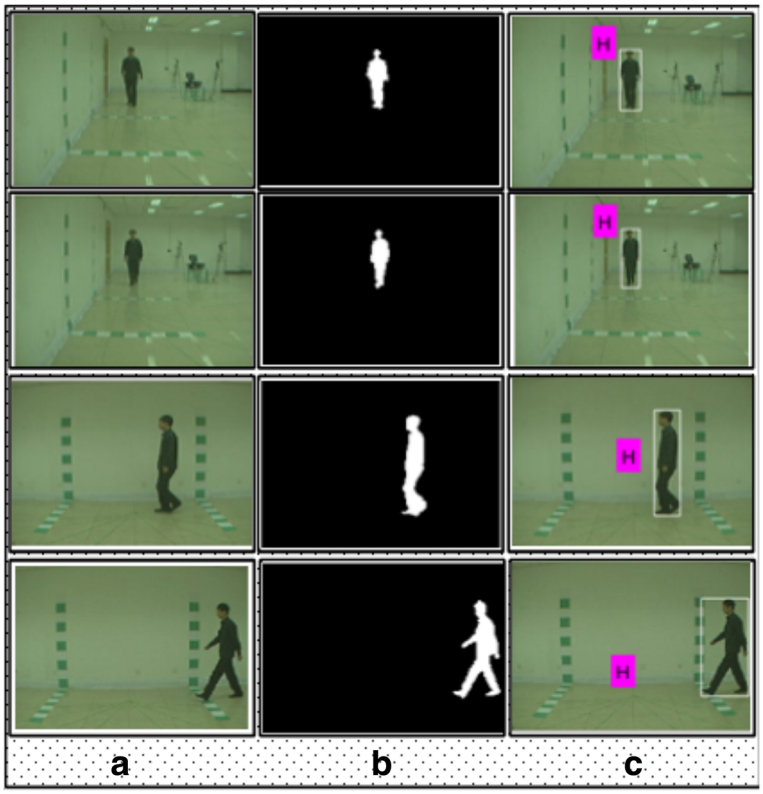

Fig. 10 Classification results for CASIA dataset

In EPFL [50] video dataset, we selected ten videos of two different views namely campus and passageway. This dataset consists of several types of video sequences and we selec the campus and passageway sequences for experiments. Table 1 detailed the performance of the proposed algorithm on EPFL video sequences compared with two different classifiers. Also, the average results of this dataset are described in Table 2. The multi-class SVM performs significantly well and obtains maximum classification accuracy of 99.56% on Computer Vision Laboratory (CVLAB) campus video sequences and CVLAB passageway video sequences. The processing time of the proposed algorithm using this dataset is one frame per second (FP/S), and we also improved this execution time by using GPU.

