Multimedia Tools and Applications (2020) 79:21037–21072
https://doi.org/10.1007/s11042-020-08902-w

Orientation Invariant Skeleton Feature (OISF): a new feature for Human Activity Recognition

Neelam Dwivedi[1] · **Dushyant Kumar Singh**[1] · **Dharmender Singh Kushwaha**[1]

Received: 26 March 2019 / Revised: 7 February 2020 / Accepted: 27 March 2020 /
Published online: 30 April 2020

Abstract
Human Activity Recognition is the process of identifying the activity of a person by analyzing continuous frames of a video. In many application areas, human activity identification is either a direct goal or it is a key segment of a bigger objective. Some of the examples are surveillance system, elder healthcare monitoring system, abnormal activity detection systems such as fight detection, theft detection etc. Robust and accurate activity recognition is a challenging task due to diverse reasons, such as changing ambient illumination, noise, background turbulence, camera placements etc. Existing literatures discuss some techniques for identifying human activity but these approaches are restricted to the case of videos recorded from static camera. The aim of the proposed approach is to fill this gap. In this proposed method, a new skeleton based feature for human activity recognition- "Orientation Invariant Skeleton Feature ($OISF$)"- is introduced and used to train Random Forest (RF) classifier for Human Activity Recognition. Efficiency of newly introduced feature $OISF$ is analyzed for the videos recorded with multiple cameras positioned at two different slant angles. Experimental results reveal that the newly introduced feature OISF has minimal dependency on variations of camera orientation. Accuracy achieved is $\approx 99.30\%$ with ViHASi dataset, $\approx 96.85\%$ with KTH dataset and $\approx 98.34\%$ with in-house dataset which is higher than those achieved by other researches with existing features. The improved result of human activity recognition in terms of accuracy proves the appropriateness of the proposed research in being used commercially.

Keywords Human activity recognition · Video surveillance · Skeleton · FV1 · OISF

✉ Neelam Dwivedi
neelamdw@gmail.com

Dushyant Kumar Singh
dushyant@mnnit.ac.in

Dharmender Singh Kushwaha
dsk@mnnit.ac.in

[1] Motilal Nehru National Institute of Technology Allahabad, Allahabad, India

1 Introduction

Human Activity Recognition aims to recognize the physical activity of one or more persons from a sequence of observations using sensors or video logs. It is a difficult task for real time videos because of background clutter, variations in illumination, varying viewpoints, partial occlusions and camera movements. Recognition of human activity is thus an open and challenging research area in the field of computer vision, machine vision, image processing [1] etc. A robust human activity recognition system can greatly enhance efficacy of surveillance system [7], human-computer interaction [21], gesture interpretation [39], life-care system [14, 23], sports analysis [32], smart homes [15] and other application domains. Due to its wide applicability in real life, many researchers have been working in this area since long. Recognition accuracy of such system broadly depends on the quality of extracted features and learning capability of the classifier. Different features like Motion Stable Shape (MSS) [18], Motion Scale-Invariant Feature Transformation (MoSIFT) [6], Speeded Up Robust Features Motion History Image (SURF-MHI) [35], Spatio-temporal interest point (STIP) detector [8], skeleton based features etc. have been used for Human Activity Recognition by the research community in the past.

Human Activity Recognition system is applied on the data collected through the wearable sensor or the recorded video (recorder from CCTV camera, Surveillance camera, etc.). Wearable sensor [3, 26, 39] based approach first collects the data from the sensors attached to the different parts of human body. Thereafter, physical activity of human is recognized by analyzing these data. This type of approach can be used only when a person is always wearing sensors which is not always feasible in a real life scenario. To mitigate this limitation, computer vision based human activity recognition is the preferred method. In the vision based approach, features like MoSIFT, SURF-MHI, etc. are extracted from the video data which are further used to train the classifier. These features work well for the videos recorded with static cameras, but not for the videos recorded with moving cameras. Also, processing time is high for extracting these features because feature extraction requires processing of all pixels in the image. To minimize these limitations, skeleton based features are proposed by many researchers in literature for human activity recognition. Feature selection plays an important role in vision based human activity recognition [9, 40] system because multiple features may be required to train the system for improving its accuracy. One can select large number of features to increase the accuracy, but this results in increased time for feature extraction & learning and testing of the system. Hence, skeleton-based features prove to be a viable alternative.

Skeleton features [15, 32, 37] are extracted by analyzing the skeleton pose in a sequence of frames. The skeletal representation of the human body at any instant reflects original pose of that person. A human pose in a video frame contains lots of pixels, while its skeletal representation contains comparatively smaller number of pixels thus reducing processing time for feature extraction. Human Activity Recognition using skeleton involves following steps:

1. Extraction of skeletons from video frames
2. Extraction of skeleton features
3. Construction of feature vector
4. Training and testing of the classifier

The skeleton feature usually contains joint information of different body parts like end and mid points of hands, legs, and position of head. Feature vectors are then constructed by fusing the features extracted from the person being tracked, in few consecutive frames. Finally, a supervised machine learning based classifier is used to identify the physical

activity of the tracked person based on the extracted features. Neural trees, SVM, LDA and other classifiers are used in the existing literature for the activity recognition task. Since, human activity recognition is a multi-class classification problem, tree based classifier is a preferable approach. This observation motivates us to explore the suitability of Random Forest classifier for human activity recognition. The aim of this work is to propose skeleton features for human activity recognition with the following goals:

- To improve the recognition accuracy of the system
- To reduce the dependency on camera positions and angles
- To obtain similar recognition accuracy for the videos recorded by either static or moving cameras

In this paper, a new feature named "Orientation Invariant Skeleton Feature ($OISF$)" is introduced for human activity recognition. This newly introduced feature is examined through number of experiments on two publicly available datasets: KTH and ViHASi, and one in-house dataset which is available at https://github.com/neelamdw/Actiondataset/. The in-house dataset is developed in the lab which contains five different actions (boxing, hand clapping, hand waving, jogging and walking). Three subsets (SA_1, SA_2&SA_3) are prepared by selecting similar and dissimilar activities from the ViHASi dataset (Table 1 presented in Section 4 may be reffered). Experiment #1, experiment #2 and experiment #3 are performed on SA_1, SA_2 and SA_3. All the activities of KTH dataset are taken together to perform the experiment #4. Similarly, All the activities of in-house dataset are taken together to perform the experiment #5. Following are the major contributions of this paper:

- A new skeleton based feature named as "Orientation Invariant Skeleton Feature ($OISF$)" for human activity recognition
- Performance evaluation of proposed feature with Random Forest classifier for KTH dataset, in-house dataset and different subsets of ViHASi dataset

This paper is organized as follows: In Section 2, literature survey for human activity recognition is presented. In Section 3, methodology and architecture of the proposed approach are explained. Experimental results and their analysis are presented in Section 4. Conclusion is presented in Section 5.

2 Literature survey for human activity recognition

In this section, various human activity recognition approaches are briefly reviewed. Uddin et al. [33] utilized Independent Component Analysis (ICA) to extract the activity shape information from the body joints instead of using whole body. Further, they apply Hidden Markov Model (HMM) on extracted activity shape information to recognize the human activity. In this method, features are easy to extract from the silhouettes but it cannot recognize the difference between near and distant body parts. To overcome this limitation, Jalal et al. [13] present human activity recognition for smart homes based on R transformation applied on depth silhouettes which require more processing time. To minimize the processing time, a new model for action recognition based on the combination of mid-level representation (HoG and BoW poselets) and discriminative key frame selection is proposed by Raptis et al. [28]. These approaches require the complete sequence of frames to recognize the human activity correctly. To achieve the early detection of action, a hybrid technique is presented by Vats et al. [34] which combines the benefits of computer vision and fuzzy set theory. This approach may recognize the action, even if a partial action occurs. They use

Fuzzy BK sub-product because of its flexibility and capability to imitate the natural human behavior. In the above-mentioned approaches, large number of pixel processing is required.

To reduce the pixel processing, skeleton are extracted from the images at first. Thereafter, features are extracted by processing the skeleton sequences that reduce the pixel processing time. Anjum et al. [2] present complex human activity recognition by tracking a subset of human skeleton joints instead of tracking the whole skeleton. Skeleton joints are selected either manually or automatically from the depth videos recorded by Kinect camera. They use Multiclass Support Vector Machines (MSVMs) to classify the human activity. Weng et al. [37] present human activity recognition using length-variable edge trajectory (LV-ET) and spatio-temporal motion skeleton descriptor. The LV-ET extracted by tracking edge points across video frames based on optical flow with the aim of better descriptor with the evolution of different type of actions. A novel encoding method for trajectory clustering is proposed to extract Spatio Temporal Motion Skeleton (STMS) (also called motion skeleton). Habli et al. [11] propose skeleton-based human activity recognition for elderly monitoring systems. For this task, they use spatial and temporal coordinates for the 3D skeleton and combine both to represent each frame of a human activity. Randomised tree algorithm is used to train and validate the method on the MSR-Action3D and DailyActivity3D datasets. Manzi et al. [22] have introduced an activity recognition system for two persons using skeleton data extracted from a depth camera. They used an unsupervised clustering approach to compute the activity using a set of few basic informative postures first. Thereafter, models are created using multiclass support vector machines on the training set. An optimal number of clusters for each sample are dynamically found by the X-means algorithm during classification phase. Li et al. [19] focus on multi-view skeletal interactions for human activity recognition. For this, a multi Active Joint Interaction Graph (AJIG) model is proposed to code the spatio-temporal patterns of two-person skeletal interactions. Then AJIG kernel is used to compute the similarity between two AJIGs. Further, a multiple kernel learning approach is applied to jointly learn the optimal combination of the numerous AJIG kernels. Ofli et al. [25] present a sequence of the Most Informative Joints (SMIJ) feature to recognize a human skeleton activity. At each time instant, few skeletal joints that are most related to the current action are selected. The selection of joints is based on highly interpretable measures such as mean or variance of joint angles, maximum angular velocity of joints etc. Zhu et al. [40] presents a deep LSTM network for skeleton activity recognition. Their model facilitates the learning of skeleton joint features with the help of co-occurrence exploration mechanism. This model dropout the complex structures among the important joints using exploration mechanism. They compare their method with other deep networks for skeletal activity recognition [9]. A methodology of our proposed approach is discussed in the next section.

3 Proposed approach for human activity recognition

In this section, a skeleton based feature to identify human activity is introduced. Figure 1 depicts the architectural design of the proposed approach for human activity recognition, which consists of three phases.

In the first phase, video is converted into consecutive frames followed by Skeletonization using Algorithm 1. Figure 2 depicts the skeletonization process of an input video frame. This process involves extraction of foreground object, conversion of foreground object into binary frame, enhancement of the frame, filling of small holes, removal of islands, and repeated thinning operation. After skeletonization, Region of Interest (RoI) is obtained and

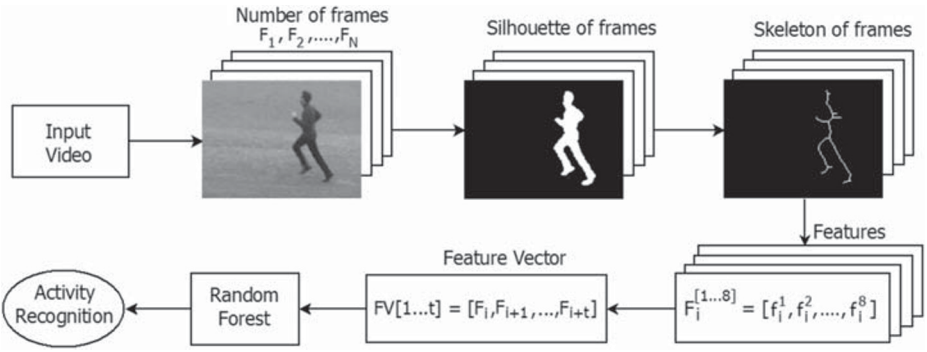


Fig. 1 Flow chart of the proposed approach

is marked by an elliptical boundary in the second phase. To draw the elliptical bounding box around skeleton: orientation, centroid, length of major axis, length of minor axis and eccentricity of the ellipse is obtained by applying a MATLAB function "$regionprops()$". This elliptical bounding box is further divided into eight symmetric regions. All of these tasks are performed by applying Algorithm 2. In the third phase, feature FV1 and the newly introduced feature OISF are extracted using Algorithm 3 and Algorithm 4 respectively. Time complexity for extraction of each of the feature: $FV1$ and $OISF$ is $\theta(m \times n)$ where, $m \times n$ is the total number of pixels in a frame. Random Forest classifier is trained with each extracted feature separately for human activity recognition. The structure of random forest classifier proposed by LEO BREIMAN [4] is used in this paper. Each forest consist of 500 trees. Once the classifier is trained, it is used to recognize the activity from a new sequence of frames.

