Multimedia Systems (2017) 23:451–467
DOI 10.1007/s00530-016-0505-x

REGULAR PAPER

Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns

Alok Kumar Singh Kushwaha[1] · **Subodh Srivastava**[1] · **Rajeev Srivastava**[1]

Received: 9 August 2015 / Accepted: 16 February 2016 / Published online: 3 March 2016

Abstract This paper addresses the problem of silhouette-based human activity recognition. Most of the previous work on silhouette based human activity recognition focus on recognition from a single view and ignores the issue of view invariance. In this paper, a system framework has been presented to recognize a view invariant human activity recognition approach that uses both contour-based pose features from silhouettes and uniform rotation local binary patterns for view invariant activity representation. The framework is composed of three consecutive modules: (1) detecting and locating people by background subtraction, (2) combined scale invariant contour-based pose features from silhouettes and uniform rotation invariant local binary patterns (LBP) are extracted, and (3) finally classifying activities of people by Multiclass Support vector machine (SVM) classifier. The rotation invariant nature of uniform LBP provides view invariant recognition of multi-view human activities. We have tested our approach successfully in the indoor and outdoor environment results on four multi-view datasets namely: our own view point dataset, VideoWeb Multi-view dataset [28], i3DPost multi-view dataset [29], and WVU multi-view human action recognition dataset [30]. The experimental results show that the proposed method of multi-view human activity recognition is robust, flexible and efficient.

Keywords Human activity recognition · Features extraction · Local binary patterns · Multiclass support vector machine (SVM)

1 Introduction

Recognition of human activities from multiple views is a popular area of research in the field of computer vision. It is the basis of many applications in video surveillance and monitoring, human–computer interactions, model-based compressions, and video retrieval in various situations [1, 2]. Although a large amount of work has been performed on activity recognition in the last few years, still it is an open and challenging problem.

The various issues and challenges involved in automatic human activity recognition from video sequences are as follows: (1) clutter backgrounds (2) stationary or non-stationary camera (3) scale variation (4) starting and ending state variation (5) individual variations in appearance and cloths of people (6) changes in light and view-point. These situations make the human activity recognition a challenging task. Most of the work on activity recognition are view dependent and deal with recognition from one fixed view. To account these problems, many activity recognition systems have been developed [1–4] and various surveys and frameworks can be found in literature [5–9].

Activity recognition methods available in the literature can broadly be categorized into two groups: sensor-based activity recognition and vision based activity recognition. In sensor-based activity recognition methods some smart sensory device is used to capture various activity signals

Communicated by Y. Zhang.

✉ Alok Kumar Singh Kushwaha
alok.rs.cse12@iitbhu.ac.in

Subodh Srivastava
subodh.cse@iitbhu.ac.in

Rajeev Srivastava
rajeev.cse@iitbhu.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi 221005, UP, India

for activity recognition. Vision-based activity recognition methods use the spatial or temporal structure of an activity to recognize it. A recent survey on vision-based action representation and recognition methods can be found in [10]. Machine learning based and Silhouette-based methods [10] are popular vision based approaches for human activity recognition in videos. The machine learning based approaches for activity recognition generally solve the problem of activity recognition as a classification problem and classify an activity into one of known activity classes. Silhouette-based methods are good options for activity recognition in video and can be easily used because of their simplicity and robustness. In silhouette-based human activity methods [8, 9], activities are recognized based on the key poses in the image. It is a local representation of activities. It decomposes the image into smaller interest regions and describes each region as a separate feature.

An immediate advantage of above-mentioned approaches is that they neither rely on explicit body part labeling, nor on explicit human detection and localization. Bobick et al. [11] used motion templates for recognizing the activities in a specific environment of aerobic exercise. They used MEI (motion energy image) for obtaining foreground and MHI (motion history image) for obtaining motion information in a view-specific environment. It does not produce good activity recognition accuracy in an outdoor environment. Moreover, this technique is capable of only identifying one activity in the scene with one actor at a time. Weinland et al. [10] extended the work of Bobick et al. [11] to 3D motion history volume in order to combine images from multiple cameras and to obtain a free-viewpoint representation. While Bobick et al. [11] used seven Hu Moments for description and classification; Weinland et al. [10] use Fourier analysis in cylindrical coordinates. Lv and Nevatia [12] exploited a similar idea, where each action is modeled as a series of synthetic 2D human poses rendered from a wide range of viewpoints, instead of using 3D explicitly. To deal the issue of Bobick et al. [11], Weinland et al. [13] proposed a method for multi-view human action recognition using Motion History Volumes (MHV) in 3D motion template. In this method, computing, aligning and comparing MHVs of different actions performed by different people in a variety of viewpoints. Weinland et al. [14] proposed another approach for multi-view human activity recognition which based on action taxonomy. In this method, an action taxonomy created using segmented sequence then apply a standard hierarchical clustering method to segment the sequence for human action recognition. Ahmad et al. [15] proposed a multi-view human action recognition system using combined local–global (CLG) optic flow based motion and shape feature for recognition of human actions in daily life. In this approach, actions are modeled by using a set of multidimensional HMMs for multiple views using the combined features of the training action videos. Iosifidis et al. [16] also presented a method for view-independent human action recognition. In this method, author used a new feature space (dyneme space) for multi-view movement representation and Fuzzy distances to represent the human body postures in the dyneme space. In this paper, view identification problem is solved by exploiting the circular shift invariance property of the Discrete Fourier Transform (DFT). Iosifidis et al. [17] proposed a method for multi-view human action recognition using neural network. In this paper, Fuzzy distances from human body posture prototypes are used to produce a time invariant action representation and Multilayer perceptron are used for human action classification. This method is slow as it is based on neural network framework. To account this problem, Iosifidis et al. [18] presented a view-independent human action recognition method that exploits a 3D action representation. In this approach, binary human body images are temporally concatenated in order to produce action volumes (AVs) which represent the action. Multi-view action representation is obtained by exploiting the circular shift invariance property of the Discrete Fourier Transform coefficients. Chaaraoui et al. [19] proposed a method which is based on sequences of key poses. In this approach, learning of key poses is modeled using *K*-means and Dynamic Time Warping for action categorization. This approach suffer the problem of fixed view. To deal this issue, Iosifidis et al. [20] presented a method for view-independent human action recognition. In this method, fuzzy vector quantization is used for determine the human body poses. Sparsity-based Learning Machine (SbLM) has been used for view-independent action video classification. Sharma et al. [21] have proposed a human activity recognition method which used motion history images and object shape information for different human activities in a video. The major disadvantage of this method is that it is view dependent and deals with activity recognition from one fixed view. Le et al. [22] proposed a method which is based on spatio-temporal features. This approach using unsupervised feature learning as a way to learn features directly from video data. This method performs well when combined with deep learning techniques such as stacking and convolution to learn hierarchical representations. Liu et al. [23] proposed an innovative RGB-D-based orientation estimation method to address the problem of activity recognition in real time. In this method, static cues (SVFH) and motion cues (SSF) are extracted based on the RGB-D superpixels. In the proposed approach, author utilize a dynamic Bayesian network system (DBNS) to effectively employ the complementary nature of both static and motion cues.

