Chapter 3 MULTI-VIEW HUMAN ACTIVITY RECOGNITION SYSTEM USING MULTIPLE FEATURES FOR VIDEO SURVEILLANCE SYSTEM

3.1. Introduction.

The recognition of human activities is a popular field of computer vision in the field of research. It builds on many applications, such as security, surveillance, biomechanical clinical applications, human interaction between manipulators, entertainment, education, training, digital libraries, video or image annotation, video conferencing, and Model coding. Awareness activities provide important clues to human behavioural analysis techniques. Although there is a lot of work that has been carried out in the recognition of activities in the past few years, it is still an open and challenging issue. The various issues and challenges involved in automatic human recognition from video sequences are as follows:

- Trajectory activity is different from different viewing directions, some body parts (partial hand, lower leg, body part, etc.), block due to changes made in the view.
- Other common problems include fixed or moving cameras, lenses with changes in moving or clutter backgrounds, changes in light point of view, size, starting and ending states, changes in appearance on the human rights of individuals and wipes, etc. Problems and circumstances make the recognition of human activities a challenging task.
- Real 3D environments where human activities are performed and cameras capture only real scenes of 2D projections. Therefore, the activities carried out by the visual analysis have only one projected actual activity on the image plane. This

predicted activity depends on the point of view, does not include all the information, the activities performed.

To overcome these problems, the idea of using information through the cameras placed at different views has been used [293]. Use of information obtained from multiple cameras at multiple views provides efficient analysis of actual human activities. Few approaches have been developed by introducing view invariant representation for multiple views [293], [294]. Exploring the information from multiple views of a scene improves recognition accuracy of human activities by extracting features from different 2D image views and achieves view invariance. The ultimate goal is to be able to perform human activity recognition applicable for video surveillance and designing an automatic HAR (Human activity recognition) system which is view-invariant robust and reliable.

We contribute to this field, by proposing an intelligent system for multi-view human activity recognition in videos, whose framework is given below:

(i) Detecting and locating people by background subtraction approach.

(ii) Extracting Contour based distance signal feature, optical flow-based motion feature, and uniform rotation local binary patterns.

(iii) Modelling activities by using a set of hidden Markov models (HMMs).

Most of the work on activity recognition are view dependent and deal with recognition from one fixed view. The task of recognizing people's activities from different views is still unsolved.

In this chapter, we have combined contour-based distance signal feature, optical feature and uniform rotation invariant local binary patterns (LBP) feature to model human activities to solve above mentioned problems. At first, statistical background model based approach is used for background subtraction. In the second step, contour based distance signal features, optical flow based motion features, and uniform rotation invariant local

binary patterns (LBP) are extracted. The contour based distance signal features find the different key poses for human activities such as bending, standing, sitting etc. Optical flow based motion features helps in representing the approximation of the moving direction of the human body and can be effectively characterized by motion rather than other cues, such as color, depth, and spatial features e.g. walking running, jogging, boxing etc. 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, the activities are modelled by using a set of HMMs. The use of a set of HMMs for modelling the activities provides view-invariant operation, deal with time-sequential data and also provide time-scale invariability in activity recognition. This overall approach has never been used before in literature for human activity recognition.

To demonstrate the effectiveness and robustness of the proposed method, we have conducted our experiments on our own viewpoint dataset and four representatives publicly available human activity recognition video datasets–KTH action recognition dataset [295], i3DPost multi-view dataset [296], MSR view-point action dataset [297] and WVU multi-view human action recognition dataset [298]. The proposed system has been compared with four existing human activity recognition methods proposed by Qian *et al.* [29], Sadek*et al.*[299], Ikizler-Cinbis & Sclaroff [300], and Ahmad *et al*. [301]. For comparison the proposed methods with other standard methods the confusion matrix and recognition accuracy (in percentage) evaluation parameters have been used. Experimental results on the above mentioned five datasets illustrate the efficiency and the effectiveness of the proposed method.

3.2. The Proposed Method

The proposed system for recognition of human activity is shown in Fig. 3.1. The forepart from the frame is extracted by using the concept of statistical background modelling. From the foreground image, we estimated contour-based distance signal feature which is further processed on the basis of the centre of mass (CM) as per the same silhouette image. These features are used to find out the different human key poses (such as standing, bending, sitting etc.) for activity recognition. The velocity of a different activity (running, jogging, walking, etc.) is estimated by using optical flow which is used as motion features from the foreground image. We use uniform rotation invariant local binary patterns (LBP) for extracting the feature from the foreground image which provides view invariant recognition of multi-view human activities.

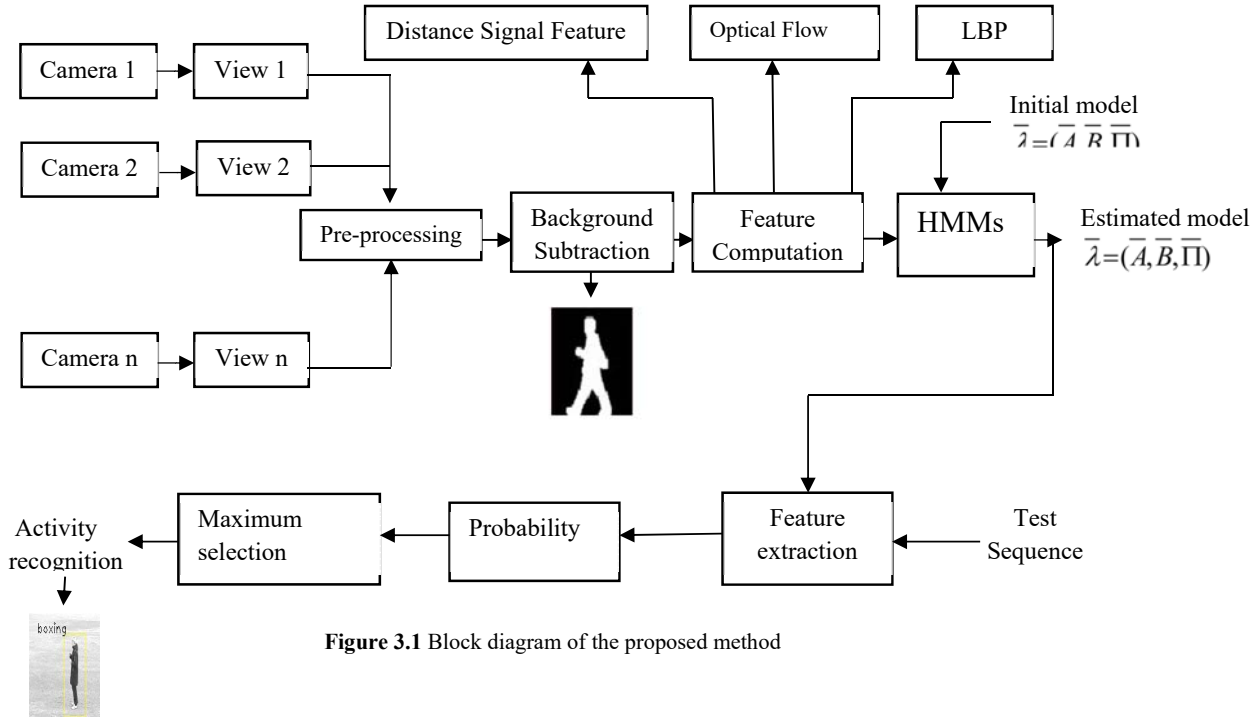


Figure 3.1 Block diagram of the proposed method

The combined features (Contour based distance signal feature, optical feature based motion feature and uniform rotation invariant local binary patterns (LBP) feature are then feed to a hidden Markov model (HMM). In HMM based human activity modelling,

matching of an unknown sequence with a model is done through the calculation of the probability and maximum likelihood estimation that hidden Markov model (HMMs) could generate the particular unknown sequence.

The block diagram of the proposed method is shown in Figure 3.1.