3.1 Phase 1 (pre-processing of video): skeletonization of input video frames

Depends on frame rate and duration of video, an input video (V) is converted into N consecutive frames (F_1, F_2, ..., F_N). For complex backgrounds of the image, foreground object is extracted by applying the background subtraction technique. Each foreground frame is then transformed into binary frame (image) in two steps:

Step 1: For each framve, two intensity thresholds (th_1 and th_2) of the frame are calculated as:

$$th1 = h_j - \text{c1}$$
$$th2 = h_j + \text{c2}$$

Here, hj is the pixel intensity value with maximum frequency in a grayscale frame. Different values of c1 & c2 have been tried through experimentation (In the experiments valuses

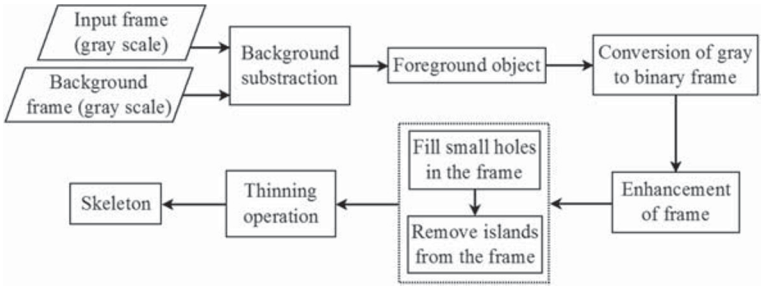


Fig. 2 Flow chart of skeletonization of a frame

taken c1 = 35 and c2 =40). These values (c1, c2) are initially adjusted and they remain constants for a dataset.

Step 2: All the intensity values lying between th_1 and th_2 in a frame are set to '1'. All other intensity values are set to $'0'$. The resulting image is a binary image.

Each binary frame is further enhanced by applying median filter. Two morphological operations namely, dilation and erosion are used to obtain a well-defined shape of the person (silhouette) from the binary frame. To perform these morphological operations a linear structuring element of length $c3$ is used. Dilation and erosion are performed $k1$ and $k2$ times respectively, on the input binary frame. Various combinations of dilation and erosion operation are tried. Here, $k1 = 2$, $k2 = 1$ & $c3 = 3$ have shown promising result in terms of silhouette extraction. Skeleton is obtained by applying repeated thinning operation on the silhouette and the silhouette is obtained after performing morphological operations. Thinning operation removes pixels in such a way that an object shrinks to a minimally connected stroke. The pseudo code of skeletonization process is presented in Algorithm 1 (detail Algorithm is presented in Appendix as Algorithm A.1).

Algorithm 1 *Construct_skeleton*().

INPUT: V(N): Video V containing N number of frames; B: Background frame
OUTPUT: skeleton[N]: N number of skeletons
Ensure:
(1) $k1$, $k2$, $k3$: dilation, erosion and thinning operations counts respectively.

```
Procedure construct_skeleton(V)
for F_i ∈ V, i = 1 ... N do
    F_i ← backgroundSubtraction(F_i, B)        ▷ F_i represents i^th frame.
    h ← histogram(F_i)
    th1 ← lowerThreshold(h)
    th2 ← higherThreshold(h)
    for pixel_value ∈ F_i do
        if th1 ≤ pixel_value ≤ th2 then
            F_i[pixel_value] ← 1
        else
            F_i[pixel_value] ← 0
        end if
    end for
    for k = 1 ... k_1 do
        F_i ← dilation(F_i, linearStructuringElement)
    end for
    for k = 1 ... k_2 do
        F_i ← erosion(F_i, linearStructuringElement)
    end for
    for k = 1 ... k_3 do
        F_i ← thinning(F_i)
    end for
    skeleton(i)
end for
return(skeleton[N])
end procedure
```

Figure 3a, b and c show the ten original frames, and the corresponding binary and skeleton frames obtained through Algorithm 1. These frames are taken from the KTH dataset for

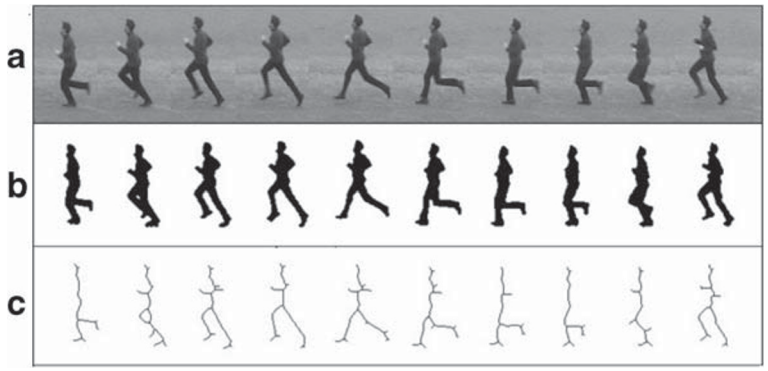


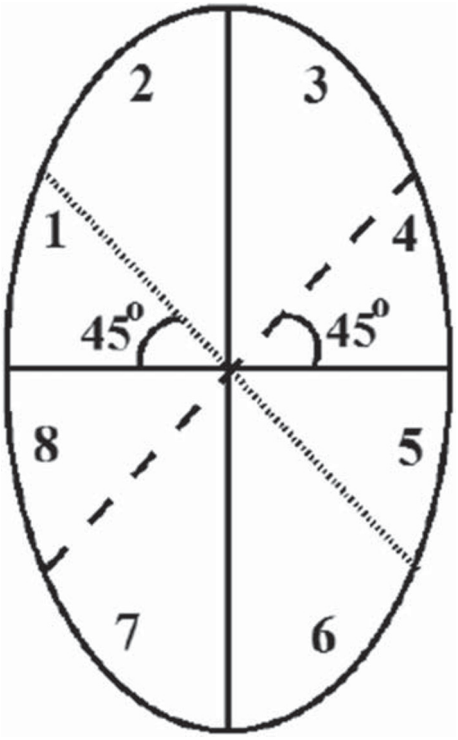
Fig. 3 (**a**) Frames obtained from input video (**b**) Extracted silhouettes from the frames (**c**) Skeletons obtained from silhouettes

'running' activity. For the sake of visual clarity, all the figures in this paper include both the complements of actual silhouettes and the corresponding skeletons.

3.2 Phase 2: Region of Interest (RoI) selection

The skeleton extracted from the silhouette (obtained in Phase1) is our Region of Interest. The elliptical boundary around the skeleton separates it from the input frame. Centroid (x_c, y_c), length of major axis $(2 \times a)$, length of minor axis $(2 \times b)$, eccentricity (e) and orientation (θ) of the ellipse are calculated for further processing. As explained before,

Fig. 4 Ellipse with 8 regions



"$regionprops()$" function is used to obtain these values. This ellipse is then divided into eight symmetric regions for in order to extract hands, legs and head features as shown in Fig. 4. To divide the ellipse into eight symmetric regions following four lines are drawn that pass through the centroid of the ellipse:

1. Major-axis
2. Minor-axis
3. Line passing through centroid and inclined at an angle of 45^0 in clockwise direction from the minor axis
4. Line passing through centroid and inclined at an angle of 45^0 in anti-clockwise direction from the minor axis

Algorithm 2 RoI selection and it's division in 8-regions.

INPUT: Skeleton S of a frame.
OUTPUT: Co-ordinates of eight symmetric regions of the ellipse.

Procedure $boundingellipse(S)$
Draw an elliptical bounding box around the skeleton S with orientation of θ.
Calculate the co-ordinates of chords $(x^1_{mj}, y^1_{mj})(x^2_{mj}, y^2_{mj})$ to draw the major axis in the ellipse.
Calculate the co-ordinates of chords $(x^1_{mn}, y^1_{mn})(x^2_{mn}, y^2_{mn})$ to draw the minor axis in the ellipse.
Calculate the co-ordinates of chords $(x^1_{ch1}, y^1_{ch1})(x^2_{ch1}, y^2_{ch1})$ to draw a line in the ellipse having orientation is of 45° anti-clockwise from the minor axis.
Calculate the co-ordinates of chords $(x^1_{ch2}, y^1_{ch2})(x^2_{ch2}, y^2_{ch2})$ to draw a line in the ellipse having orientation is of 45° clockwise from the minor axis.
return(*Co-ordinates of eight symmetric regions of the ellipse*)
end procedure

Let us assume that, (x_c, y_c) is the co-ordinate of centroid of the given ellipse. Then, equation (1) represents parametric equation of the ellipse whose orientation is θ:

$$\frac{(x_1 cos\theta + y_1 sin\theta)^2}{b^2} + \frac{(x_1 sin\theta - y_1 cos\theta)^2}{a^2} = 1 \tag{1}$$

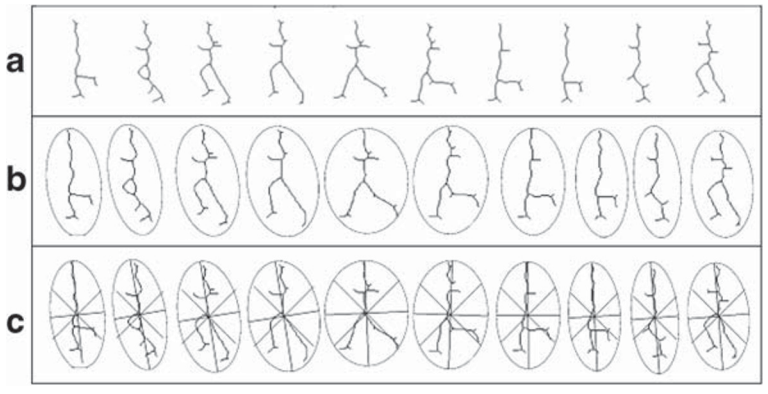


Fig. 5 (**a**) Obtained skeletons from input video frame (**b**) Region of Interest(RoI) selection (**c**) division of RoI in eight symmetric regions

where, $x_1 = x - x_c$ and $y_1 = y - y_c$

To draw these lines, coordinates are obtained by applying eight cuts on the ellipse at the angles 0° (cut 1), 45° ($cut2$), 90° ($cut3$), 135° ($cut4$), 180° ($cut5$), 225° ($cut6$), 270° ($cut7$), 315° ($cut8$). Four lines are drawn between $cut1$ & $cut5$ (minor axis), $cut2$ & $cut6$ (dotted line), $cut3$ & $cut7$ (major axis), and $cut4$ & $cut8$ (dashed line) respectively.

Obtained ROI has eight symmetric regions and each region contains the information of skeleton. Figure 5a, b and c show the skeletal representation of the frames obtained by applying Algorithm 1, ROI with elliptical bounding box and its division into 8-regions using Algorithm 2 (detail Algorithm is presented in Appendix as Algorithm A.2) respectively. Algorithms for extracting the features are explained in the Section 3.3.

3.3 Features used

Two features: $FV1$ [17] and Orientation Invariant Skeleton Feature ($OISF$) have been used for the proposed approach. To extract these features, skeletons are obtained from an input video by applying Algorithm 1. Region of Interest (ROI) is selected and bounded by an elliptical bounding box that is further divided into eight symmetric regions (Fig. 3) by applying Algorithm 2. The information in each region of the skeleton is number of pixels, coordinates of hands, legs and head. By using this information, feature vector $FV1$ and $OISF$ are calculated. Total eight features for each skeleton are obtained both for $FV1$ and $OISF$. Details of these features are discussed in Sections 3.3.1 and 3.3.2 respectively.

