Human Action Recognition via Depth Maps Body Parts of Action

Adnan Farooq [1], Faisal Farooq [1] and Anh Vu Le [2]
[1] Department of Biomedical Engineering, Kyung Hee University, Suwon, South Korea.
[e-mail: adnanfaroq86@gmail.com]
[2] Optoelectronics Research Group, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[e-mail: leanhvu@tdt.edu.vn]
*Corresponding author: Anh Vu Le

Received April 7, 2017; revised August 14, 2017; accepted December 26, 2017; published May 31, 2018

Abstract

Human actions can be recognized from depth sequences. In the proposed algorithm, we initially construct *depth, motion* maps (DMM) by projecting each depth frame onto three orthogonal Cartesian planes and add the motion energy for each view. The body part of the action (BPoA) is calculated by using bounding box with an optimal window size based on maximum spatial and temporal changes for each DMM. Furthermore, feature vector is constructed by using BPoA for each human action view. In this paper, we employed an ensemble based learning approach called Rotation Forest to recognize different actions Experimental results show that proposed method has significantly outperforms the state-of-the-art methods on Microsoft Research (MSR) Action 3D and MSR DailyActivity3D dataset.

Keywords: Human action recognition, depth motion maps, feature

1. Introduction

Human action recognition (HAR) has been one of the most active research topics in the area of computer vision and machine learning. It has been widely used for the various applications including security system, smart home systems, content-based video search, video surveillance, and health care [1]. The HAR systems have been used mainly for identifying a particular human action among several different actions [2, 3]. Until now, there are few intelligent HAR systems with a robust and efficient performance which can recognize each class of the human activities [3]. It is due to several factors, which directly affect the performance of the HAR systems, can be susceptible to the similarity of the performed actions, size, appearance and complexity of human actions [3]. To overcome these issues, researchers proposed different methods to improve the overall performance of HAR system using RGB sensor [4–6].

Action sequence generation using traditional RGB image frames is the first step for the conventional action recognition systems [4, 5]. That is, in this step, the binary silhouettes and spatiotemporal interest point's vertices are used from the RGB frames to extract the action sequences. Here, clutter background, low illumination, camera movement, and flat pixel intensity values (i.e., 0 and 1) makes it difficult to extract the human actions and eventually cause low recognition rates in the HAR systems [4–6].

Recently, low price depth sensors such as Microsoft Kinect have been adopted for consumer applications [7–10]. The Microsoft Kinect sensor is capable of capturing both depth and RGB information. The appealing properties of depth maps are: (i) it is less sensitive to illumination changes; (ii) it provides valuable information for the representation of the human body (i.e., 3D information). The depth images are used to distinguish far and near body parts, which provides more accurate information compared to the binary maps from the gray level images [6]. **Fig. 1** illustrates the comparison of both depth and binary maps for High throw action. From the figure, it is notice that depth values are changing with the movement of body parts and provides useful information about the 3D actions compared to the conventional binary maps.

A method for generating depth motion maps (DMM) has been proposed by Yang et al. [11] is based on accumulating a difference between two consecutive depth images. These depth images are projected onto the three orthogonal Cartesian planes to distinguish the motion of an action. Then, the histogram of oriented gradients (HOG) descriptor [12] is used to extract the features from each DMM view for training the support vector machine (SVM) classifier. Chen et al. [13] modified the method of Yang et al. [11] by taking an absolute difference between two consecutive depth maps without thresholding and then stack the motion variations to form a DMM. In [14], DMMs are divided into overlapped blocks, and local binary pattern (LBP) [15] is applied to each block to calculate LBP histogram. These LBP histograms were further use to make feature vector that belongs to different actions. However, the proposed system in [11, 13, 14] fails to perform when there is significant temporal variation which leads to difficulties in discriminating the actions. This implies that considering the whole body as a feature vector, which may include unnecessary information that belongs to those of body parts that are not related to a particular action, may degrade the overall performance of HAR systems.