Algorithm: Multi-View Recognition System for Human Activity Based on Multiple Features for Video Surveillance System

Input: Video captured from multiple camera

Output: Classification of input video sequence in recognised activity

1. Accumulate all frames staring from sFrame to eFrame in an array $\eta_i(\phi,\psi)$
2. for each frame in the array $\eta_i(\phi,\psi)$
   - 2.1. Perform Steps 2.1.1 – 2.1.6 for pre-processing of the frames
     - 2.1.1. Calculate variance ξ^2 using equation – 3.1
     - 2.1.2. Covariance between frames α and β can be calculated using equation – 3.2.
     - 2.1.3. The learning based on the variance and covariance to estimate absolute and relative changes in a pixel's intensity is stored in a reference image $R(\phi,\psi)$.
     - 2.1.4. Calculate difference between subsequent video frames and the reference image.
     - 2.1.5. Perform Binary thresholding on the image obtained from Step-5, to obtain the binary – segmented image
     - 2.1.6. Calculate updated background reference model $R_{new}(\phi,\psi)$ using equation – 3.3.
   - 2.2. Perform Steps 2.2.1 – 2.2.5 for calculating Distance Signal Feature
     - 2.2.1. Calculate silhouette A={ a_1, a_2… a_n} of the binary image obtained in the previous step
     - 2.2.2. Now we calculate centre of mass Cm = ($\bar{x}$, $\bar{y}$) using equation – 3.4.
     - 2.2.3. Now we calculate distance signal D= {d_1, d_2, d_3……d_n} using equation 3.5.
     - 2.2.4. To obtain the scale invariance, Fix the distance signal D and sub-sample the feature size to a constant length L using equation –3.6
     - 2.2.5. Value obtained can be normalized using equation – 3.7
   - 2.3. Perform Steps 2.3.1 – 2.3.2 for calculating Optical Flow Feature
     - 2.3.1. With the help of calculate normalized optical flow $v_{nx}(x,t)$ at any instance of time using equation – 8 and equation – 3.9.
     - 2.3.2. Calculated absolute optical flow of the activity boundary at any time K using value obtained in step – 2.3.1, using equation – 3.10.
   - 2.4. Perform Steps 2.4.1 – 2.4.2 for calculating LBP feature.
     - 2.4.1. Calculate overall feature vector $LBP_{P,R}$ using equation – 3.11.
     - 2.4.2. Now Calculate Uniform LBP feature using equation 3.14.
3. Let I={I_1, I_2,..., I_T}, be the time sequential frames and f_i is the feature vector, generated using steps – 2, from each input frame I_i, where $f_i \in R^n$, (i = 1, 2,…,T & n is the dimension of the feature space R^n).
   - 3.1. Generate a hidden Markov model(HMM) to transform f_i into a symbol O_i.
   - 3.2. Generate Markov chain of symbol sequence O = {O_1, O_2,…,O_T} from the model.
4. Now optimise the model parameter (A, B, π) to maximize the probability of observation sequence
   - 4.1. Calculate maximum likelihood estimation P(O/ λ) using equation – 3.15 to obtain the recognition result.

3.2.1. Preprocessing

After receiving videos from multiple cameras from different views, 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 in the next subsection:

3.2.2. Background Subtraction

The Statistical model for the background is constructed by learning the variance and covariance of pixels in a video sequence. The variance is used to model the absolute variations in a pixel's statistics while covariance is used to model the relative variations. Based on this model, a reference image for the background is created. Frame differencing and thresholding is performed to obtain the segmented video frames. The steps of the algorithm used are given as follows:

Step 1: The frames, starting from index sFrame and ending at index eFrame, are accumulated in array $\eta_i(\phi,\psi)$. The method learns the variance and covariance of each pixel as its model of the background. If $\bar{\eta}$ is the mean of for all η_i samples where $0 \le i < eFrame$ then variance ξ^2 is given as follows:

$$\xi^2 = \left(\frac{1}{eFrame} \sum_{i=0}^{eFrame\text{-}1} (\eta_i - \bar{\eta})^2 \right) \tag{3.1}$$

The covariance between frames α and β is calculated to estimate the variation in a pixel's intensity relative to the other pixels.

$$\mathrm{cov}(\alpha,\beta) = \left(\frac{1}{eFrame} \sum_{i=0}^{eFrame-1} \alpha_i \beta_i \right) - \left(\frac{1}{eFrame} \sum_{i=0}^{eFrame-1} \alpha_i \right) \left(\frac{1}{eFrame} \sum_{j=0}^{eFrame-1} \beta_j \right) \tag{3.2}$$

Step 2: The learning based on the variance and covariance to estimate absolute and relative changes in a pixel's intensity is stored in a reference image $R(\phi,\psi)$.

Step 3: For object segmentation, the subsequent video frames are differenced with the reference image and binary thresholding is performed for obtaining grayscale segmented frame.

Step 4: A temporal updation of the background model is needed in order to adapt the changes in background and in lighting conditions. A counter ρ can be used to track when to update background model. When the value of counter exceeds a threshold number ξ, the background model $R(\phi,\psi)$ is updated. The background model is updated using the following equation:

$$\mathrm{R}_{\mathrm{new}}(\phi,\psi) = \tau * \mathrm{R}(\phi,\psi) \; + \; (1-\tau) * frame_{\rho}(\phi,\psi) \tag{3.3}$$

Where τ is the updating speed, $frame_{\rho}(\phi,\psi)$ is the current video frame and $\mathrm{R}_{\mathrm{new}}(\phi,\psi)$ is the updated background reference model. Figure 3.2 shows the threshold segmented image obtained after background substraction for some activities of our test dataset.

Running Walking Sitting

Figure 3.2: Threshold segmented image obtained after background sabstraction for running, walking, and sitting activities of KTH, i3DPost and own dataset.

3.2.3. Feature Extraction

We use the distance signal feature, local binary pattern (LBP) and optical flow based motion feature for activities representation and classification. The foreground image sequence (which is obtained in section 3.2.1) is used to extract the distance signal feature, local binary pattern (LBP) and motion flow features.

i. Distance Signal 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 section 3.2.2 by human silhouette extraction techniques, e.g. background subtraction. We have chosen Dedeog˘lu *et al.* [302] approach for contour-based distance signal feature extraction (see Fig. 3.2) for different key poses (sitting, sleeping etc.), which is described briefly in the following.

Figure 3.3 Sequence of key poses of several activities (walking, jogging, running) to obtain contour based distance feature in some selected frames (KTHDB). [285]

At first, the contour points A= $\{a_1, a_2 \ldots a_n\}$ of the silhouette need to be obtained. For this purpose, contour extraction is applied on the border using Suzuki *et al.* [303] approach. Secondly, the centre of mass Cm = $\left(\bar{x}, \bar{y}\right)$ of the silhouette's contour points is calculated with respect to the n number of points:

$$\bar{x} = \frac{\sum_{w=1}^{n} x_w}{n}, \quad \bar{y} = \frac{\sum_{w=1}^{n} y_w}{n} \tag{3.4}$$

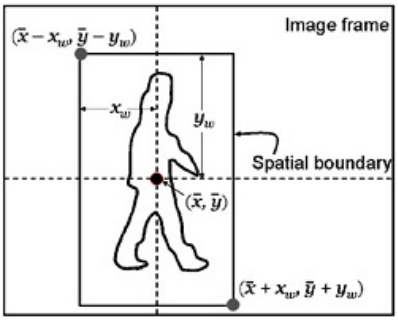


Figure 3.4 Activity boundary definitions [301]

Thirdly, the distance signal D= $\{d_1, d_2, d_3 \ldots\ldots d_n\}$ is generated by determining the Euclidean distance between each contour point and the centre of mass (see fig 3.3.).

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.

$$d_i = \left\| C_m - a_i \right\| , \ \forall i \in [1.....n] \tag{3.5}$$

Finally, scale-invariance is obtained by fixing the size of the distance signal D, sub-sampling the feature size to a constant length L and normalizing its values to unit sum.

$$\hat{D}[i] = D\left\lceil i * \frac{n}{L} \right\rceil , \ \forall i \in [1.....L] \tag{3.6}$$

$$\overline{D}[i] = \frac{\hat{D}[i]}{\sum_{i=1}^{L} \hat{D}[i]} , \ \forall i \in [1.....L] \tag{3.7}$$

ii. Optical Flow Features (Motion Features)

In this chapter, we used optical flow to describe motion feature. From the results of the background subtraction, we obtain a region of interest (ROI) whose example is showed in Figure2 &3. Then, we compute the optical flow based motion feature using method proposed in [301], [304] which is described briefly as follows:

The following notations are used in to describe the concepts of optical flow:

$v = [v_x, v_y, 1]^T$ at pixel (x, y)

Velocity gradient: $\nabla v = \left| \nabla v_x \right|^2 + \left| \nabla v_y \right|^2$,

intensity gradient : $\nabla_3 I = (I_x, I_y, I_t)^T$,

and motion tensor : $J_I(\nabla_3 p) = K_I * (\nabla_3 I \nabla_3 I^T)$

where K_I is smoothing kernel

Then the optical flow function is given by

$$E_{optical\ flow} = \int_{video} (v^T J_I(\nabla_3 \rho) v + \alpha \left| \nabla_3 v \right|^2) dx dy dt \tag{3.8}$$

Where α is the smoothing constant and $|\nabla_3 v|^2 = |\nabla_3 v_x|^2 + |\nabla_3 v_y|^2$. $|\nabla_3 v|^2$ minimizes the optical flow field $v(x,t) = (v_x(x,t), v_y(x,t))$ using Euler–Lagrange equations [305].