3.3.1 *FV1* feature extraction

Figure 6 shows the flow chart for extracting the $FV1$ feature value of one frame. Skeletons sk_1, sk_2, ..., sk_N are obtained from the input video V by applying Algorithm 1. For each skeleton, an elliptical boundary is drawn and divided into eight symmetric regions by applying Algorithm 2. sk_i^j contains the total number of white pixels of i^{th} skeleton in the j^{th} ($1 \leq j \leq 8$) region of the ellipse. For each region, one feature value is extracted as follows:

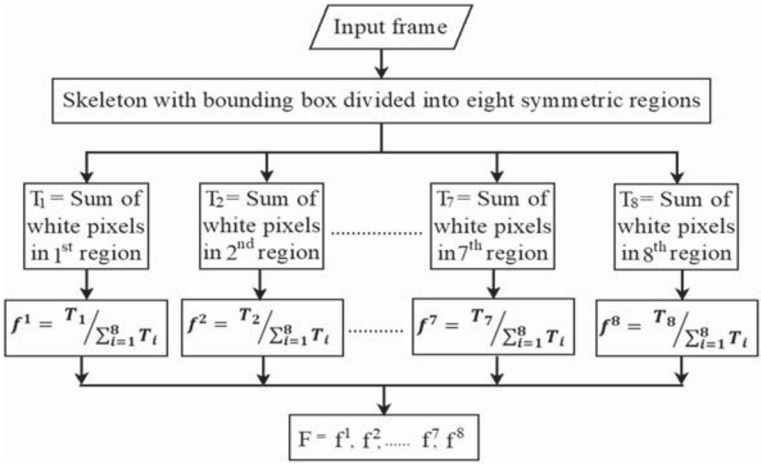


Fig. 6 Flow chart of extraction of FV1 feature in a frame

Step 1: Compute the sum of pixels of j^{th} region of i^{th} skeleton sk_i:

$$P_i^j = sum(sk_i^j)$$

Step 2: Compute the sum of total pixels (p^t) for the i^{th} skeleton sk_i:

$$p^t = \sum_{j=1}^{8} P_i^j$$

Thus, feature corresponding to j^{th} region is given by the following expression:

$$f_i^j = \frac{P_i^j}{p^t}$$

Here, $f_i^{[1...8]} = [f_i^1, f_i^2..., f_i^8]$ is the feature corresponding to i^{th} skeleton sk_i. Thus, a feature vector of size $(N \times 8)$ is obtained for the input video V. Final feature vector $FV1$ is generated by fusing features of ten consecutive frames, since an activity is characterized by analyzing sequence of frames. This feature vector $FV1$ is used to train and test the classifier for human activity recognition. Size of final feature matrix $FV1$ for N frames is $(\frac{N}{10} \times 80)$. Motivation for and the process of extraction of newly introduced feature $OISF$ are discussed in the next section.

Algorithm 3 FV1 feature extraction.

INPUT: skeleton[N]
OUTPUT: Feature Matrix FV1 $[\frac{N}{10}][80]$

Procedure $feature1(V)$
for $i = 1 ... N$ **do**
  $region\_coordinates[8] \leftarrow boundingEllipse(skeleton[i])$
  $p^t \leftarrow \sum_{skeleton[i]} pixel\_value$ ▷ p^t represents total pixels in frame.
  for $l = 1 ... 8$ **do**
    $p^r \leftarrow \sum_{region\_coordinates(l)} pixel\_value$ ▷ p^r represents total pixels in region.
    $f_l^i \leftarrow \frac{p^r}{p^t}$
  end for
  $f1(i)[1...8] \leftarrow [f_1^i, f_2^i, ..., f_l^i]$
end for
create feature vector FV1 by concatenating features of 10 consecutive skeletons as a single feature
return$(Feature\ Matrix FV1[\frac{N}{10}][80])$
end procedure

3.3.2 *OISF* feature extraction

In this section, algorithm and the motivation behind introducing a new feature, "Orientation Invariant Skeleton Feature ($OISF$)" is discussed. In the literature, it is observed that existing features for human activity recognition are dependent on the orientation and positioning of the cameras that reduces the recognition accuracy for the videos recorded by moving

camera. Motivation of introducing $OISF$ feature is to overcome this limitation. Application of $OISF$ feature improves the recognition accuracy of human activity recognition system for the videos recorded with moving as well as static cameras. This is because, $OISF$ characterizes human actions with respect to the relative movements of hands and legs along the x & y axes separately in each frame. The features corresponding to each skeleton sk_i is obtained as follows:

1. x and y coordinates of first white pixel (having least value of x coordinate) in the first, second, seventh and eighth regions are determined.
2. x and y coordinates of last white pixel (having maximum value of x coordinate) in the third, fourth, fifth and sixth regions are determined.
3. Absolute differences of x and y coordinates of first & fourth, second & third, fifth & eighth, sixth & seventh regions are calculated and taken as eight features of skeleton sk_i.

Algorithm 4 OISF feature extraction.

INPUT: skeleton[N]
OUTPUT: $OISF[\frac{N}{10}][80]$: Orientation Invariant Skeleton Feature (OISF)
Ensure:
(1) f_x^1 to f_x^4 and f_y^1 to f_y^4 are the OISF features of one frame w.r.t. x and y coordinates separately.

```
Procedure feature2(V)
for i = 1 ... N do
    region_coordinates[8] ← boundingEllipse(skeleton[i])
    for l = 1 ... 8 do
        if (l ≡ 1st region or 2nd region or 7th region or 8th region) then
            (X_l, Y_l) ← x and y coordinates of first white pixel
        else
            (X_l, Y_l) ← x and y coordinates of last white pixel
        end if
    end for
    f_x^1 ← absoluteDifference(X_1, X_4), f_y^1 ← absoluteDifference(Y_1, Y_4)
    f_x^2 ← absoluteDifference(X_2, X_3), f_y^2 ← absoluteDifference(Y_2, Y_3)
    f_x^3 ← absoluteDifference(X_5, X_8), f_y^3 ← absoluteDifference(Y_5, Y_8)
    f_x^4 ← absoluteDifference(X_6, X_7), f_y^4 ← absoluteDifference(Y_6, Y_7)
    f2(i)[1...8] ← [f_x^1, f_x^2, f_x^3, f_x^4, f_y^1, f_y^2, f_y^3, f_y^4]
end for
create feature vector OISF by concatenating features of 10 consecutive skeletons as a single feature
return(OISF[N/10][80])
end procedure
```

Example 1 : For any skeleton sk_i, let (x_1, y_1), (x_2, y_2), (x_7, y_7) and (x_8, y_8) represent the coordinates of first white pixel of first, second, seventh and eighth regions respectively, while (x_3, y_3), (x_4, y_4), (x_5, y_5) and (x_6, y_6) represent the coordinates of last white pixel of third, fourth, fifth and sixth regions, respectively. First, second, third, fourth, fifth, sixth, seventh and eighth feature values of the skeleton sk_i are calculated as $abs(x_1 - x_4)$, $abs(x_2 -$

Table 1 Three subsets of ViHasi dataset

	SA_1		SA_2		SA_3	
S. No.	Action class	Action name	Action class	Action name	Action class	Action name
1.	C2	Jump Get On Bar	C1	Hang On Bar	C1	Hang On Bar
2.	C3	Jump Over Object	C2	Jump Get On Bar	C9	Hero Smash
3.	C4	Jump From Object	C3	Jump Over Object	C10	Hero Door Slam
4.	C5	Run Pull Object	C4	Jump From Object	C12	Knockout
5.	C6	Run Push Object	C5	Run Pull Object	C13	Granade
6.	C7	Run Turn90 Left	C6	Run Push Object	C14	Collapse
7.	C8	Run Turn90 Right	C11	Knockout Spin	C15	StandLookAround
8.	C18	Walk	C12	Knockout	C16	Punch
9.	C19	Walk Turn180	C16	Punch	C17	Jump Kick
10.	C20	Run	C17	Jump Kick	C18	Walk

$x_3)$, $abs(x_5 - x_8)$, $abs(x_6 - x_7)$, $abs(y_1 - y_4)$, $abs(y_2 - y_3)$, $abs(y_5 - y_8)$ and $abs(y_6 - y_7)$ respectively.

Normally, activities cannot be distinguished by analyzing a single frame. Therefore, features of ten consecutive frames are combined to generate the $OISF$ feature vector for the classification of human activities. In the next section, experimental set-up and datasets used are discussed along with the analysis of the proposed method using various performance metrics.

4 Experiments and their analysis

To evaluate the performance of the proposed approach, five experiments have been conducted. All these experiments have been conducted using MATLAB R2017a in core i7 processor with 4GB RAM. For experiments #1, #2, & #3, frames are taken from the Virtual Human Action Silhouette (ViHASi) dataset [27]. For experiment #4, frames are taken from the KTH dataset [31]. For experiment #5, frames are taken from the in-house dataset. The ViHASi dataset contains synthetic videos of 20 action classes and are recorded by 9 actors. These videos used a maximum of 40 synchronized perspective camera views. This 40 synchronized perspective camera views are divided into two sets, each consists of 20 cameras views. The two sets of cameras are fixed at slant angles of 27° and 45° with the horizontal plane respectively. Angular difference between the cameras is 18° in both the sets. In the videos of KTH dataset, there are six actions (boxing, hand clapping, hand waving, jogging, running and walking). These actions are performed by 25 persons in two different scenarios that are indoor and outdoor scenarios with different scale variations. All the video sequences are taken over homogeneous background with static camera. As explained in Section 1, the videos of in-house dataset contains five different actions. These actions are performed by 2 actors in indoor scenario. All the videos are recorded by the static camera with low-resolution, complex background, and variations in illumination.

To create maximum possible combination of activities, twenty action classes of ViHASi dataset have been divided into three subsets for experiments #1, #2 & #3. These subsets are categorized on the basis of similarity of actions and named as SA_1, SA_2 and SA_3. Table 1 presents actions of all sub activities. SA_1 contains almost similar actions like running; walking etc. SA_2 contains combination of similar (run pull object, run push object etc.) and dissimilar (knockout, punch etc.) actions. SA_3 contains dissimilar actions like hang on bar, granade etc. Actions of SA_1, SA_2 and SA_3 are used in experiments #1, #2 & #3, respectively. These three experiments are conducted in two separate scenarios:

1. Videos recorded by the first set of cameras fixed at slant angle of 27°
2. Videos recorded by the second set of cameras fixed at slant angle of 45°

In all the experiments, input videos are pre-processed using Algorithm 1 and, $FV1$ and $OISF$ features are extracted using Algorithm 3 (detail Algorithm is presented in Appendix as Algorithm A.3) and Algorithm 4 (detail Algorithm is presented in Appendix as Algorithm A.4) respectively. A separate human activity recognition model is created by training Random Forest classifier for each of the following cases:

Case 1: Training with $FV1$ only
Case 2: Training with $OISF$ only

By doing this, two separate Random Forest models are developed for each of the experiments. Random Forest classifier is a supervised machine learning based classifier and uses an "ensemble learning method" for the classification. The general method of Random Forest was first proposed by Ho in 1995 [12]. Each classification tree of this classifier uses two well-known methods, named boosting [30] and bagging [5]. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. A weighted vote is taken from all the decision trees to predict the class of the new frame. For example, if any classification problem has $n1$ classes, then samples of all the classes are selected randomly for the training. If each sample contains K variables, then k ($k < K$) randomly selected variables are specified at each node. Each decision tree grows up to its maximum extent without any pruning. By considering maximum vote from all the decision trees, a new data

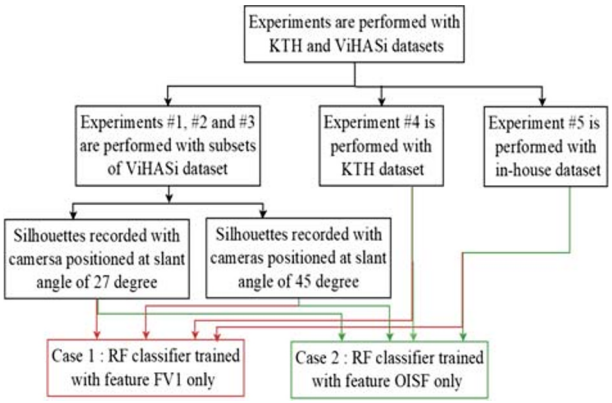


Fig. 7 Summary of all the experiments

Table 2 Classification results of a classifier

	Predicted			
Frame	Activity 1	Activity 2	Activity 3	Activity 4
Activity 1 (*Actual*)	N_1^1	N_1^2	N_1^3	N_1^4
Activity 2 (*Actual*)	N_2^1	N_2^2	N_2^3	N_2^4
Activity 3 (*Actual*)	N_3^1	N_3^2	N_3^3	N_3^4
Activity 4 (*Actual*)	N_4^1	N_4^2	N_4^3	N_4^4

is classified. Here, features are randomly selected to split the nodes. The structure of random forest classifier proposed by LEO BREIMAN [4] is used in this paper. Total number of 500 trees are used to construct a forest. Figure 7 summarizes all the experiments. To measure the performance of the proposed approach of the system $Confusion\ Matrix$ and five performance metrics: $Precision$, $Recall$, $Specificity$, $F_1\ score$ and $Accuracy$ are used and discussed in the next section.