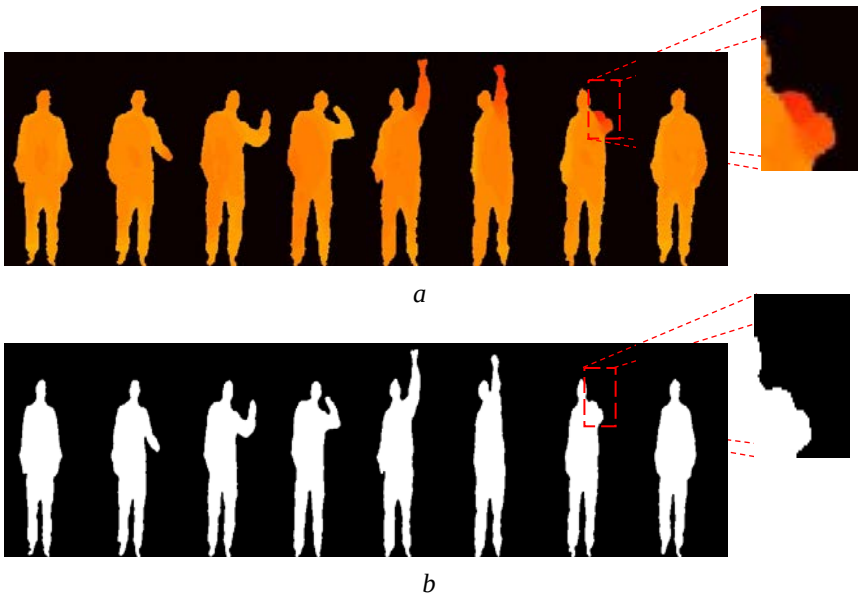

Fig. 1. High throw action from MSR Action3D dataset a) depth map sequence b) binary map sequence

Based on the aforementioned problems, the current work is aiming to improving the recognition rate of the HAR system by exploring large spatial and temporal variations. Our literature review show that the information from the spatial and temporal variation is not yet addressed yet for the recognition of human actions. Therefore, exploring large spatial and temporal is the primary focus of this paper. Furthermore, the processing time required to identify an action could also be an important aspect when considering the performance of the system [16]. The depth maps of the actions have been captured by using the Kinect sensor [7]. Moreover, an optimal size of window encapsulated in the moving body parts can be identified in the depth maps [17]. That is, the body part of the action (BPoA) is identified by a window, where the window is determined to which includes the maximum depth variation from each DMM. Here, the DMM is generated by stacking motion energy of all the depth maps onto three Cartesian planes (i.e., front view, side view and top view). Each action category has its distinct morphology (appearance and shape) which is entirely based on the accumulation of spatial and temporal motion variations. The proposed system, which is based upon BPoA, contains the salient information belongs to a maximum change in spatial and temporal domains. However, body parts of non-action (BPoNA) provides static or slowly varying regions in spatial and temporal domains. The difference between the BPoA and BPoNA is shown in **Fig. 2**. In the end, the feature vector was feeds into a non-linear tree based Rotation Forest (RF) classifier to classify all the actions.

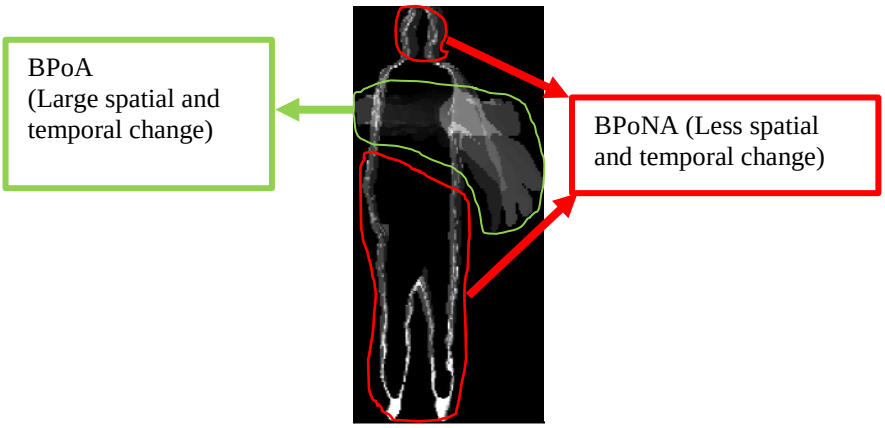

Fig. 2. DMM of horizontal wave action with BPoA (green line) and BPoNA (red line)

The rest of the paper is structure as follows: section 2, comprises of literature review on action recognition using RGB and depth sensors. In section 3, includes an introduction to proposed action recognition method, which relies on the BPoA of the DMM. Furthermore, we discuss the proposed RF approach for classification of different human actions. The details of the experimental procedure and evaluation results are shown in Section 4, and a conclusion is given in Section 5.