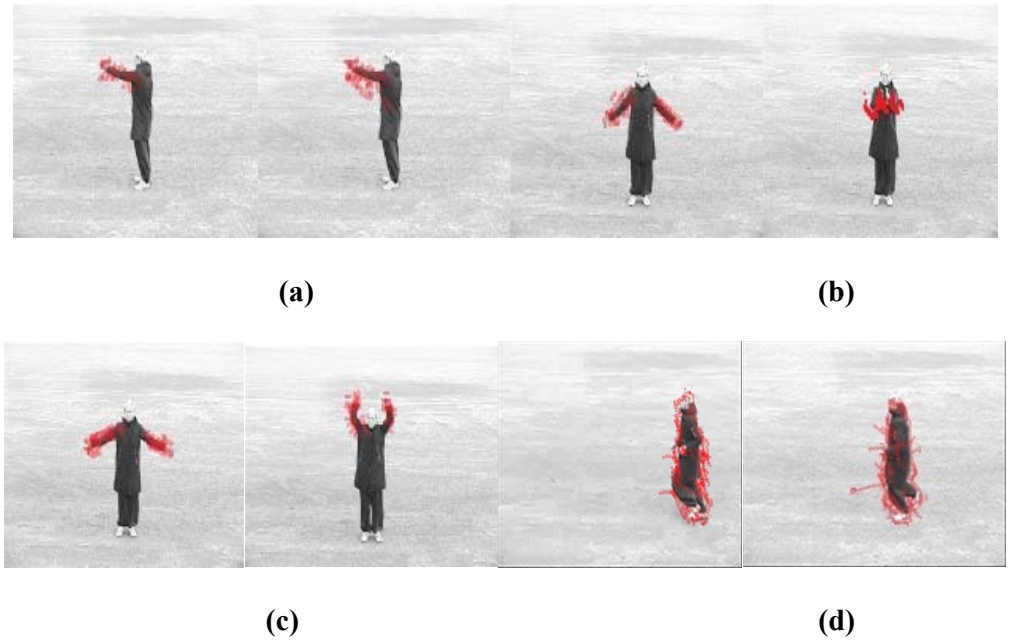


Figure 3.5 Optical flow velocity of several activities in some selected frames (KTHDB) [285]. (a) Boxing; (b) hand clapping; (c) hand waving; (d) jogging

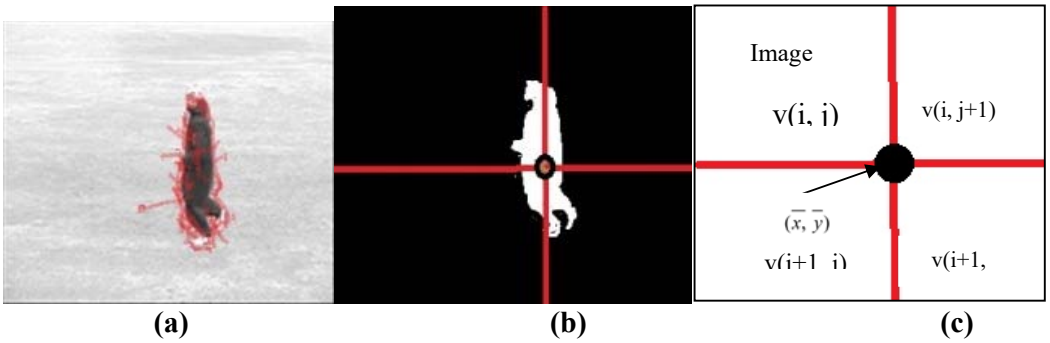


Figure 3.6 Optical flow motion features extraction. (a) Optical flow show in the quadrant regions. (b) The small circle represents the Centre of Mass for (a). (c) Four quadrant blocks from the Centre of Mass

Fig. 3.4 shows the optical flow velocity overlapping on the image of several activities. For example, when the person performs the "hand waving", motion only involves the hand. Similarly, when the person conducts the "running", then motion involves the whole body. For consistency of further analysis, the optical flow vector are normalized at any instant of time, by

$$v_{nx}(x,t)=\begin{cases} x-flow:\dfrac{v_x(x,t)-v_{x.\min}(t)}{(v_{x.\max}(t)-v_{x.\min}(t))} \\ y-flow:\dfrac{v_y(x,t)-v_{y.\min}(t)}{(v_{y.\max}(t)-v_{y.\min}(t))} \end{cases} \quad (3.9)$$

where $v_{nx}(x,t)$ represents the normalized optical flow (the x-component or y-component velocity)in the spatial activity boundary. Moreover, $v_{x.\max}$ and $v_{x.\min}$ represent the maximum and minimum motion of $v_x(x,t)\in \mathrm{H}_a$ where H_a is the spatial boundary of an activity. Similarly, $v_{y.\max}$ and $v_{y.\min}$ represent the maximum and minimum motion of $v_y(x,t)\in \mathrm{H}_a$. In order to extract the features from normalized flow, we partition the spatial activity boundary into four quad-rant blocks, $_{S(k)}$ of equal size, as shown in Fig. 3.5. The four quadrants are described by (i) $\left\{(\bar{x}-x_w,\bar{y}-y_w),(\bar{x},\bar{y})\right\}$ (ii) $\left\{(\bar{x},\bar{y}-y_w),(\bar{x}+x_w,\bar{y})\right\}$ (iii) $\left\{(\bar{x}-x_w,\bar{y}),(\bar{x},\bar{y}+y_w)\right\}$ and (iv) $\left\{((\bar{x},\bar{y}),(\bar{x}+x_w,\bar{y}+y_w)\right\}$ (see figure 3). The point $(\bar{x},\bar{y})$ denotes the centre of mass, x_w is the width, and y_w is the height of the block of the current silhouette image. Therefore, the optical flow feature vectors are extracted at each block with n_S number of pixels using

$$v_{kl,t}=\begin{cases} x-flow:\dfrac{1}{n_S(pix>0)} & \sum\limits_{x\in S(k)} v_{nx}(x,t) \\ y-flow:\dfrac{1}{n_S(pix>0)} & \sum\limits_{x\in S(k)} v_{ny}(x,t) \\ abs.flow:\dfrac{1}{n_S(pix>0)} & \sum\limits_{x\in S(k)} \sqrt{v^2{}_{nx}(x,t)+v^2{}_{ny}(x,t)} \end{cases} \quad (3.10)$$

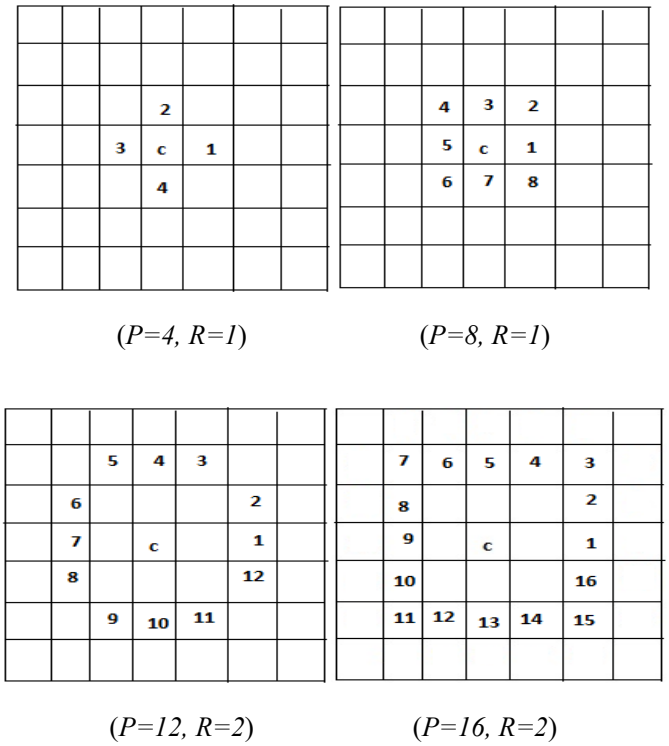
Here, the vector $v_{kl,t}$ represents absolute optical flow of the activity boundary at any time, k denotes the number of blocks, pix represents the nonzero pixel value in the spatial boundary, and n_S is the number of motion pixels at any block.

iii. Uniform Rotation Invariant Local Binary Patterns (LBP) Based Feature Extraction

In this chapter, we extract the Uniform rotation invariant local binary patterns (LBP) features from the background-subtracted video, which is obtained in section 3.3.2. 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.

Local Binary Patterns (LBP)

A local binary pattern (LBP) feature can be constructed for a specific circular pixel neighbourhood of radius *R*. The intensities of the *P* sample pixel points are compared in the circular neighbourhood with the centre pixel in clockwise or anticlockwise direction (see Fig. 3.6).



(*P=4, R=1*) (*P=8, R=1*)

(*P=12, R=2*) (*P=16, R=2*)

Figure 3.7 Circularly symmetric neighbour sets for different (P, R) (here anti-clockwise) .

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 $LBP_{P,R}$, is given as below:

$$LBP_{P,R}(x,y) = \sum_{P=0}^{P-1} s(g_P - g_c)2^P \tag{3.11}$$

Where *(x,y)* is the location of the centre pixel, g_c represent intensity of centre pixel, g_p represent intensity of neighbourhood pixel and $s(u)$ is defined as

$$s(u) = \begin{Bmatrix} 1, & u \geq 0 \\ 0, & u < 0 \end{Bmatrix} \tag{3.12}$$

Now, the feature vector $LBP_{P,R}$ 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 histograms can be merged into one histogram of size $M \cdot 2^P$.

Rotation Invariance

Several modified versions of LBP [306] 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 $LBP_{P,R}$ may be computed. To remove the effect of rotation, the modified version with rotation invariance is defined as follows

$$LBP_{P,R}^{ri}(x,y) = \min\left\{ ROR(LBP_{P,R}, i) \mid i = 0,1,\ldots,R-1 \right\} \tag{3.13}$$

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

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.