4.1 Parameters used for performance measurement

Precision, Recall, F_1 score, Accuracy and Confusion Matrix are most important parameters that are being used to evaluate the performance of the proposed approach. Assume a classifier classifies input video frames as shown in Table 2, where N_i^j denotes the number of i^{th}

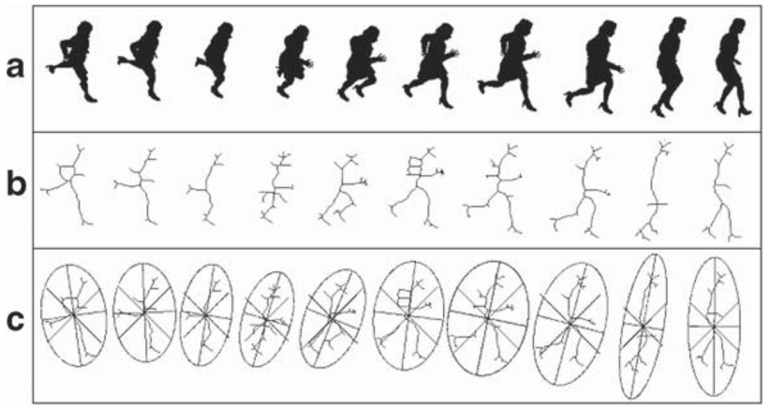


Fig. 8 (**a**) Silhouettes of RunTurn90Left activity (**b**) skeletons obtained from the silhouettes (**c**) symmetric 8-regions division of RoI

Table 3 Confusion Matrix of Experiment #1 for the silhouettes captured using cameras with a slant angle of 27°

		C2	C3	C4	C5	C6	C7	C8	C18	C19	C20
Case 1	C2	**1.00**	0	0	0	0	0	0	0	0	0
	C3	0	**0.98**	0	0	0.02	0	0	0	0	0
	C4	0	0	**1.00**	0	0	0	0	0	0	0
	C5	0	0	0	**0.98**	0	0	0	0.02	0	0
	C6	0	0	0	0	**0.98**	0	0	0.02	0	0
	C7	0	0	0	0	0	**0.88**	0	0	0.02	0.10
	C8	0	0	0	0	0	0	**1.00**	0	0	0
	C18	0.02	0	0	0	0.02	0	0.02	**0.95**	0	0
	C19	0	0	0.03	0	0	0	0	0	**0.97**	0
	C20	0.02	0	0	0	0	0.07	0	0	0	**0.92**
Case 2	C2	**1.00**	0	0	0	0	0	0	0	0	0
	C3	0	**0.97**	0	0	0.03	0	0	0	0	0
	C4	0	0	**1.00**	0	0	0	0	0	0	0
	C5	0	0	0	**1.00**	0	0	0	0	0	0
	C6	0.02	0	0	0	**0.98**	0	0	0	0	0
	C7	0	0.02	0	0	0	**0.90**	0	0	0	0.8
	C8	0	0	0.02	0	0	0	**0.98**	0	0	0
	C18	0	0.03	0	0	0	0	0	**0.95**	0.02	0
	C19	0	0	0	0	0	0	0	0.02	**0.98**	0
	C20	0	0	0	0	0	0.08	0	0	0	**0.90**

Table 4 Confusion Matrix of Experiment #1 for the silhouettes captured using cameras with a slant angle of 45°

		C2	C3	C4	C5	C6	C7	C8	C18	C19	C20
Case 1	C2	**0.97**	0.03	0	0	0	0	0	0	0	0
	C3	0	**0.95**	0.01	0.01	0	0	0	0	0.03	0
	C4	0.02	0.03	**0.92**	0.01	0	0	0	0	0.03	0
	C5	0	0.02	0.01	**0.93**	0	0	0	0.02	0.02	0.02
	C6	0.01	0.02	0	0.01	**0.95**	0	0	0	0.01	0.01
	C7	0	0	0.01	0.03	0	**0.87**	0	0	0	0.09
	C8	0.02	0.01	0	0	0	0.01	**0.90**	0.01	0.05	0.01
	C18	0	0.01	0.01	0	0	0.01	0.03	**0.88**	0.06	0.01
	C19	0.01	0	0.03	0	0	0	0	0.02	**0.95**	0
	C20	0.02	0	0	0	0	0.06	0	0.01	0.02	**0.92**
Case 2	C2	**0.99**	0	0.01	0	0	0	0	0	0	0
	C3	0	**1.00**	0	0	0	0	0	0	0	0
	C4	0.01	0	**0.98**	0	0	0	0	0	0.01	0
	C5	0.01	0.01	0	**0.92**	0	0	0	0.04	0.01	0.02
	C6	0.01	0	0	0	**0.98**	0	0	0	0	0.01
	C7	0	0	0	0	0	**0.89**	0.03	0	0	0.08
	C8	0	0.01	0	0	0.01	0.01	**0.94**	0.03	0	0.01
	C18	0.01	0	0.02	0.01	0	0	0.01	**0.90**	0.03	0.03
	C19	0	0	0.02	0	0	0	0.01	0	**0.98**	0
	C20	0	0	0	0	0	0.05	0.02	0	0	**0.93**

activity classified as j^{th} activity by the classifier. Performance metrics for this classification results are calculated as follows:

$$Precision_i = \frac{N_i^i}{\sum_{j=1}^{4} N_j^i}$$

$$Recall_i / Sensitivity_i = \frac{N_i^i}{\sum_{j=1}^{4} N_i^j}$$

$$Specificity_i = \frac{\sum_{j,k=1}^{4} N_j^k;\ j,k \neq i}{\sum_{j=1,i\neq j}^{4} N_j^i + \sum_{j,k=1}^{4} N_j^k;\ j,k \neq i}$$

$$F_1\ score_i = 2 \times \frac{Precision_i \times Recall_i}{Precision_i\ +\ Recall_i}$$

$$Accuracy_i = \frac{N_i^i + \sum_{j,k=1}^{4} N_j^k;\ j,k \neq i}{\sum_{j,k=1}^{4} N_j^k}$$

4.2 Experiment #1

To perform this experiment, ten similar activities (C2: JumpGetOnBar, C3: JumpOverObject, C4: JumpFromObject, C5: RunPullObject, C6: RunPushObject, C7: RunTurn90Left, C8: RunTurn90Right, C18: Walk, C19: WalkTurn180, and C20: Run) grouped

Table 5 Performance Metrics of Experiment #1 for the silhouettes captured using cameras with a slant angle of 27°

		TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	C2	120	04	96.77	100.0	99.63	98.36	99.67
	C3	118	00	100.00	98.33	100.00	99.16	99.83
	C4	120	04	96.77	100.0	99.63	98.36	99.67
	C5	118	00	100.00	98.33	100.00	99.16	99.83
	C6	118	04	96.72	98.33	99.63	97.52	99.50
	C7	106	08	92.98	88.33	99.26	90.60	98.16
	C8	120	02	98.36	100.0	99.81	99.17	99.83
	C18	114	04	96.61	95.00	99.63	95.80	99.17
	C19	116	02	98.31	96.67	99.81	97.48	99.50
	C20	110	12	90.16	91.67	98.89	90.91	98.17
Case 2	C2	120	02	98.36	100.0	99.81	99.17	99.83
	C3	116	06	95.08	96.67	99.44	95.87	99.17
	C4	120	02	98.36	100.0	99.81	99.17	99.83
	C5	120	00	100.00	100.0	100.00	100.00	100.0
	C6	118	04	96.72	98.33	99.63	97.52	99.50
	C7	108	10	91.53	90.00	99.07	90.76	98.17
	C8	118	00	100.00	98.33	100.00	99.16	99.83
	C18	114	02	98.28	95.00	99.81	96.61	99.33
	C19	118	02	98.33	98.33	99.81	98.33	99.67
	C20	110	10	91.67	91.67	99.07	91.67	98.33

in SA_1 (Table 1) for the jump, walk and run categories are taken from the ViHASi dataset. Figure 8a, b and c show ten samples silhouettes of $RunTurn90Left$ activity recorded from the second set of cameras, skeletons obtained using Algorithm 1 and 8-regions symmetrical division of elliptical bounding box obtained using Algorithm 2 respectively. To measure the effectiveness of the proposed approach, confusion matrix, precision, recall, specificity, F_1 score, and accuracy are used as discussed in Section 4.1.

Tables 3 and 4 show the confusion matrices that are obtained in the first experiment for both the cases (Case 1 and Case 2) on two different sets of camera angles. The activities with similar body movements are misclassified in some of the instances for both of the cases. For example, most of the misclassified instances of Run (C20) activity are classified as RunTurn90Left (C8) activity and vice versa. It happens because both the actions have similar body movements.

Tables 5 and 6 list the $Precision$, $Recall$, $Specificity$, $F1\ score$ and $Accuracy$ that are obtained in the first experiment for both the cases (Case 1 and Case 2) on two different sets of camera angles (27° *and* 45°). By analysing the results presented in Tables 5 and 6, it is observed that average $Precision$ is 96.67% & 96.83%, average $Recall$ is 96.67% & 96.83%, average $Specificity$ is 99.63% & 99.65%, average $F_1\ score$ is 96.65% & 96.83% and average $Accuracy$ is 99.33% & 99.37% with first set of cameras. On the other hand, with second set of cameras, average $Precision$ is 92.30% & 95.25%, average $Recall$ is 92.00% & 95.17%, average $Specificity$ is 99.11% & 99.46% , average $F1\ score$ is 92.04%

Table 6 Performance Metrics of Experiment #1 for the silhouettes captured using cameras with a slant angle of 45°

		TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	C2	116	08	93.55	96.67	99.26	95.08	99.00
	C3	114	13	89.76	95.00	98.80	92.31	98.42
	C4	110	07	94.02	91.67	99.35	92.83	98.58
	C5	111	07	94.07	92.50	99.35	93.28	98.67
	C6	114	00	100.00	95.00	100.00	97.44	99.50
	C7	104	09	92.04	86.67	99.17	89.27	97.92
	C8	108	04	96.43	90.00	99.63	93.10	98.62
	C18	105	06	94.59	87.50	99.44	90.91	98.25
	C19	114	26	81.43	95.00	97.59	87.69	97.33
	C20	108	16	87.10	90.00	98.52	88.52	97.67
Case 2	C2	119	04	96.75	99.17	99.63	97.94	99.58
	C3	120	02	98.36	100.0	99.81	99.17	99.83
	C4	118	05	95.94	98.33	99.54	97.12	99.42
	C5	110	01	99.10	91.67	99.91	95.24	99.08
	C6	118	01	99.16	98.33	99.91	98.74	99.75
	C7	107	07	93.86	89.17	99.35	91.45	98.33
	C8	113	07	94.17	94.17	99.35	94.17	98.33
	C18	108	08	93.10	90.00	99.26	91.53	98.33
	C19	117	05	95.90	97.50	99.54	96.69	99.33
	C20	112	18	86.15	93.33	98.33	89.60	97.83

& 95.17% and average $Accuracy$ is 98.40% & 99.03%. Following conclusions can be made from this experiment:

- Average precision, recall, specificity, F_1 score and accuracy are higher in Case 2 (when Random Forest classifier is trained with $OISF$ feature) in comparison to Case 1 (when Random Forest classifier is trained with $FV1$). It confirms the effectiveness of the proposed $OISF$ feature.
- If the difference calculated between two sets of camera angles (slant angle of cameras is 27° and 45°) in Case 1 and Case 2, the minimum variation is observed in the accuracy of Case 2. This observation further confirms that $OISF$ is least dependent on camera positioning.
- High F_1 score and high accuracy confirms the applicability of the proposed work for human activity recognition.

4.3 Experiment #2

To perform this experiment, combination of ten similar and dissimilar activities (C1: HangOnBar, C2: JumpGetOnBar, C3: JumpOverObject, C4: JumpFromObject, C5:

Table 7 Confusion Matrix of Experiment #2 for the silhouettes captured using cameras with a slant angle of 27°

		C1	C2	C3	C4	C5	C6	C11	C12	C16	C17
Case 1	C1	**1.00** 0	0	0	0	0	0	0	0	0	
	C2	0	**1.00**	0	0	0	0	0	0	0	0
	C3	0	0	**0.95**	0	0	0	0	0.03	0.02	0
	C4	0	0	0	**1.00**	0	0	0	0	0	0
	C5	0	0	0	0	**0.97**	0.02	0	0	0.02	0
	C6	0	0	0	0	0	**1.00**	0	0	0	0
	C11	0	0	0	0	0	0	**0.98**	0	0.02	0
	C12	0.02	0	0.02	0	0	0	0	**0.95**	0.02	0
	C16	0	0	0	0	0	0	0	0	**1.00**	0
	C17	0	0	0	0	0	0.02	0	0.02	0	**0.97**
Case 2	C1	**1.00**	0	0	0	0	0	0	0	0	0
	C2	0	**0.97**	0.02	0	0	0	0	0	0.02	0
	C3	0	0	**1.00**	0	0	0	0	0	0	0
	C4	0	0	0	**1.00**	0	0	0	0	0	0
	C5	0	0	0.02	0	**0.98**	0	0	0	0	0
	C6	0	0	0	0	0	**1.00**	0	0	0	0
	C11	0	0	0	0.02	0	0	**0.98**	0	0	0
	C12	0	0.02	0	0.02	0	0	0	**0.97**	0	0
	C16	0	0	0	0	0	0	0	0	**1.00**	0
	C17	0	0	0	0	0	0	0	0	0	**1.00**

Table 8 Confusion Matrix of Experiment #2 for the silhouettes captured using cameras with a slant angle of 45°

