Chapter 5 : MULTI-VIEW HUMAN ACTIVITY RECOGNITION SYSTEM BASED ON SPATIO-TEMPORAL TEMPLATE

In this chapter, an efficient view invariant framework for the recognition of human activities from an input video sequence is presented. The proposed framework is composed of three consecutive modules: (i) detecting and locating people by background subtraction, (ii) view invariant spatio-temporal template creation for different activities (iii) and finally template matching is performed for view invariant activity recognition. The foreground objects present in a scene are extracted by using change detection and background modeling. The view invariant templates are constructed using the motion history images and object shape information for different human activities in a video sequence. For matching the spatio-temporal templates for various activities; the moment invariants and mahalanobis distance is used. The proposed approach is tested successfully in our own viewpoint dataset, KTH action recognition dataset [163], i3DPost multi-view dataset [164], MSR view-point action dataset [165], Video-Web Multi-view dataset [166], and WVU multi-view human acti`on recognition dataset [167]. From the experimental results and analysis over the chosen datasets it is observed that the proposed framework is robust, flexible and efficient with respect to multiple views activity recognition, scale and phase variations.

5.1. Introduction

Human activity recognition is a popular area of research in the field of computer vision. It is the basis of many applications such as security, surveillance, clinical applications, biomechanical applications, human robot interaction, entertainment, education, training, digital libraries, video or image annotations, video conferencing and model based coding [147-148]. Recognition of activities provides important cues for human behavior analysis techniques. Although a large amount of work has been performed on activity recognition in the last few years yet still it is an open and challenging problem. The various issues and challenges involved in automatic human recognition from video sequences are as follows:

- The trajectory of activities from different viewing directions is different and some of the body parts (part of hand, lower part of leg, part of body, etc.) are occluded due to view changes.
- The other common issues include fixed or moving cameras, scenes having moving or clutter backgrounds, changes in light and view-point, variations in scale, starting and ending state, variations in appearance of individuals and cloths of human etc. These issues and situations make the human activity recognition a challenging task.
- Human activities are performed in real 3D environment and cameras only capture the 2D projection of the real scene. Therefore, visual analyses of activities carried out in the image plane are only a projection of the real activities. This projection of the activities depends on the viewpoint and do not contain full information about the performed activities.

Most of the work on activity recognition are view dependent and deal with recognition from one fixed view. Recognizing human activities from multiple views has been a challenging task for researchers around the globe and needs a lot of improvement. To account for these problems, many activity recognition systems have been developed [31, 149-153] and various surveys and frameworks can be found in literature [24, 44, 154-156]. 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 for activity recognition. Vision based activity recognition methods use the spatial or temporal structure of an activity in order to recognize it. A recent survey on vision-based action representation and recognition methods can be found in [44]. Machine learning based and template based methods [44] 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. Holte *et al.* [79] proposed a machine learning based approach to detect motion of the actors by computing optical flow in video data. The video data was captured by a multi-view camera setup for combining optical flow into 3D motion vector fields for human recognition. The authors have used 3D Motion Context (3D-MC) and Harmonic Motion Context (HMC) to represent the 3D motion vector fields efficiently and in a view invariant manner. However, the HMC descriptor stabilizes more slowly at a higher number of views, and also causes confusion between bend and sit-stand-up, walk and run, and the combined actions like *walk-sit* and *run-jump-walk* are confused. Junejo *et al.* [80] have proposed a self-similarity based descriptor for view independent human action recognition. In this paper, an action descriptor has been developed by the authors that capture the structure of temporal similarities and dissimilarities within an action sequence. The drawbacks of machine learning based methods are the long training time, slow operation, constrained accuracy and difficulty to include a new activity. Template based methods are good options for activity recognition in video and can be easily used because of their simplicity and robustness. Weinland *et al.* [44] provided a good survey of template matching based human activity recognition. The template matching based techniques can broadly be classified into three categories: body template based methods, feature template based methods and image template based methods. Body template based methods [157] represent the spatial structure of activities with respect to the human body. In each frame of the observed video sequence, the posture of a human body is reconstructed from a variety of available image features. The activity recognition is performed based on these posture estimations. This is an intuitive and biologically-plausible approach for activity recognition and supported by psychophysical work on visual interpretation of biological motion. However, in body model based representations, the resulting interest regions are linked to certain body-parts or even image coordinates. This imposes certain restrictions on recognition of different activities. In feature template based methods [50, 158], activities are recognized based on the statistics of sparse features 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 these approaches is that they neither rely on explicit body part labeling, nor on explicit human detection and localization. The image template based methods [45, 58, 62, 66, 159-161] are simple than the above described methods and can be computed efficiently. Motion history images (MHI) and motion energy images (MEI) can be used to determine the location and type of activities in the scene. In an outdoor environment, where variations in lighting conditions and change in background produce noise, a robust background modeling is required. MHI and MEI prove to be good solutions of this problem. The approach proposed by Bobick *et al.* [45], used motion templates for recognizing the activities in a specific environment of aerobic exercise. They used MEI for obtaining segmented foreground and MHI for obtaining motion information in a view-specific environment. However, their technique does not give good activity recognition accuracy in an outdoor environment. Moreover, it is capable of only identifying one activity in the scene with one actor at a time. Our work is an extension of the work done by Bobick *et al.* [45]. The proposed method presents a spatio-temporal template based activity recognition. This approach considers the shape information along with the motion history for performing an activity. For obtaining the good foreground segmentation a robust change detection based background model is constructed. The technique can recognize the static activities like standing and sleeping as well as dynamic activities like walking, jogging, etc. In the proposed approach, covariance based matching is applied to recognize static activities and moment invariants [45, 162] are used to recognize dynamic activities. This technique can recognize the activities of no motion such as standing and sleeping, along with those with motion such as walking, jumping.

To demonstrate the effectiveness and robustness of the proposed method, we have conducted our experiments on our own viewpoint dataset and five publicly available human activity recognition video datasets, namely, KTH action recognition dataset [163], i3DPost multi-view dataset [164], MSR view-point action dataset [165], VideoWeb Multi-view dataset [166] , and WVU multi-view human action recognition dataset [167]. The proposed system has also been compared with six existing human activity recognition methods proposed by Qian *et al.* [168], Bobick *et al.* [45], Ikizler-Cinbis & Sclaroff [169], Holte *et al.* [79], Junejo *et al.* [80], and Ahmad *et al.* [52]. To compare the proposed method with the six mentioned standard methods, the confusion matrix and recognition accuracy (in percentage) evaluation parameters have been used. Experimental results on the above mentioned six datasets illustrate the efficiency and the effectiveness of the proposed method.

The rest of the chapter is organized as follows: section 5.2 describes the concept of motion history images (MHI), section 5.3 gives the detailed methodology of the proposed technique used for human activity recognition, section 5.4 shows various experimental results and section 5.5 gives the conclusions.

5.2. Motion History Images (MHI)

The motion history image (MHI) captures the motion of a single or multiple object silhouettes over a period of time. The intensity values in the MHI show the time of the pixels where last motion happened or the presence of the object. Pixel intensity H_t, can be considered as a function of the temporal history of motion at that point of time and MHIs can be calculated for the time interval 0-t for each motion extracted from video frames. Although MHI is a representation of the history of pixel-level changes, yet the advantage is that only one previous frame needs to be stored. A simple replacement and decay operator for MHIs:

$$H_\tau(u,v,t)=\left\{\begin{matrix}\tau & if\ D(u,v,t)=1 \\ \max(0,H_\tau(u,v,t-1)-1) & otherwise\end{matrix}\right\} \tag{5.1}$$

where $H_\tau\,(u,v,t)$ is a motion history template image H_τ at position (u,v) and time t and $D(u,v,t)$ is its corresponding video frame.

Object silhouette motion history template images are shown in Fig. 5.1. for activities like boxing, handclapping, hand waving, running and walking.

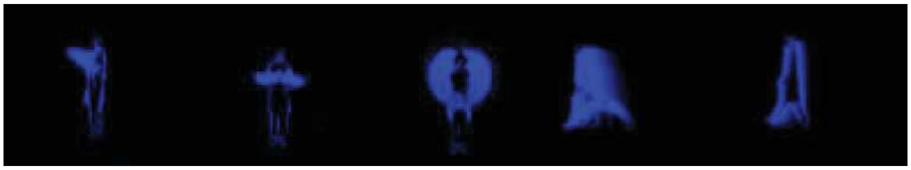

(a) (b) (c) (d) (e)

Figure 5.1: Object silhouette motion history in a video frame for activities (a) Boxing (b) Handclapping (c) Hand waving (d) Running and (e) Walking

As the object moves it leaves behind a motion history of its movements. With the passage of time the old motion histories of object should be discarded to capture the new motion patterns. So the silhouette MHI needs to be updated.

5.3. The Proposed Method

The proposed multi-view human activity recognition system is shown in Fig. 5.2. In our system, the foreground is extracted by using frame difference and background modeling.

From the foreground image, templates are constructed using the motion history images and object shape information for different human activities in a video sequence. These templates are matched using 7 Hu moment invariants [162] and mahalanobis distance.

The steps of the proposed method are as follows:

(1) Preprocessing

(2) Background Subtraction

(3) Activity template creation for different activities

(4) Template matching and activity recognition

The block diagram of the proposed method is shown in Fig. 5.2.