Uniform Local Binary Patterns (LBP) for Feature Extraction

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

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{P=0}^{P-1} s(g_P - g_C), \ if \ \mathrm{U}(\mathrm{LBP}_{P,R}) \le 2 \\ P+1, \qquad otherwise \end{cases} \tag{3.14}$$

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

and $s(u) = \begin{cases} 1, & u \ge 0 \\ 0, & u < 0 \end{cases}$ (given in equation 3.12)

g_c = centre pixel of background subtraction image (which is obtained in section 3.2.2)

and g_p = neighbourhood pixel of background subtraction image (which is obtained in section 3.2.2).

3.2.4. Activity Modeling and Classification using Hidden Markov Model (HMM)

Hidden Markov model (HMM) is a stochastic state-space transit model which is robust against temporal, spatial & view-point variations. Moreover, HMM can deal with time-sequential data and also provide time-scale invariability in recognition. Hence HMM is

mostly used as classifier for activity recognition [307], [308]. In this chapter, we have used HMM for modeling and testing activities. The detailed explanation of the HMM can be found in [309]. Before modeling human activity, we review the basic hidden Markov model (HMM) notations as follows:

- T = length of the observation sequence.
- $Q = \{q_1, q_2,\ldots,q_N\}$ – set of N states of model.
- N = number of states in the model.
- $V = \{v_1,v_2,\ldots..,v_M\}$ – set of M output symbols.
- $A = \{a_{ij}\}$, is the N×N transition matrix whose elements $a_{ij} = P(q_t+1= S_j | q_t = S_i)$ are transition probabilities.
- $B = \{b_j(O_k)\}$ is the N×M emission matrix of emitting symbol, where $\{b_j(O_k)= P(O_k= v_k q_t = S_j)\}$ is the probability of emitting v_k at time t by state S_j.
- $\pi = \{\pi_i | \pi_i= P (S_1= q_i\}$, Initial state probability.
- $\lambda = \{A, B, \pi\}$ complete parameter set of the model.

Using this model, transitions are described as follows:

- $S = \{S_t\}$, t = 1, 2,…, T: State S_t is the t_{th} state (unobservable).
- $O= \{O_1, O_2,\ldots,O_T\}$: Observed symbol sequence.

The basic concept of hidden Markov model (HMMs) is shown in Fig.3.7.There is three states in this example. Each state stochastically outputs a symbol v_k with a probability of$b_j(k)$. If there are M symbols, $b_j(k)$ becomes a matrix of N×M. The hidden Markov model (HMMs) outputs the symbol sequence $O = \{O_1, O_2,\ldots\ldots,O_T\}$ from time 1 to T. We can observe the symbol sequence but cannot the HMM states. The initial state of HMM is also determined stochastically by the initial state probability π. A complete HMM is defined by $\lambda = \{A, B, \pi\}$.

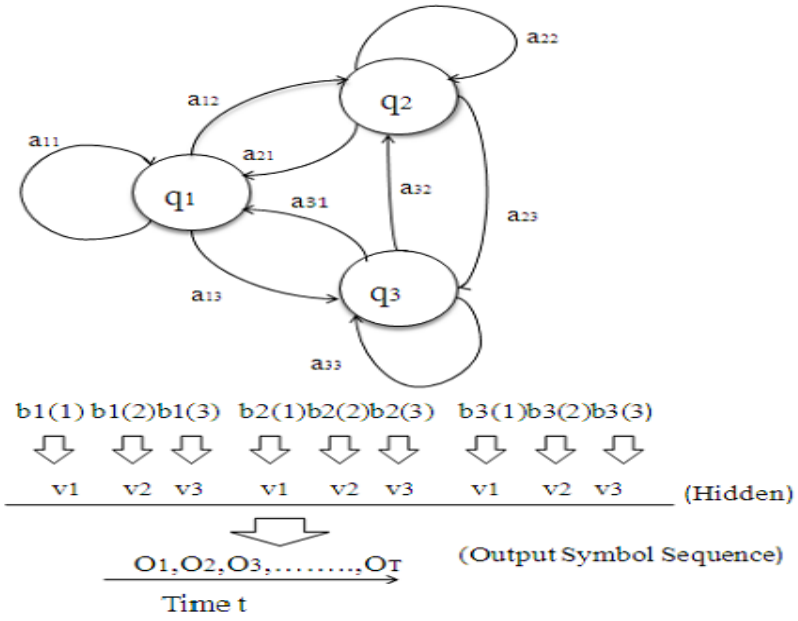


Figure 3.8 Left–right HMM structure for an activity

There are two steps of activity recognition using HMM: Learning and Training.

Step 1: Learning of HMM

We learn the different features (which are obtained in section 3.2.3) using hidden Markov model (HMMs). The input to the feature extraction stage are the time-sequential video frames $I = \{I_1, I_2,..., I_T\}$. The feature extraction stage extract feature vector f_i from each input frame I_i where $f_i \in R^n$, $(i = 1, 2,\ldots,T$ &n is the dimension of the feature space R^n).In the learning phase, hidden Markov model (HMMs) is generated which transforms each feature vector f_i into a symbol O_i. Thus a Markov chain of symbol sequence $O = \{O_1, O_2,\ldots,O_T\}$ are generated from the model.

Step 2: Training of HMM

Once the hidden Markov model (HMMs) learning phase is completed it is trained to recognize the human activities into different classes. In the training phase, the model parameters (A, B, π) are optimized to maximize the probability of observation sequence $P(O/\lambda)$.The forward–backward algorithm or the Viterbi algorithm can be used to classify the activity and find the $P(O/\lambda)$.

$$P(O/\ \lambda)\ =\ \arg\max_{i}\ \left(P(\lambda i / O)\right) \qquad \text{(for i = 1 to T)} \qquad (15)$$

Where P(O/ λ) is called the maximum likelihood estimation of the model parameters. This Maximum likelihood is selected as the recognition result. O represents the unknown feature vector sequence of an unknown activity and λi represents the set of all known activity.

3.3. Experimental Results

This section deals with the various concepts for recognition of human activities based on hidden Markov model (HMM), Moreover implementation is done on image sequences of different activities in different viewing directions. We have presented our own viewpoint dataset result and also for four publicly available video datasets–KTH action recognition dataset [285], i3DPost multi-view dataset[286], MSR view-point action dataset [310] and WVU multi-view human action recognition dataset[288] for the purpose of activity recognition. These datasets have the different rotation angle views of the images. The workstation Open CV 2.4.9 environment with Intel® Core™ i3 2.53 GHz having 4 GB RAM is used for experimentation purpose.

As described in Section 3.2, first of all the training video is taken then background subtraction is applied. After that, the discussed approaches have been implemented for extracting features using contour-based distance signal feature, optical flow-based motion feature and uniform rotation invariant LBP. Lastly, hidden Markov model (HMMs) are used for classification of different activities in videos. At last we have discussed five case studies consisting of our own viewpoint dataset, KTH action recognition dataset [285], i3DPost multi-view dataset [286], MSR view-point action dataset [310] and WVU multi-view human action recognition dataset [288]. In all case studies, we have compared and tested the proposed method with the other standard methods proposed by Sadek *et al.*

[299], Qian*et al.* [29], Ikizler-Cinbis &S claroff [300], and Ahmad *et al.* [301]. For the quantitative analysis of the proposed models, various performance measures as discussed in section 2.6 have been used. The proposed framework has been compared with different state of the arts frameworks in terms of accuracy, error, recall, specificity, precision and f-score.

3.3.1. Experiment 1

Robustness of the proposed method is demonstrated in this very first experiment, we demonstrate the robustness of the proposed method for different rotational movement activities. Testing of the proposed method for activity recognition from a different perspective from our own activity database is done.

Results for own database is shown in the fig. 3.8. This database contains the static human activities of the video i.e. seating and 6 dynamic activities namely walking, running, jogging, boxing, slap, jogging in different directions. These videos are taken in a substantial indoor environment. From the observation of this graph, it is clear that the proposed method is good and can recognize these static and dynamic activities. In addition, there are some small activities in each of the activities that constitute an infringement of the object still cannot be for all the time. Every human object in the direction also changes in different frames. Therefore, the proposed method is to constitute an incomprehensible chapter and a positive view is not necessary, for the object of identification and object of litigation recognition with positive and side views. The suggested method is to be able to recognize whether these activities are correct at different viewing angles, and the proposed method is different for the powerful rotational movement activities.

(a) Boxing

(b) Clapping

(c) Jogging

(d) Running

(e) Sitting

(f) Walking

(g) Hand-waving

Figure 3.9 Recognition of Activities in our own database (a) Boxing (b) Clapping (c) Jogging (d) Running(e) Sitting (f)Walking(g) Hand-waving in different views.

We have shown qualitative results of the proposed method on different 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 Qian *et al.* [29], Sadek *et al.* [299], Ikizler-Cinbis and Sclaroff [300], and Ahmad *et al.* [301].