		C1	C2	C3	C4	C5	C6	C11	C12	C16	C17
Case 1	C1	**0.99**	0	0	0.01	0	0	0	0	0	0
	C2	0.03	**0.95**	0.02	0.01	0	0	0	0	0	0
	C3	0	0.01	**0.92**	0	0.01	0	0	0.03	0.02	0.03
	C4	0	0.03	0.01	**0.91**	0	0	0	0	0.05	0
	C5	0.03	0	0.01	0.01	**0.90**	0.02	0	0.01	0.01	0.03
	C6	0.01	0.03	0	0.01	0.01	**0.89**	0.01	0.01	0.03	0.01
	C11	0	0	0	0	0	0	**0.95**	0.02	0.03	0.01
	C12	0	0	0	0.01	00.1	0	0	**0.97**	0	0.02
	C16	0	0.01	0	0	0	0	0	0.01	**0.98**	0
	C17	0	0	0	0.02	0	0	0	0	0	**0.98**
Case 2	C1	**1.00**	0	0	0	0	0	0	0	0	0
	C2	0.01	**0.99**	0	0	0	0	0	0	0	0
	C3	0	0	**1.00**	0	0	0	0	0	0	0
	C4	0	0.02	0.01	**0.97**	0	0	0	0	0	0.01
	C5	0	0.01	0	0	**0.97**	0.03	0	0	0	0
	C6	0	0.03	0	0	0	**0.95**	0	0	0	0.02
	C11	0.01	0	0	0.02	0	0	**0.91**	0.07	0	0
	C12	0	0	0	0.02	0	0	0.07	**0.92**	0	0
	C16	0	0	0	0	0	0	0	0	**1.00**	0
	C17	0	0	0	0	0	0.01	0	0	0	**0.99**

Table 9 Performance Metrics of Experiment #2 for the silhouettes captured using cameras with a slant angle of 27°

		TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	C1	120	02	98.36	100.0	99.81	99.17	99.83
	C2	120	00	100.00	100.0	100.00	100.00	100.0
	C3	114	02	98.28	95.00	99.81	96.61	99.33
	C4	120	00	100.00	100.0	100.00	100.00	100.0
	C5	116	00	100.00	96.67	100.00	98.31	99.67
	C6	120	04	96.77	100.0	99.63	98.36	99.67
	C11	118	00	100.00	98.33	100.00	99.16	99.83
	C12	114	06	95.00	95.00	99.44	95.00	99.00
	C16	120	08	93.75	100.0	99.26	96.77	99.33
	C17	116	00	100.00	96.67	100.00	98.31	99.67
Case 2	C1	120	00	100.00	100.0	100.00	100.00	100.0
	C2	116	02	98.31	96.67	99.81	97.48	99.50
	C3	120	04	96.77	100.0	99.63	98.36	99.67
	C4	120	04	96.77	100.0	99.63	98.36	99.67
	C5	118	00	100.00	98.33	100.00	99.16	99.83
	C6	120	00	100.00	100.0	100.00	10000	100.0
	C11	118	00	100.00	98.33	100.00	99.16	99.83
	C12	116	00	100.00	96.67	100.00	98.31	99.67
	C16	120	02	98.36	100.0	99.81	99.17	99.83
	C17	120	00	100.00	100.0	100.00	100.00	100.0

RunPullObject, C6: RunPushObject, C11: KnockoutSpin, C12: Knockout, C16: Punch and C17: JumpKick) grouped in SA_2 (Table 1) for jump, run and knockout categories are taken from the ViHASi dataset.

Tables 7 and 8 present the confusion matrices that are obtained in the second experiment for both the cases (Case 1 and Case 2) on two different sets of camera angles. Tables 9 and 10 show the $Precision$, $Recall$, $Specificity$, $F_1\ score$ and $Accuracy$ that are obtained in the second experiment for both the cases (Case 1 and Case 2) on two different sets of camera angles. After evaluating the results of experiment #2, it is observed that average $Precision$ is 98.22% & 99.02%, average $Recall$ is 98.17% & 99.00%, average $Specificity$ is 99.79% & 99.89%, average $F1\ score$ is 98.16% & 98.99%, and average $Accuracy$ is 99.63% & 99.80% with first set of cameras, whereas average $Precision$ is 94.60% & 96.92%, average $Recall$ is 94.42% & 96.92%, average $Specificity$ is 99.37% & 99.66%, average $F_1\ score$ is 94.41% & 96.90% and average $Accuracy$ is 98.99% & 99.38% with second set of cameras. In Case 2, the average precision, recall, specificty, F_1 score and accuracy are relatively high when compared to that of Case 1. This again proves the effectiveness of $OISF$ feature.

Table 10 Performance Metrics of Experiment #2 for the silhouettes captured using cameras with a slant angle of 45°

		TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	C1	119	07	94.44	99.17	99.35	96.75	99.33
	C2	114	09	92.68	95.00	99.17	93.83	98.75
	C3	110	04	96.49	91.67	99.63	94.02	98.83
	C4	109	07	93.97	90.83	99.35	92.37	98.50
	C5	108	03	97.30	90.00	99.72	93.51	98.75
	C6	107	02	98.17	89.17	99.81	93.45	98.75
	C11	114	01	99.13	95.00	99.91	97.02	99.42
	C12	116	08	93.55	96.67	99.26	95.08	99.00
	C16	118	16	88.06	98.33	98.52	92.91	98.50
	C17	118	10	92.19	98.33	99.07	95.16	99.00
Case 2	C1	120	02	98.36	100.0	99.81	99.17	99.83
	C2	119	07	94.44	99.17	99.35	96.75	99.33
	C3	120	01	99.17	100.0	99.91	99.59	99.92
	C4	116	04	96.67	96.67	99.63	96.67	99.33
	C5	116	00	100.00	96.67	100.00	98.31	99.67
	C6	114	04	96.61	95.00	99.63	95.80	99.17
	C11	109	08	93.16	90.83	99.26	91.98	98.42
	C12	110	08	93.22	91.67	99.26	92.44	98.50
	C16	120	00	100.00	100.0	100.00	100.00	100.0
	C17	119	03	97.54	99.17	99.72	98.35	99.67

Table 11 Confusion Matrix of Experiment #3 for the silhouettes captured using cameras with a slant angle of 27°

		C1	C9	C10	C12	C13	C14	C15	C16	C17	C18
Case 1	C1	**1.00**	0	0	0	0	0	0	0	0	0
	C9	0	**1.00**	0	0	0	0	0	0	0	0
	C10	0	0	**1.00**	0	0	0	0	0	0	0
	C12	0	0	0	**0.98**	0.02	0	0	0	0	0
	C13	0	0	0	0	**1.00**	0	0	0	0	0
	C14	0	0	0	0.02	0.03	**0.95**	0	0	0	0
	C15	0	0	0	0.02	0	0	**0.97**	0	0.02	0
	C16	0.02	0	0	0	0	0	0	**0.98**	0	0
	C17	0	0	0	0	0	0	0	0	**1.00**	0
	C18	0	0.02	0	0	0	0	0	0	0	**0.98**
Case 2	C1	**1.00**	0	0	0	0	0	0	0	0	0
	C9	0	**1.00**	0	0	0	0	0	0	0	0
	C10	0	0	**1.00**	0	0	0	0	0	0	0
	C12	0	0	0	**1.00**	0	0	0	0	0	0
	C13	0	0	0	0	**1.00**	0	0	0	0	0
	C14	0	0.03	0	0	0	**0.95**	0.02	0	0	0
	C15	0	0	0	0	0	0	**1.00**	0	0	0
	C16	0	0	0	0	0	0	0	**1.00**	0	0
	C17	0	0	0	0	0	0	0	0	**1.00**	0
	C18	0	0	0	0	0	0.03	0	0	0	**0.97**

4.4 Experiment #3

To perform this experiment, ten dissimilar activities of different categories (C1: HangOnBar, C9: HeroSmash, C10: HeroDoorSlam, C12: Knockout, C13: Granade, C14: Collapse, C15: StandLookAround, C16: Punch, C17: JumpKick and C18: Walk) grouped in SA_3 (Table 1) are taken from the ViHASi dataset.

Tables 11 and 12 show the confusion matrices that are obtained in the third experiment for both the cases (Case 1 and Case 2) on two different sets of camera angles. By comparing all the confusion matrices obtained in experiment #1, experiment #2, and experiment #3, following conclusions can be drawn:

- Probability of misclassification among similar activities is higher than the probability of misclassification among dissimilar activities.
- Misclassification rate of the activities captured by the second set of cameras is higher than the activities captured by first set of cameras.
- Average misclassification rate in Case 2 is less than the average misclassification rate in Case 1. It proves the appropriateness of $OISF$ feature for human activity recognition.

Tables 13 and 14 list the values of $Precision$, $Recall$, $Specificity$, $F_1\ score$ and $Accuracy$ that are obtained in the third experiment for both the cases (Case 1 and Case 2)

Table 12 Confusion Matrix of Experiment #3 for the silhouettes captured using cameras with a slant angle of 45°

		C1	C9	C10	C12	C13	C14	C15	C16	C17	C18
Case 1	C1	**0.99**	0	0	0	0	0	0	0	0	0.01
	C9	0.01	**0.94**	0	0.01	0.02	0.01	0	0	0.02	0
	C10	0	0	**0.97**	0	0	0.03	0	0	0	0.01
	C12	0	0	0	**0.91**	0	0.02	0.04	0.03	0.01	0
	C13	0	0	0	0	**0.96**	0.01	0	0	0.03	0
	C14	0	0.02	0	0	0	**0.93**	0.03	0	0	0.02
	C15	0.01	0	0.01	0.01	0	0.04	**0.91**	0	0	0.03
	C16	0.03	0.01	0	0	0	0	0.01	**0.92**	0	0.03
	C17	0	0	0	0.01	0.03	0	0	0	**0.97**	0
	C18	0.01	0	0	0	0	0	0.01	0.01	0	**0.98**
Case 2	C1	**1.00**	0	0	0	0	0	0	0	0	0
	C9	0	**0.99**	0	0	0	0.01	0	0	0	0
	C10	0	0	**0.99**	0	0	0	0	0.01	0	0
	C12	0	0	0	**0.98**	0	0.02	0	0	0	0
	C13	0	0	0	0	**0.96**	0	0	0.03	0	0.01
	C14	0	0.01	0	0.03	0	**0.95**	0	0	0.01	0.01
	C15	0.01	0	0.01	0	0	0	**0.98**	0.01	0	0
	C16	0	0	0	0	0	0	0	**1.00**	0	0
	C17	0.02	0	0	0	0	0	0	0	**0.98**	0
	C18	0.02	0	0.01	0	0	0	0	0.01	0	**0.97**

Table 13 Performance Metrics of Experiment #3 for the silhouettes captured using cameras with a slant angle of 27°

		TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	C1	120	02	98.36	100.0	99.81	99.17	99.83
	C9	120	02	98.36	100.0	99.81	99.17	99.83
	C10	120	00	100.00	100.0	100.00	100.00	100.00
	C12	118	04	96.72	98.33	99.63	97.52	99.50
	C13	120	06	95.23	100.0	99.44	97.56	99.50
	C14	114	00	100.00	95.00	100.00	97.44	99.50
	C15	116	00	100.00	96.67	100.00	98.31	99.67
	C16	118	00	100.00	98.33	100.00	99.16	99.83
	C17	120	02	98.36	100.0	99.81	99.17	99.83
	C18	118	00	100.00	98.33	100.00	99.16	99.83
Case2	C1	120	00	100.00	100.0	100.00	100.00	100.0
	C9	120	04	96.77	100.0	99.63	98.36	96.67
	C10	120	00	100.00	100.0	100.00	100.00	100.0
	C12	120	00	100.00	100.0	100.00	100.00	100.0
	C13	120	00	100.00	100.0	100.00	100.00	100.0
	C14	114	04	96.61	95.00	99.63	95.80	99.17
	C15	120	02	98.36	100.0	99.81	99.17	99.83
	C16	120	00	100.00	100.0	100.00	100.00	100.0
	C17	120	00	100.00	100.0	100.00	100.00	100.0
	C18	116	00	100.00	96.67	100.00	98.31	99.67

Table 14 Performance Metrics of Experiment #3 for the silhouettes captured using cameras with a slant angle of 45°

		TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	C1	119	07	94.44	99.17	99.35	96.75	99.33
	C9	113	03	97.41	94.17	99.72	95.76	99.17
	C10	116	01	99.15	96.67	99.91	97.89	99.58
	C12	109	03	97.32	90.83	99.72	93.97	98.83
	C13	115	05	95.83	95.83	99.54	95.83	99.17
	C4	112	12	90.32	93.33	98.89	91.80	98.33
	C15	109	11	90.83	90.83	98.98	90.83	98.17
	C16	110	04	96.49	91.67	99.63	94.02	98.83
	C17	116	07	94.31	96.67	99.35	95.47	99.08
	C18	117	11	91.41	97.50	98.98	94.35	98.83
Case 2	C1	120	05	96.00	100.0	99.54	97.96	99.58
	C9	119	01	99.17	99.17	99.91	99.17	99.83
	C10	119	02	98.35	99.17	99.81	98.76	99.75
	C12	118	03	97.52	98.83	99.72	97.93	99.58
	C13	115	00	100.00	95.83	100.00	97.87	99.58
	C4	114	03	97.44	95.00	99.72	96.20	99.25
	C15	117	00	100.00	97.50	100.00	98.73	99.75
	C16	120	07	94.45	100.0	99.35	97.16	99.42
	C17	118	01	99.16	98.33	99.91	98.74	99.75
	C18	116	02	98.31	96.67	99.81	97.48	99.50

on two different sets of camera angles. By analysing the results presented in Tables 13 and 14, it is observed that average $Precision$ is 98.70% & 99.17%, average $Recall$ is 98.67% & 99.17%, average $Specificity$ is 99.85% & 99.91% , average $F_1\ score$ is 98.66% & 99.16%, and average $Accuracy$ is 99.73% & 99.83% with first set of cameras. It is also observed from this table that average $Precision$ is 94.75% & 98.04%, average $Recall$ is 94.67% & 98.00%, average $Specificity$ is 99.40% & 99.78% , average $F_1\ score$ is 94.67% & 98.00%, and average $Accuracy$ is 98.93% & 99.60% with second set of cameras. From this experiment, it can be concluded that the average accuracy is more than 99% for all the dissimilar activities taken from ViHASi dataset. This shows that utilizing $OISF$ feature for human activity recognition gives effective results in terms of activity classification both for similar and dissimilar activities.