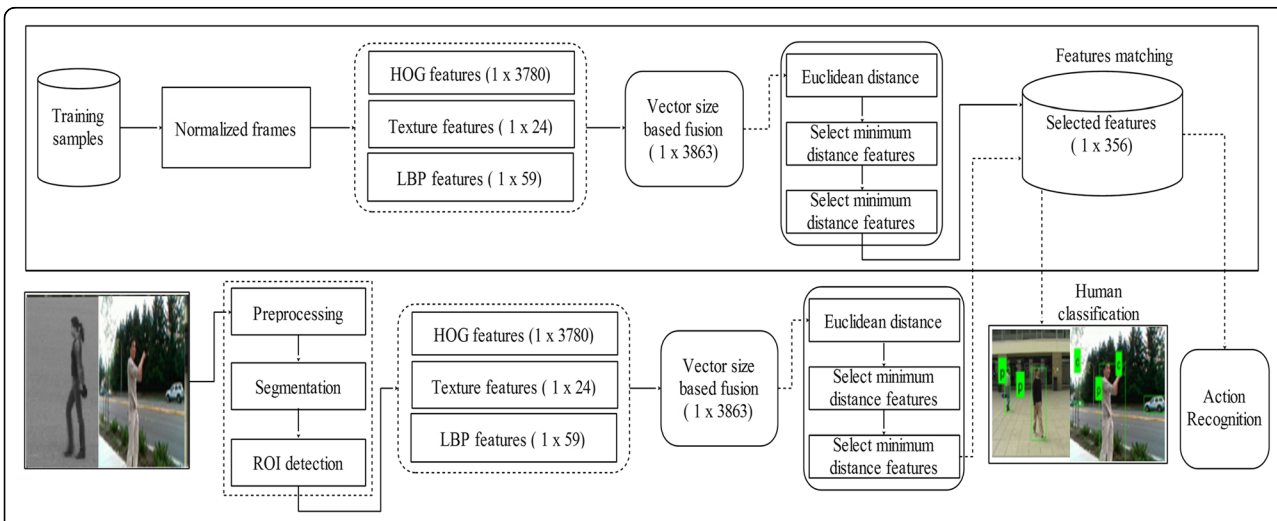

Fig. 11 Training and testing procedure using a fusion of three types of features

Table 1 Proposed algorithm results using EPFL dataset

Dataset	Classifier	Sequence	ACC (%)	TPR	FPR	Recall (%)	FDR (%)	FNR (%)	AUC
CVLAB campus sequences	M-class SVM	S1:Cv0	100	1.00	0.00	100	0.00	0.00	1.00
		S1:Cv1	100	1.00	0.00	100	0.00	0.00	1.00
		S2:Cv2	98.70	0.98	0.01	98.00	2.00	1.30	0.99
	KNN	S1:Cv0	98.20	0.98	0.01	98.25	1.75	1.80	0.98
		S2:Cv1	99.00	0.99	0.01	99.15	0.85	1.00	0.98
		S2:Cv2	96.90	0.96	0.03	97.00	3.00	3.10	0.96
	EBT	S1:Cv0	99.90	0.99	0.00	100	0.00	0.10	1.00
		S2:Cv1	99.70	0.99	0.00	100	0.00	0.30	0.99
		S2:Cv2	98.10	0.98	0.01	98.20	1.80	1.90	0.99
CVLAB passageway sequences	M-class SVM	S1:Cv0	100	1.00	0.00	100	0.00	0.00	1.00
		S2:Cv1	100	1.00	0.00	100	0.00	0.00	1.00
		S2:Cv2	98.70	0.98	0.01	98.00	2.00	1.90	0.99
	KNN	S1:Cv0	98.20	0.98	0.01	98.25	1.70	1.80	0.98
		S2:Cv1	99.00	0.98	0.01	99.15	0.85	1.00	0.99
		S2:Cv2	96.90	0.96	0.03	97.00	3.00	3.10	0.96
	EBT	S1:Cv0	100	1.00	0.00	100	0.00	0.00	1.00
		S2:Cv1	99.80	0.99	0.00	100	0.00	0.20	0.99
		S2:Cv2	98.10	0.98	0.01	98.00	2.00	1.90	0.99

B. **MSRA Action dataset**

MSR Action dataset [22] is originally designed for detection and classification of moving object in the background. This dataset contains 16 video sequences having three action classes. The resolution of each video sequence is 320×240 in static camera environment. To validate the performance of the proposed algorithm, tenfold cross-validation is held, and their result is described in Table 3. Also, the average results of tested video sequences are depicted in Table 4. The proposed algorithm has a maximum classification rate of 100% on tested videos. Each video sequence is processed in the frames, and the execution time of one frame is 0.50 FP/S.

C. **CASIA dataset**

The CASIA action dataset [23] is originally designed for human gait recognition. This dataset is a collection of various video sequences of distinct angle and views. There are 1446 video sequences containing eight actions of a single person. Each action is performed by 24 subjects. To check the performance of the proposed algorithm on this dataset, tenfold cross-validation is performed. Table 5 described the results of the proposed algorithm using CASIA dataset having a maximum accuracy of 98.70%, and fast positive-region reduction (FPR) is 0.01. The frame execution time using CASIA dataset is four frames per second. The algorithm performs better on this dataset as compared to EPFL and MSR Action dataset.

D. **Own recorded videos**

To test the performance of the proposed algorithm, we recorded our own video sequences in a complex background. The videos are recorded in front of a university cafeteria and in a terrace. The total video

Table 2 Average results of proposed algorithm using EPFL dataset

Dataset	Classifier	ACC (%)	TPR	FPR	Recall (%)	FDR (%)	FNR (%)	AUC
CVLAB campus sequences	M-class SVM	99.56	0.99	0.01	99.33	0.66	0.43	0.99
	KNN	98.03	0.98	0.02	98.13	1.86	1.96	0.97
	EBT	99.23	0.99	0.01	99.40	0.60	0.76	0.99
CVLAB passageway sequences	M-class SVM	99.56	0.99	0.01	99.33	0.66	0.63	0.99
	KNN	98.03	0.97	0.03	98.13	1.85	1.96	0.97
	EBT	99.30	0.99	0.01	99.33	0.66	0.70	0.99