Table 3.1 Confusion matrices for the proposed and other methods over own dataset

Recognized Instances → **Total Instances ↓**	**Boxing**	**Clapping**	**Jogging**	**Running**	**Sitting**	**Walking**	**Hand-waving**
For the proposed method							
Boxing	.99	.01	0	0	0	0	0
Clapping	0	1	0	0	0	0	0
Jogging	0	0	.98	.02	0	0	0
Running	0	0	.01	.99	0	0	0
Sitting	0	0	0	0	1	0	0
Walking	0	0	0	0	0	1	0
Hand-waving	0	0	0	0	0	0	1
Ikizler-Cinbis and Sclaroff [300]							
Boxing	0.71	0.10	0	0.10	0.05	0	0.04
Clapping	0.15	0.68	0	0.12	0.03	0.02	0
Jogging	0.12	0.10	0.73	0	0	0	0.05
Running	0.05	0.08	0.15	0.70	0.01	0.01	0
Sitting	0.12	0.15	0.05	0	0.65	0.02	0.01
Walking	0	0.15	0.05	0.05	0	0.62	0.13

Hand-waving	0.10	0.14	0.08	0	0.01	0	0.67
Qian*et al.* [29]							
Boxing	0.39	0	0	0.39	0	0	0.22
Clapping	0.15	0.56	0.26	0	0	0.03	0
Jogging	0.02	0.15	0.55	0.1	0.18	0	0
Running	0.03	0.33	0.03	0.58	0.03	0	0
Sitting	0.05	0.15	0.18	0	0.47	0.15	0
Walking	0	0	0	0	0.23	0.77	0
Hand-waving	0.02	0.28	0.03	0	0	0	0.67
Ahmad *et al.* [301]							
Boxing	0.71	0.15	0.10	0.02	0	0	0.02
Clapping	0.12	0.76	0	0.08	0.04	0	0
Jogging	0.10	0.05	0.75	0	0.05	0.03	0.02
Running	0.18	0.03	0	0.72	0.02	0.05	0
Sitting	0.06	0.11	0	0	0.78	0	0.05
Walking	0	0.11	0.03	0.03	0	0.73	0.10
Hand-waving	0	0	0.15	0.03	0.03	0.05	0.74
Sadek *et al.* [299]							
Boxing	0.52	0.18	0.06	0.01	0.20	0.03	0
Clapping	0.22	0.50	0	0.25	0	0.01	0.02
Jogging	0.01	0.19	0.45	0	0.20	0.05	0.10
Running	0.25	0.18	0.02	0.41	0.05	0	0.09
Sitting	0.30	0.10	0.10	0	0.44	0.03	0.03
Walking	0	0.02	0.07	0.21	0.03	0.49	0.18
Hand-waving	0.18	0	0.10	0.15	0.15	0	0.42

Table 3.2 Recognition results over the Own dataset

Method	**Accuracy (%)**	**Error (%)**	**Recall (%)**	**Specificity (%)**	**Precision (%)**	**F-Score (%)**
Ikizler-Cinbis & Sclaroff[300]	68.00	32.00	68.00	94.68	71.44	69.67
Qian *et al.* [29]	57.43	42.57	57.43	92.83	58.81	58.11
Ahmad *et al.* [301]	74.14	25.86	74.14	95.69	75.34	74.74
Sadek *et al.* [299]	46.16	53.84	46.14	91.02	49.34	47.69
Proposed	99.43	0.57	99.43	99.90	99.43	99.43

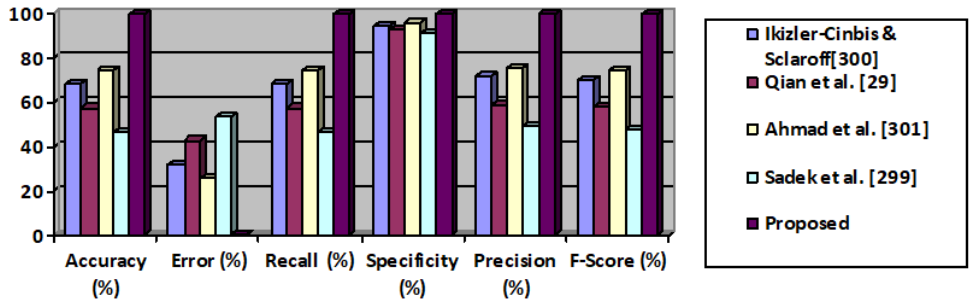
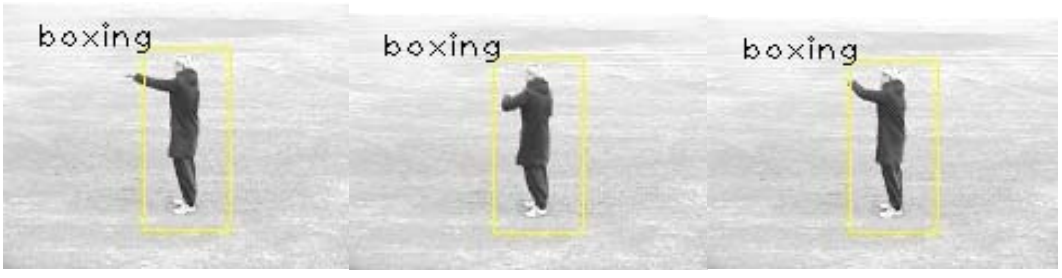


Figure 3.10 Comparison chart over the Own dataset

Different methods of different activities of this confusion matrix are set out in table 3.1. After observing these values, we can see that the value of the diagonal is the highest for the proposed method, in each case. Several different methods for comparing the recognition result in terms of accuracy, error, recall, specificity, precision and f-score have been listed in Table 3.2; above parameters have been calculated using performance measures described in section 2.6. From the results of these confusion matrices, recognition results and bar chart in figure 3.10, it can be observed that the performance of the proposed method is better than the current other methods. The method of identifying the accuracy of the proposed method is better than other methods.

3.3.2. Experiment 2

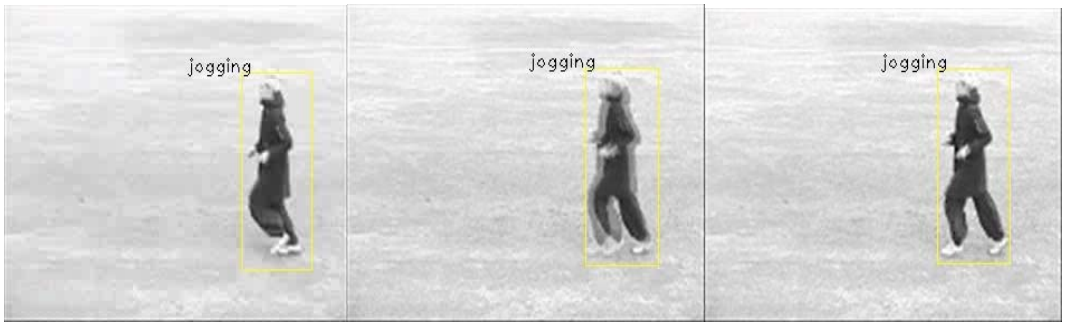
In this section, we will demonstrate the proposed method for identifying KTHDB actions in the database [285]. KTHDB is one of the largest databases with sequences of human actions taking a variety of different scenarios [285]. This data set contains 6 types of human behaviour's (walking, jogging, running, boxing, hand waving, one-hand pat) that have been performed several times by 25 people in four different scenes. The database contains 2391 sequences. The image sequence has a spatial resolution of 160*120 pixels and has an average of 4 seconds in length.



(a) Boxing

(b) Handclapping

(c) Hand Waving



(d) Jogging

(e) Running

Figure 3.11 Recognition of Activities in KTH database [285](a) Boxing (b) Handclapping (c) Hand Waving (d) Jogging (e) Running.

In Figure 3.9, we have shown activity recognition with the standard KTH database [285]. This database contains 6 activities (such as boxing, handclapping, hand waving, jogging, running and walking. The database also suggests that the proposed method performs well. In addition, the database contains not only activities involving leg movements (eg, jogging, running). And walking) but it also contains activities involving hand movements (such as boxing, handclapping and hand waving.) Jogging and running as well as jogging and walking occur between the most confusing, although it varies in different situations but it is proposed It is easy to handle these scenarios. Now, quantitative results have been shown for KTHDB dataset [285] in Tables 3.