To deal with the issues mentioned in [11–21], in this paper, we have combined contour-based pose features from silhouettes and uniform rotation invariant local binary patterns (LBP) feature to model human activities. At first,

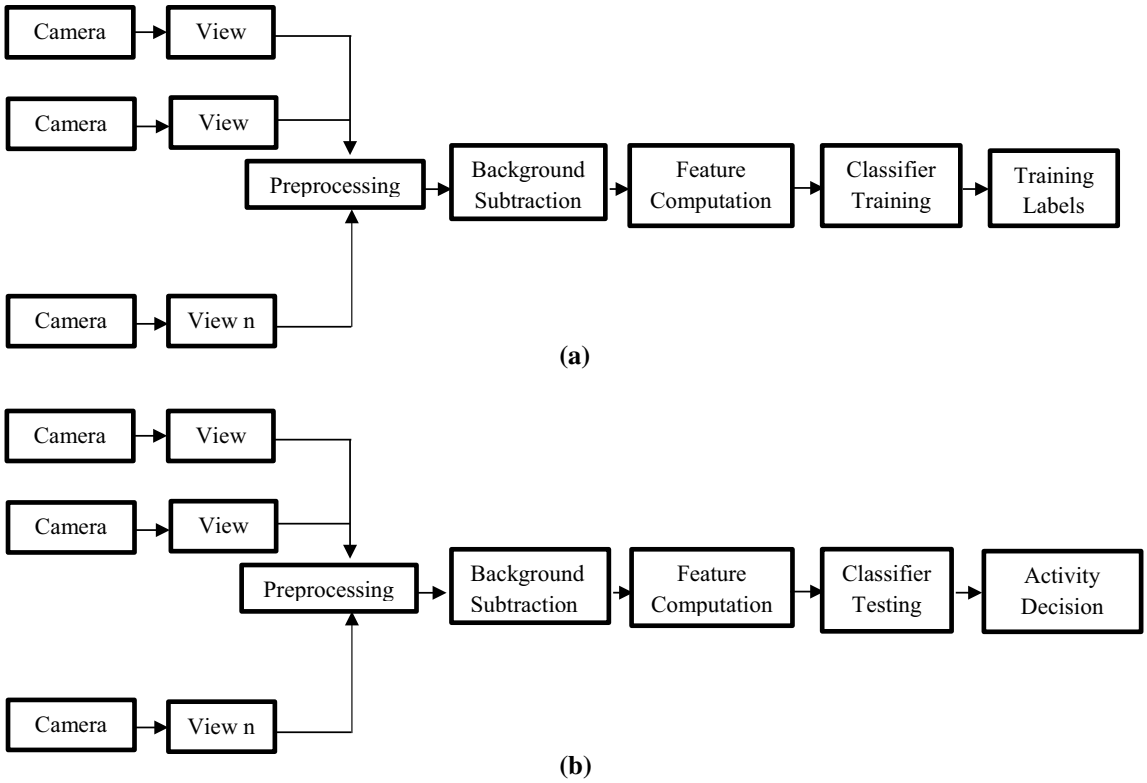

Fig. 1 The block diagram of the proposed human activity recognition system. **a** Training phase and **b** Testing phase

change detection based approach is used for background subtraction. In the second step, combined contour-based pose features from silhouettes and uniform rotation invariant local binary patterns (LBP) are extracted. The contour-based pose features from silhouettes find the different key poses for human activities such as bending, standing, sitting etc. To solve view depend (e.g. single or fixed view) problem, we have used uniform rotation invariant local binary patterns (LBP) feature. The uniform rotation invariant local binary patterns (LBP) feature provides view-independent analysis of human activities and it possess good discriminating ability, therefore they are better suited for distinguishing different activities. Finally, in a third step, this feature has been classified using Multiclass support vector machine (SVM) classifier.

The rest of the paper is organized as follows: Sect. 2 briefly explains the proposed method. Section 3 presents experimental results and discussions. Finally, conclusions are drawn in Sect. 4.

2 Methods and models

The proposed multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns consists of three steps applied on given video frames which include: (1) detecting and locating people by background subtraction, (2) combined scale invariant contour-based pose features from silhouettes and uniform rotation invariant local binary patterns (LBP) are extracted, and (3) finally classifying activities of people by Multiclass Support vector machine (SVM) classifier. The proposed method consists of two phases testing and training as depicted in block diagram shown in Fig. 1a, b. In the training phase, we have given video data and apply above-mentioned steps and finally training labels are given to the classifier. In the testing phase, testing video is performed with the help of training labels (Fig. 2).

Fig. 2 Sequence of key poses of walking activity in some selected frames

The algorithm for the proposed framework is as follows:
Algorithm

Step 1 Captured video from the multiple cameras
Step 2 Pre-processing of the captured video using background model creation (using Eqs. 2 and 3)
Step 3 Feature Extraction:
Computation of contour-based pose features from silhouettes for find out the different human key poses (using Eqs. 12 and 13)
Computation of uniform rotation local binary patterns feature for view invariant recognition of multi-view human activities (using Eq. 17)
Step 4 Feature modeling and activity classification by using SVM classifier (using Eq. 20).

The various stages of the proposed method are discussed as follows:

2.1 Pre-processing

In the pre-processing steps, we extract foreground from the background. We then define the boundary from the foreground image sequence. Briefly, these are explained below:

2.1.1 Background subtraction

The proposed background subtraction method is based on the change detection and background modeling. The steps of background subtraction are as follows:

2.1.1.1 Frame difference The difference between the current frame and the previous frame is calculated using change detection. Let f_n and f_{n-1} be the current frame and the previous frame at location (i, j). Instead of assigning a fixed a priori threshold $V_{\text{th},d}$ to each frame difference, this paper uses the fast Euler number computation technique [24] to automatically determine $V_{\text{th},d}$ from the video frame. The fast Euler numbers algorithm calculates the Euler number for every possible threshold with a single raster of the frame difference image using following equation:

$$E(i) = \frac{1}{4}\left[(q_1(i) - q_3(i) - 2q_d(i))\right] \tag{1}$$

where q_1, q_3, and q_d is the quads (quad is a 2×2 masks of bit cells) contained in the given image.

The output of the algorithm is an array of Euler numbers: one of each threshold value. The Zero Crossings find out the optimal threshold. Detailed algorithms for the fast Euler number computation method can be found in [24].

The frame differences $WD_n(i,j)$ for respective frames are computed as:For every pixel location (i, j) in the coordinate of frame

$$WD_n(i,j) = \left\{ \begin{array}{ll} 1 & \text{if } |f_n(i,j) - f_{n-1}(i,j)| > V_{\text{th},WD} \\ 0 & \text{otherwise} \end{array} \right\} \tag{2}$$

2.1.1.2 Background model creation for segmentation For background modeling, we have used frame difference, background registration, background difference, and background difference mask. The background modeling step is divided into five major phases. The first phase calculates the frame difference mask $WD_n(i,j)$ which is obtained by difference between two consecutive frames as follows:

$$WD_n(i,j) = \left\{ \begin{array}{ll} 1 & \text{if } |Wf_n(i,j) - Wf_{n-1}(i,j)| \geq V_{\text{th},WD} \\ 0, & \text{if } |Wf_n(i,j) - Wf_{n-1}(i,j)| < V_{\text{th},WD} \end{array} \right\} \tag{3}$$

$V_{\text{th},WD}$ is a threshold determined automatically from the video frame by the fast Euler number computation method as explained in [24].

The second phase of dynamic background modeling maintains an up-to-date background buffer as well as background registration mask indicating whether the background information of a pixel is available or not. According to the frame difference mask of the past several frames, pixels that are not moving for a long time are considered as reliable background and registered in the background buffer. The background registration process uses the following two equations:

$$S_n(i,j) = \left\{ \begin{array}{ll} S_n(i,j) + 1 & \text{if } WD_n(i,j) = 0 \\ 0 & \text{if } WD_n(i,j) = 1 \end{array} \right\} \tag{4}$$

$$\mu_n(i,j) = \left\{ \begin{array}{ll} f_n(i,j) & \text{if } S_n(i,j) \geq N_f \\ \text{Undefined} & \text{if } S_n(i,j) < N_f \end{array} \right\} \tag{5}$$

where $S_n(i, j)$ is a stationary index and $\mu_n(i, j)$ is the background buffer value of a pixel with position (i, j) in the nth frame. The initial values of $S_n(i, j)$ and $\mu_n(i, j)$ are set to 0 and $f_n(i,j)$, respectively. If a pixel is masked as stationary for N_f successive frames (i.e., if the accumulated value in registration stationary index exceeds N_f), then that pixel is classified as part of the background region. Here, N_f is set to 30 experimentally. According to our experiments, N_f may be set at a larger value for fast moving object.