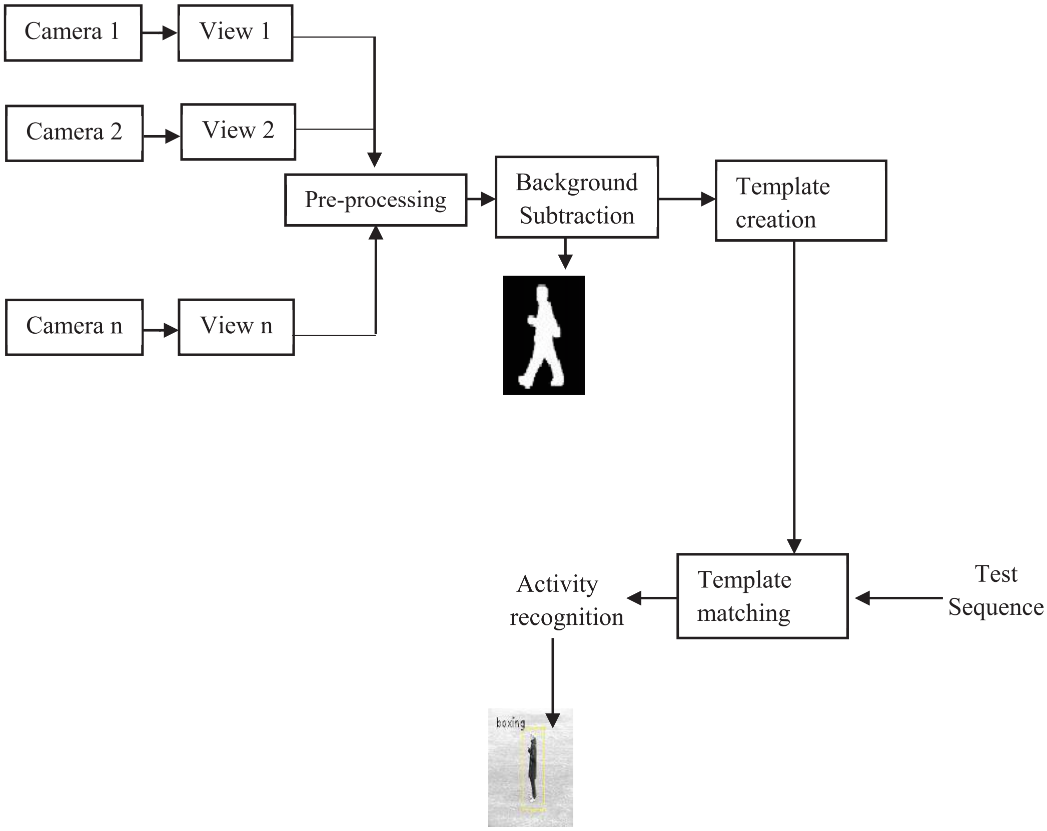


Figure 5.2: Block diagram of the proposed method

5.3.1. Input video

Input training video which is a sequence of frames, where number of frames can vary from *1….n*. Therefore,

$$V = \left(F_i\right)_{i=1}^{n} \tag{5.2}$$

where the video, *V*, is represented as a sequence of frames F_i and *i* ranges from *1 to n. n* is the total number of frames in the video.

5.3.2. Preprocessing

The video frames are preprocessed to normalize color and size because different videos can exist in different color formats and frame size. The normalized video is represented as

$$|V| = \left(|F_i|\right)_{i=1}^{n} \tag{5.3}$$

5.3.3. Background Subtraction

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

A. 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 $V_{th,WD}$ to each frame difference, this method uses the fast Euler number computation technique [126] to automatically determine $V_{th,WD}$ from the video frame. The stable Euler number technique is one of the most effective algorithms for determining thresholds for change differences. However, its high computational complexity has always precluded its employment in real-time applications. A fast Euler number computation method was proposed in [126] to overcome the high computational complexity of the stable Euler number method.

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:

$$E(i)=\frac{1}{4}[(q_1(i)-q_3(i)-2q_d(i))] \qquad (5.4)$$

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 [126].

The frame differences $WD_n(i,j)$ for respective frames are computed as:

for every pixel location (i, j) in the co-ordinate of frame

$$WD_n(i,j)=\begin{Bmatrix} 1 & if & |f_n(i,j)-f_{n-1}(i,j)|>V_{th,WD} \\ 0 & otherwise & \end{Bmatrix} \qquad (5.5)$$

B. 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:

$$WD_n(i,j)=\begin{Bmatrix} 1 & if & |f_n(i,j)-f_{n-1}(i,j)|\geq V_{th,WD} \\ 0, & if & |f_n(i,j)-f_{n-1}(i,j)|<V_{th,WD} \end{Bmatrix} \qquad (5.6)$$

$V_{th,WD}$ is a threshold determined automatically from the video frame by the fast Euler number computation method as explained in [126].

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:

$$S_n(i,j)=\begin{cases} S_n(i,j)+1 & \text{if } \ WD_n(i,j)=0 \\ 0 & \text{if } \ WD_n(i,j)=1 \end{cases} \tag{5.7}$$

$$\mu_n(i,j)=\begin{cases} f_n(i,j) & \text{if } \ S_n(i,j)\geq N_f \\ \mathit{Undefined} & \text{if } \ S_n(i,j)< N_f \end{cases} \tag{5.8}$$

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.

if $$|f_n(i,j)-\mu_n(i,j)|<2\sigma_n(i,j) \tag{5.9}$$

then $$\begin{cases} \mu_n(i,j)=\chi\mu_{n-1}(i,j)+(1-\chi)f_n(i,j) \\ \sigma_n^2(i,j)=\chi\sigma_{n-1}^2(i,j)+(1-\chi)(f_n(i,j)-\mu_n(i,j))^2 \end{cases} \tag{5.10}$$

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:

$$BD_n(i,j)=\left|f_n(i,j)-\mu_n(i,j)\right| \tag{5.11}$$

$$BDM_n(i,j)=\begin{cases}1, & BD_n(i,j)\geq V_{th,WD}\\ 0, & BD_n(i,j)<V_{th,WD}\end{cases} \tag{5.12}$$

where $BD_n(i,j)$ is the background difference and $BDM_n(i,j)$ 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 [126].

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

5.3.4. Activity Template Creation:

We construct the spatio-temporal templates for every activity like walking, standing, sleeping, jumping, bending etc. are modeled using MHIs collections on the segmented video frames obtained from section 5.3.3 (see Fig. 5.3 & 5.4). The MHIs are used for creating spatio-temporal templates for each activity in an activity set A (a_1, a_2, a_3…a_k) and can be constructed in the following manner:

For each a_i in A do

a) Initialize the motion history image MHI for activity a_i with the initial pose of the actor in order to include spatial information in the MHI.

b) Measure the minimum and maximum durations, $\tau_{\min}$ and $\tau_{\max}$ that a movement may take.

c) At each time, a new MHI, $H_{\tau}(u,v,t)$ is computed setting $\tau = \tau_{\max}$, where $\tau_{\max}$ is the longest time window we want the system to consider. We choose

$$\Delta\tau = \frac{(\tau_{\max} - \tau_{\min})}{(n-1)} \qquad (5.13)$$

where $\Delta\tau$ is the time difference, and n is the number of temporal integration windows to be considered. A simple thresholding of MHI values less than $(\tau - \Delta\tau)$ generates $H_{\tau-\Delta\tau}$ as below:

$$H_{\tau-\Delta\tau}(u,v,t) = \begin{cases} (H_{\tau}(u,v,t) - \Delta\tau) & H_{\tau}(u,v,t) > \Delta\tau \\ 0 & otherwise \end{cases} \Bigg\} \qquad (5.14)$$

where $H_{\tau-\Delta\tau}$ defines the MHI values for time duration $\tau - \Delta\tau$. We store the MHI values for each activity in order to match them against the actual actions.

d) The direction of motion in the video is computed using the gradient orientation. Gradients of the MHI can be calculated by convolution with separable Sobel filters in the X and Y directions which yields $\mathrm{F_x}(\mathrm{x,y})$ and $\mathrm{F_y}(\mathrm{x,y})$. Where, $\mathrm{F_x}(\mathrm{x,y})$ and $\mathrm{F_y}(\mathrm{x,y})$ represent the derivatives in x and y directions. At each pixel, gradient orientation, $\phi(\mathrm{x,y})$ can be calculated as follows:

$$\phi(\mathrm{x,y}) = \arctan\frac{\mathrm{F_y}(\mathrm{x,y})}{\mathrm{F_x}(\mathrm{x,y})} \qquad (5.15)$$

The gradient orientation is calculated only for the pixels inside the MHI, where the intensity values for these pixels are non-zero.