Table 3.3 Confusion matrix for the proposed method over the KTH action recognition dataset

Recognized Instances → **Total Instances ↓**	**Boxing**	**Hand-clapping**	**Jogging**	**Hand-waving**	**Running**
For the proposed method					
Boxing	1	0	0	0	0
Hand-clapping	0	1	0	0	0
Jogging	0	0	1	0	0
Hand-waving	0	0	0	1	0
Running	0	0	0	0	1
Ikizler-Cinbis & Sclaroff [300]					
Boxing	0.74	0.10	0.10	0.03	0.03
Hand-clapping	0.10	0.76	0.05	0.05	0.04
Jogging	0.05	0.06	0.81	0.05	0.03
Hand-waving	0.10	0.04	0.03	0.83	0
Running	0.05	0.05	0.12	0	0.78
Qian *et al.* [29]					
Boxing	0.83	0.10	0.07	0	0
Hand-clapping	0.07	0.81	0.02	0.10	0
Jogging	0.10	0.10	0.79	0.01	0
Hand-waving	0.02	0.05	0.05	0.78	0.10
Running	0	0.10	0.06	0.04	0.80
Ahmad *et al.* [301]					
Boxing	0.88	0.05	0.07	0	0
Hand-clapping	0.04	0.92	0	0.03	0.01
Jogging	0.10	0	0.87	0.03	0
Hand-waving	0.05	0	0.06	0.86	0.03
Running	0.03	0.	0.02	0.05	0.90
Sadek *et al.* [299]					
Boxing	0.81	0.05	0.05	0.09	0
Hand-clapping	0.10	0.79	0.05	0.05	0.01
Jogging	0.12	0.10	0.74	0.04	0
Hand-waving	0.14	0.13	0.01	0.71	0.01
Running	0	0.10	0.10	0.03	0.77

Table 3.4 Recognition results over the KTH action recognition dataset [285]

Method	Accuracy (%)	Error (%)	Recall (%)	Specificity (%)	Precision (%)	F-Score (%)
Ikizler-Cinbis & Sclaroff [300]	78.40	21.60	78.40	94.60	78.89	78.64
Qian *et al.* [29]	80.20	19.80	80.20	95.05	80.75	80.47
Ahmad *et al.* [301]	88.60	11.40	88.60	97.15	88.91	88.75
Sadek *et al.* [299]	76.40	23.60	76.40	94.10	77.86	77.12
Proposed	100	0	100	100	100	100

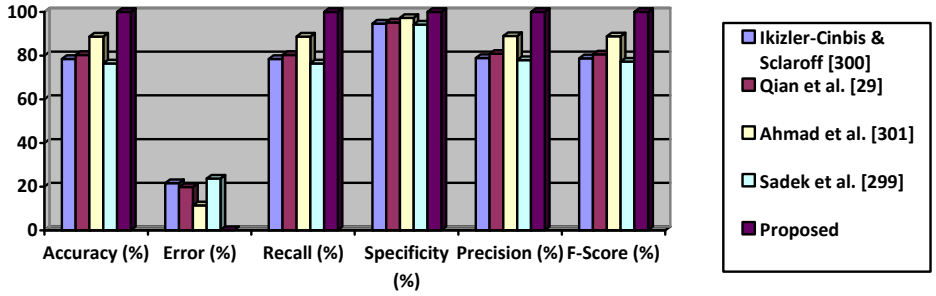


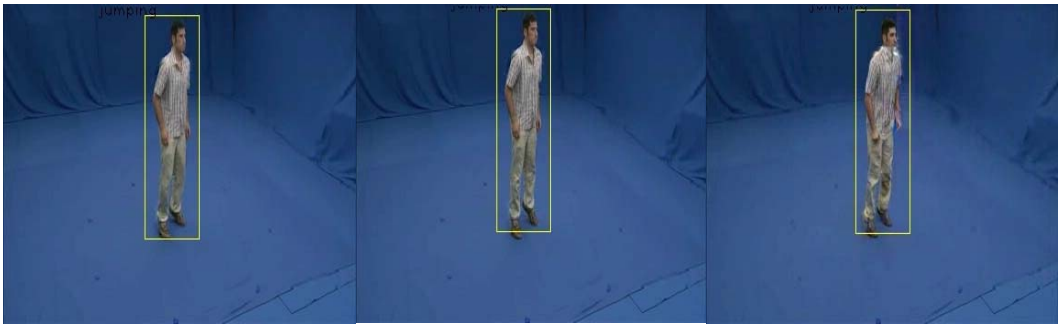
Figure 3.12 Comparison result over the KTH action recognition dataset

From these confusion matrices and recognition results in Tables 3.3, one can find that the accuracy of the proposed method is better than other existing methods. Each confusion matrix shows the performance of a particular method for this dataset. Diagonal values indicate the correct recognition rate for this purpose which are far better in case of the proposed method in Table 3.3. Comparison of recognition result in terms of accuracy, error, recall, specificity, precision and f-score of different method with the proposed method has been shown in Table 3.4 and figure 3.12 (calculated using performance measures described in section 2.6). It shows that the performance of the proposed method is better than other methods.

3.3.3. Experiment 3

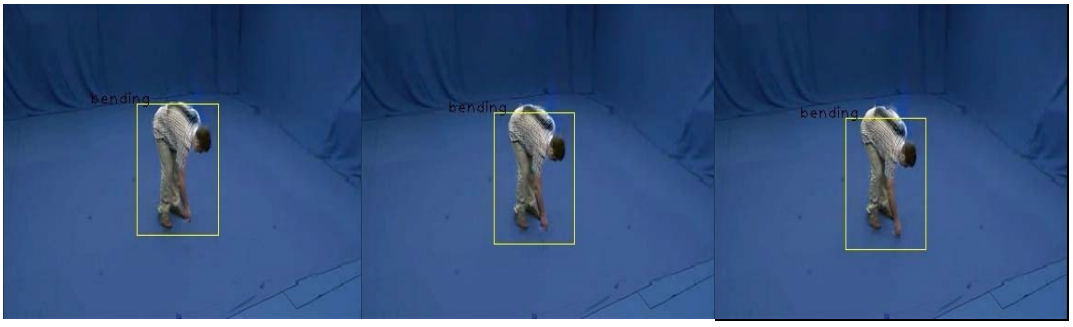
Now, we have selected i3DPost dataset, which is a multi-view dataset [286] for view-invariant human activity recognition. In this dataset, 8 people performing 13 actions (walking, running, jumping, bending, hand-waving, jumping in place, sitting-stand up, running-falling, walking-sitting, running-jumping-walking, handshaking, pulling, and

facial-expressions) each one. The actors have different body sizes, clothing and are of different sex, nationality, etc. According to the authors of this dataset [286], it was expected that full view invariant action recognition, robust to occlusion, would be much more feasible through algorithms based on multi-view videos or 3D posture model sequences. Qualitative recognition results are shown in Figure 10 which shows correct results.

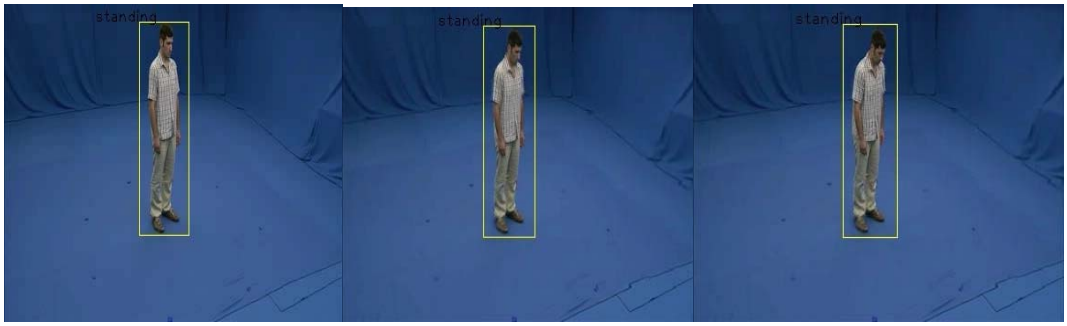


(a) Jumping

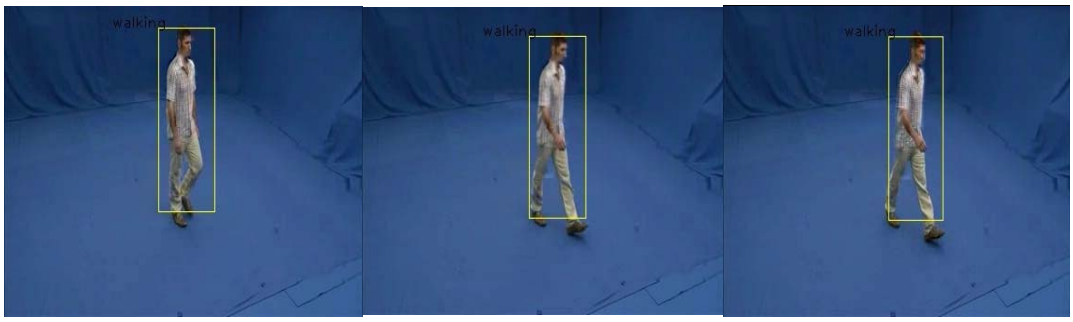
(b) Running



(c) Bending



(d) Standing



(e) Walking

(f) Sitting

Figure 3.13 Recognition of Activities in i3DPost multi-view dataset (a) Jumping (b) Running (c) Bending (d) Standing (e) Walking (f) Sitting (g) Walking

In fig.3.10, 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 provides proper recognition results for this set of videos also. Now, we present quantitative results for i3DPost multi-view dataset [286].