2. Related Work

HAR systems have been focused on using spatial-temporal features, space-time volumes, and trajectories from video sequences via traditional RGB sensors. Temporal information is obtained by employing global descriptors such as "motion history image" (MHI) and "motion energy image" (MEI) to represent action sequences [4]. However, the MHI and MEI descriptors heavily rely on the performance of background subtraction. Furthermore, these methods are very sensitive to camera movement and dynamic environments, which leads to the failure for recognizing the complex actions [18]. Image features can also be extracted using spatial and temporal methods. Gaussian kernel and Gabor filter methods have been employed to extract features in the spatial and temporal domain in [5]. This method is usually based on convolution operation. Thus it is easy to implement. However, the efficiency of these methods is reduced in the case of high video resolution, and it is also difficult to extract the features such as optical flow and silhouettes [16].

A hierarchical structure to model the spatial-temporal context information using SIFT has been proposed in [19] to compute trajectories by matching SIFT descriptors between two successive frames. Their model consists of point-level context, intra-trajectory context, and inter-trajectory context. However, this method is based on finding a fixed-dimensional velocity description using the Markov chain velocity model [20], and the method relies on large sets of fully labeled training data. Later on, a method based on dividing an entire sequence of frames into a bag of features (BoF), which is used to obtain spatial-temporal histograms, has been proposed by Laptev et al. [21]. However, the drawback of the color-based action recognition systems is that they are very sensitive to brightness changes [13].

Several problems related to human pose estimation and the RGB-D sensor has solved recognition. This sensor captures the real-time depth maps, which have been used to analyze human actions [7-10] and [22–24]. A method calculating the difference between the joints in both temporal and spatial domains has been proposed by Yang et al. [8]. Furthermore, principal component analysis (PCA) is employed to obtain eigen joints features for action recognition system. A system which is based on both the RGB and depth information to recognize human actions has been proposed by Sung et al. [9]. In this system, the features belonging to body pose, hand position, and motion information are modeled by skeleton joints. The HOG features are extracted by calculating region of interest (ROI) in gray images and depth maps are used to distinguish their appearances. In [22], skeleton joint information is used to model body pose and motion information. Then, motion and geometry cues based on a histogram of normal orientation (i.e., 4D depth, time and spatial coordinates) are successfully processed to recognize the human activities. An invariant posture representation is used via a histogram of 3D joint locations, which has been proposed in [25], where the joint points are then transferred to a modified spherical coordinate system to achieve view-invariance. However, the applications using body joints are very limited, as the joints information is not available and hence not reliable for many real applications [11].

In [26], Kamal et al. represented the human skeleton by considering joint information (i.e., relative joint positions, temporal movement of joints and offset of the joints) with respect to the depth intensity values to evaluate recognition performance. In [27], joints plus body features based method is proposed which extract joints information from skin color detection and depth pixels values from multi-views body shape. These features were combined to make a concatenated feature vector. Furthermore, these features are then trained and recognized human activities via self-organized map. A method based on pairwise relative positions of the joints has been proposed by Wang et al. [28]. The proposed method characterize the spatial relationship of joints. However, these methods completely ignored the temporal information which leads to uncertainty in the description of action sequence. More complex skeleton representations have been proposed by [29, 30]. The sequence of the skeleton was represented as a curve in the Lie group. This is done by defining the relative geometry between two rigid body parts as a rigid rotation and translation transformation.

A global feature based on space-time occupancy patterns (STOP) has been proposed in [10], which maintains both spatial and temporal information to recognize different human actions using depth map sequences. In [7], bag-of-3D-points are obtained by sampling the points from the body surface in each frame of performed action. These methods are evaluated on MSR Action3D dataset [7], which shows unsatisfactory results, especially for the cross-subject test. Furthermore, sampling the 3D points from a whole body requires a large number of the dataset; thereby high computation is needed for training the classes. The DMM has been proposed in [11, 13], which reduces the computational complexity and improves the results as compare to [7, 8, 25] but their results are still unsatisfactory probably due to large intra-class variations. A real-time processing of the DMM is also conducted to calculate the processing time of each component of the proposed system for action recognition in [13].