In the third phase of background modeling, a registered background buffer pixel is updated using the following equation.

$$\text{if} \quad |f_n(i,j) - \mu_n(i,j)| < 2\sigma_n(i,j) \tag{6}$$

$$\text{then} \quad \left\{ \begin{array}{l} \mu_n(i,j) = \chi\mu_{n-1}(i,j) + (1-\chi)f_n(i,j) \\ \sigma_n^2(i,j) = \chi\sigma_{n-1}^2(i,j) + (1-\chi)(f_n(i,j) - \mu_n(i,j))^2 \end{array} \right. \tag{7}$$

where $\sigma_n(i,j)$ is the standard deviation of a pixel with position (i, j) in the nth frame and χ is the predefined constant and we considered four different sequences, and recorded 10 different observations over 800 frames for each of the sequences. This resulted in 50 samples of size 800 each.

The test statistic was calculated for each of the samples and the value of χ is set to 0.7

In the fourth phase of background modeling, we find the background difference mask with the help of background difference which distinguishes moving objects from the background, and its operation are shown as follows:

$$\mathrm{BD}_{n,\mathrm{LL}}(i,j) = |f_n(i,j) - \mu_n(i,j)| \tag{8}$$

$$\mathrm{BDM}_n(i,j) = \begin{cases} 1, & \mathrm{BD}_n(i,j) \geq V_{\mathrm{th},WD} \\ 0, & \mathrm{BD}_n(i,j) < V_{\mathrm{th},WD} \end{cases} \tag{9}$$

where $\mathrm{BD}_{n,\mathrm{LL}}(i,j)$ is the background difference and $\mathrm{BDM}_{n,\mathrm{LL}}(i,j)$ is the background difference mask of a pixel with position (i, j) in the nth frame. The threshold value $V_{th,WD}$ is also automatically determined by the fast Euler number computation method [24].

In the fifth phase of background modeling, a background model is constructed using the frame difference, background registration, background difference, and background difference mask.

2.2 Multi-view features extraction

We use contour-based pose features and local binary pattern (LBP) feature for activities representation and classification. The foreground image sequence (which is obtained in Sect. 2.1) is used to extract the contour-based pose features and local binary pattern (LBP).

2.2.1 Contour-based pose feature

In this section, we find out the distance signal feature using contour points of the silhouette for different key poses (sitting, standing, sleeping etc.). We obtained a binary silhouette in Sect. 2.1 by human silhouette extraction techniques, e.g. background subtraction.

Let H be the binary silhouette image of an object. We determine its center of mass $C_{\mathrm{m}} = (\bar{x}, \bar{y})$ of the silhouette's contour points using [25] where

$$\bar{x} = \frac{\sum_{w=1}^{n} x_w}{n}, \quad \bar{y} = \frac{\sum_{w=1}^{n} y_w}{n} \tag{10}$$

and n is the number of silhouette pixels.

The distance signal $D = \{d_1, d_2, d_3, \ldots, \mathrm{d}_n\}$ is generated by determining the Euclidean distance between each contour point and the centre of mass (see Fig. 3). Contour points should be considered always in the same order. For instance, the set of points can start at the most left point with equal y-axis value as the centre of mass, and follow a clockwise order.

$$d_i = \|C_m - a_i\|, \quad \forall i \in [1, \ldots, n] \tag{11}$$

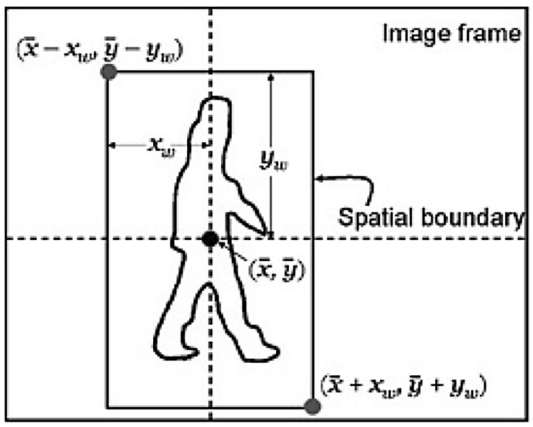

Fig. 3 Activity boundary definitions

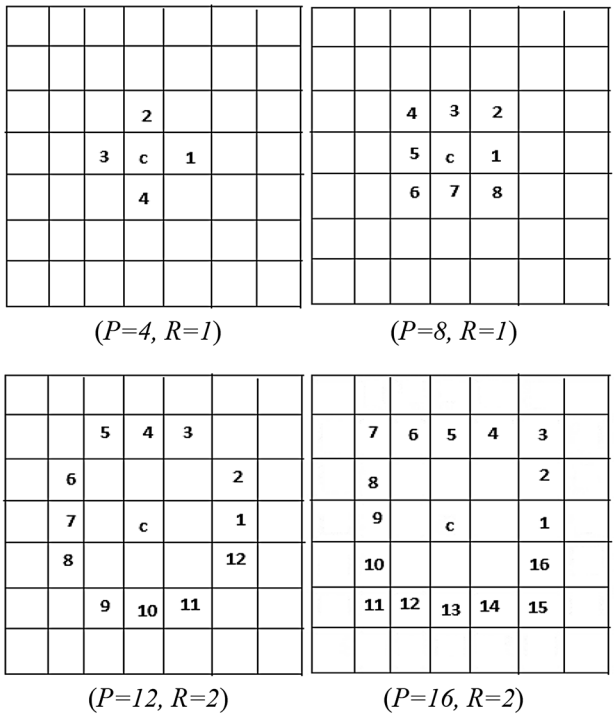

Fig. 4 Circularly symmetric neighbor sets for different (P, R) (here anti-clockwise)

In order to provide a uniform representation for varying image sizes and shapes, d is scaled to a constant size L such that

$$\hat{D}[i] = d\left\lceil i \times \frac{n}{L} \right\rceil, \quad \forall i \in [1, \ldots, L] \tag{12}$$

where $\lceil \cdot \rceil$ is the ceiling function.

Finally, the scaled distance vector $\hat{D}$ is normalized to have unit sum:

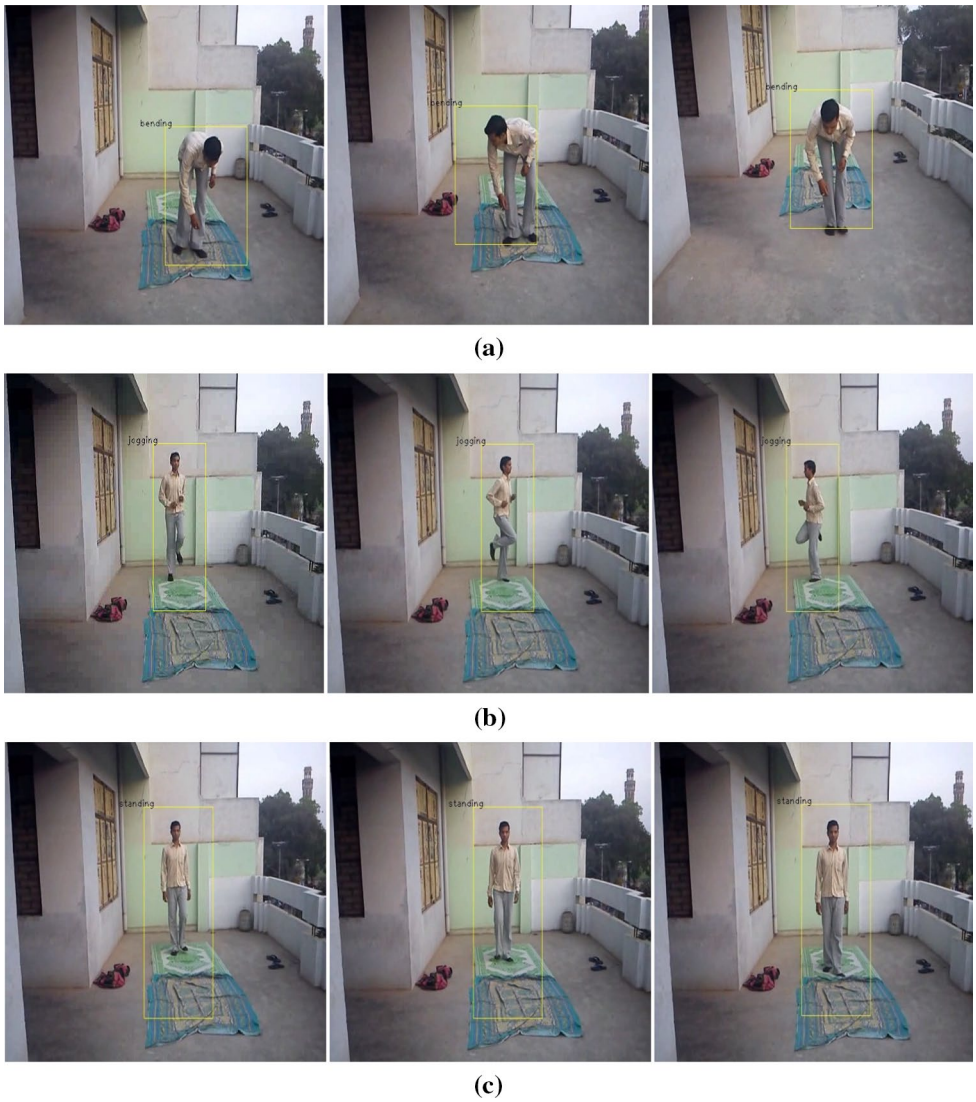

Fig. 5 Recognition of activities in our own database. **a** Bending. **b** Jogging. **c** Standing in different views

$$\bar{D}[i] = \frac{\hat{D}[i]}{\sum_{i=1}^{L} \hat{D}[i]}, \quad \forall i \in [1, \ldots, L] \tag{13}$$

2.2.2 Uniform rotation invariant local binary patterns (LBP)

In this paper, we extract the Uniform rotation invariant local binary patterns (LBP) features from the background subtracted video which is obtained in Sect. 2.1. Uniform rotation invariant local binary patterns (LBP) provide view invariant recognition of multi-view human activities. Using uniform patterns instead of all the possible patterns has produced better recognition results for human activity. The basic description of local binary patterns (LBP) is given below.

2.2.2.1 Local binary patterns (LBP) A local binary pattern (LBP) feature can be constructed for a specific circular pixel neighborhood of radius R. The intensities of the P sample pixel points are compared in the circular neighborhood with the centre pixel in clockwise or anticlockwise direction (see Fig. 4).

After extracting LBP of each sample point in the image, value of each pixel in the image is replaced by a binary pattern. With the help of these considerations, the overall feature vector of the whole image, denoted by $\mathrm{LBP}_{P,R}$, is given as below:

$$\mathrm{LBP}_{P,R}(x, y) = \sum_{P=0}^{P-1} s(g_\mathrm{P} - g_\mathrm{c}) 2^P \tag{14}$$

where (x, y) is the location of the centre pixel, g_c represent intensity of centre pixel, g_P represent intensity of neighborhood pixel and $s(u)$ is defined as

$$s(u) = \left\{ \begin{array}{ll} 1, & u \geq 0 \\ 0, & u < 0 \end{array} \right\} \tag{15}$$

Table 1 Confusion matrices for the proposed and other methods

Recognized instances / Total instances	Bending	Jogging	Walking
For the proposed method			
Bending	0.98	0.02	0
Jogging	0	0.99	0.01
Walking	0	0	1
Chaaraoui et al. [19]			
Bending	0.85	0.03	0.12
Jogging	0.15	0.80	0.05
Walking	0.02	0.10	0.82
Bobick et al. [11]			
Bending	0.75	0.10	0.15
Jogging	0.15	0.70	0.15
Walking	0.12	0.11	0.77
Ahmad et al. [15]			
Bending	0.79	0.20	0.01
Jogging	0.10	0.82	0.08
Walking	0.12	0	0.88
Iosifidis et al. [20]			
Bending	0.94	0.06	0
Jogging	0.03	0.95	0.02
Walking	0.02	0	0.98
Weinland et al. [13]			
Bending	0.82	0.18	0
Jogging	0.10	0.84	0.06
Walking	0.14	0	0.86
Iosifidis et al. [16]			
Bending	0.90	0.05	0.05
Jogging	0.06	0.94	0
Walking	0.09	0	0.91
Weinland et al. [14]			
Bending	0.78	0.12	0.10
Jogging	0.20	0.80	0
Walking	0.22	0	0.78
Iosifidis et al. [17]			
Bending	0.96	0.04	0
Jogging	0	0.96	0.04
Walking	0.02	0.02	0.94
Iosifidis et al. [18]			
Bending	0.98	0.02	0
Jogging	0.02	0.94	0.02
Walking	0	0.05	0.95

Now, the feature vector $\mathrm{LBP}_{P,R}$ of the image is a histogram of the LBP of different pixels in the image. The starting size of the histogram is 2^P because each possible LBP has been assigned a separate bin. Suppose, there are M regions in an image, then all histogram scan be merged into one histogram of size $M{\cdot}2^P$.

Table 2 Recognition results over own activity recognition dataset

Method	Accuracy (%)
Proposed method	99
Chaaraoui et al. [19]	82
Bobick et al. [11]	74
Ahmad et al. [15]	83
Iosifidis et al. [20]	95.66
Weinland et al. [13]	84
Iosifidis et al. [16]	91.66
Weinland et al. [14]	78.66
Iosifidis et al. [17]	95.33
Iosifidis et al. [18]	95.66

2.2.2.2 Rotation invariance Several modified versions of LBP [26] have been proposed for achieving rotation invariance and reducing the histogram dimension of the LBP. When the image is rotated, the gray value g_P will correspondingly move along the perimeter of the circle, so different $\mathrm{LBP}_{P,R}$ may be computed. To remove the effect of rotation, the modified version with rotation invariance is defined as follows:

$$\mathrm{LBP}^{ri}_{P,R}(x,y) = \min\{\mathrm{ROR}(\mathrm{LBP}_{P,R}, i) | i = 0,\ 1, \ldots,\ R-1\} \tag{16}$$

where $\mathrm{ROR}(\mathrm{LBP}_{P,R}, i)$ performs a circular bit-wise right shift on the R-bit number $\mathrm{LBP}_{P,R}$ for i times. $\mathrm{LBP}^{ri}_{P,R}$ can have 36 different values when $R = 8$, and the histogram dimension of $\mathrm{LBP}^{ri}_{P,R}$ over an image region is 36.

2.2.2.3 Uniform patterns The uniform LBP are those LBP which have very few spatial transitions. Formally, uniform LBP have maximum two circular transitions between 0 and 1. For example, patterns 00000001 and 11111011 have only one and two transitions between 0 and 1 respectively, therefore they are uniform patterns.