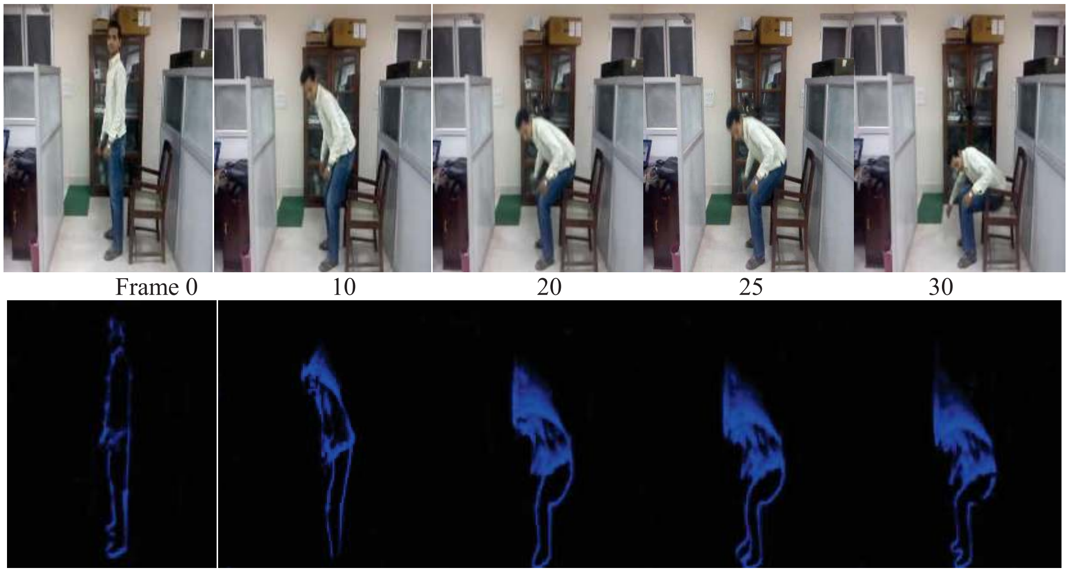


Figure 5.3: The spatio-temporal MHI based template formation of the sitting activity in proposed technique. Background segmentation is achieved using change detection approach described in section 5.3.3.

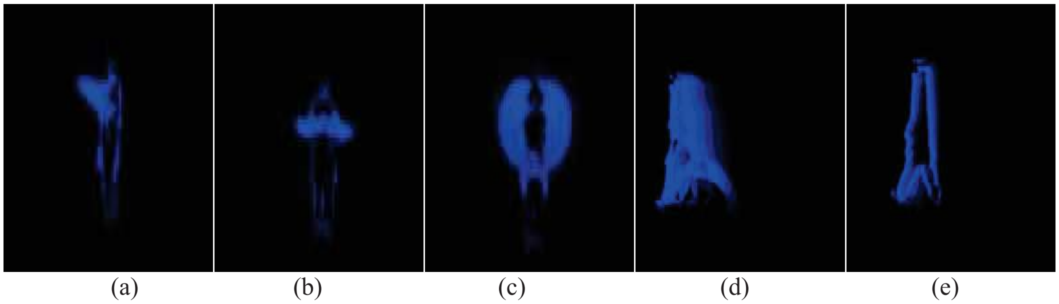


Figure 5.4: MHIs of different activities by the proposed method; (a) boxing, (b) clapping, (c) hand waving, (d) running, and (e) walking.

5.3.5. Template Matching and Activity Recognition

The technique used for matching the spatio-temporal templates to recognize the activities is rotation, scale, and translation invariant [45]. The training of actions is performed considering different views of activity performance. For each view of each action a statistical model of the moments using variance and covariance is generated for MHIs. The 7 moment invariants [162] are used as activity descriptors. To recognize an input action, a mahalanobis distance is calculated between the moment description of the input and each of the known actions using equation 5.16. The distance matrix so obtained is analyzed in terms of separation distances for different actions.

$$\text{mahal}(p) = (p - m)^T K^{-1} (p - m) \tag{5.16}$$

where, p is a moment feature vector, m is the mean value of vector p and K^{-1} represents the inverse covariance matrix of the feature vectors.

If more than one match is found for an activity then the match with the smaller mahalanobis distance is chosen.

The algorithm for human activity recognition is summarized as given below:

Algorithm for Human Activity Recognition

Input: Sequence of frames; Output: Detected multi-view Human Activity

1. Load the sequence of image frames (see Eq. (5.3)).

2. Apply background subtraction using frame differences and background model creation.

2a. Compute frame differences $WD_n(i, j)$ for respective frames using change detection (see Eq. (5.5)).

2b. Create background model by using frame differencing, background registration, background difference, and background difference mask (see Eq. 5.6-5.12).

3. Construct the spatio-temporal activity templates using MHIs collections of each activity performed on the segmented video frames obtained from step 2 (see Eq. (5.13-5.15)).

4. To match the spatio-temporal templates for various activities created in step 3 proceed as follows.

4a. Use similar moment invariants for as scale, translation and rotation as used in [45].

4b. Calculate Mahalanobis distance for matching the input action with the stored templates trained in step 3 using Eq. (5.16).

5. If more than one match is found for an activity then chose the match with the smaller mahalanobis distance.

5.4. Experimental Results

In this section, we present results for our own viewpoint dataset and five publicly available human activity recognition video datasets [163-167]. Videos in these datasets have been captured at different rotation angle for multiple viewpoints. The experiments have been performed in Open CV 2.4.9 environment on an Intel® Core™ i3 2.53 GHz machine with 4 GB RAM.

In our implementation, first we take the videos and apply background subtraction according to the method described in Section 5.3.3. After that, activity templates are created for different activities using MHI. Lastly, Template matching is performed. Six case studies of our own viewpoint dataset, KTH action recognition dataset [163], i3DPost multi-view dataset [164], MSR view-point action dataset [165], VideoWeb Multi-view dataset [166], and WVU multi-view human action recognition dataset [167] are discussed here one by one. In all case studies, we have illustrated and tested the proposed method with the other standard methods proposed by Qian *et al.* [168], Bobick *et al.* [45], Ikizler-Cinbis & Sclaroff [169], Holte *et al.* [79], Junejo *et al.* [80], and Ahmad *et al.* [52]. 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:

$$CRR = \frac{N_c}{N_a} \times 100 \text{ (in percentage)} \tag{5.17}$$

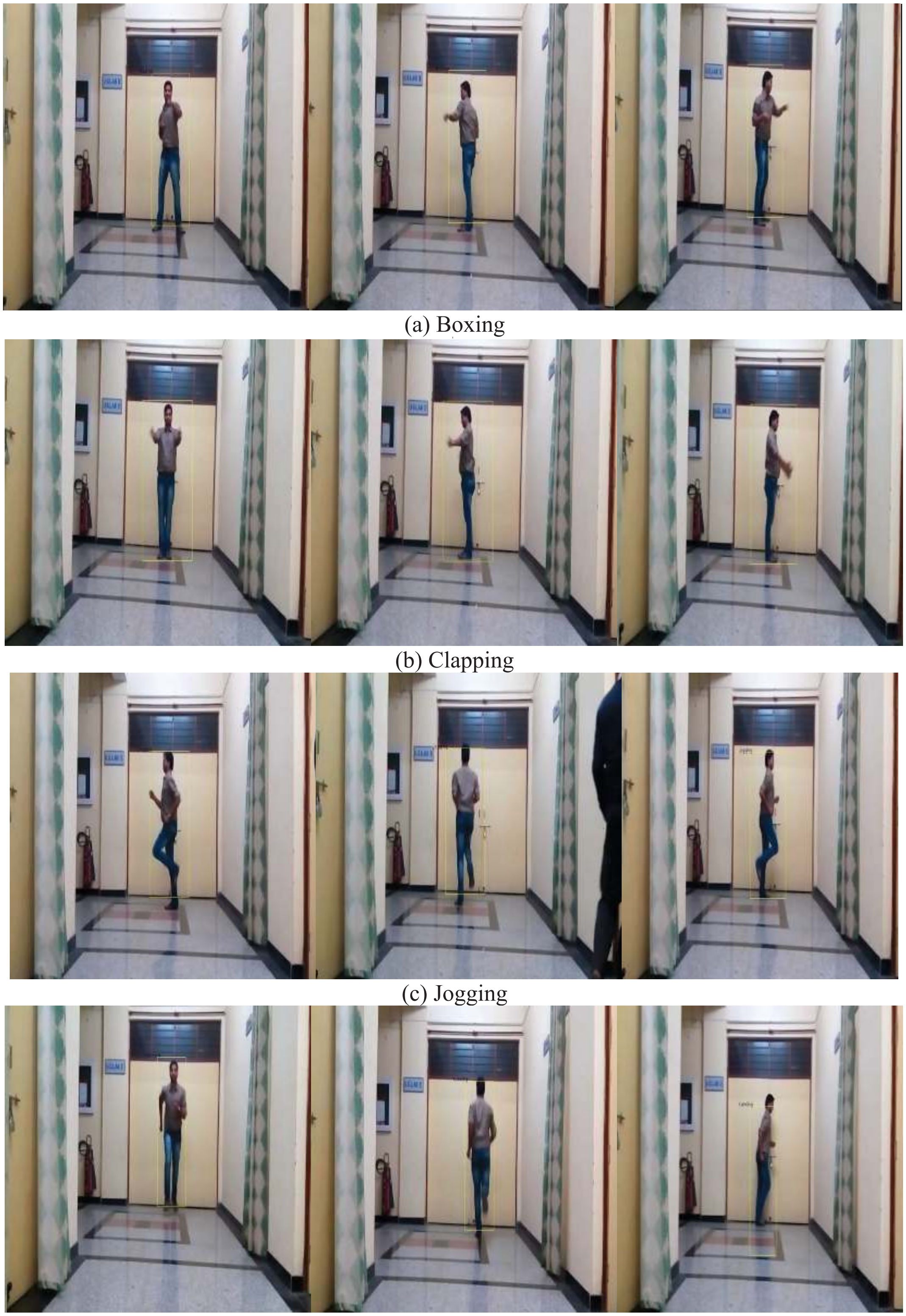
where Nc is the total number of correct recognition sequences while Na is the number of total activity sequences.

Experiment 1

In Fig. 5.5, we have shown results on our own created database. This database contains video of static human activities, namely sitting and six dynamic activities, namely walking, running, jogging, boxing, clapping and jogging in the different view direction. These videos are taken in a real indoor 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 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 hence, 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 is robust towards different rotations of the activity.