In this work, we propose a method to recognize the actions based on BPoA features. The features are obtained by extracting the body part, which has the maximum depth variation and discards the regions, which have a small change in spatial and temporal domains from each DMM. The features obtained by the proposed method are then fed into the classifier. To evaluate the performance of proposed method, we use publically available MSR-Action3D and MSR DailyActivity3D dataset. The results illustrate approximately 5% increases in performance compared to the state-of-the-art methods for action recognition system.

Furthermore, the processing time of the system using the proposed method is also evaluated and compared to existing state-of-the-art methods in the literature. Though the processing time of our proposed method is minutely longer, yet the recognition accuracy is 5% more in cross-validation experiment than the one in primitive methods. That is why, the proposed method is robust and efficient compared to conventional methods, and hence, the BPoA based method is a reliable candidate for the better HAR systems.

3. Methodology

The proposed HAR system consists of input depth maps from the depth sensor and processes them for extracting BPoA features from each DMM. These features are then fed into RF classifier to recognize the action. The performance of the proposed approach heavily depends on two components, i.e., a robust feature extraction method and a machine learning method which can correctly identify the actions. To evaluate the proposed feature extraction method, we use Microsoft Research (MSR) Action-3D dataset [7], consisting of a sequence of depth maps captured using depth sensor. The feature extraction stage has the following feed-forward steps:

- Calculating the DMMs for the front view, side view and top view, each of which is a projected view of an action in 3D depth maps. The reason to use these projected views is that DMMs are reliable and capable of providing useful information to improve the performance of HAR systems [11].
- Finding maximum depth variation in each DMM and localizing it in an optimal size window for extracting BPoA.

The feature matrix is then fed into RF classifier to recognize each action. The overall flow of our approach is illustrated in **Fig. 3**, and each step of the block diagram is explained in this section.

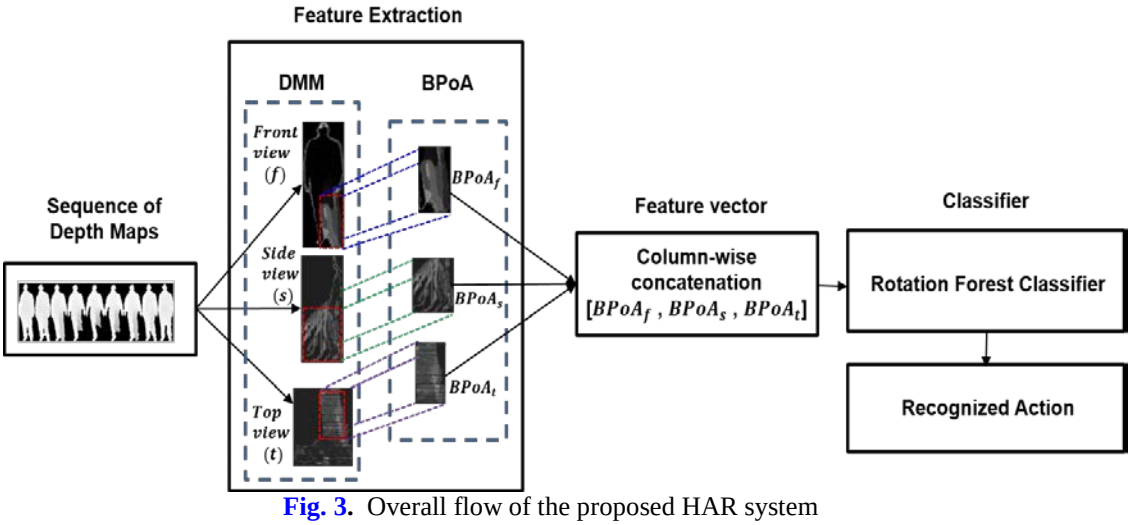

Fig. 3. Overall flow of the proposed HAR system

3.1 Feature Extraction

One of the critical components of HAR system is the feature extraction. Our feature extraction method considers depth variations to extract the features from the body part of the action (namely BPoA), which are computed by calculating maximum depth change in each DMM.

a) Depth motion maps:

Motion information in the depth video can be treated readily by projecting depth motion maps onto the three Cartesian planes to generate three 2D-view images to analyze the motions at different views. In particular, 2D projected maps of front view (f), side view (s) and top view (t) are generated from depth video $D_{map} = \{D_{map}^{q} \,|q = 1,2,3, \dots, Q\}$ where Q is the total number of depth sequences in a video shot and D_{map}^{q} represents the qth frame of the depth maps. That is, each 2D view denoted by D_{map_v}, where $v \in \{f, s, t\}$, is generated by projecting D_{map}^{q} onto each view v. Then, the depth motion map for the view$v$, DMM_v, is obtained by accumulating the absolute differences between two consecutive depth maps for each view v. So, DMM_v of each projected view is presented as:

$$DMM_v = \sum_{q=2}^{Q} \left| D_{map_v}^{q} - D_{map_v}^{q-1} \right| \tag{1}$$

where q is the index of each depth map, and $D_{map_v}^{q}$ is the projected map of qth frame under the projection view v. For example, DMM_v of High wave action are generated from depth video sequences and illustrated in **Fig. 4**.

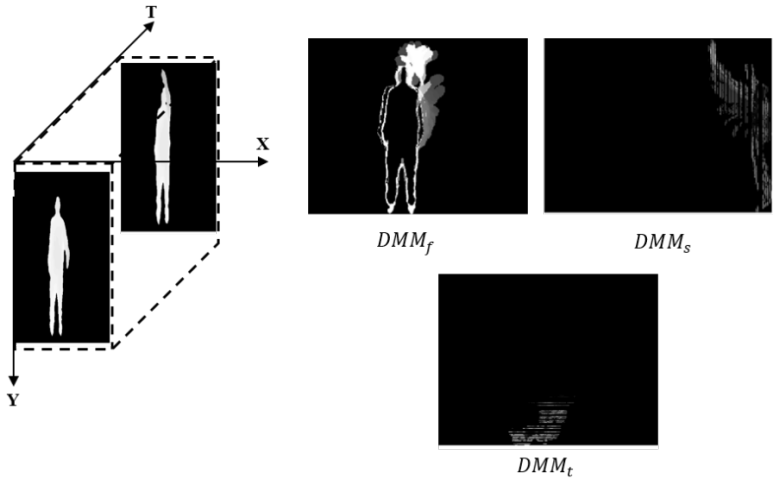

Fig. 4. DMM_v generated from depth sequences of High wave action

The bounding box denoted as BDMM (bounded depth motion maps), which tightly encapsulates the area of non-zero motion activities for all $\{DMM_v\}$, is determined by four points $\{A, B, C,$ and $D\}$ as illustrated in **Fig. 5**. That is, for each $N_v \times M_v$ frame of $DMM_v = \{I(i,j) | \, 1 \le i \le N_v, \; 1 \le j \le M_v\}$, we can find a tightest bounding box with four vertices (i.e., A, B, C, and D) as in **Fig. 5**. This can be easily done by discarding the background regions with the lines of S_1, S_2, S_3, and S_4 in **Fig. 5**.

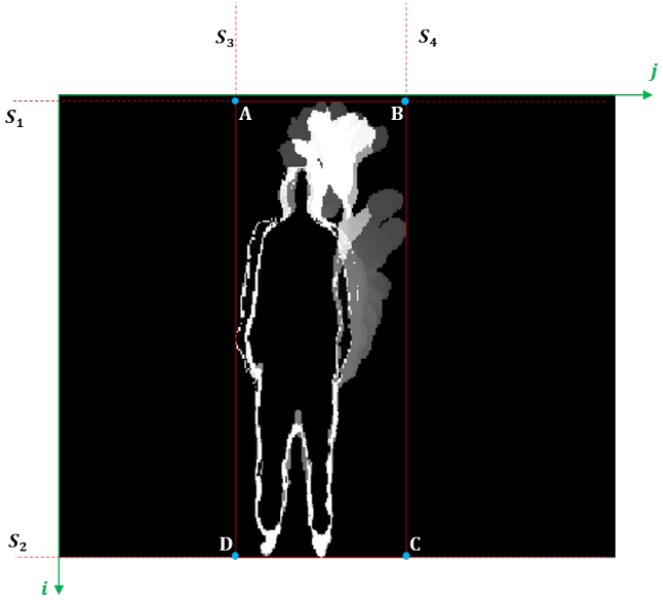

Fig. 5. Encapsulation of the non-zero region by a bounding box with four vertices

b) Detecting BPoA in a bounding box:

The differentiation between the BPoA and the BPoNA can be made by finding calculating maximum depth variation in each BDMM. We seek a window of an optimal size iteratively with an initial window size of $(2w_r + 1) \times (2w_c + 1)$ centered at $C = (C_r, C_c)$ for each BDMM_v as shown in **Fig. 6**.