2.2.2.4 Uniform local binary patterns (LBP) for feature extraction The rotation invariant uniform local binary patterns (LBP) for feature extraction is defined as

$$\mathrm{LBP}^{riu2}_{P,R} = \begin{cases} \sum_{P=0}^{P-1} s(g_P - g_C), & \text{if } \mathrm{U}(\mathrm{LBP}_{P,R}) \le 2 \\ P+1, & \text{otherwise} \end{cases} \tag{17}$$

where $U(\mathrm{LBP}_{P,R}) = |s(g_{P-1} - g_C) - s(g_0 - g_C)| + \sum_{P=1}^{P-1} |s(g_P - g_C) - s(g_{P-1} - g_C)|$ is a rotation invariant operator with uniform patterns having at most two transitions between 0 and 1 bits. In a circularly symmetric neighborhood of P pixels, $P + 1$ uniform pattern can be found. Each pattern assigns a unique label to each pixel.

Fig. 6 Recognition of activities in VideoWeb multi-view dataset [28]. **a** Boxing. **b** Clapping. **c** Jogging. **d** Running. **e** Walking

and $s(u) = \begin{cases} 1, & u \geq 0 \\ 0, & u < 0 \end{cases}$ (given in Eq.15)

g_c = centre pixel of background subtraction image (which is obtained in Sect. 2.1) and g_p = neighborhood pixel of background subtraction image (which is obtained in Sect. 2.1).

Table 3 Confusion matrices for the proposed and other methods

Recognized instances / Total instances	Boxing	Clapping	Jogging	Running	Walking
For the proposed method					
Boxing	1	0	0	0	0
Clapping	0	0.97	0.02	0	0.01
Jogging	0	0.02	0.98	0	0
Running	0	0	0	1	00
Walking	0	0	0	0	1
Bobick et al. [11]					
Boxing	0.65	0.10	0.10	0	0.15
Clapping	0.18	0.61	0.01	0.20	0
Jogging	0.15	0.06	0.67	0	0.12
Running	0.16	0	0	0.64	0.20
Walking	0.10	0	0.25	0	0.65
Chaaraoui et al. [19]					
Boxing	0.78	0	0.15	0	0.07
Clapping	0.10	0.80	0	0.10	0
Jogging	0.12	0.07	0.81	0	0
Running	0.13	0	0.11	0.76	0
Walking	0	0.15	0	0.05	0.80
Ahmad et al. [15]					
Boxing	0.86	0.04	0	0.10	0
Clapping	0.12	0.88	0	0	0
Jogging	0.02	0.02	0.84	0.12	0
Running	0	0.11	0	0.89	0
Walking	0.16	0	0	0	0.84
Iosifidis et al. [20]					
Boxing	0.98	0.02	0	0	0
Clapping	0.05	0.95	0	0	0
Jogging	0	0	0.95	0.05	0
Running	0.01	0	0	0.93	0.06
Walking	0	0.04	0	0	0.96
Weinland et al. [13]					
Boxing	0.88	0.12	0	0	0
Clapping	0	0.82	0.10	0.08	0
Jogging	0.13	0	0.87	0	0
Running	0	0	0	0.88	0.12
Walking	0.12	0.04	0	0	0.84
Iosifidis et al. [16]					
Boxing	0.91	0.09	0	0	0
Clapping	0.06	0.94	0	0	0
Jogging	0	0	0.91	0.09	0
Running	0	0	0.05	0.90	0.05
Walking	0	0	0.09	0	0.91
Weinland et al. [14]					
Boxing	0.77	0.20	0.03	0	0
Clapping	0.20	0.80	0	0	0
Jogging	0	0	0.82	0.18	0
Running	0.18	0	0	0.82	0
Walking	0.12	0	0.10	0	0.78
Iosifidis et al. [17]					
Boxing	0.94	0.06	0	0	0
Clapping	0.03	0.97	0	0	0
Jogging	0	0	0.90	0.05	0.05
Running	0.09	0	0	0.91	0
Walking	0	0	0	0.10	0.90
Iosifidis et al. [18]					
Boxing	0.97	0.03	0	0	0
Clapping	0	0.96	0	0.04	0
Jogging	0	0	0.98	0	0.02
Running	0	0	0.03	0.97	0
Walking	0.07	0	0	0	0.93

2.3 Classifier training and testing

After having computed features from a video, the classifier is trained and tested with these video. To model and classify activities, we have used multi-class SVM classifiers. First training of classifier is performed. The obtained training labels are supplied into the classifier then after testing is performed. The activities in the testing video are performed with the help of training labels. Finally, different test labels have been obtained for the test video of human activities. Consider the pattern recognition problem of training samples (x_1, y_1), (x_2, y_2),…,(x_l, y_l), where x_i, $i = 1, 2,\ldots,l$ is a vector and $y_i \in \{1, 2,\ldots,k\}$ represents the class of samples. The multi-class support vector machines (SVM) [27] require the solution of the following optimization problem:

Table 4 Recognition results over our the VideoWeb action recognition dataset [28]

Method	Accuracy (%)
Proposed method	99
Chaaraoui et al. [19]	64.4
Bobick et al. [11]	79
Ahmad et al. [15]	86.2
Iosifidis et al. [20]	95.4
Weinland et al. [13]	85.8
Iosifidis et al. [16]	91.4
Weinland et al. [14]	79.8
Iosifidis et al. [17]	92.4
Iosifidis et al. [18]	96.2