Table 3 Proposed algorithm results using MSR Action dataset

Dataset	Classifier	Sequence	ACC (%)	TPR	FPR	PPV (%)	FDR (%)	FNR (%)	AUC
ssMSR Action dataset	M-class SVM	S1:Cv1	100	1.00	0.00	100	0.00	0.00	1.00
		S1:Cv14	100	1.00	0.00	100	0.00	0.00	1.00
		S1:Cv16	100	0.99	0.00	100	0.00	0.00	0.99
	K-Nearest neighbor	S1:Cv1	97.30	0.97	0.03	97.30	2.70	2.70	0.97
		S1:Cv14	92.50	0.94	0.05	91.15	8.85	2.70	0.94
		S1:Cv16	97.30	0.97	0.02	97.00	3.05	2.70	0.97
	Ensemble boosted tree	S1:Cv1	100	0.99	0.00	100	0.00	0.10	0.99
		S1:Cv14	100	1.00	0.00	100	0.00	0.00	1.00
		S1:Cv16	99.80	0.99	0.00	100	0.00	0.20	0.99

sequences are 120 containing two persons, and each person recorded 60 videos. The videos are recorded by 10 MP cameras. Five measures are implemented to check the algorithm performance including accuracy, FNR, recall, FPR, and AUC. The testing results of the proposed algorithm on our own videos are described in Table 6, and the maximum accuracy rate is 99.60% and FPR is 0.01. Due to high-resolution videos, the proposed algorithm performs a little bit slower as compared to standard datasets. The execution time of one frame is approximately 1.5 s.

E. **INRIA person dataset**

INRIA person dataset [24] is considered to be one of the most comprehensive and flexible datasets containing 4754 images divided into testing and training samples under multiple scenarios of positive and negative samples. The number of positive and negative training samples are 2416 and 912, respectively. For test image samples, it includes 1126 positive and 300 negative samples. The classification results of this dataset are described in Table 7 having a maximum accuracy of 98.80%.

5.2.2 Action recognition results

The action recognition results are validated on four publicly available datasets including Weizmann, KTH, UIUC, and Muhavi. Five performance measures are calculated for analyzing the performance of proposed algorithm including accuracy, FPR, FNR, AUC, and recall rate. The execution time on these datasets is similar, and in 1 second, ten frames are performed. The detailed description of each dataset is described below.

Table 4 Average results of proposed algorithm using MSR Action dataset

Dataset	Classifier	ACC (%)	TPR	FPR	PPV (%)	FDR (%)	FNR (%)	AUC
MSR Action dataset	M-class SVM	100	0.99	0.00	99.99	0.00	0.01	0.99
	K-NN	95.70	0.96	0.03	95.15	4.86	4.30	0.96
	EBT	99.93	0.99	0.01	99.80	0.02	0.06	0.99

A. **Weizmann dataset**

The Weizmann dataset [23] consists of 90 human action videos. It contains ten human action classes performed by nine actors. For validation of proposed algorithm, 50:50 strategies are performed. It means, 50% of the videos are utilized for training the classifier and 50% are utilized for testing the proposed algorithm. The results are tested in two scenarios: (a) without feature selection algorithm and (b) utilizing feature selection algorithm. The results are depicted in Table 9 having a maximum recognition rate of 95.80% and FNR 4.20%. The goal of both scenarios is to show how feature selection algorithm effect on the recognition. The proposed algorithm is also compared with an existing action recognition method in Table 14, which shows that the proposed method performs significantly better compared to other methods.

B. **KTH dataset**

The KTH dataset [23] consists of 599 video sequences of six human action classes, where each action is performed by 25 persons. The videos are captured by outdoor, indoor, different variations, and

Table 5 Results of proposed algorithm compared using four classifiers on CASIA dataset

Method	AUC	FPR	FNR (%)	Recall (%)	Accuracy (%)
M-class SVM	0.99	0.01	1.3	98.65	98.70
DT	0.98	0.02	2.1	97.90	97.90
LDA	0.99	0.01	1.8	98.15	98.20
KNN	0.98	0.02	1.9	98.05	98.10
EBT	0.97	0.07	7.3	93.55	92.70

Table 6 Results of the proposed algorithm compared using four classifiers on our recorded videos

Method	AUC	FPR	FNR (%)	Recall (%)	Accuracy (%)
M-class SVM	0.99	0.01	0.4	100	99.60
DT	0.96	0.06	3.3	96.40	96.70
LDA	0.99	0.01	0.6	99.65	99.40
KNN	0.99	0.01	0.6	99.70	99.40
EBT	0.99	0.02	1.1	99.00	98.90

distinct clothes with lighting variations. For the validation of proposed algorithm, 50:50 strategies are done for training and testing the proposed algorithm. From Table 10, the results of the proposed algorithm with and without utilizing feature selection algorithm are presented. The results are significantly better when feature selection algorithm is utilized and obtain maximum recognition rate of 99.30%. Also, the proposed algorithm is compared with existing methods in Table 14, which shows the authenticity of the proposed method.

C. **UIUC dataset**

The UIUC dataset [22] consists of 532 high-resolution video sequences of 14 human action classes, and every action is performed by eight persons. All the video sequences are recorded indoor scenes. For the validation, we make a strategy of 50:50 for training and testing of the proposed algorithm. The proposed method is validating in two scenarios as without feature selection algorithm and feature selection algorithm. The results are depicted in Table 11 having a maximum recognition rate (RR) of 99% on feature selection algorithm and also reduces the FNR up to 1%. Also, a comparison of the proposed algorithm with existing methods is shown in Table 14, which shows that proposed method performs significantly well as compared to existing methods.