4.5 Experiment #4

To perform this experiment, all the six activities (Boxing, Hand clapping, Hand waving, Jogging, Running and Walking) are taken from the KTH dataset. Figure 9a, b and c show the ten sample frames of silhouettes of $Hand\ clapping$ activity, skeletons obtained using Algorithm 1 and 8-regions division of elliptical bounding box using Algorithm 2 respectively.

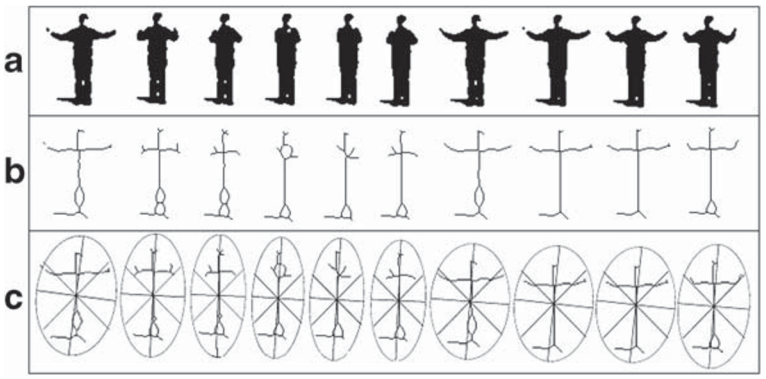


Fig. 9 (**a**) Extracted silhouettes from the frames of Hand clapping activity (**b**) skeletons obtained from silhouettes (**c**) symmetric 8-regions division of RoI

Table 15 shows the confusion matrix of the fourth experiment. Activities that have similar type of body movements such as Running, Jogging and Walking are misclassified in both cases. Table 16 lists the values of $Precision$, $Recall$, $Specificity$, $F_1\ score$ and $Accuracy$ that are obtained in the fourth experiment for Case 1 and Case 2. By examining the results illustrated in Table 16, it is observed that average $Precision$ is 88.79% & 90.81%, average $Recall$ is 88.67% & 90.78%, average $Specificity$ is 97.73% & 98.15% , average $F_1\ score$ is 88.68% & 90.74%, and average $Aaccuracy$ is 96.22% & 96.85% for Case 1 and Case 2 respectively. From this experiment, it is concluded that average precision, recall and accuracy increases in Case 2 of the experiment with respect to Case 1 of the experiment. It can be further concluded from these results that Random Forest classifier when trained with newly proposed feature $OISF$ performs well for all types of activity (similar or dissimilar).

Table 15 Confusion Matrix of the Experiment #4

		Boxing	H.Clapping	H.Waving	Jogging	Runnig	Walking
Case 1	Boxing	**0.97**	0	0	0.01	0	0.01
	H.Clapping	0	**0.89**	0.07	0.01	0.01	0.01
	H.Waving	0.01	0.05	**0.90**	0.01	0.01	0.01
	Jogging	0.02	0	0	**0.86**	0.07	0.05
	Running	0.01	0.01	0.01	0.05	**0.81**	0.11
	Walking	0	0	0.01	0.05	0.05	**0.89**
Case 2	Boxing	**1.00**	0	0	0	0	0
	H.Clapping	0.01	**0.96**	0.03	0	0	0
	H.Waving	0.03	0.03	**0.93**	0	0	0
	Jogging	0	0.01	0.01	**0.84**	0.05	0.09
	Running	0	0.01	0.01	0.09	**0.83**	0.07
	Walking	0	0.01	0	0.08	0.03	**0.89**

Table 16 Performance Metrics of the Experiment #4

	Activities	TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	Boxing	146	06	96.05	97.33	99.20	96.69	98.89
	HandClapping	133	09	93.66	88.67	98.80	91.10	97.11
	HandWaving	135	14	90.60	90.00	98.13	90.30	96.78
	Jogging	129	20	86.58	86.00	97.33	86.29	95.44
	Running	122	22	84.72	81.33	97.07	82.99	94.44
	Walking	133	31	81.10	88.67	95.87	84.71	94.67
Case 2	Boxing	150	07	95.54	100.0	99.07	97.72	99.22
	HandClapping	144	08	94.74	96.00	98.93	95.36	98.44
	HandWaving	140	07	95.24	93.33	99.07	94.28	98.11
	Jogging	126	26	82.90	84.00	96.53	83.44	94.44
	Running	124	12	91.18	82.67	98.40	86.71	95.78
	Walking	133	23	85.26	88.67	96.92	86.93	95.11

4.6 Experiment #5

To perform this experiment, all the five activities (Boxing, Hand clapping, Hand waving, Jogging, and Walking) are taken from the in-house dataset. Figure 10a, b, c and d show the ten input sample frames of *Boxing* activity, respective silhouettes and skeletons obtained by using Algorithm 1, and 8-regions division of elliptical bounding box obtained by using Algorithm 2.

Table 17 shows the confusion matrix of the fifth experiment. In the complex background also, our method results in high accuracy and minimum false classification for similar types of activities. Table 18 presents the values of $Precision$, $Recall$, $Specificity$, $F_1\ score$ and $Accuracy$ that are obtained in the fifth experiment for Case 1 and Case 2. By examining the results illustrated in Table 18, it is observed that average $Precision$ is 94.09% & 95.91%,

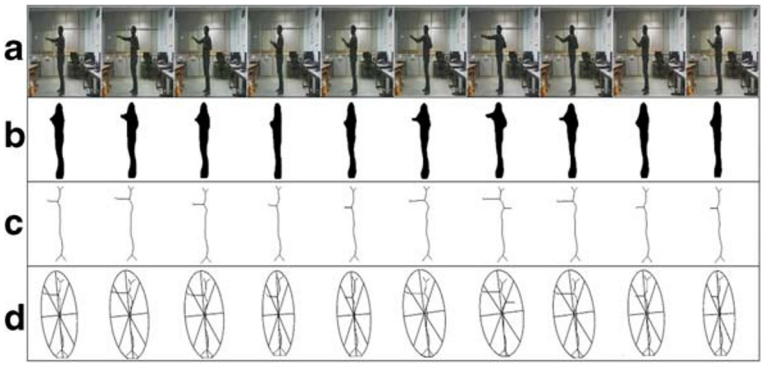


Fig. 10 (**a**) Input frames of boxing activity (**b**) extracted silhouettes from the frames (**c**) skeletons obtained from silhouettes (**d**) symmetric 8-regions division of RoI

Table 17 Confusion Matrix of the Experiment #5

		Boxing	HandClapping	HandWaving	Jogging	Walking
Case 1	Boxing	**1.00**	0	0	0	0
	HandClapping	0.01	**0.90**	0.03	0.01	0.06
	HandWaving	0	0.01	**0.98**	0	0.01
	Jogging	0	0.03	0	**0.90**	0.07
	Walking	0	0	0	0.08	**0.92**
Case 2	Boxing	**1.00**	0	0	0	0
	HandClapping	0	**0.98**	0	0.01	0.01
	HandWaving	0	0	**0.99**	0	0.01
	Jogging	0	0.01	0	**0.91**	0.08
	Walking	0	0	0	0.08	**0.92**

average $Recall$ is 94.00% & 95.86%, average $Specificity$ is 98.50% & 98.97% , average $F_1\ score$ is 94.01% & 95.88%, and average $Accuracy$ is 97.60% & 98.35% for Case 1 and Case 2 respectively. From this experiment, it is concluded that average precision, average recall and average accuracy increases in Case 2 of the experiment with respect to Case 1 of the experiment. It can be further concluded from these results that Random Forest classifier when trained with newly proposed feature $OISF$ performs well for all types of activity (similar or dissimilar) even in the complex background.

4.7 Effectiveness analysis of the proposed approach

The proposed approach has been tested on three datasets having different characteristics such as videos recorded with a low-resolution camera, complex background, variation in illumination, outdoor (with varying variations of scale) and indoor scenarios in day vision, and different view angles. Figure 11 depicts the average accuracy obtained in the experi-

Table 18 Performance Metrics of the Experiment #5

	Activities	TP	FP	Precision (%)	Recall (%)	Specificity (%)	F_1 score (%)	Accuracy (%)
Case 1	Boxing	150	01	99.34	100.00	99.83	99.67	99.86
	HandClapping	135	06	95.75	90.00	99.00	92.78	97.20
	HandWaving	147	04	97.35	98.00	99.33	97.67	99.06
	Jogging	135	12	91.22	90.00	97.83	90.60	96.27
	Walking	138	21	86.79	92.00	96.50	89.32	95.60
Case 2	Boxing	150	00	100.00	100.00	100.00	100.00	100.00
	HandClapping	147	02	98.66	98.00	99.67	98.33	99.33
	HandWaving	148	00	100.00	98.67	100.00	99.33	99.73
	Jogging	136	13	91.28	90.67	97.83	90.97	96.40
	Walking	138	16	89.61	92.00	97.33	90.79	96.26

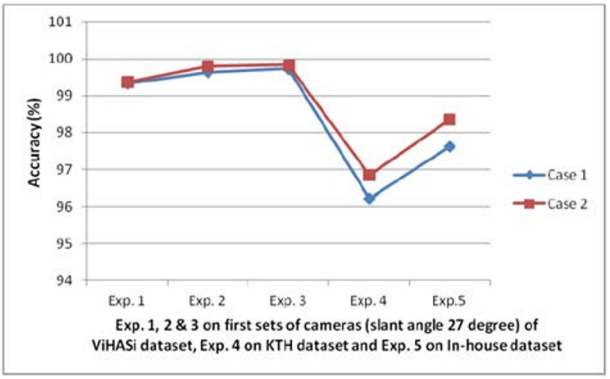


Fig. 11 Recognition accuracy (%) of experiments #1, #2 & #3 of ViHASi dataset at slant angle 27°, experiment #4 of KTH dataset and experiment #5 of in-house dataset

ments #1, #2 & #3 performed on ViHASi dataset recorded from the cameras of slant angle 27°. This figure also shows the average accuracy obtained in experiment #4 on KTH dataset and experiment #5 on in-house dataset. Figure 12 shows the average accuracy obtained in the first three experiments with cameras slant angle 45°.

The x-axis and y-axis of these graphs represent experiment numbers and average accuracy, respectively. It can be observed from the graphs shown in Figs. 11 and 12 that average accuracy greatly varies when $FV1$ feature is used to train the Random Forest classifier whereas, it remains consistent with the use of $OISF$ feature. Figure 13 depicts absolute difference between the two sets of camera angles (27° and 45°).

The x-axis and y-axis of the graph in Fig. 13 represents experiment numbers and absolute difference in accuracies, respectively. Minimum variation in the accuracy may be observed in Case 2 for the experiments #1, #2 & #3 which shows that $OISF$ feature is invariant towards the orientation of the camera. Apart from all these comparisons, when this feature is used to train Random Forest classifier for human activity recognition $\approx$ 97% of accuracy is achieved on both static and moving cameras.