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.* [168], Bobick *et al.* [45], Ikizler-Cinbis & Sclaroff [169], Holte *et al.* [79], Junejo *et al.* [80], and Ahmad *et al.* [52].

The confusion matrix of different activities for different methods have been shown in Table 5.1. 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 has been shown in Table 5.2 (calculate using Eq. 5.17). 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.

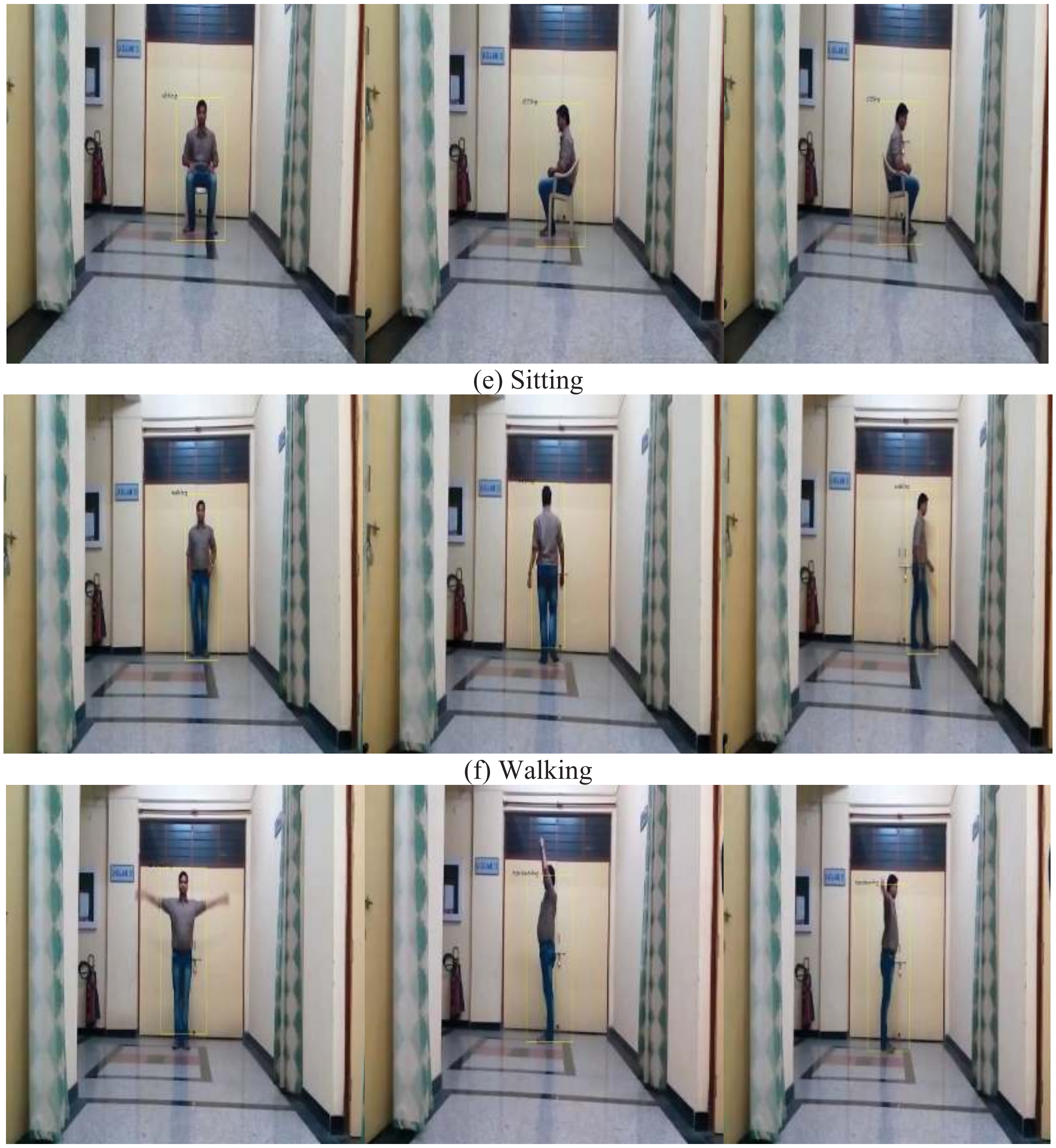


(a) Boxing

(b) Clapping

(c) Jogging

(d) Running



(e) Sitting

(f) Walking

(g) Hand-waving

Figure 5.5: 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.

Table 5.1: Confusion matrices for the proposed and other methods

Recognized Instances → **Total Instances ↓**	**Boxing**	**Clapping**	**Jogging**	**Running**	**Sitting**	**Walking**	**Hand-waving**
Proposed method							
Boxing	0.98	0.02	0	0	0	0	0
Clapping	0	1	0	0	0	0	0
Jogging	0	0	0.99	0.01	0	0	0
Running	0	0	0.02	0.98	0	0	0
Sitting	0	0	0	0	0.99	0	0.01
Walking	0	0	0	0	0	1	0
Hand-waving	0	0	0	0	0	0	1
Qian *et al.* [168]							
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	0.05	0.45	0.10	0.18	0.22	0
Running	0.03	0.33	0.03	0.58	0.03	0	0.22
Sitting	0.05	0.15	0.21	0	0.44	0.15	0
Walking	0	0	0	0	0.23	0.77	0
Hand-waving	0	0	0	0	0	0	0.67
Ikizler-Cinbis and Sclaroff [169]							
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	0.01	0	0	0.04
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
Bobick *et al.* [45]							
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
Ahmad *et al.* [52]							
Boxing	0.70	0.15	0.11	0.02	0	0	0.02
Clapping	0.22	0.66	0	0.08	0.04	0	0
Jogging	0.10	0.05	0.70	0	0.05	0.08	0.02
Running	0.18	0.06	0	0.75	0.01	0	0

Sitting	0.12	0.11	0	0	0.72	0	0.05
Walking	0	0.11	0.03	0.03	0	0.68	0.15
Handwaving	0	0	0.15	0.07	0.03	0.05	0.70
Holte e*t al.* [79]							
Boxing	0.81	0.05	0	0.10	0	0.04	0
Clapping	0.05	0.84	0.05	0	0.04	0	0.02
Jogging	0.10	0	0.80	0	0	0.10	0
Running	0.12	0	0	0.83	0.05	0	0
Sitting	0.08	0	0.05	0	0.81	0	0.06
Walking	0.10	0	0.03	0.02	0	0.85	0
Handwaving	0.12	0.01	0	0	0.04	0	0.83
Junejo e*t al.* [80]							
Boxing	0.91	0.04	0	0.03	0	0	0.02
Clapping	0.05	0.89	0	0.05	0.01	0	0
Jogging	0	0.05	0.92	0	0	0.03	0
Running	0	0	0.05	0.91	0.04	0	0
Sitting	0.04	0	0	0	0.88	0.08	0
Walking	0	0	0	0	0	0.90	0.10
Handwaving	0.05	0	0.05	0	0.02	0	0.88

Table 5.2: Recognition results over our own action recognition dataset

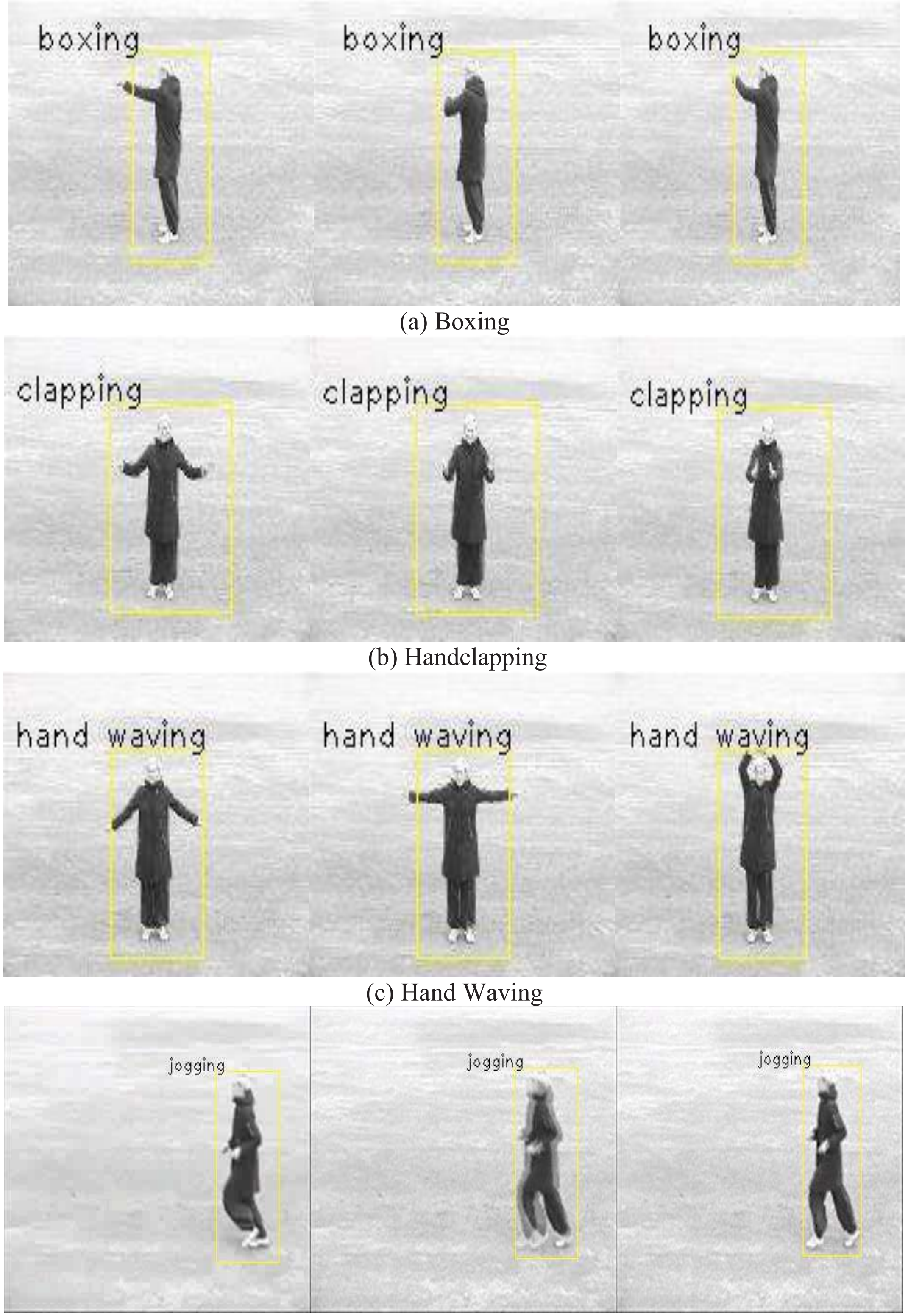
Method	**Accuracy (%)**
Qian *et al.* [168]	55.14
Ikizler-Cinbis & Sclaroff [169]	68.00
Bobick *et al.* [45]	42.71
Ahmad *et al.* [52]	74.14
Holte *et al.* [79]	82.42
Junejo *et al.* [80]	89.85
Proposed Method	99.14