Table 7 Results on INRIA person dataset using eight classification methods (decision tree (DT), linear discriminant analysis (LDA), cubic-SVM (C-SVM), logistic regression (LR))

Method	AUC	FPR	FNR (%)	Recall (%)	Accuracy (%)
DT	0.930	0.064	5.6	93.6	94.4
LDA	0.976	0.073	6.8	92.6	93.2
C-SVM	0.991	0.019	1.4	98.4	98.6
LR	0.913	0.084	5.9	91.5	94.1
W-KNN	0.973	0.105	7.1	89.6	92.6
C-KNN	0.972	0.117	7.9	88.4	92.1
EBT	0.993	0.143	3.1	96.85	96.9
M-class SVM	0.990	0.015	1.2	98.7	98.8

Table 8 Comparison of the proposed human classification algorithm with existing methods

Method	Accuracy (%)
Beipping et al. [28]	90.23
Xia et al. [51]	98.40
Ye. Q et al. [11]	99.01
Conde et al. [35]	99.13
Ped et al. [49]	98.94
Seeraj et al. [52]	87.46
Liang et al. [32]	97.60
Proposed	99.48

D. **Muhavi dataset**

Muhavi action dataset [25] consists of 1904 videos of 17 human action classes. Each action is performed by 17 actors on eight different cameras. Each actor performs one action three times in each video sequence. For the validation of the proposed algorithm, 50:50 strategies are done for training and testing. Seven action classes are selected for testing including ClimbLadder, CrownOnKnees, JumpOverGap, Kick, Punch, RunStop, and walking. The proposed method is tested in two scenarios (i.e., utilizing feature selection algorithm; without utilizing feature selection algorithm). The results are depicted in Table 12 having a maximum accuracy of 99.40% and FNR 0.60, which is confirmed by their confusion matrix given in Table 13. Also, a comparison of the proposed method with existing methods is done in Table 14, which shows the authenticity of the proposed method.

5.3 Discussion

Our proposed methodology, in general, is the conjunction of two primary phases: (a) human detection and classification and (b) human action recognition. Each phase is the amalgamation of the series of steps as shown in Figs. 2 and 11. In the first phase, human is

Table 9 Recognition results on Weizmann dataset

Method	Algorithm		Measures			
	FS	WFS	PPV (%)	AUC	FNR (%)	RR (%)
W-KNN	✓		91.96	0.993	7.20	92.80
		✓	86.10	0.980	11.1	88.90
SDA	✓		82.70	0.984	15.80	84.20
		✓	60.10	0.920	38.80	61.20
LR	✓		71.70	0.933	26.10	73.90
		✓	61.70	0.931	37.1	62.90
M-class SVM	✓		95.30	0.973	4.20	95.80
		✓	91.20	0.983	7.9	92.10

Table 10 Recognition results on KTH dataset

Method	Algorithm		Measures			
	FS	WFS	PPV (%)	AUC	FNR (%)	RR (%)
W-KNN	✓		96.30	0.998	2.70	97.30
		✓	82.70	0.961	15.80	84.20
SDA	✓		82.90	0.951	15.70	84.30
		✓	70.78	0.947	23.50	76.50
LR	✓		77.30	0.945	21.0	79.0
		✓	68.81	0.928	25.20	74.80
M-class SVM	✓		98.93	0.993	0.7	99.30
		✓	92.1	0.978	6.70	93.30

segmented from the given video sequences using a combination of EM and uniform segmentation techniques prior to feature extraction of ROI. The primary reason for segmentation here is to ideally extract the object/s (human in our case) by ignoring the background. Therefore, we use multiple segmentation methods. The crux is, EM technique is considering foreground and background pixels based on random distribution of pixels' intensity level for making clusters. This is the core reason, EM binary images have a lot of noise factor (Fig. 4). On the other hand, uniform distribution is considering only those values with an approximately the same range of mean and variance. This is the reason the output is maximumly differentiable (foreground and background) in a binary image. The other advantage of uniform segmentation is the avoidance of clusters' overlapping. Additionally, EM works well with multiple clusters, but uniform works well for less number of clusters. Therefore, in proposed, both techniques are embedded so that approach works better for different kinds of images (with less and with more number of clusters). Technically, EM and uniform segmentations are fused, based on multiplication law of probability. In the next step, features are selected with the proposed method and fed into multi-class SVM for classification. The EPFL, MSRA Action, CASIA, INRIA,

Table 11 Recognition results on UIUC dataset

Method	Algorithm		Measures			
	FS	WFS	PPV (%)	AUC	FNR (%)	RR (%)
W-KNN	✓		98.18	0.999	2.10	97.90
		✓	83.62	0.980	14.6	85.40
SDA	✓		93.20	0.980	5.3	94.70
		✓	85.25	0.965	14.9	85.10
LR	✓		95.43	0.990	2.3	97.70
		✓	83.98	0.981	16.0	84.0
M-class SVM	✓		98.6	0.994	1.0	99.0
		✓	96.1	0.990	2.1	97.90