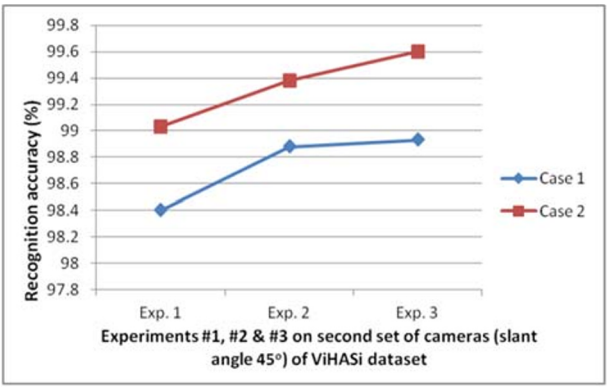


Fig. 12 Recognition accuracy (%) of experiments #1, #2 & #3 of ViHASi dataset at slant angle 45° for both cases

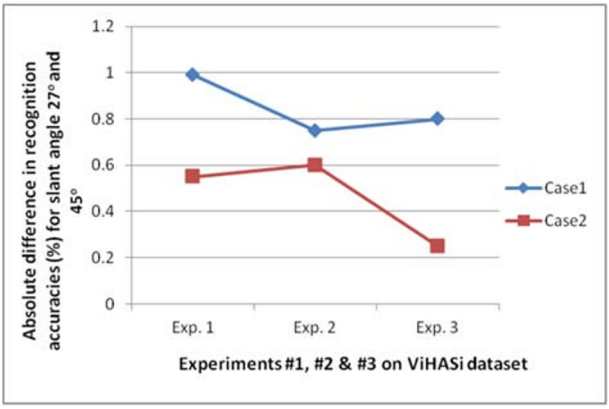


Fig. 13 Difference in recognition accuracies with slant angle 27° and 45° for all the first three experiments of ViHASi dataset

Average accuracy for all the experiments varies from $\approx$ 97% to $\approx$ 99%. Through these results, we can say that the proposed method is capable to deal with scenarios like low resolution, complex background, etc. From experimental results, it is observed that variation in recognition accuracy is small ($\approx$ 2%) for all of the experiments, which confirms the robustness of our proposed method.

FV1 and OISF feature extraction time are shown in Table 19 for all the five experiments. From Table 8, it is observed that the average feature extraction rate of FV1 and OISF features are 38 frames per second (fps) and 34 frames per second (fps) respectively. Training and Testing time of the random forest model are also recorded but it is found to be static and very less as compared to the feature extraction time. These experiments show that even if the size of a frame is in the range of 480 $\times$ 640, it can be used for the real-time activity recognition task.

5 Comparison of proposed approach with existing state-of-the-art approaches

In this section, average accuracy of the proposed approach is compared with state-of-the-art approaches performed on KTH dataset for human activity recognition.

Table 19 Time analysis of all the experiments

Experiment number	*Number of frames*	*Size of frame*	*FV1 feature extraction time (sec)*	*OISF feature extraction time (sec)*
Experiment #1	12000	480 × 640	302.60	372.10
Experiment #2	12000	480 × 640	332.64	397.55
Experiment #3	12000	480 × 640	355.32	422.90
Experiment #4	9000	120 × 160	185.54	176.32
Experiment #5	7500	120 × 160	178.78	170.96

Table 20 lists the average accuracy of different state-of-the-art approaches and proposed approach. Average accuracy of state-of-the-art approaches is about 94%. Accuracy achieved with $FV1$ feature and $OISF$ feature is 96.22% and 96.85% respectively. The average accuracy achieved by $OISF$ feature is on an average 2.5% higher than the others which validate our proposed method. It can be concluded from this comparison that maximum accuracy can be achieved with $OISF$ feature for human activity recognition.

6 Conclusion

An attempt has been made in this work to introduce a novel and efficient feature termed as $OISF$ that is tested for Human Activity Recognition. To check the robustness of this feature for moving camera, silhouettes from ViHASi dataset that contain videos recorded by different cameras at different angles are taken. Average recognition accuracy of the proposed approach for experiments #1, #2 & #3 are 99.20%, 99.59% and 99.72% respectively. Small variations in recognition accuracy confirm the robustness of the newly proposed $OISF$ feature towards the nature of activities (similar activities or combination of similar and dissimilar activities or dissimilar activities). The recognition accuracy of newly proposed feature $OISF$ is superior to the existing approaches in case of videos for moving camera while its performance is at par with the existing feature in the case of static camera. Experimentally, it is found that overall recognition accuracy of the proposed approach with ViHASi dataset is $\approx$ 99.30%, for KTH dataset is $\approx$ 96.85% and for in-house dataset is $\approx$ 98.34%. In this proposed approach, skeletons are used to extract the features which reduces the processing time of feature extraction. Average feature extraction rate of FV1 and OISF features are 38 frames per second (fps) and 34 frames per second (fps) respectively. Higher accuracies obtained in both the cases prove that the proposed approach is applicable for real-life activities recognition such as patient monitoring, fight detection between persons, etc.

Table 20 Accuracy of state-of-the-art approaches and proposed approach

S. No.	*Name of Author's*	*Feature/Technique used*	*Accuracy achieved(%)*
1.	Lu et al. [20]	Local spatio-temporal distribution descriptor fused with HOG3D	91.50
2.	Naveed et al. [24]	LBP, HOG, Haar wavelets, SIFT, velocity and displacement	92.29
3.	Sadek et al. [29]	Affine-Invariant shape feature	93.50
4.	Wang et al. [36]	Dense trajectories based on motion boundary histograms	94.20
5.	Gilbert et al. [10]	Mined Dense spatio-temporal corner feature	94.50
6.	Kovashka et al. [16]	Space-time neighborhood feature	94.53
7.	Xu et al. [38]	Pose points selection method	95.80
8.	Case 1	Feature $FV1$	**96.22**
9.	Proposed Approach (Case 2)	$OISF$	**96.85**

Appendix: Details of Algorithms

Algorithm A.1 *Construct_skeleton*().

INPUT: V(N): Video V containing N number of frames; B: Background frame
OUTPUT: skeleton[N]: N number of skeletons
Ensure:
(1) $c1$ & $c2$ are two constant depend on the dataset.
(2) $c3$ & θ are the length and the angle of the line from horizontal axis respectively.
(3) F, bw, ew, tw are two dimensional matrix represents image.
(4) $k1$, $k2$ & $k3$ times dilation, erosion & thinning operations counts respectively.
(5) $\hat{(.)}$ is used for complement operation.

Procedure $constructskeleton(V)$
$F_1, F_2, ... F_N$ = video2frame(V) ▷ Converting input video into sequence of frames
B = rgb2gray(B) ▷ Converting background frame from rgb scale to gray scale
for $i \leftarrow 1$ to N **do**
  F_i = rgb2gray(F_i) ▷ Converting input frame from rgb scale to gray scale
  F_i = backgroundSubtraction(F_i, B) ▷ Extracting forground object
  $h[freq][0...255] = imhist(g)$ ▷ Finding the histogram of gray frame
  $[h_m, h_j] = max(h)$ ▷ h_m and h_j contains maximum value within h and respective index
  $th1 = h_j$ - $c1$
  $th2 = h_j + c2$
  [m,n]=size(F_i) ▷ It returns size of matrix F
  for $l1 \leftarrow 1$ to m **do**
    for $l2 \leftarrow 1$ to n **do**
      if ($th1 \leq g[l1][l2] \leq th2$) **then** ▷ Converting gray scale to binary scale
        $g[l1][l2] \leftarrow 1$
      else
        $g[l1][l2] \leftarrow 0$
      end if
    end for
  end for
  $b \leftarrow g$ ▷ b is binary image
  bw = medfilt2(b) ▷ Apply median filter to remove noise
  $bw = \hat{bw}$
  $se = strel('line', c3, \theta)$ ▷ Creating a linear structuring element
  for $k \leftarrow 1$ to $k1$ **do**
    dw = imdilate(bw,se) ▷ Expand binary image bw by structuring element se
  end for
  for $k \leftarrow 1$ to $k2$ **do**
    ew = imerode(dw,se) ▷ Shrink binary image dw by using structuring element se
  end for
  for $k \leftarrow 1$ to $k3$ **do**
    tw = thinimage(ew) ▷ Apply thinning operation to extract skeletion.
    $ew \leftarrow tw$
  end for
  $skeleton(i) \leftarrow ew$
end for
return($skeleton(N)$)
end procedure

Algorithm A.2 RoI selection and it's division in 8-regions.

INPUT: Skeleton S of a frame.
OUTPUT: Co-ordinates of eight symmetric regions of ellipse.
Ensure:
(1) (x_c, y_c) is the centroid co-ordinate of the ellipse.
(2) a is the semi-major axis length of the ellipse.
(3) b is the semi-minor axis length of the ellipse.
(4) θ is the orientation of the ellipse.
(5) x_q^i, y_q^j represent x and y co-ordinates of i^{th} region of the ellipse.

Procedure $boundingEllipse(S)$
Draw an elliptical bounding box around the skeleton S
$x_{mj}^1 = x_c + a \times cos(0°) \times cos(\theta)$ - $b \times sin(0°) \times sin(\theta)$ ▷ Co-ordinates of chords
$x_{mj}^2 = x_c + a \times cos(180°) \times cos(\theta)$ - $b \times sin(180°) \times sin(\theta)$
$y_{mj}^1 = y_c + a \times cos(0°) \times sin(\theta) + b \times sin(0°) \times cos(\theta)$
$y_{mj}^2 = y_c + a \times cos(180°) \times sin(\theta) + b \times sin(180°) \times cos(\theta)$
$line([x_{mj}^1, y_{mj}^1], [x_{mj}^2, y_{mj}^2])$ ▷ Drawing major axis in the ellipse
$x_{mn}^1 = x_c + a \times cos(90°) \times cos(\theta)$ - $b \times sin(90°) \times sin(\theta)$
$x_{mn}^2 = x_c + a \times cos(270°) \times cos(\theta)$ - $b \times sin(270°) \times sin(\theta)$
$y_{mn}^1 = y_c + a \times cos(90°) \times sin(\theta) + b \times sin(90°) \times cos(\theta)$
$y_{mn}^2 = y_c + a \times cos(270°) \times sin(\theta) + b \times sin(270°) \times cos(\theta)$
$line([x_{mn}^1, y_{mn}^1], [x_{mn}^2, y_{mn}^2])$ ▷ Drawing minor axis in the ellipse
$x_{ch1}^1 = x_c + a \times cos(45°) \times cos(\theta)$ - $b \times sin(45°) \times sin(\theta)$
$x_{ch1}^2 = x_c + a \times cos(225°) \times cos(\theta)$ - $b \times sin(225°) \times sin(\theta)$
$y_{ch1}^1 = y_c + a \times cos(45°) \times sin(\theta) + b \times sin(45°) \times cos(\theta)$
$y_{ch1}^2 = y_c + a \times cos(225°) \times sin(\theta) + b \times sin(225°) \times cos(\theta)$
$line([x_{ch1}^1, y_{ch1}^1], [x_{ch1}^2, y_{ch1}^2])$ ▷ Drawing line whose orientation is 45° anticlock-wise from the minor axis in the ellipse
$x_{ch2}^1 = x_c + a \times cos(135°) \times cos(\theta)$ - $b \times sin(135°) \times sin(\theta)$
$x_{ch2}^2 = x_c + a \times cos(315°) \times cos(\theta)$ - $b \times sin(315°) \times sin(\theta)$
$y_{ch2}^1 = y_c + a \times cos(135°) \times sin(\theta) + b \times sin(135°) \times cos(\theta)$
$y_{ch2}^2 = y_c + a \times cos(315°) \times sin(\theta) + b \times sin(315°) \times cos(\theta)$
$line([x_{ch2}^1, y_{ch2}^1], [x_{ch2}^2, y_{ch2}^2])$ ▷ Drawing line whose orientation is 45° clock-wise from the minor axis in the ellipse
$[x_q^1, y_q^1] = ([x_c, x_{mn}^1, x_{ch2}^1], [y_c, y_{mn}^1, y_{ch2}^1])$
$[x_q^2, y_q^2] = ([x_c, x_{ch2}^1, x_{mj}^2], [y_c, y_{ch2}^1, y_{mj}^2])$
$[x_q^3, y_q^3] = ([x_c, x_{mj}^2, x_{ch1}^2], [y_c, y_{mj}^2, y_{ch1}^2])$
$[x_q^4, y_q^4] = ([x_c, x_{ch1}^2, x_{mn}^2], [y_c, y_{ch1}^2, y_{mn}^2])$
$[x_q^5, y_q^5] = ([x_c, x_{mn}^2, x_{ch2}^2], [y_c, y_{mn}^2, y_{ch2}^2])$
$[x_q^6, y_q^6] = ([x_c, x_{ch2}^2, x_{mj}^1,], [y_c, y^2ch2, y_{mj}^1])$
$[x_q^7, y_q^7] = ([x_c, x_{mj}^1, x_{ch1}^1], [y_c, y_{mj}^1, y_{ch1}^1])$
$[x_q^8, y_q^8] = ([x_c, x_{ch1}^1, x_{mn}^1,], [y_c, y_{ch1}^1, y_{mn}^1])$
return$([x_q^1 ... x_q^8], [y_q^1 ... y_q^8])$
end procedure

Algorithm A.3 FV1 feature extraction.