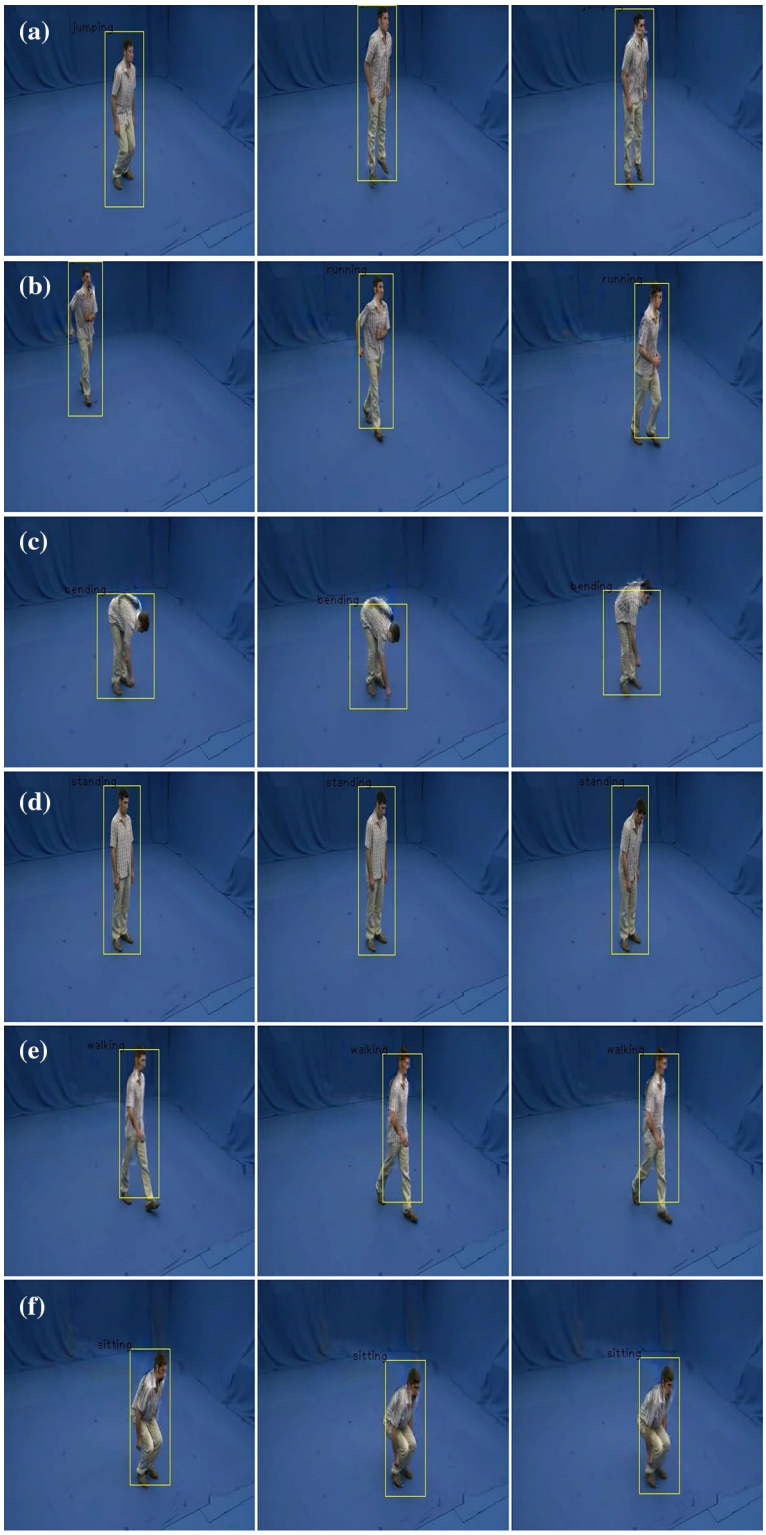

Fig. 7 Recognition of activities in i3DPost multi-view dataset [29]. **a** Jumping. **b** Running. **c** Bending. **d** Standing. **e** Walking. **f** Sitting

minimize

$$\phi(\omega, \xi) = \frac{1}{2} \sum_{m=1}^{k} \omega_m \times \omega_m + C \sum_{i=1}^{l} \sum_{m \neq y} \xi_i^m \tag{18}$$

with constraints

$$\begin{aligned} &(\omega_{y_i} \times x_i) + b_{y_i} \geq (\omega_m \times x_i) + b_m + 2 - \xi_i^m, \\ &\xi_i^m \geq 0, \quad i = \{1, 2 \ldots, l\}, \quad m \in \{1, 2, \ldots, k\}/y_i \end{aligned} \tag{19}$$

Table 5 Confusion matrix for the proposed method and other methods

Recognized instances Total instances	Jumping	Running	Bending	Standing	Walking	Sitting
For the proposed method						
Jumping	0.98	0.01	0	0.01	0	0
Running	0	1	0	0	0	0
Bending	0	0	0.99	0	0.01	0
Standing	0.02	0	0	0.98	0	0
Walking	0	0	0	0	1	0
Sitting	0	0.01	0	0	0.01	0.98
Chaaraoui et al. [19]						
Jumping	0.80	0.10	0.05	0	0.05	0
Running	0.10	0.85	0.05	0	0	0
Bending	0.08	0.05	0.82	0	0.05	0
Standing	0	0	0.10	0.84	0	0.06
Walking	0	0	0	0.10	0.80	0.10
Sitting	0	0.10	0	0.01	0.08	0.81
Bobick et al. [11]						
Jumping	0.75	0.05	0.10	0.05	0.05	0
Running	0.10	0.73	0.10	0.04	0.03	0
Bending	0.06	0.05	0.76	0.03	0	0.10
Standing	0.05	0.05	0.08	0.72	0.10	0
Walking	0.01	0.08	0.08	0.10	0.71	0.02
Sitting	0.05	0.07	0.10	0.03	0.05	0.70
Ahmad et al. [15]						
Jumping	0.89	0.11	0	0	0	0
Running	0.10	0.90	0	0	0	0
Bending	0	0	0.94	0.06	0	0
Standing	0.05	0.05	0	0.90	0	0
Walking	0	0	0	0.12	0.88	0
Sitting	0.11	0	0	0	0	0.89
Iosifidis et al. [20]						
Jumping	0.98	0.02	0	0	0	0
Running	0	0.97	0	0.03	0	0
Bending	0	0	0.99	0	0.01	0
Standing	0	0	0	0.95	0	0.05
Walking	0.02	0	0	0	0.98	0
Sitting	0	0	0.04	0	0	0.96
Weinland et al. [13]						
Jumping	0.80	0.10	0	0	0	0
Running	0	0.86	0.04	0	0	0.10
Bending	0	0	0.88	0	0.12	0
Standing	0	0	0.17	0.83	0	0
Walking	0	0	0	0.18	0.82	0
Sitting	0.16	0	0	0	0	0.84
Iosifidis et al. [16]						
Jumping	0.90	0	0	0.10	0	0
Running	0.06	0.94	0	0	0	0
Bending	0	0	0.93	0	0.07	0
Standing	0	0	0	0.92	0	0.08
Walking	0.05	0	0	0	0.95	0
Sitting	0	0	0	0.09	0	0.91

Table 5 continued

Recognized instances / Total instances	Jumping	Running	Bending	Standing	Walking	Sitting
Weinland et al. [14]						
Jumping	0.76	0.12	0.12	0	0	0
Running	0	0.80	0	0	0	0.20
Bending	0	0	0.82	0.18	0	0
Standing	0	0	0.20	0.79	0.01	0
Walking	0	0	0	0	0.80	0.20
Sitting	0	0	0	0	0.19	0.81
Iosifidis et al. [17]						
Jumping	0.96	0	0	0.04	0	0
Running	0	0.95	0	0	0.05	0
Bending	0.06	0	0.94	0	0	0
Standing	0	0	0.04	0.96	0	0
Walking	0	0	0	0	0.95	0.05
Sitting	0	0.04	0	0	0	0.96
Iosifidis et al. [18]						
Jumping	0.98	0.02	0	0	0	0
Running	0	0.96	0.04	0	0	0
Bending	0	0	0.99	0	0.01	0
Standing	0	0	0	0.95	0.05	0
Walking	0.06	0	0	0	0.94	0
Sitting	0	0	0	0.08	0	0.92

where C is the penalty parameter, l is the number of training data, k is the number of classes, yi is the class of the ith training data ω points perpendicular to the separating hyper plane, b is the offset parameter to increase the margin, and ξ is the degree of misclassification of the datum x_i. This gives the decision function:

$$f(x) = \arg\max_{m=1,\ldots,k}[\omega_m \times x + b_m] \tag{20}$$

3 Results, analysis and discussion

In this section, we perform the experiments and show results of the proposed method. We implemented human activity recognition method as described in Sect. 2 and tested on several datasets of human activity videos. Here, we present results for six representative publicly available human activity recognition video datasets—our own view point dataset, VideoWeb Multi-view dataset [28], i3DPost multi-view dataset [29], WVU multi-view human action recognition dataset [30], MSR action recognition database [31] and i3DPost multi-view dataset [29] for multiple human. Videos in these datasets have been captured at different rotation angle for multiple viewpoints. The experiments have been performed in Matlab R2013a window 7 environments on an Intel® Core™ i3 2.27 GHz machine with 4 GB RAM.

In our implementation, first we take the training videos and apply background subtraction according to the method described in Sect. 2.1. Secondly, scale invariant contour-based pose feature from silhouettes have been extracted. After that, extracting uniform rotation invariant local binary patterns (LBP) feature. Since it is rotation invariant, therefore provides robust results towards viewpoint changes. Uniform patterns provide good discriminating power. Lastly, multiclass SVM classifier has been employed for classification of activities in videos.

Four case studies of our own view point dataset, VideoWeb Multi-view dataset [28], i3DPost multi-view dataset [29], and WVU multi-view human action recognition dataset [30] are discussed here one by one. In all case studies, we have illustrated and tested the proposed method in comparison to Chaaraoui et al. [19], Bobick et al. [11], Ahmad et al. [15], Iosifidis et al. [20], Weinland et al. [13], Iosifidis et al. [16], Weinland et al. [14], Iosifidis et al. [17], and Iosifidis et al. [18]. For quantitative analysis of the proposed method and its comparative analysis with other methods correct recognition rate (CRR) is calculated which is defined as follows:

$$\text{CRR} = \frac{N_c}{N_a} \times 100 \text{ (in percentage)} \tag{21}$$

where N_c is the total number of correct recognition sequences while N_a is the number of total activity sequences.