Experiment 2

In this section, we demonstrate results of the proposed method for KTHDB action recognition database [163].The KTHDB is one of the largest databases with sequences of human actions taken over different scenarios [163]. This dataset contains six types of human actions (walking, jogging, running, boxing, hand waving, and hand clapping) performed several times by 25 people in six different scenarios. The database contains 2391 sequences. The image sequences have the spatial resolution of 160 * 120 pixels and have a length of six seconds in average.

In Fig.5.6, we have shown activity recognition with standard KTH database [163]. This database includes six activities like boxing, handclapping, hand-waving, jogging, running and walking. For this database also, the proposed method performs well. Moreover, this database not only contains activities involving leg motion (like jogging, running and walking) but it also contains activities involving hand motion (like boxing, handclapping and hand-waving). The most confusion occurs between jogging and running as well as jogging and walking, although it varies at different scenarios but proposed method handles these scenarios very easily. Quantitative results have been shown for KTHDB dataset [163] in Tables 5.3 and 5.4.

From these confusion matrices and recognition results in Tables 5.3, 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 correct recognition rate for this purpose which are far better in case of the proposed method in Table 5.3. Comparison of recognition accuracy of different method with the proposed method has been shown in Table 5.4 (calculate using Eq. 5.17). It shows that performance of the proposed method is better than other methods.

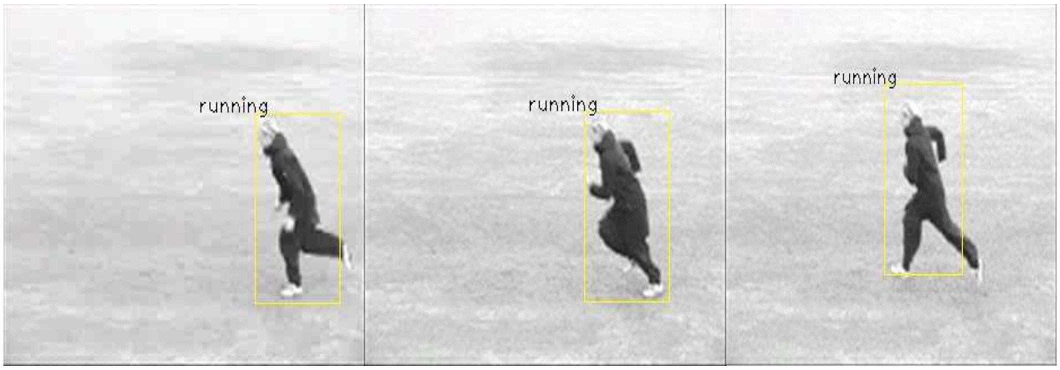


(a) Boxing

(b) Handclapping

(c) Hand Waving

(d) Jogging



(e) Running

Figure 5.6: Recognition of Activities in KTH database [163] (a) Boxing (b) Handclapping (c) Hand Waving (d) Jogging (e) Running.

Table 5.3: Confusion matrix for the proposed method and other method**s**

Recognized Instances → **Total Instances ↓**	**Boxing**	**Hand-clapping**	**Jogging**	**Hand-waving**	**Running**
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
Qian *et al.* [168]					
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
Ikizler-Cinbis & Sclaroff [169]					
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
Bobick *et al.* [45]					
Boxing	0.85	0.05	0.09	0	0.01
Hand-clapping	0.05	0.85	0.05	0.05	0
Jogging	0.10	0.10	0.80	0	0
Hand-waving	0.14	0.10	0	0.75	0.01
Running	0.04	0.10	0.08	0.03	0.75
Ahmad *et al.* [52]					
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
Holte e*t al.* [79]					
Boxing	0.91	0.03	0	0.06	0
Hand-clapping	0.05	0.89	0.02	0	0.04
Jogging	0.05	0	0.93	0.02	0
Hand-waving	0	0.10	0	0.88	0.02
Running	0.04	0	0.04	0.01	0.91
Junejo *et al.* [80]					
Boxing	0.95	0	0.03	0	0.02
Hand-clapping	0	0.96	0	0.04	0
Jogging	0	0.05	0.93	0	0.02
Hand-waving	0.05	0	0	0.95	0
Running	0	0.06	0.02	0	0.92

Table 5.4: Recognition results over the KTH action recognition dataset [163]

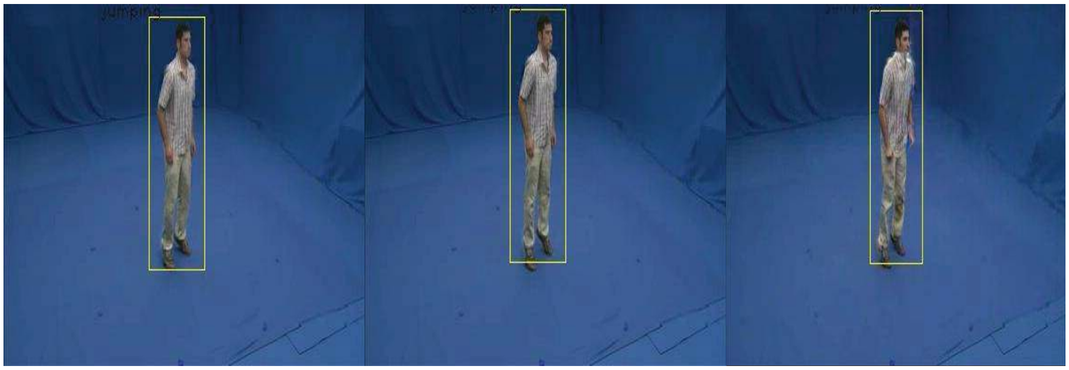
Method	**Accuracy (%)**
Qian *et al.* [168]	80.20
Ikizler-Cinbis & Sclaroff [169]	78.40
Bobick *et al.* [45]	80.00
Ahmad *et al.* [52]	88.60
Holte *et al.* [79]	90.40
Junejo *et al.* [80]	94.20
Proposed Method	100

Experiment 3

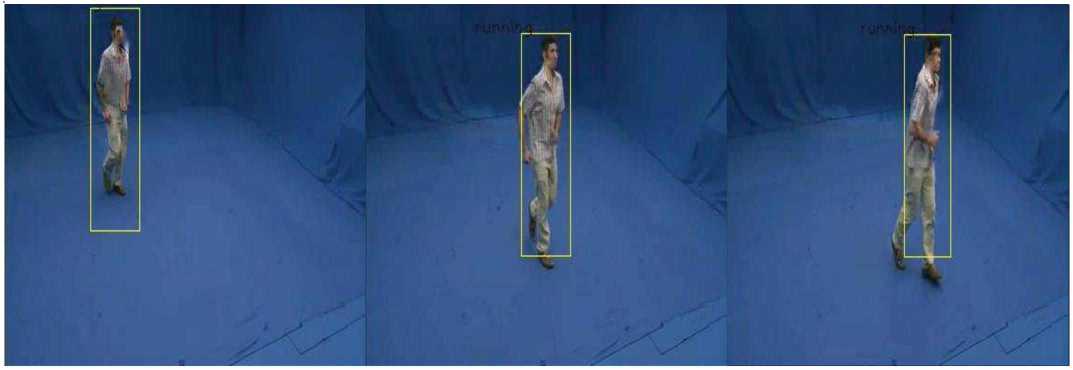
Now, we have selected i3DPost dataset which is a multi-view dataset [164] for view-invariant human activity recognition. In this dataset,8 people performing 13 actions (walking, running, jumping, bending, hand-waving, jumping in place, sitting-stand up, running-falling, walking-sitting, running-jumping-walking, 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 [164], 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 Fig. 5.7 which shows correct results.

In Fig.5.7, 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, we present quantitative results for i3DPost multi-view dataset [164].