Fig. 6. A window of $(2w_r + 1) \times (2w_c + 1)$ within a bounding box

An alternate optimization process using (2) and (3) can be applied to find the optimal window size (w_r, w_c) and the center of the window (C_r, C_c). However, BDMM_v may not be appropriate to calculate the maximum depth change as it may contain cluttered motions. To solve this problem, firstly, a threshold T_g is applied to each pixel in BDMM_v to find the maximum depth change.

$$(C_r^{(n)}, C_c^{(n)}) = \underset{1\le i\le N_r, 1\le j\le N_c}{argmax} \left\{ \sum_{a=-w_r^{(n-1)}}^{w_r^{(n-1)}} \sum_{b=-w_c^{(n-1)}}^{w_c^{(n-1)}} \mathrm{BDMM}_v(i+a, j+b) \right\} \quad (2)$$

$$(w_r^{(n)}, w_c^{(n)}) = \underset{\substack{w_{min}\le w_r\le w_{max} \\ w_{min}\le w_c\le w_{max}}}{argmax} \left\{ \mathrm{BDMM\_B}^{motion}_{w_r, w_c, C_r^{(n)}, C_c^{(n)}} \chi\left(\mathrm{BDMM\_B}^{non-motion}_{w_r, w_c, C_r^{(n)}, C_c^{(n)}}\right) < T_l \right\} \quad (3)$$

Where

$$\mathrm{BDMM\_B}^{motion}_{w_r, w_c, C_r^{(n)}, C_c^{(n)}} = \sum_{a=-w_r}^{w_r} \sum_{b=-w_c}^{w_c} \chi(\mathrm{BDMM}_v\left((C_r^{(n)}+a, (C_c^{(n)}+b)\right) > T_g \quad (4)$$

$$\mathrm{BDMM\_B}^{non-motion}_{w_r, w_c, C_r^{(n)}, C_c^{(n)}} = \sum_{a=-w_r}^{w_r} \sum_{b=-w_c}^{w_c} \chi(\mathrm{BDMM}_v\left((C_r^{(n)}+a, (C_c^{(n)}+b)\right) < T_g \quad (5)$$

$$\chi(\varphi) = \begin{Bmatrix} 1, if\ \varphi\ is\ True \\ 0, otherwise \end{Bmatrix}.$$

where T_l and T_g are the pre-determined thresholds to define maximum number of gradient pixels in the estimated optimal windows and the gradient magnitude, respectively. The w_{min} and w_{max} are the minimum and maximum window sizes. n is the iteration number and $N_r \times N_c$ is the size of the BDMM. That is, for the current window size $(w_r^{(n-1)}, w_c^{(n-1)})$, we find the center of the widow $(C_r^{(n)}, C_c^{(n)})$ using (2). Then, based on the updated center point, the maximum depth changes within the window are examined for all the possible window sizes using (3) to update the window to $(w_r^{(n)}, w_c^{(n)})$. Algorithm 1 provides the summary to determine a window size and BPoA location.

Algorithm 1: Determination of the window size and its center to encapsulate BPoA.
Input: Initial window size $(w_r^{(0)}, w_c^{(0)})$ while $((w_r^{(n-1)}, w_c^{(n-1)}) \neq (w_r^{(n)}, w_c^{(n)}))$ Calculate $(C_r^{(n)}, C_c^{(n)})$ using (4) Update $(w_r^{(n)}, w_c^{(n)})$ according to (5) (increase n by 1) end Output (C_r, C_c) (w_r, w_c)

The red windows shown in **Fig. 7** are the successfully detected BPoA_v using the 2D projected maps of the High throw action. Where BPoA_f shows the maximum depth variation in the front view and maximum depth change in the side and top view is illustrates in BPoA_s and BPoA_t respectively. The features extracted from BPoA are expected to improve the overall performance of HAR system compared to the features extracted from the whole human body used in [11, 13, 14]. In order to reduce large intra-class variability the size of each BPoA_v is kept fixed to the mean value of all BPoA_v under the same projection view [31]. In order to generate the features of an action sequence we arrange BPoA matrix (i.e. $(2w_r^v + 1) \times (2w_c^v + 1)$) to a single column vector (i.e. $1 \times (2w_r^v + 1)(2w_c^v + 1)$) in raster-scan order for each view, $\mathrm{BPoA}_f \in \{f_1, f_2, \dots, f_{n_f}\}$; $\mathrm{BPoA}_s \in \{s_1, s_2, \dots, s_{n_s}\}$; $\mathrm{BPoA}_t \in \{t_1, t_2, \dots, t_{n_t}\}$where the size of n_f is $(2w_r^f + 1) \times (2w_c^f + 1)$, n_s is $(2w_r^s + 1) \times (2w_c^s + 1)$ and n_t is $(2w_r^t + 1) \times (2w_c^t + 1)$.