Table 12 Recognition results on Muhavi dataset

Method	Algorithm		Measures			
	FS	WFS	PPV (%)	AUC	FNR (%)	RR (%)
W-KNN	✓		98.42	0.990	1.60	98.40
		✓	93.92	0.950	5.0	95.00
SDA	✓		98.74	0.990	1.30	98.70
		✓	88.85	0.865	10.90	89.10
LR	✓		88.80	0.970	12.0	88.0
		✓	82.98	0.940	17.0	83.0
M-class SVM	✓		99.80	0.999	0.60	99.40
		✓	96.40	0.990	3.1	96.90

and own recorded videos are used for human detection and classification. In EPFL dataset, two types of video sequences are collected as mentioned in section A, and we obtained average correct classification results of 99.56% as presented in Tables 1 and 2. Secondly, the MSRA Action dataset is used for classification which is more complex compared to EPFL, where we obtained 100% accuracy with M-SVM as presented in Tables 3 and 4. Thirdly, the CASIA, INRIA person, and own recorded videos are utilized for testing, and the classification results are 98.70, 98.80, and 99.60%, respectively, presented in Tables 5, 6, and 7. Finally, the proposed classification results are compared with the recent algorithms, as described in Table 8, having improved performance. The graphical human segmentation, detection, and classification results are shown in Figs. 4, 5, 6, 7, 8 and 10.

In the second phase, the action recognition results are obtained using proposed feature selection algorithm, and we also compared it with WFS (without feature selection) algorithm. The primary reasons to opt feature selection strategy are (1) it reduces the overfitting problems due to a decrease in redundancy, as noise values are ignored in this complete process; (2) it improves the accuracy, as redundant data has already been skipped, so only relevant information is considered for a

Table 13 Confusion matrix of Muhavi dataset (L (ClimbLadder), N (CrownOnKnees), J (JumpOverGap), K (Kick), P (Punch), R (RunStop) and W (walking))

Class	L	N	J	K	P	R	W
L (%)	99.8					0.2	
N (%)		100					0.1
J (%)			100				
K (%)				100			
P (%)					100	0.1	
R (%)		0.1		0.1	0.9	98.2	0.7
W		0.1			0.1	0.2	100

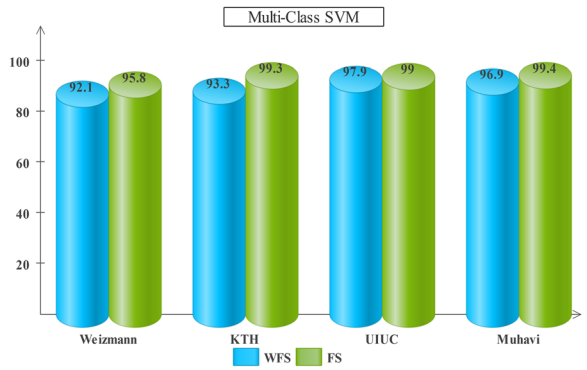

Fig. 12 Comparison of proposed feature selection method with WFS

decision; and (3) it decreases the computational time, as irrelevant and redundant information has already been removed, so only valuable and salient information is going to be used for training and finally testing. To validate the results, a tenfold cross-validation is being performed for the proposed algorithm where 50:50 strategy is adopted for training and testing. Four publicly available datasets (i.e., Weizmann, KTH, UIUC, and Muhavi) are utilized for the validation of the proposed method. The recognition results are presented in Tables 9, 10, 11, 12 for Weizmann, KTH, UIUC, and Muhavi dataset, which show the results from FS and WFS. The maximum WFS recognition results are 92.10, 93.30, 97.90, and 96.90%, whereas, the proposed feature selection results are 99.30, 99.0, 99.0, and 99.40%, respectively, for Weizmann, KTH, UIUC, and Muhavi dataset. The recognition results on Muhavi dataset are confirmed by their confusion matrix given in Table 13, which shows the authenticity of the proposed method. Also, the WFS results are compared with proposed feature selection in Fig. 12, which shows that the proposed feature selection method is better compared to WFS on M-SVM. The proposed recognition results are compared with existing methods as presented in Table 14, which explains the recognition rates (RR) of the proposed method to be 95.80, 99.30, 99, and 99.40% on Weizmann, KTH, UIUC, and Muhavi dataset, respectively. It clearly shows that the proposed method performs significantly better compared to other methods as depicted in Table 14.