Table 6 Recognition results over the i3DPost multi-view dataset [29]

Method	Accuracy (%)
Proposed method	98.83
Chaaraoui et al. [19]	82
Bobick et al. [11]	72.83
Ahmad et al. [15]	90
Iosifidis et al. [20]	97.16
Weinland et al. [13]	83.83
Iosifidis et al. [16]	92.5
Weinland et al. [14]	79.66
Iosifidis et al. [17]	95.33
Iosifidis et al. [18]	95.66

3.1 Experiment 1

In Fig. 5, we have shown results our own created database from different viewpoints. This database contains video of static human activities namely standing and two dynamic activities namely bending and jogging in different view direction. These videos are taken in real outdoor environment. From the observation of this figure, it is clear that the proposed method is well capable of recognizing these static and dynamic activities. Moreover, there is some little movement in each activity, i.e. pose of human object does not remain still for all the time. Direction of each human object also changes in different frames. Therefore, the proposed method is pose invariant and frontal view is not necessary for recognition of objects and suits for recognition of objects with frontal as well as side view. The proposed method is capable of recognizing the activity at these different viewing angles correctly and the proposed method is robust towards different rotations of the activity.

We have shown qualitative results of the proposed method on our own datasets. Now, we show quantitative results of the proposed method and compare them with other existing methods in terms of confusion matrix. The other methods are Chaaraoui et al. [19], Bobick et al. [11], Ahmad et al. [15], Iosifidis et al. [20], Weinland et al. [13], Iosifidis et al. [16], Weinland et al. [14], Iosifidis et al. [17], and Iosifidis et al. [18].

The confusion matrix of different activity for different methods is shown in Table 1. After observing these tables, we see that the diagonal values are the highest for the proposed method in each case. A comparison of recognition accuracy of different methods is shown in Table 2 (calculate using Eq. 21). Higher the value, higher will be the recognition accuracy. From these confusion matrices and recognition results, it can be observed that the performance of the proposed method is better in comparison to other existing methods. The recognition accuracy of the proposed method is greater than other methods.

3.2 Experiment 2

We have demonstrated results of the proposed method for VideoWeb Multi-view dataset [28]. VideoWeb dataset involves up to 10 actors interacting in various ways (with each other, with vehicles or with facilities). The activities are: waving, boxing, clapping, jogging, running and walking. It consists of about 2.5 h of video recorded from a minimum of 4 and a maximum of 8 cameras. Each video is recorded by a camera network whose number of cameras depends on the type of scene.

From Fig. 6, it can be observed that the person is performing different activity such as boxing, clapping, jogging, running, and walking at different viewing angles. From Fig. 6, it is concluded that the pose of human object does not remain still for all the time. Direction of each human object also changes in different frames. Therefore, the proposed method is pose invariant and frontal view is not necessary for recognition of objects and suits for recognition of objects with frontal as well as side view. These visual results show that the obtained results are accurate and the proposed method provide proper recognition results for VideoWeb Multi-view dataset [28].

Now, we show quantitative results of the proposed method and compare them with other existing methods in terms of confusion matrix. The other methods are Chaaraoui et al. [19], Bobick et al. [11], Ahmad et al. [15], Iosifidis et al. [20], Weinland et al. [13], Iosifidis et al. [16], Weinland et al. [14], Iosifidis et al. [17], and Iosifidis et al. [18].

The confusion matrix of different activity for different methods is shown in Table 3. After observing these tables, we see that the diagonal values are the highest for the proposed method in each case. A comparison of recognition accuracy of different methods is shown in Table 4. Higher the value, higher will be the recognition accuracy. From these confusion matrices and recognition results, it can be observed that the performance and recognition accuracy of the proposed method is better in comparison to other existing methods.

3.3 Experiment 3

We have shown activity recognition with standard i3DPost dataset which is a multi-view dataset [29]. In this dataset, 8 people performing 13 actions (walking, running, jumping, bending, hand-waving, jumping in place, sitting-stand up, running-falling, walking-sitting,

Fig. 8 Recognition of Activities in WVU multi-view human action recognition dataset [30]: **a** hand waving,and **b** hand clapping and **c** walking

running-jumping-walking, handshaking, pulling, and facial expressions) each one. In Fig. 7, six different activities have been performed on multi-view. These activities have been performed with the help of 5 cameras placed at different viewing angles and activities have been captured simultaneously with these cameras. These visual results show that the obtained results are accurate and the proposed method provide proper recognition results for this set of videos also.

Now, quantitative results have been shown for i3DPost multi-view dataset [29] in Tables 5 and 6.

These confusion matrices and recognition results in Tables 5 and 6 indicate that the proposed method performs better than other methods.

3.4 Experiment 4

In Fig. 8, we have shown activity recognition with WVU multi-view human action recognition dataset [30]. This database includes different activities hand waving, clapping, jumping, jogging, bowling, throwing, pickup, and kicking. WVU multi-view human action recognition dataset [30] has been sorted based on the eight views. For each view, action sequences performed by different subjects are provided. In Fig. 8, it is easily concluded that the proposed method is invariant with respect to pose of the human object and also a frontal view is not necessary for recognition of objects and gives satisfactory results for human objects with frontal as well as side view.

Now, quantitative results have been shown WVU multi-view human action recognition dataset [30] in Tables 7 and 8.

These confusion matrices and recognition results presented in Tables 7 and 8 show that the accuracy of the proposed method is better than the other existing methods. Each confusion matrix shows the performance of a particular method for the chosen dataset. Comparison of recognition accuracy of different method with the proposed method is shown in Table 8.

In Table 9, average computation time (second/frame) and memory consumption for different methods for a video of frame size 480 $\times$ 320 with 100 frames [31] are shown. From the Table 9, it can be observed that the proposed method is faster than Chaaraoui et al. [19], Bobick

Table 7 Confusion matrix for the proposed method and other methods

Recognized instances Total Instances	Hand waving	Hand-clapping	Walking
For the proposed method			
Hand waving	1	0	0
Hand-clapping	0	1	0
Walking	0	0.02	0.98
Chaaraoui et al. [19]			
Hand waving	0.72	0.28	0
Hand-clapping	0.30	0.70	0
Walking	0.30	0.01	0.69
Bobick et al. [11]			
Hand waving	0.79	0.21	0
Hand-clapping	0	0.81	0.19
Walking	0.18	0	0.82
Ahmad et al. [15]			
Hand waving	0.88	0.02	0.10
Hand-clapping	0.16	0.84	0
Walking	0	0.15	0.85
Iosifidis et al. [20]			
Hand waving	0.97	0	0.03
Hand-clapping	0.07	0.93	0
Walking	0	0.08	0.92
Weinland et al. [13]			
Hand waving	0.88	0.12	0
Hand-clapping	0	0.85	0.15
Walking	0	0.13	0.87
Iosifidis et al. [16]			
Hand waving	0.93	0.07	0
Hand-clapping	0.09	0.91	0
Walking	0	0.07	0.93
Weinland et al. [14]			
Hand waving	0.78	0	0.22
Hand-clapping	0.09	0.81	0.10
Walking	0	0.18	0.82
Iosifidis et al. [17]			
Hand waving	0.94	0.06	0
Hand-clapping	0.05	0.95	0
Walking	0	0.08	0.92
Iosifidis et al. [18]			
Hand waving	0.97	0.03	0
Hand-clapping	0.05	0.95	0
Walking	0.05	0	0.95

et al. [11], Ahmad et al. [15], Iosifidis et al. [20], Weinland et al. [13], Iosifidis et al. [16], Weinland et al. [14], Iosifidis et al. [17], and Iosifidis et al. [18]. Also from Table 9, the proposed method consumes only 3.90 megabytes of RAM which is the least in comparison with the other methods discussed [11, 13–20]. Therefore, it can be concluded that the time required for the execution of the proposed method is faster to other methods and consumes less amount of the system memory.