Table 3.5 Confusion matrix for the proposed method over the i3DPost multi-view dataset

Recognized Instances → / **Total Instances ↓**	**Jumping**	**Running**	**Bending**	**Standing**	**Walking**	**Sitting**
For the proposed method						
Jumping	1	0	0	0	0	0
Running	0	1	0	0	0	0
Bending	0	0	1	0	0	0
Standing	0	0	0	1	0	0
Walking	0	0	0	0	1	0
Sitting	0	0	0	0	0	1
Ikizler-Cinbis & Sclaroff [300]						
Jumping	0.80	0.05	0.05	0.05	0.05	0
Running	0	0.83	0.10	0.04	0.03	0
Bending	0.06	0.05	0.86	0.03	0	0
Standing	0.05	0.05	0.08	0.82	0	0
Walking	0.01	0.08	0.08	0	0.81	0.02
Sitting	0.05	0.07	0	0.03	0.05	0.80
Qian*et al.* [29]						
Jumping	0.76	0.10	0.09	0.01	0.04	0
Running	0.10	0.71	0.10	0.05	0	0.04
Bending	0.10	0.05	0.80	0	0.05	0
Standing	0.09	0.01	0.10	0.77	0	0.03
Walking	0	0.05	0.05	0.10	0.74	0.06
Sitting	0	0.10	0	0.10	0.02	0.78
Ahmad *et al.* [301]						
Jumping	0.87	0.05	0.05	0.03	0	0
Running	0.06	0.90	0	0.04	0	0
Bending	0.10	0	0.86	0	0.02	0.02
Standing	0.05	0.07	0	0.84	0.02	0.02
Walking	0	0.05	0.05	0.01	0.88	0.01
Sitting	0	0.05	0	0.06	0.02	0.87
Sadek *et al.* [299]						
Jumping	0.78	0.10	0	0.10	0	0.02
Running	0.07	0.82	0.08	0	0.03	0
Bending	0	0.06	0.81	0.08	0.05	0
Standing	0	0.10	0.07	0.78	0.05	0
Walking	0.04	0.02	0.03	0.07	0.84	0
Sitting	0	0.05	0	0.06	0.02	0.87

Table 3.6 Recognition results over the i3DPost multi-view dataset [286].

Method	**Accuracy (%)**	**Error (%)**	**Recall (%)**	**Specificity (%)**	**Precision (%)**	**F-Score (%)**
Ikizler-Cinbis & Sclaroff [300]	82	18	82	96.4	82.95	82.47
Qian *et al.* [29]	76	24	76	95.20	76.62	76.31
Ahmad *et al.* [301]	87	13	87	97.40	87.40	87.20
Sadek *et al.* [299]	81.67	18.33	81.67	96.33	82.49	82.08
Proposed	100	0	100	100	100	100

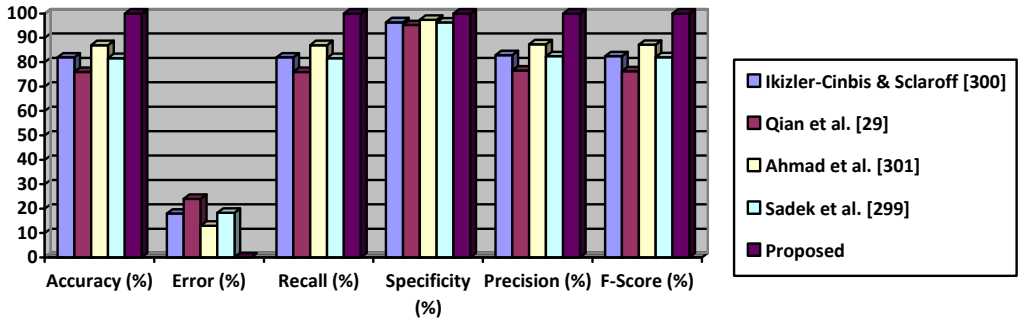


Figure 3.14 Comparison chart over the i3DPost multi-view dataset

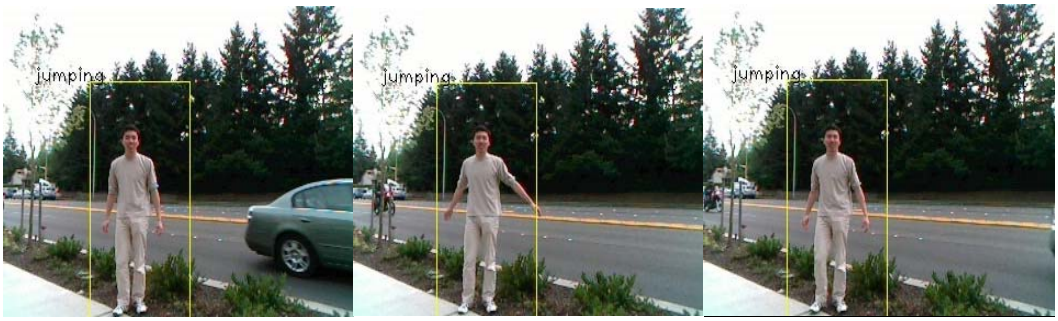
These confusion matrices, recognition results and comparison chart in Tables 3.5, 3.6 and figure 3.14 indicate that the proposed method performs better than other methods.

3.3.4. Experiment 4

In this section, we demonstrate results of the proposed method for MSR action recognition database [310]. MSR Action dataset contains 16 video sequences and has in total 63 actions: 14 hand clapping, 24 hand-waving and 25 boxing, performed by 10 subjects. Each sequence contains multiple types of actions. Some sequences contain actions performed by different people. There are both indoor and outdoor scenes. All of the video sequences are captured with clutter and moving backgrounds. Each video is of low resolution 320 x 240 and frame rate 15 frames per second. Their lengths are between 32 to 76 seconds. Qualitative recognition results are shown in Figure 3.11, which shows correct results.

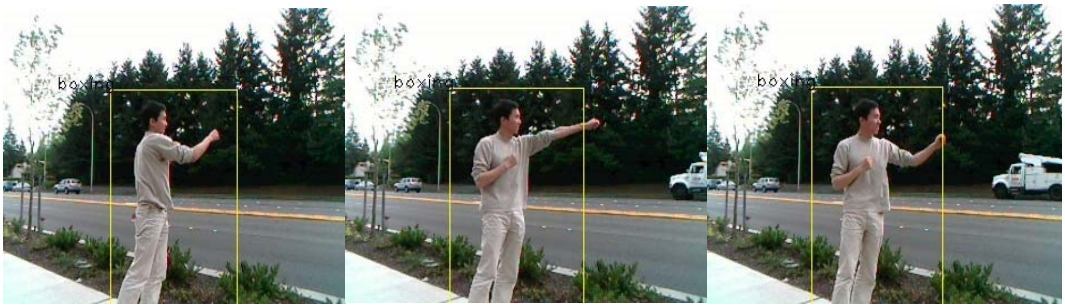
(a) Standing

(b) Handwaving



(c) Jumping

(d)Hand-clapping



(e)Boxing

Figure 3.15 Recognition of Activities with MSR action recognition database [310] (a) Standing (b) Hand-waving (c) Jumping (d) Hand-clapping (e) Boxing

From Figure 11, it can be observed that the person is performing "standing" activity at different viewing angles. From Figure 3.11, it is also clear that the proposed method is well capable of recognizing 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. Hence, one can get correct visual results by using the proposed method. It is capable of recognizing the activity at these different viewing angles correctly and the proposed method is robust towards different rotations of the activity.

Table 3.7 Confusion matrix for the proposed method over the MSR view-point action dataset

Recognized Instances → / **Total Instances ↓**	**Standing**	**Handwaving**	**Jumping**	**Handclapping**	**Boxing**
For the proposed method					
Standing	0.96	0.02	0.02	0	0
Handwaving	0	1	0	0	0
Jumping	0	0	1	0	0
Handclapping	0	0	0	1	0
Boxing	0	0	0	0	1
Qian *et al.* [29]					
Standing	0.70	0.10	0.10	0.10	0
Handwaving	0.04	0.80	0.06	0.05	0.05
Jumping	0.10	0.02	0.76	0.10	0.02
Handclapping	0.10	0.08	0.01	0.81	0
Boxing	0.05	0.12	0.07	0.02	0.74
Ikizler-Cinbis & Sclaroff [300]					
Standing	0.81	0.06	0.08	0.05	0
Handwaving	0.10	0.84	0.06	0	0
Jumping	0.14	0.06	0.78	0.01	0.01
Handclapping	0.10	0.10	0.04	0.76	0
Boxing	0	0.11	0	0.08	0.81
Sadek *et al.* [299]					
Standing	0.75	0.10	0.05	0.05	0.05
Handwaving	0.10	0.71	0.10	0.05	0.04
Jumping	0.14	0.11	0.73	0.02	0
Handclapping	0.10	0.10	0.05	0.70	0.05
Boxing	0.06	0.04	0.12	0	0.78
Ahmad *et al.* [301]					
Standing	0.88	0.06	0	0.04	0.02
Handwaving	0.03	0.95	0.02	0	0
Jumping	0.05	0.03	0.90	0.02	0