6 Conclusion

We have presented a novel approach by improving the human detection and action recognition. The proposed method is based on two major steps. In the first step, multiple humans are detected in the given video sequences by fusion novel uniform and EM segmentation. Then, we extracted the texture and shape features from given sequences and fused them based on vector dimensions. A novel Euclidian distance and joint entropy-PCA method are also implemented for best feature selection from the fused vector. The selected features are given to the classifiers for human classification and action recognition. The proposed method is tested on several datasets such as EPFL, MSRA Action, CASIA, INRIA person, own recorded videos, Weizmann, KTH, UIUC, and Muhavi datasets, and

Table 14 Comparison of recognition results

Weizmann dataset		
Method	Year	Recognition (%)
[53]	2013	95.45
[54]	2014	95.56
[55]	2015	95.10
[56]	2016	88.10
Proposed	2017	95.80
KTH Dataset		
[57]	2014	95.0
[52]	2014	95.21
[58]	2015	96.50
[59]	2016	97.10
[60]	2017	94.92
Proposed	2017	99.30
UIUC Dataset		
[61]	2012	98.84
[62]	2013	98.10
[63]	2014	98.30
[64]	2015	98.87
Proposed	2017	99.0
Muhavi		
[65]	2015	99.26
[66]	2016	96.36
[67]	2017	93.75
Proposed	2017	99.40

we obtained the accuracy of 99.56, 100, 98.70, 98.80, 99.60, 99.30, 99.0, 99.0, and 99.40%, respectively (Fig. 9).

The proposed algorithm has proven to be accurate both visually and empirically. There were a few major challenges which we have handled in this research including accurate identification of moving regions while ignoring all those regions which show minor changes due to light and intensity variations and the areas of high and low brightness.

In this research, occlusions are not handled, so as a future work, this problem needs to be tackled. Additionally, another possible direction would be to introduce saliency to improve the segmentation accuracy. Training and testing is the main part of improving performance, so we will increase the datasets for both training and testing to make our approach more robust. These days, deep approaches are playing their role, which is data hungry methodologies, so one of our targets is to implement deep convolutional neural networks.

Acknowledgements
This work was supported by the Machine Learning Research Group, Prince Sultan University, Riyadh, Saudi Arabia [RG-CCIS-2017-06-02]. The authors are grateful for this financial support. We also thank Mr. Badar Suleman, the Chair Person of Simulation Lab (CIIT WAH) and HOD of CS Department, for giving us full supports.

Availability of data and materials
Not applicable

Authors' contributions
MS generated this idea and developed a classification design and also identified the sole application. MAK performed the simulations by developing different patches of code with full integration. He is also responsible for this complete write-up. Different accuracy criteria are finalized and also simulated by this author. TA finalized the simulations. Additionally, due to his research in Gaussian Mixture models, he has solely developed the code for Expectation Maximization. MYJ has given a complete shape to this article and identified several issues and helped the primary authors to overcome all those shortcomings. TS is responsible for the final proofreading along with the technical support in the classification step due to her research major. AR provided technical support in different sections which include feature extraction and fusion along with the issues raised in the development of entropy-PCA-based feature selection. All authors read and approved the final manuscript.
Ethics approval and consent to participate
"Not applicable".

Funding
Not applicable

Consent for publication
"Not applicable".

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]COMSATS Institute of Information technology, Wah Cant 40470, Pakistan. [2]Department of Computer Science and Engineering, HITEC University, Museum Road, Taxila, Pakistan. [3]College of Computer and Information Science Prince Sultan University, Riyadh 11586, Saudi Arabia. [4]College of Computer and Information Systems Al-Yamamah University, Riyadh 11512, Saudi Arabia.

Received: 22 August 2017 Accepted: 29 November 2017
Published online: 19 December 2017