$$F_1 = \left[\overbrace{f_1, f_2, \ldots, f_{n_f}}^{\text{Front view}}, \overbrace{s_1, s_2, \ldots, s_{n_s}}^{\text{Side view}}, \overbrace{t_1, t_2, \ldots, t_{n_t}}^{\text{Top view}}\right]^{\mathrm{T}} \tag{6}$$

In (6) each view is concatenated to make a single feature vector F_1 for a certain action sequence and here T shows the matrix transpose. The size of feature vector F_1 is $(n_f + n_s + n_t) \times 1$. Hence, the final feature matrix F comprises of feature vectors from all the action sequences. Each feature vector is arranged column-wise $F = [F_1, F_2, \ldots, F_p] \in \mathbb{R}^{(n_f+n_s+n_t)\times p}$, where p is the total number of action sequences in a dataset.

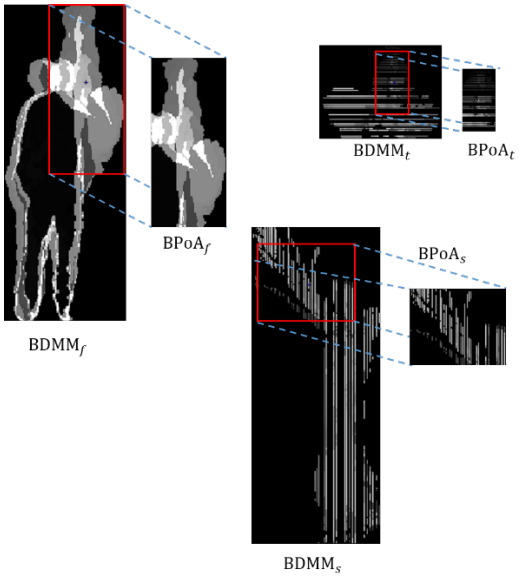

Fig. 7. Determined optimal windows from each view ($BDMM_f$, $BDMM_s$ and $BDMM_t$)

3.2 Rotation Forest

The classification performance of HAR systems is also important. Note the accuracy of the action classification is to be improved by considering the conventional algorithms together. To achieve high classification accuracy, a group of classifiers (GoC) has been used such as Rotation Forest (RF) [32] instead of a single classifier. An efficient GoC system mainly consists of accurate and diverse base classifiers. Thus, a sample, which is misclassified by one base classifier, will be corrected by other ones and the final classification after all the GoC is more accurate than the individual best classifier [33].

The RF approach has been proposed by Rodriguez et al. [34], and the concept of RF is to encourage both diversity and individual accuracy simultaneously within the GoC. In the RF, each classifier is independently constructed and trained on the training samples in a rotated feature space, which is derived from the PCA transformation. The RF performs much better than previous ensemble methods [33, 34]. In particular, RF is a newly proposed multi-classifier scheme, and the overall flow of the algorithm is described as follows:

For each base classifier, training dataset F_T is randomly split into K subsets. Consider F_T be the input training samples of size $\left((n_f + n_s + n_t) \times p_T\right)$ matrix, where p_T is the total number of training samples from all the action sequences. Y_T be the corresponding labels with dimensionality $(p_T \times 1)$ is defined as the class labels $\{1, \ldots, l\}$, where l is the total number of

classes. We split F_T into K disjoint subsets randomly, by assuming that each feature subset contains $M = p_T/K$ features. The steps for training the base classifiers E_i, where $i = 1, \ldots, Q$ and Q is the number of base classifiers, are explained in the following [33].