INPUT: skeleton[N]
OUTPUT: Feature Matrix FV1 $[\frac{N}{10}][80]$
Ensure:
(1) size (.) is a function that returns height (h) and width (w) of input matrix
(2) sum(.) is a function that returns addition of all elements in the vector

Procedure $feature1(skeleton[N])$
for $i \leftarrow 1$ to N **do**
  $[x_i^1 \ldots x_i^8],[y_i^1 \ldots y_i^8] = boundingEllipse(skeleton(i))$
  $[h, w] = size(skeleton(i))$
  $p^t = 0$
  for $j \leftarrow 1$ to h **do**
    for $k \leftarrow 1$ to w **do**
      $p^t = p^t + skeleton(i)[j][k]$
    end for
  end for
  for $l \leftarrow 1$ to 8 **do**
    $bw[h][w] = poly2mask(x_i^l, y_i^l, h, w)$ ▷ Construct binary mask
    for $j \leftarrow 1$ to h **do**
      for $k \leftarrow 1$ to w **do**
        $bw_m(j)(k) = bw(j)(k) \times skeleton(i)[j][k]$
      end for
    end for
    $p^r = sum(bw_m)$
    $f1(i)(l) = \frac{p^r}{p^t}$
  end for
end for
$count1 \leftarrow 0$
$count2 \leftarrow 1$
for $i \leftarrow 1$ to $\frac{N}{10}$ **do**
  for $j \leftarrow 1$ to 10 **do**
    $FV1(count2)(count1 \times 8 + 1 ... (count1 + 1) \times 8) \leftarrow f1(count1)(1...8)$
    $count1 \leftarrow count1 + 1$
  end for
  $count2 \leftarrow count2 + 1$
end for
return$(Feature\ Matrix FV1[\frac{N}{10}][80])$
end procedure

Algorithm A.4 OISF feature extraction

INPUT: skeleton[N]
OUTPUT: $OISF[\frac{N}{10}][80]$: Orientation Invariant Skeleton Feature (OISF)
Ensure:
(1) size(.) is a function that returns height (h) and width (w) of an image matrix
(2) abs is a function that returns absolute value of a number
(3) f_x^1 to f_x^4 and f_y^1 to f_y^4 are the OISF features of one frame w.r.t. x and y coordinates separately.

Procedure $feature2(skeleton[N])$
for $i \leftarrow 1$ to N **do**
  $[x_i^1 \ldots x_i^8],[y_i^1 \ldots y_i^8] = boundingellipse(skeleton(i))$
  $[h,w] = size(skeleton(i))$
  for $l \leftarrow 1$ to 8 **do**
    $bw[h][w] = poly2mask(x_i^l, y_i^l, h, w)$ ▷ Construct binary mask
    for $j \leftarrow 1$ to h **do**
      for $k \leftarrow 1$ to w **do**
        $bw_m(j)(k) = bw(j)(k) \times skeleton(i)[j][k]$ ▷ Masking skeleton except its l^{th} region
      end for
    end for
    for $j \leftarrow 1$ to w **do**
      for $k \leftarrow 1$ to h **do**
        if $((1 \leq l \leq 2)$ || $(7 \leq l \leq 8))$ **then**
          if $(bw_m(k)(j)$ equals to 1) **then**
            $xd(l) \leftarrow j$
            $yd(l) \leftarrow k$
            break;
          end if
        else
          if $(bw_m(k)(w-j)$ equals to 1) **then**
            $xd(l) \leftarrow w - j$
            $yd(l) \leftarrow k$
            break;
          end if
        end if
      end for
    end for
  end for
  $f_x^1 = abs(xd(1) - xd(4))$
  $f_x^2 = abs(xd(2) - xd(3))$
  $f_x^3 = abs(xd(5) - xd(8))$
  $f_x^4 = abs(xd(6) - xd(7))$
  $f_y^1 = abs(yd(1) - yd(4))$
  $f_y^2 = abs(yd(2) - yd(3))$
  $f_y^3 = abs(yd(5) - yd(8))$
  $f_y^4 = abs(yd(6) - yd(7))$
  $f2(i)[1...8] = [f_x^1, f_x^2, f_x^3, f_x^4, f_y^1, f_y^2, f_y^3, f_y^4]$
end for
$count1 \leftarrow 0$
$count2 \leftarrow 1$
for $i \leftarrow 1$ to $\frac{N}{10}$ **do**
  for $j \leftarrow 1$ to 10 **do**
    $FV1(count2)(count1 \times 8 + 1 \ldots (count1 + 1) \times 8) \leftarrow f2(count1)(1...8)$
    $count1 \leftarrow count1 + 1$
  end for
  $count2 \leftarrow count2 + 1$
end for
return$(Feature\ Matrix\ OISF[\frac{N}{10}][80])$
end procedure

References

1. Agarwal JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv (CSUR) 43(3): 1–43
2. Anjum ML, Rosa S, Bona B (2017) Tracking a subset of skeleton joints: an effective approach towards complex human activity recognition. Journal of Robotics
3. Bächlin M, Forster K, Troster G (2009) SwimMaster: a wearable assistant for swimmer. In: Proceedings of the 11th international conference on ubiquitous computing, pp 215–224
4. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
5. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
6. Chen MY, Hauptmann A (2009) Mosift: recognizing human actions in surveillance videos. Citeseer
7. Singh DK, Kushwaha DS (2016) Tracking movements of humans in a real-time surveillance scene. In: Proceedings of fifth international conference on soft computing for problem solving, pp 491–500
8. Dawn DD, Shaikh SH (2016) A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis Comput Springer 32(3):289–306
9. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1110–1118
10. Gilbert A, Illingworth J, Bowden R (2009) Fast realistic multi-action recognition using mined dense spatio-temporal features. In: 2009 IEEE 12th international conference on computer vision, pp 925–931
11. Hbali Y, Hbali S, Ballihi L, Sadgal M (2017) Skeleton-based human activity recognition for elderly monitoring systems. IET Comput Vis 12(1):16–26
12. Ho TK (1995) Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol 1. IEEE, pp 278–282
13. Jalal A, Uddin MZ, Kim JT, Kim TS (2012) Recognition of human home activities via depth silhouettes and R transformation for smart homes. Indoor Built Environ 21(1):184–190
14. Jalal A, Kamal S, Kim D (2017) A depth video-based human detection and activity recognition using multi-features and embedded hidden Markov models for health care monitoring systems. Int J Interact Multimed Artif Intell 4:4
15. Jalaland A, Kamal S (2014) Real-time life logging via a depth silhouette-based human activity recognition system for smart home services. In: 2014 11th IEEE International conference on advanced video and signal based surveillance (AVSS), pp 74–80
16. Kovashka A, Grauman K (2010) Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 2046–2053
17. Kumar S, Kumar S, Raman B, Sukavanam N (2011) Human action recognition in a wide and complex environment. Real-Time Image Video Process 7871:78710I
18. Lassoued I, Zagrouba E (2018) Human actions recognition: an approach based on stable motion boundary fields. Multimed Tools Appl 77(16):20715–20729
19. Li M, Leung H (2016) Multiview skeletal interaction recognition using active joint interaction graph. IEEE Trans Multimed 18(11):2293–2302
20. Lu M, Zhang L (2014) Action recognition by fusing spatial-temporal appearance and the local distribution of interest points. In: International conference on future computer and communication engineering (ICFCCE 2014)
21. Manresa C, Varona J, Mas R, Perales FJ (2005) Hand tracking and gesture recognition for human-computer interaction. ELCVIA Electron Lett Comput Vis Image Anal 5(3):96–104
22. Manzi A, Fiorini L, Limosani R, Dario P, Cavallo F (2017) Two-person activity recognition using skeleton data. IET Comput Vis 12(1):27–35
23. Min W, Cui H, Rao H, Li ZZ, Yao L (2018) Detection of human falls on furniture using scene analysis based on deep learning and activity characteristics. IEEE Access 6:9324–9335
24. Naveed H, Khan G, Khan AU, Siddiqi A, Khan MUG (2019) Human activity recognition using mixture of heterogeneous features and sequential minimal optimization. Int J Mach Learn Cybern 10(9):2329–2340
25. Ofli F, Chaudhry R, Kurillo G, Vidal R, Bajcsy R (2014) Sequence of the most informative joints (SMIJ): a new representation for human skeletal action recognition. J Vis Commun Image Represent 25(1):24–38
26. Quaid MAK, Jalal A (2019) Wearable sensors based human behavioral pattern recognition using statistical features and reweighted genetic algorithm. Multimed Tools Appl, 1–23
27. Ragheb H, Velastin S, Remagnino P, Ellis T (2008) ViHASi: virtual human action silhouette data for the performance evaluation of silhouette-based action recognition methods. In: Second ACM/IEEE international conference on distributed smart cameras. IEEE, pp 1–10

28. Raptis M, Sigal L (2013) Poselet key-framing: a model for human activity recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 2650–2657
29. Sadek S, Al-Hamadi A, Gerald K, Michaelis B (2013) Affine-invariant feature extraction for activity recognition. ISRN Mach Vis, 2013
30. Schapire RE, Freund Y, Bartlett P, Lee WS (1998) Boosting the margin: a new explanation for the effectiveness of voting methods. Annals Stat 26(5):1651–1686
31. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th international conference on pattern recognition (ICPR). IEEE, pp 32–36
32. Shah H, Chokalingam P, Paluri B, Pradeep N, Raman B (2007) Automated stroke classification in tennis. In: International conference image analysis and recognition, pp 1128–1137
33. Uddin MZ, Lee JJ, Kim TS (2010) Independent shape component-based human activity recognition via hidden Markov model. Appl Intell 33(2):193–206
34. Vats E, Chan CS (2016) Early detection of human actions—a hybrid approach. Appl Soft Comput 46:953–966
35. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp 3551–3558
36. Wang H, Kläser A, Schmid C, Lin-Cheng L (2011) Action recognition by dense trajectories. In: CVPR 2011-IEEE conference on computer vision & pattern recognition, pp 3169–3176
37. Weng Z, Guan Y (2018) Action recognition using length-variable edge trajectory and spatio-temporal motion skeleton descriptor. EURASIP J Image Video Process 2018(1):8
38. Xu K, Jiang X, Sun T (2015) Human activity recognition based on pose points selection. In: 2015 IEEE International conference on image processing (ICIP), pp 2930–2834
39. Zhu C, Sheng W (2011) Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. IEEE Trans Syst Man Cybern-Part A: Syst Humans 41(3):569–573
40. Zhu W, Lan C, Xing J, Zeng W, Li Y, Shen L, Xie X (2016) Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: AAAI Conference on artificial intelligence, p 8

Neelam Dwivedi is pursuing Ph.D. (CSE) from Motilal Nehru National Institute of Technology Allahabad (MNNIT), Prayagraj, India. She has received the M.C.A. degree from JEC, Jabalpur, India in 2009. On the year 2011, she completed her M.Tech. in Computer Science & Engineering from DAVV Indore, India. She has more than 3 years of teaching experience. Her area of interest is Image Processing and currently working in the areas of Gender Recognition, Human Activity Recognition and Fight detection.

Dr. Dushyant Kumar Singh received the B.Tech. and M.Tech. degree in computer science and engineering from AMU Aligarh, India in 2007 and 2010. He received the Ph.D. degree in computer science and engineering from Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India in 2017. Since 2012 he is working as Assistant Professor with Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, India. He has published more than 24 research papers in various International Journals and Conferences. His research interests are in the field of Computer Vision, Image Processing, AI and Machine Learning for Vision, Embedded System Design. Web Page/Home Page: http://www.mnnit.ac.in/images/csedfp/dushantcsed/

Prof. Dharmender Singh Kushwaha received the B.E. degree in computer engineering from University of Pune, India in 1990. He received the M.Tech. and Ph.D. degree in computer science and engineering from Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India in 2007. He was recipient of Gold Medal for his Masters degree. Since 2018 he is working as Professor with Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, India. He is the author of a book titled "Data Structures: A Programming Approach with C" published by PHI, and has published more than 140 research papers in various International Journals and Conferences. His research interest includes Distributed Systems, Service Oriented Architecture, Software Engineering, Data Structure and Image Processing. Web Page/Home Page: http://www.mnnit.ac.in/dsk/basic.php