References
1. Y Xu et al., Detection of sudden pedestrian crossings for driving assistance systems. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **42**(3), 729–739 (2012)
2. W Fernando et al., in *Information and Automation for Sustainability (ICIAfS), 2014 7th International Conference on*. Object identification, enhancement and tracking under dynamic background conditions (IEEE, 2014)
3. D Thombre, J Nirmal, D Lekha, in *Intelligent Agent and Multi-Agent Systems, 2009*. Human detection and tracking using image segmentation and Kalman filter (IAMA 2009. International Conference on, 2009) IEEE
4. C Li, L Guo, Y Hu, in *Image and Signal Processing (CISP), 2010 3rd International Congress on*. A new method combining HOG and Kalman filter for video-based human detection and tracking (IEEE, 2010)
5. A Fakharian, S Hosseini, T Gustafsson, in *Control and Automation (MED), 2011 19th Mediterranean Conference on*. Precise hybrid motion detection and tracking in dynamic background (IEEE, 2011)
6. W Choi, C Pantofaru, S Savarese, A general framework for tracking multiple people from a moving camera. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1577–1591 (2013)
7. R Krerngkamjornkit, M Simic, in *Telecommunication in Modern Satellite, Cable and Broadcasting Services (TELSIKS), 2013 11th International Conference on*. Enhancement of human body detection and tracking algorithm based on Viola and Jones framework (IEEE, 2013)
8. J Liu et al., in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. Real-time human detection and tracking in complex environments using single RGBD camera (IEEE, 2013)
9. R Xu, Y Guan, Y Huang, Multiple human detection and tracking based on head detection for real-time video surveillance. Multimedia Tools and Applications **74**(3), 729–742 (2015)
10. W-C Cheng, D-M Jhan, A self-constructing cascade classifier with Adaboost and SVM for pedestriandetection. Eng. Appl. Artif. Intell. **26**(3), 1016–1028 (2013)
11. Q Ye et al., Human detection in images via piecewise linear support vector machines. IEEE Trans. Image Process. **22**(2), 778–789 (2013)
12. Z Lin, LS Davis, Shape-based human detection and segmentation via hierarchical part-template matching. IEEE Trans. Pattern Anal. Mach. Intell. **32**(4), 604–618 (2010)
13. Q Zhu et al., in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Fast human detection using a cascade of histograms of oriented gradients (IEEE, 2006)
14. N Dalal, B Triggs, in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Histograms of oriented gradients for human detection (IEEE, 2005)
15. A Satpathy, X Jiang, H-L Eng, Human detection by quadratic classification on subspace of extended histogram of gradients. IEEE Trans. Image Process. **23**(1), 287–297 (2014)
16. S Zhang, C Bauckhage, AB Cremers, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Informed haar-like features improve pedestrian detection (2014)
17. P Viola, MJ Jones, D Snow, in *null*. Detecting pedestrians using patterns of motion and appearance (IEEE, 2003)
18. WR Schwartz et al., in *Computer Vision, 2009 IEEE 12th International Conference on*. Human detection using partial least squares analysis (IEEE, 2009)
19. W Gao, H Ai, S Lao, in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Adaptive contour features in oriented granular space for human detection and segmentation (IEEE, 2009)
20. B Leibe, A Leonardis, B Schiele, Robust object detection with interleaved categorization and segmentation. Int. J. Comput. Vis. **77**(1–3), 259–289 (2008)
21. Zhang, Lun, Rufeng Chu, Shiming Xiang, Shengcai Liao, Stan Z. Li. "Face detection based on multi-block lbp representation." In International Conference on Biometrics, pp. 11–18. Springer, Berlin, Heidelberg, 2007.