Table 8 Recognition results over the WVU action recognition dataset [30]

Method	Accuracy (%)
Proposed method	99.33
Chaaraoui et al. [19]	70.33
Bobick et al. [11]	80.66
Ahmad et al. [15]	85.66
Iosifidis et al. [20]	94
Weinland et al. [13]	86.66
Iosifidis et al. [16]	92.33
Weinland et al. [14]	80.33
Iosifidis et al. [17]	93.66
Iosifidis et al. [18]	95.66

To see the more qualitative and quantitative experiments of the proposed method please visit the following link: https://sites.google.com/site/alokkushwaha1581988/home/experiments.

4 Conclusions

In this paper, we have proposed a multi-view human activity recognition system. This system is based on three consecutive modules. These are (1) background subtraction (2) feature extraction and (3) classification. The background subtraction has been performed using change detection method. The contour-based pose features from silhouettes find the different key poses for human activities such as bending, standing, sitting etc. After that uniform rotation invariant LBP descriptor has been computed. Its rotation invariant nature provides view invariant recognition of multi-view human activities and uniform patterns facilitate good discriminating capabilities. Multiclass SVM classifier has been applied for recognition of different activities. This approach has been performed on four multi-view human activity video datasets: own view point dataset, VideoWeb Multi-view dataset [28], i3DPost multi-view dataset [29], and WVU multi-view human action recognition dataset [30]. Qualitative and quantitative experimental results demonstrate the robustness of the proposed method against different viewpoints. The proposed method has been compared with Chaaraoui et al. [19], Bobick et al. [11], Ahmad et al. [15], Iosifidis et al. [20], Weinland et al.

Table 9 Computational time and consumption memory for MSR view-point action dataset [31]

S. no.	Methods	Computational time (in frame/s)	Memory consumption (MB)
1	Proposed method	1.232	3.90
2	Chaaraoui et al. [19]	1.722	22.92
3	Bobick et al. [11]	1.443	9.40
4	Ahmad et al. [15]	1.376	24.95
5	Iosifidis et al. [20]	1.912	8.64
6	Weinland et al. [13]	1.753	7.08
7	Iosifidis et al. [16]	1.325	11.37
8	Weinland et al. [14]	1.687	13.35
9	Iosifidis et al. [17]	1.912	30.92
10	Iosifidis et al. [18]	1.412	17.62

[13], Iosifidis et al. [16], Weinland et al. [14], Iosifidis et al. [17], Iosifidis et al. [18] and found better than them.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2), 90–126 (2006)
2. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **34**(3), 334–352 (2004)
3. Gavrila, D.: The visual analysis of human movement: a survey. Comput. Vis. Image Underst. **73**(1), 82–98 (1999)
4. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Comput. Vis. Image Underst. **81**(3), 231–268 (2001)
5. Haritaoglu, I., Harwood, D., Davis, L.S.: W4: real-time surveillance of people and their activities. IEEE Trans. Pattern Anal. Mach. Intell **22**, 809–830 (2000)
6. Olson, T., Brill, F.: Moving object detection and event recognition algorithms for smart cameras. In: Proc. DARPA Image Understanding Workshop, pp. 159–175 (1997)
7. Lipton, A.J., Fujiyoshi, H., Patil, R.S.: Moving target classification and tracking from real-time video. In: Proc. IEEE Workshop Applications of Computer Vision, pp. 8–14 (1998)
8. Srinivasan, K., Porkumaran, K., Sainarayanan, G.: Intelligent human body tracking, modeling, and activity analysis of video surveillance system: a survey. Int. J. Converg. Eng. Technol. Sci. **1**, 1–8 (2009)
9. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Syst. Man Cybern **34**(3), 334–352 (2006)
10. Weinland, D., Ronfard, R.: A survey of vision based methods for action representation, segmentation, and recognition. Comput. Vis. Image Underst. **115**(2), 529–551 (2011)
11. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans Pattern Anal. Mach. Intell **23**(3), 257–267 (2001)
12. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
13. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Comput. Vis. Image Underst. **104**(2–3), 249–257 (2006)
14. Weinland, R., Ronfard, R., Boyer, E.: Automatic discovery of action taxonomies from multiple views. IEEE Conf. Comput. Vis. Pattern Recognit **2**, 1639–1645 (2006)
15. Ahmad, M., Lee, S.: Human action recognition using shape and CLG-motion flow from multi-view image sequences. Pattern Recogn. **41**(7), 2237–2252 (2008)
16. Iosifidis, A., Tefas, A., Nikolaidis, N., Pitas, I.: Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis. Comput. Vis. Image Underst. **116**, 347–360 (2011)
17. Iosifidis, A., Tefas, A., Pitas, I.: View-invariant action recognition based on artificial neural networks. IEEE Trans. Neural Netw. Learn. Syst **23**(3), 412–424 (2012)
18. Iosifidis, A., Tefas, A., Pitas, I.: Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis. Sig. Process. **93**, 1445–1457 (2013)
19. Chaaraoui, A.A., Climent-P´erez, P., Fl´orez-Revuelta, F.: Silhouette-based human action recognition using sequences of key poses. Pattern Recogn. Lett. **34**(15), 1799–1807 (2013)
20. Iosifidis, A., Tefas, A., Pitas, I.: Learning sparse representations for view-independent human action recognition based on fuzzy distances. Neurocomputing **121**, 334–353 (2013)
21. Sharma, C.M., Kushwaha, A.K.S., Nigam, S., Khare, A.: Automatic human activity recognition in video using background modeling spatio-temporal template matching based technique. In: Proceedings of ACM International Conference on Advances in Computing and Artificial Intelligence, pp. 97–101 (2011)
22. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3361–3368 (2011)
23. Liu, W., Zhang, Y., Tang, S., Tang, J., Hong, R., Li, J.: Accurate estimation of human body orientation from RGB-D sensors. IEEE Trans. Cybern. **43**(5), 1442–1452 (2013)
24. Snidaro, L., Foresti, G.L.: Real-time thresholding with Euler numbers. Pattern Recognit. Lett. **24**(9/10), 1533–1544 (2003)
25. Dedeoglu, Y., Toereyin, B.U., Gueduekbay, U, Cetin, A.E.: Silhouette-based method for object classification and human action recognition. In: Proceedings of the ECCV workshop on HCI (2006)
26. Pietikäinen, M.: Computer Vision Using Local Binary Patterns, vol. 40. Springer, Berlin (2011)

27. Westons, J., Wtkins, C.: Support vector machines for multiclass pattern recognition. In: Proceedings of the 7th European Symposium on Artificial Neural Networks, pp. 219–224 (1999)
28. Denina, G., Bhanu, B., Nguyen, H., Ding, C., Kamal A., Ravishankar, C., Roy-Chowdhury, A., Ivers, A., Varda, B.: VideoWeb Dataset for multi-camera activities and non-verbal communication. In: Distributed Video Sensor Networks, Springer (2010)
29. University of Surrey and CERTH-ITI, i3dpost multi-view human action datasets, January 2012. http://kahlan.eps.surrey.ac.uk/i3dpostaction/
30. Kulathumani, V.: WVU Multi-view action recognition dataset available on: http://csee.wvu.edu/~vkkulathumani/wvu-action.html#download2
31. Yuan, J., Liu Z, Wu, Y.: Discriminative subvolume search for efficient action detection. In: IEEE Conf. on Computer Vision and Pattern Recognition (2009). http://www.ece.northwestern.edu/~jyu410/index_files/actiondetection.html