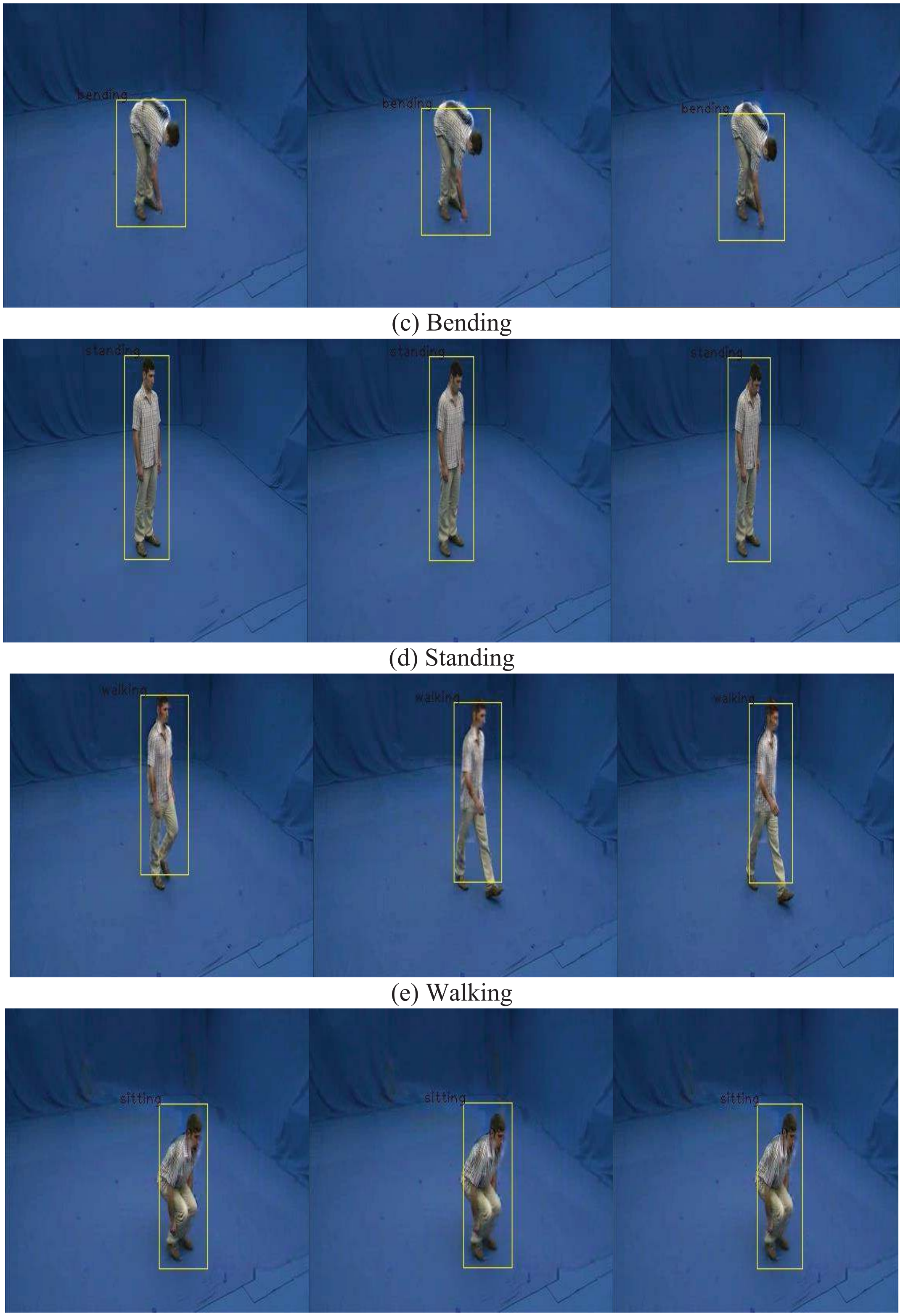
These confusion matrices and recognition results in Tables 5.5 & 5.6 indicate that the proposed method performs better than other methods.



(a) Jumping



(b) Running



(c) Bending

(d) Standing

(e) Walking

(f) Sitting

Figure 5.7: Recognition of Activities in i3DPost multi-view dataset [164] (a) Jumping (b) Running (c) Bending (d) Standing (e) Walking (f) Sitting

Table 5.5: Confusion matrix for the proposed method and other methods

Recognized Instances → Total Instances ↓	**Jumping**	**Running**	**Bending**	**Standing**	**Walking**	**Sitting**
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
Qian *et al.* [168]						
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
Ikizler-Cinbis & Sclaroff [169]						
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
Bobick *et al.* [45]						
Jumping	0.75	0.10	0.03	0.10	0	0.02
Running	0.07	0.78	0.12	0	0.03	0
Bending	0	0.10	0.85	0	0.05	0
Standing	0	0.10	0.07	0.80	0.01	0.02
Walking	0.04	0.02	0.07	0.07	0.80	0
Sitting	0.12	0	0	0.04	0	0.84
Ahmad *et al.* [52]						
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
Holte *et al.* [79]						
Jumping	0.85	0.10	0	0	0.05	0
Running	0.05	0.88	0	0.05	0	0.02
Bending	0.04	0	0.86	0	0.10	0
Standing	0.06	0	0	0.88	0	0.06
Walking	0	0.10	0	0.03	0.87	0

Sitting	0.07	0	0.05	0	0.03	0.85
Junejo *et al.* [80]						
Jumping	0.92	0.04	0	0	0	0.04
Running	0.05	0.93	0	0.02	0	0
Bending	0	0	0.89	0	0.08	0.03
Standing	0.03	0.05	0	0.92	0	0
Walking	0	0	0.06	0	0.91	0.03
Sitting	0.05	0.06	0	0	0	0.89

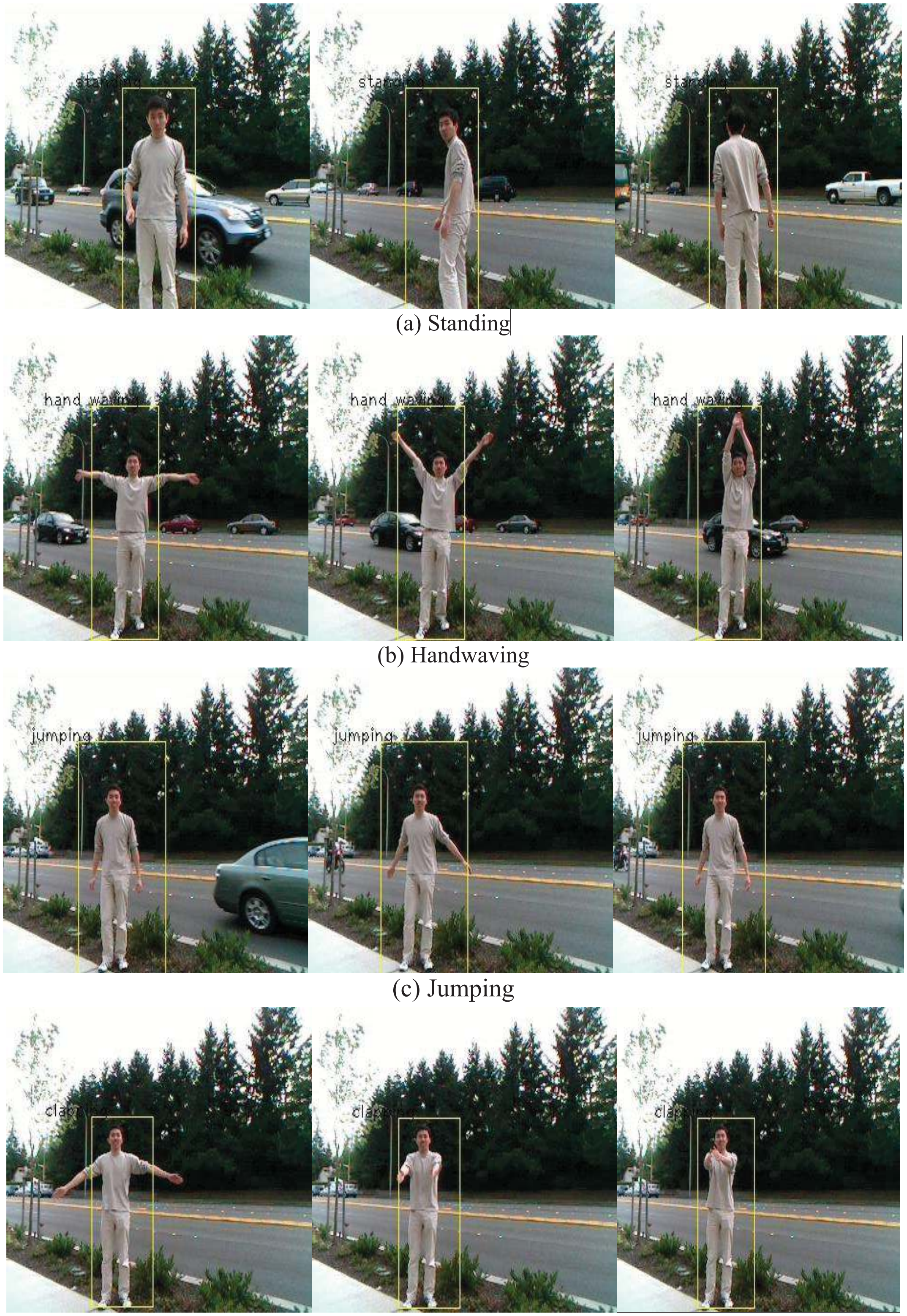
Table 5.6: Recognition results over the i3DPost multi-view dataset [164]

Method	**Accuracy (%)**
Qian *et al.* [168]	76
Ikizler-Cinbis & Sclaroff [169]	82
Bobick *et al.* [45]	80.33
Ahmad *et al.* [52]	87
Holte *et al.* [79]	86.50
Junejo *et al.* [80]	91
Proposed Method	100

Experiment 4

In this section, we demonstrate results of the proposed method for MSR action recognition database [165].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 Fig. 5.8.