Handclapping	0.01	0.01	0.02	0.96	0
Boxing	0.02	0.03	0	0.01	0.94

Table 3.8 Recognition results over the MSR view-point action dataset [310]

Method	Accuracy (%)	Error (%)	Recall (%)	Specificity (%)	Precision (%)	F-Score (%)
Ikizler-Cinbis & Sclaroff [300]	80	20	80	95	81.30	80.64
Qian *et al.* [29]	76.20	23.80	76.20	94.05	76.90	76.55
Ahmad *et al.* [301]	92.60	7.40	92.60	98.15	92.74	92.67
Sadek *et al.* [299]	73.40	26.60	73.40	93.35	74.37	73.88
Proposed	99.20	0.80	99.20	99.80	99.22	99.21

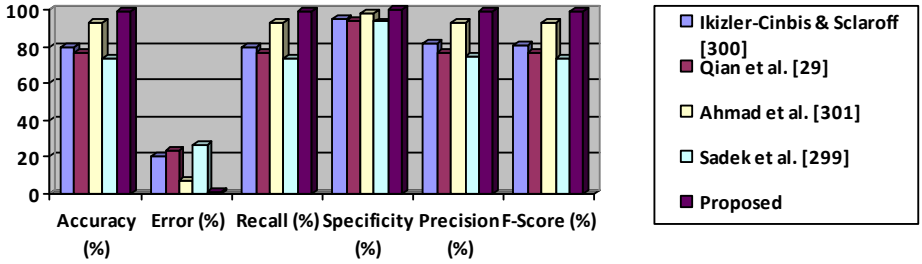
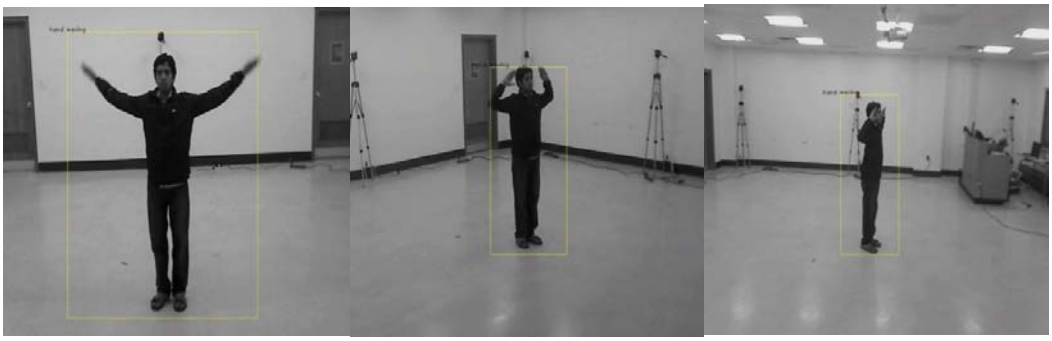


Figure 3.16 Comparison chart over the MSR view-point action dataset

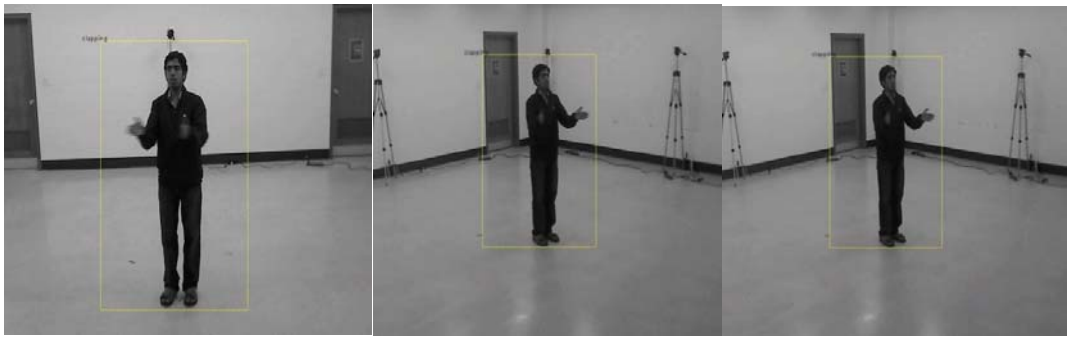
These confusion matrices, recognition results and comparison chart in Tables 3.7, 3.8 and Figure 3.16 indicate that the proposed method performs better than other methods.

3.3.5. Experiment 5

WVU multi-view human action recognition dataset [288] has been sorted based on the 8 views. This dataset includes different activities hand waving, clapping, jumping, jogging, bowling, throwing, pickup, and kicking. For each view, action sequences performed by different subjects are provided. In Fig. 3.12, we have shown activity recognition with WVU multi-view human action recognition dataset [288].



(a) Hand Waving



(b) Hand Clapping

(c) Walking

Figure 3.17 Recognition of Activities in WVU multi-view human action recognition dataset [288] (a) Hand waving (b) Hand Clapping (c) Walking

Now, quantitative results have been shown WVU multi-view human action recognition dataset [288] in Tables 3.9 - 3.10.

Table 3.9 Confusion matrix for the proposed method and other methods over the WVU action recognition dataset

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
Qian *et al.* [29]			
Hand Waving	0.72	0.28	0
Hand-clapping	0.30	0.70	0
Walking	0.30	0.01	0.69
Ikizler-Cinbis & Sclaroff [300]			
Hand Waving	0.79	0.21	0
Hand-clapping	0	0.81	0.19
Walking	0.18	0	0.82
Sadek *et al.* [299]			
Hand Waving	0.88	0.02	0.10
Hand-clapping	0.16	0.84	0

Walking	0	0.15	0.85
Ahmad *et al.* [301]			
Hand Waving	0.97	0	0.03
Hand-clapping	0.07	0.93	0
Walking	0	0.08	0.92

Table 3.10 Recognition results over the WVU action recognition dataset [288]

Method	Accuracy (%)	Error (%)	Recall (%)	Specificity (%)	Precision (%)	F-Score (%)
Ikizler-Cinbis & Sclaroff [300]	80.67	19.33	80.67	90.33	80.68	80.67
Qian *et al.* [29]	70.33	29.67	70.33	85.17	75.08	72.63
Ahmad *et al.* [301]	94.00	6.00	94.00	97.00	94.06	94.03
Sadek *et al.* [299]	85.67	14.33	85.67	92.83	85.75	85.71
Proposed	99.33	0.67	99.33	99.67	99.35	99.34

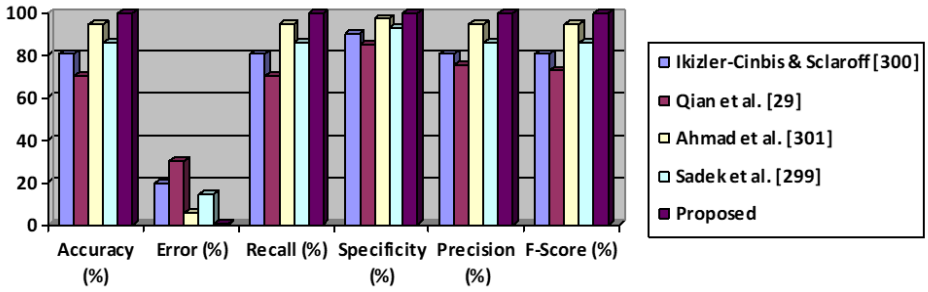


Figure 3.18 Comparison chart over the WVU action recognition dataset

Each confusion matrix shows the performance of a particular method for the chosen dataset. Comparison of the recognition results in terms of accuracy, error, recall, specificity, precision and f-score of the different method with the proposed method has been shown in Table 3.10 and Figure 3.18.

These confusion matrices and recognition results presented in Tables 3.9 and 3.10 show that the accuracy of the proposed method is better than the other existing methods.

3.4. Conclusion

In this chapter, a multi-view human activity recognition system by using combined Contour based distance signal feature, uniform rotation invariant LBP feature and motion flow based feature has been proposed. To represent each activity from multiple views or each scenario, we were used uniform rotation invariant LBP descriptor. Its rotation invariant nature provides view invariant recognition of multi-view human activities and

uniform patterns facilitate good discriminating capabilities. This system is based on three consecutive modules. These are (i) background subtraction (ii) feature extraction and (iii) classification. We also considered some sources of variability that affect human activity recognition. This variability includes the view-directional variation and phase change of different activities. The proposed approach is different from other shape-based, motion-based or combined shape and motion approaches where analysis is done at a single level. We included the features of silhouettes and original images when a person performs activity in different speed variation, change of phase variation, i.e. the starting and ending phase variations of activity. This enforces the robustness of the activity recognition. Based on the combined features, a set of HMMs was built for the mentioned activities. This approach has been performed on five multi-view human activity video datasets: Our own view point dataset, KTH action recognition dataset [285], i3DPost multi-view dataset [286], MSR view-point action dataset [310] and WVU multi-view human action recognition dataset [288]. Qualitative and quantitative experimental results demonstrate the robustness of the proposed method against different viewpoints. The proposed method has been compared with methods proposed by Qian *et al.* [29], Sadek *et al.* [299], Ikizler-Cinbis & Sclaroff [300], and Ahmad *et al.* [301], and found better than these.