22. J Berclaz et al., Multiple object tracking using k-shortest paths optimization. IEEE Trans. Pattern Anal. Mach. Intell. **33**(9), 1806–1819 (2011)
23. S Yu, D Tan, T Tan, in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition (IEEE, 2006)
24. MJA Patwary, S Parvin, S Akter, Significant HOG-histogram of oriented gradient feature selection for human detection. International Journal of Computer Applications **132**(17) (2015)
25. M Bregonzio, S Gong, T Xiang, in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Recognising action as clouds of space-time interest points (IEEE, 2009)
26. D Tran, A Sorokin, Human activity recognition with metric learning. Computer Vision–ECCV **2008**, 548–561 (2008)
27. S Singh, SA Velastin, H Ragheb, in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods (IEEE, 2010)
28. H Beiping, Z Wen, Fast human detection using motion detection and histogram of oriented gradients. JCP **6**(8), 1597–1604 (2011)
29. Q Ye, J Liang, J Jiao, Pedestrian detection in video images via error correcting output code classification of manifold subclasses. IEEE Trans. Intell. Transp. Syst. **13**(1), 193–202 (2012)
30. D Li et al., Integrating a statistical background-foreground extraction algorithm and SVM classifier for pedestrian detection and tracking. Integrated Computer-Aided Engineering **20**(3), 201–216 (2013)
31. J Marin et al., Occlusion handling via random subspace classifiers for human detection. IEEE transactions on cybernetics **44**(3), 342–354 (2014)
32. C-W Liang, C-F Juang, Moving object classification using local shape and HOG features in wavelet-transformed space with hierarchical SVM classifiers. Appl. Soft Comput. **28**, 483–497 (2015)
33. T Barbu, Pedestrian detection and tracking using temporal differencing and HOG features. Computers and Electrical Engineering **40**(4), 1072–1079 (2014)
34. V-D Hoang, M-H Le, K-H Jo, Hybrid cascade boosting machine using variant scale blocks based HOG features for pedestrian detection. Neurocomputing **135**, 357–366 (2014)
35. C Conde et al., HoGG: Gabor and HoG-based human detection for surveillance in non-controlled environments. Neurocomputing **100**, 19–30 (2013)
36. K Bhuvaneswari, HA Rauf, in *Control, Automation, Communication and Energy Conservation, 2009. INCACEC 2009. 2009 International Conference on*. Edgelet based human detection and tracking by combined segmentation and soft decision (IEEE, 2009)
37. P Viola, M Jones, in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Rapid object detection using a boosted cascade of simple features (IEEE, 2001)
38. D Kim, B Jun, in *Theory and Applications of Smart Cameras*. Accurate face and human detection using hybrid local transform features (Springer, 2016), pp. 157–185
39. Q Li, Y Yan, H Wang, Discriminative weighted sparse partial least squares for human detection. IEEE Trans. Intell. Transp. Syst. **17**(4), 1062–1071 (2016)
40. K-D Lee et al., Context and profile based cascade classifier for efficient people detection and safety care system. Multimedia Tools and Applications **63**(1), 27–44 (2013)
41. D Qichang, A Tallha, D Pan, W Xiaogang, Visual saliency detection using information contents weighting. In Optik **127**(19), 7418–7430 (2016)
42. Gonzalez, R.C.E., et al., Digital Image Processing Using MATLAB. 2004.
43. JL Barron et al., in *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*. Performance of optical flow techniques (IEEE, 1992)
44. K Suresh, in *Communications and Signal Processing (ICCSP), 2014 International Conference on*. HOG-PCA descriptor with optical flow based human detection and tracking (IEEE, 2014)
45. C Carson et al., Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Trans. Pattern Anal. Mach. Intell. **24**(8), 1026–1038 (2002)
46. M Hao et al., Unsupervised change detection with expectation-maximization-based level set. IEEE Geosci. Remote Sens. Lett. **11**(1), 210–214 (2014)
47. S Li et al., Pixel-level image fusion: a survey of the state of the art. information Fusion **33**, 100–112 (2017)
48. H Qian et al., Recognition of human activities using SVM multi-class classifier. Pattern Recogn. Lett. **31**(2), 100–111 (2010)
49. Y Said, Y Salah, M Atri, in *Image Processing, Applications and Systems Conference (IPAS), 2014 First International*. Pedestrian detection using covariance features (IEEE, 2014)
50. L Cao, Z Liu, TS Huang, in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. Cross-dataset action detection (IEEE, 2010)
51. L Xia, C-C Chen, JK Aggarwal, in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. Human detection using depth information by kinect (IEEE, 2011)
52. M Sreeraj, in *Advances in Computing and Communications (ICACC), 2015 Fifth International Conference on*. Multi-posture human detection based on hybrid HOG-BO feature (IEEE, 2015)
53. G Goudelis, K Karpouzis, S Kollias, Exploring trace transform for robust human action recognition. Pattern Recogn. **46**(12), 3238–3248 (2013)
54. JA Nasiri, NM Charkari, K Mozafari, Energy-based model of least squares twin support vector machines for human action recognition. Signal Process. **104**, 248–257 (2014)
55. J Jiang et al., Human action recognition via compressive-sensing-based dimensionality reduction. Optik-International Journal for Light and Electron Optics **126**(9), 882–887 (2015)
56. S Zhang, W Zhang, Y Li, in *Proceedings of 2016 Chinese Intelligent Systems Conference*. Human action recognition based on multifeature fusion (Springer, 2016)
57. L Shao et al., Spatio-temporal Laplacian pyramid coding for action recognition. IEEE Transactions on Cybernetics **44**(6), 817–827 (2014)
58. J Yang, Z Ma, M Xie, Action recognition based on multi-scale oriented neighborhood features. International Journal of Signal Processing, Image Processing and Pattern Recognition **8**(1), 241–254 (2015)
59. S Cheng et al., Action recognition based on spatio-temporal log-Euclidean covariance matrix. International Journal of Signal Processing, Image Processing and Pattern Recognition **9**(2), 95–106 (2016)
60. H Liu et al., in *Human Motion Sensing and Recognition*. Study of human action recognition based on improved spatio-temporal features (Springer, 2017), pp. 233–250
61. KM Chathuramali, R Rodrigo, in *Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on*. Faster human activity recognition with SVM (IEEE, 2012)
62. H Wang et al., Dense trajectories and motion boundary descriptors for action recognition. Int. J. Comput. Vis. **103**(1), 60–79 (2013)
63. Y-Y Lin et al., in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Depth and skeleton associated action recognition without online accessible rgb-d cameras (2014)
64. Z Zhang et al., Robust relative attributes for human action recognition. Pattern. Anal. Applic. **18**(1), 157–171 (2015)
65. F Murtaza, MH Yousaf, SA Velastin, in *Frontiers of Information Technology (FIT), 2015 13th International Conference on*. Multi-view human action recognition using histograms of oriented gradients (HOG) description of motion history images (MHIs) (IEEE, 2015)
66. F Murtaza, MH Yousaf, SA Velastin, Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description. IET Comput. Vis. **10**(7), 758–767 (2016)
67. S Maity, D Bhattacharjee, A Chakrabarti, A novel approach for human action recognition from silhouette images. IETE J. Res. **63**(2), 160–117 (2017)