From Fig. 5.8, it can be observed that the person is performing "standing" activity at different viewing angles. From Fig.5.8, 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.

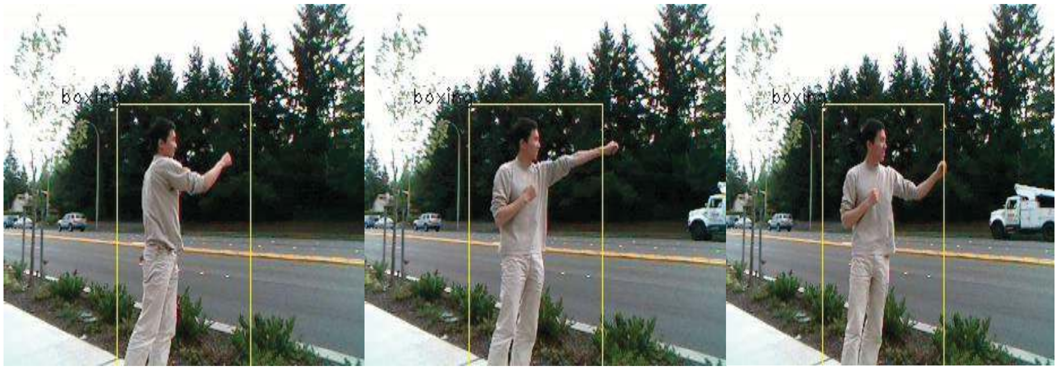


(a) Standing

(b) Handwaving

(c) Jumping

(d)Hand-clapping



(e)Boxing

Figure 5.8: Recognition of Activities with MSR action recognition database [165] (a) Standing (b) Hand-waving (c) Jumping (d) Hand-clapping (e) Boxing

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.

These confusion matrices and recognition results in Tables 5.7 & 5.8 indicate that the proposed method performs better than other methods.

Table 5.7: Confusion matrix for the proposed method and other methods

Recognized Instances → **Total Instances** ↓	**Standing**	**Handwaving**	**Jumping**	**Handclapping**	**Boxing**
Proposed method					
Standing	1	0	0	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.* [168]					
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 [169]					
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
Bobick *et al.* [45]					
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.* [52]					
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
Holte *et al.* [79]					
Standing	0.88	0.06	0	0.04	0.02
Handwaving	0.05	0.90	0.05	0	0
Jumping	0.07	0	0.91	0.02	0
Handclapping	0	0.08	0	0.90	0.02
Boxing	0.04	0	0.04	0	0.92
Junejo *et al.* [80]					
Standing	0.95	0	0.03	0.02	0
Handwaving	0.04	0.93	0	0	0.03
Jumping	0	0	0.96	0.04	0
Handclapping	0.05	0.01	0	0.94	0
Boxing	0	0	0.05	0	0.95

Table 5.8: Recognition results over the MSR view-point action dataset [165]

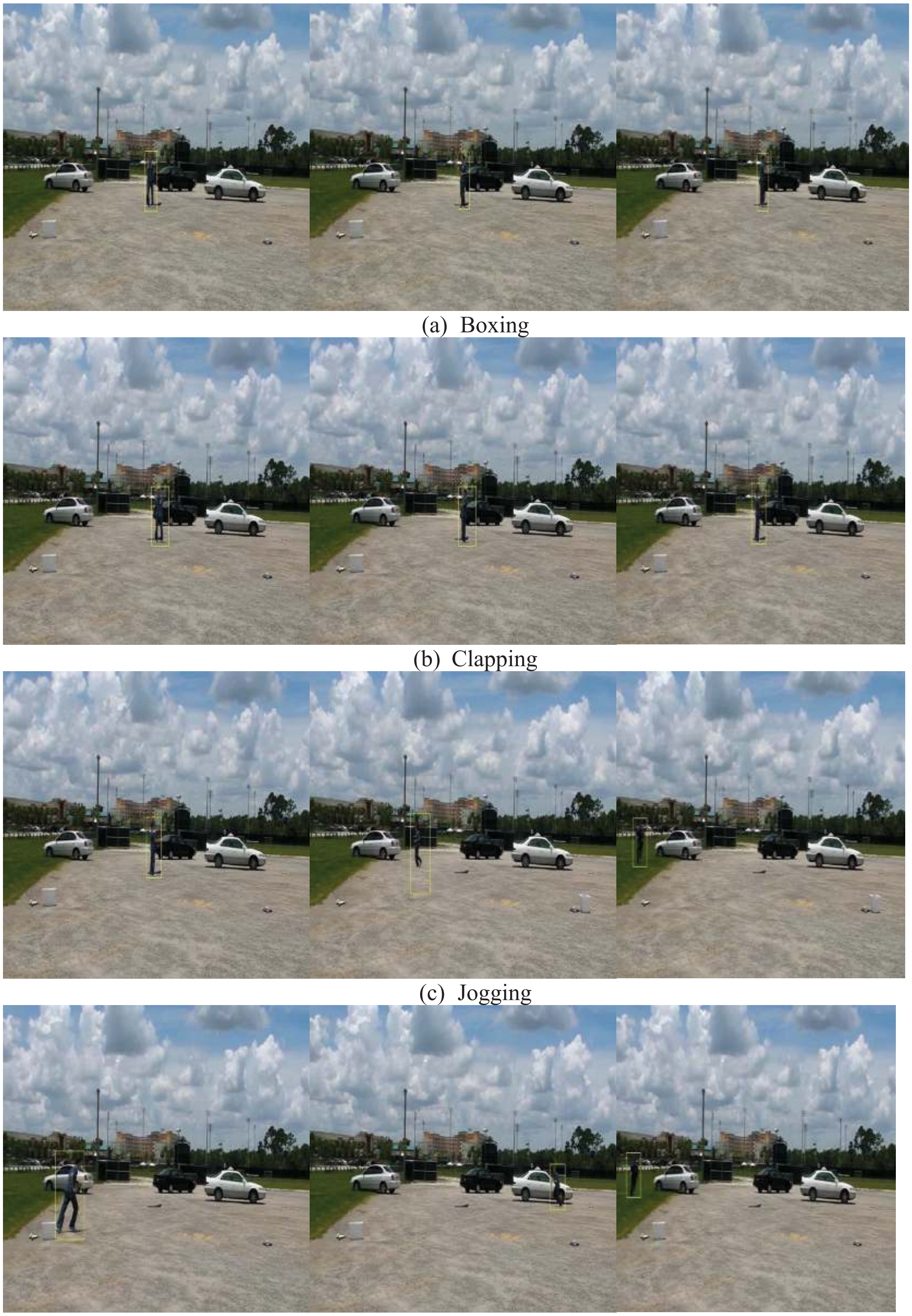
Method	**Accuracy (%)**
Qian *et al.* [168]	76.2
Ikizler-Cinbis & Sclaroff [169]	80
Bobick *et al.* [45]	73.4
Ahmad *et al.* [52]	92.6
Holte *et al.* [79]	90.2
Junejo *et al.* [80]	94.6
Proposed Method	100

Experiment 5

We have demonstrated results of the proposed method for VideoWeb Multi-view dataset [166]. 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 hours 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. 5.9, 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. 5.9, 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 [166].

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.* [168], Bobick *et al.* [45], Ikizler-Cinbis & Sclaroff [169], Holte *et al.* [79], Junejo *et al.* [80], and Ahmad *et al.* [52].

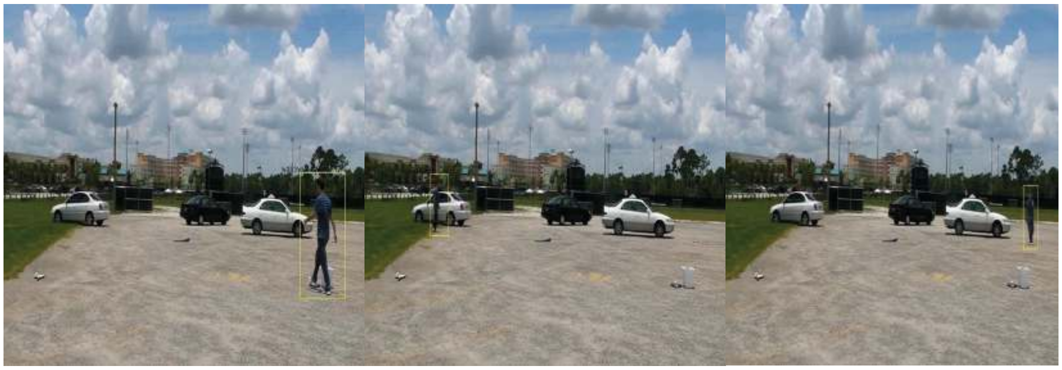


(a) Boxing

(b) Clapping

(c) Jogging

(d) Running



(e) Walking

Figure 5.9: Recognition of Activities in VideoWeb Multi-view dataset [166] (a) Boxing (b) Clapping (c) Jogging (d) Running (e) Walking

The confusion matrix of different activity for different methods has been shown in Table 5.9. 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 has been shown in Table 5.10 (calculated using Eq. 5.17). 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.

Table 5.9: Confusion matrices for the proposed and other methods

Recognized Instances ⟶ **Total Instances** ↓	**Boxing**	**Clapping**	**Jogging**	**Running**	**Walking**
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
Qian *et al.* [168]					
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
Ikizler-Cinbis and Sclaroff [169]					
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
Bobick *et al.* [45]					
Boxing	0.72	0.10	0	0.18	0
Clapping	0	0.74	0.20	0	0.06
Jogging	0.15	0	0.70	0.12	0.03
Running	0.10	0.10	0.09	0.71	0
Walking	0.15	0	0	0.15	0.70
Ahmad *et al.* [52]					
Boxing	0.86	0.10	0	0	0.04
Clapping	0.03	0.84	0.12	0.01	0
Jogging	0.15	0	0.83	0	0.02
Running	0.10	0.10	0	0.80	0
Walking	0	0	0.15	0	0.85
Holte *et al.* [79]					
Boxing	0.88	0	0.08	0	0.04
Clapping	0.10	0.90	0	0	0
Jogging	0.03	0.10	0.87	0	0
Running	0	0.02	0	0.88	0.10
Walking	0.10	0	0.01	0	0.89
Junejo *et al.* [80]					
Boxing	0.93	0	0.05	0	0.02
Clapping	0	0.92	0	0.08	0
Jogging	0.05	0	0.90	0	0.05
Running	0.05	0.02	0	0.93	0
Walking	0	0	0.05	0.05	0.90

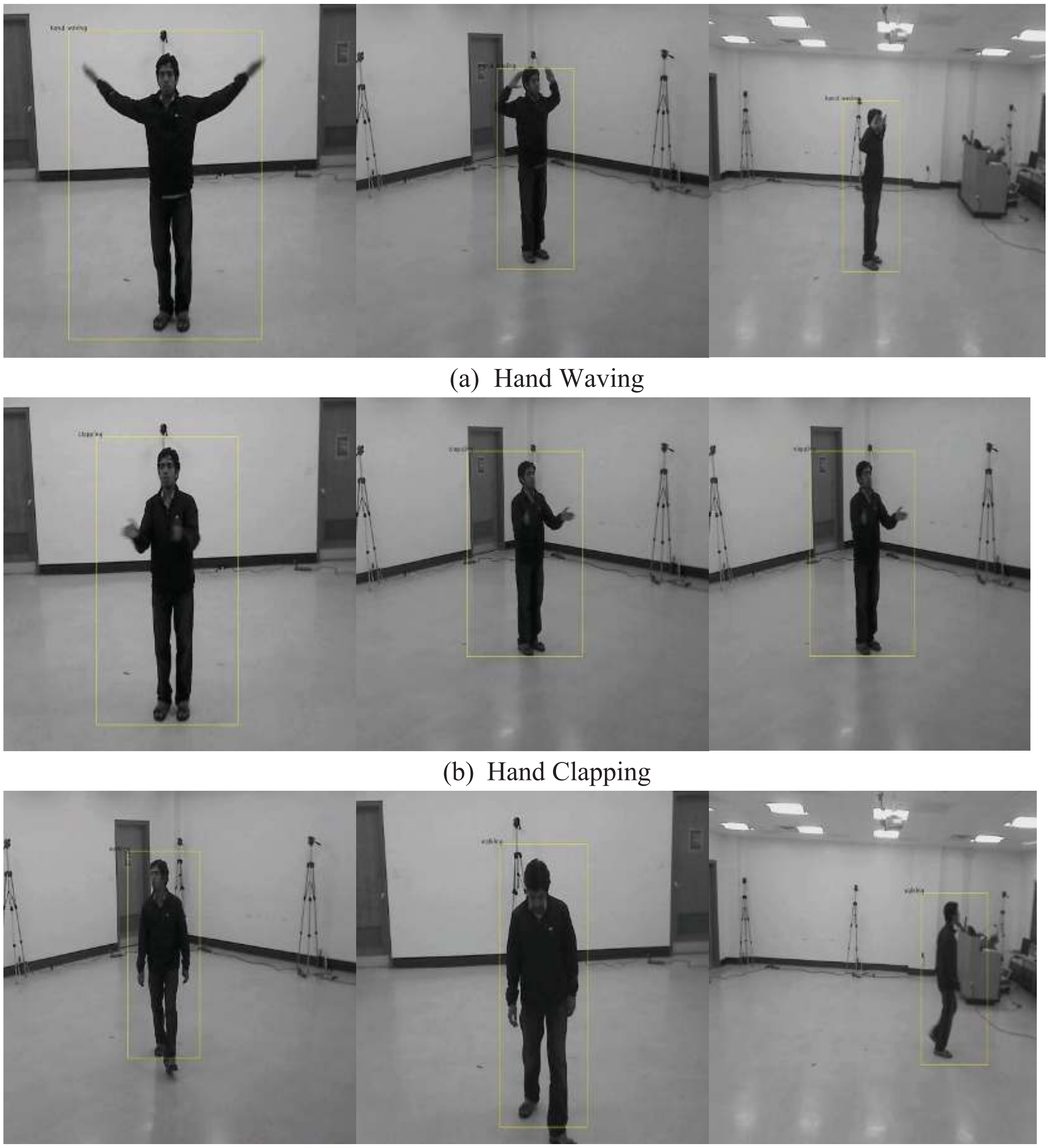
Table 5.10: Recognition results over the Video Web action recognition dataset [166]

Method	**Accuracy (%)**
Qian *et al.* [168]	64.4
Ikizler-Cinbis & Sclaroff [169]	79
Bobick *et al.* [45]	71.4
Ahmad *et al.* [52]	83.6
Holte *et al.* [79]	88.4
Junejo *et al.* [80]	91.6
Proposed Method	99

Experiment 6

In Fig.5.10, we have shown activity recognition with WVU multi-view human action recognition dataset [167]. This database includes different activities hand waving, clapping, jumping, jogging, bowling, throwing, pickup, and kicking. WVU multi-view human action recognition dataset [167] has been sorted based on the 8 views. For each view, action sequences performed by different subjects are provided. In Fig.5.10, 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 [167] in Tables 5.11-5.12.



(a) Hand Waving

(b) Hand Clapping

(c) Walking

Figure 5.10: Recognition of Activities in WVU multi-view human action recognition dataset [167] (a) Hand waving (b) Hand Clapping (c) Walking

These confusion matrices and recognition results presented in Tables 5.11 and 5.12 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 has been shown in Table 5.12 (calculate using Eq. 5.17).

Table 5.11: Confusion matrix for the proposed method and other methods

Recognized Instances → **Total Instances** ↓	**Hand Waving**	**Hand-clapping**	**Walking**
Proposed method			
Hand Waving	1	0	0
Hand-clapping	0	1	0
Walking	0	0.02	0.98
Qian *et al.* [168]			
Hand Waving	0.72	0.28	0
Hand-clapping	0.30	0.70	0
Walking	0.30	0.01	0.69
Ikizler-Cinbis & Sclaroff [169]			
Hand Waving	0.79	0.21	0
Hand-clapping	0	0.81	0.19
Walking	0.18	0	0.82
Bobick *et al.* [45]			
Hand Waving	0.74	0.26	0
Hand-clapping	0.24	0.76	0
Walking	0	0.23	0.77
Ahmad *et al.* [52]			
Hand Waving	0.84	0.16	0
Hand-clapping	0	0.81	0.19
Walking	0.07	0.10	0.83
Holte e*t al.* [79]			
Hand Waving	0.88	0	0.12
Hand-clapping	0.05	0.85	0.10
Walking	0.12	0.02	0.86
Junejo *et al.* [80]			
Hand Waving	0.92	0.03	0.05
Hand-clapping	0.11	0.89	0
Walking	0.07	0.03	0.90

Table 5.12: Recognition results over the WVU action recognition dataset [167]

Method	**Accuracy (%)**
Qian *et al.* [168]	70.33
Ikizler-Cinbis & Sclaroff [169]	80.66
Bobick *et al.* [45]	75.66
Ahmad *et al.* [52]	85
Holte *et al.* [79]	86.33
Junejo *et al.* [80]	90.33
Proposed Method	99.33

5.5. Conclusions

In this chapter, we have proposed a multi-view human activity recognition system based on temporal template matching which uses motion history images and spatial pose information to construct the activity templates. The experimental results demonstrate that the proposed method: (i) accurately recognizes different activities in various video frames, (ii) is suitable for static activities (like sitting, sleeping, standing, bending) as well as for dynamic activities (like jogging, walking), (iii) is pose invariant, frontal view is not necessary, (iv) can recognize activities in real outdoor and indoor environment both, (v) is suitable for operation in outdoor environment in the presence of shadow, (iv) is suitable for activities not only involving leg motion (like jogging, running, walking) but also for activities involving hand motion (like boxing, hand-clapping, hand- waving). This approach has been performed on six multi-view human activity video datasets: our own viewpoint dataset, KTH action recognition dataset [163], i3DPost multi-view dataset [164], MSR view-point action dataset [165], VideoWeb Multi-view dataset [166], and WVU multi-view human action recognition dataset [167]. 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.* [168], Bobick *et al.* [45], Ikizler-Cinbis & Sclaroff [169], Holte *et al.* [79], Junejo *et al.* [80], and Ahmad *et al.* [52], and has fared better than these.