- Split F_T into K disjoint subsets randomly, by assuming that each feature subset contains $M = p_T/K$ features.
- Let $F_{T_{i,j}}$ be the $j^{th}, j = 1, \ldots, K$, subset of features for the base classifier E_i. Also, let $F_{T_{i,j}}'$ be a new training dataset, which is selected from $F_{T_{i,j}}$ with the 75% size using bootstrap algorithm. Then, a linear transformation is applied as the PCA on $F_{T_{i,j}}'$ to get the coefficients $u_{i,j}^{(1)}, \ldots\ldots, u_{i,j}^{(M_j)}$, the size of each principal component $u_{i,j}$ is $M \times 1$.
- Construct a sparse rotation matrix R_i with the obtained coefficients as follow (7):
- $$R_i = \begin{bmatrix} u_{i,1}^{(1)}, u_{i,1}^{(2)} \ldots\ldots, u_{i,1}^{(M_1)} & [0]_{1\times M_2} & \ldots & [0]_{1\times M_K} \\ [0]_{1\times M_1} & u_{i,2}^{(1)}, u_{i,2}^{(2)} \ldots\ldots, u_{i,2}^{(M_2)} & \ldots & [0]_{1\times M_K} \\ \vdots & \vdots & \ddots & \vdots \\ [0]_{1\times M_1} & [0]_{1\times M_2} & \ldots & u_{i,K}^{(1)}, u_{i,K}^{(2)} \ldots\ldots, u_{i,K}^{(M_K)} \end{bmatrix} \quad (7)$$
- The columns of R_i is rearranged with respect to the original feature set to obtain rotation matrix R_i^u. Then, the training set will become $F_T R_i^u$. In this case, all classifiers are train in parallel. For a given test sample $\mathcal{X}$, let $d_{n,m}(\mathcal{X}R_i^u)$ is the probability generated by the classifier E_i to the hypothesis that $\mathcal{X}$ belongs to class c the confidence is calculated for each class by the average combination method as (8):

$$\mu_j(\mathcal{X}) = \sum_{i=1}^{Q} d_{i,j}(\mathcal{X}R_i^u) \; where \; j = 1, \ldots, l \quad (8)$$

Finally, $\mathcal{X}$ will be assigned to the class with the largest confidence. The success of RF relies on the base classifier and the rotation matrix created by the transformation methods. Here, we selected k-nearest neighbour as the base classifier [35]**.** In [36], the authors compare the performance of different transformation algorithms (e.g., PCA, NDA, and RP) and found that PCA produced the best results. For more details about RF classifier reader may refer to [34].

4. Experimental result

4.1 Experimental setup

In this study, we evaluate our feature extraction method using MSR Action-3D dataset [21] and MSR DailyActivity3D dataset [28].

The MSR Action-3D dataset is publically available and comprises of depth sequences, which are captured using depth sensor. The dataset has been recorded on ten volunteer subjects facing towards the camera for 20 different actions. During the recordings, each action was repeated 2-three times for each subject, thus, having significant intra-class variation. The size of each depth map is 320x240. To compare the proposed method with conventional ones [7, 8, 10, 11, 13, 14, 25] we follow the same experimental protocols as [7, 18, 22] by dividing the 20 action types into 3 action subsets (i.e., AS1, AS2, and AS3) as shown in **Table 1**. Based on these subsets, three different kinds of tests (i.e., Test I, Test II, and Test III) are considered to analyse the overall performance using the proposed method. In Test I, 1/3 of the samples in each action is used for training while the rest are used for testing. In Test II, 2/3 of the samples in each action is used for training, and the rest are used for testing. In Test III (i.e., cross

subject test), 1/2 of the subjects in each action are used for training whereas the remaining 1/2 subjects are used for testing. In Test I and Test II, actions performed by each subject in test data are seen in training data whereas in Test III actions performed by each subject in test data is not observed in training data. Moreover, using these three tests we perform under three experiments entitled as fixed test experiment (FTE), random test experiment (RTE) and cross-validation experiment. Furthermore, to test the recognition rate of proposed method with on 20 actions of MSR-Action3D dataset with [28-30], the protocol [28] is used.

Another hand, the publically available MSRDailyActivity3D dataset contains sixteen activities. All subjects performed an activity into two different poses (i.e., stand pose and sitting on sofa pose). The MSR DailyActivity3D dataset is one of the challenging dataset in which it is very difficult to clear the background objects.

Table 1. Three subsets of actions in MSR-Action3D dataset

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal wave (HW)	High wave (HW)	High throw (HT)
Hammer (HAM)	Hand catch (HC)	Forward kick (FK)
Forward punch (FP)	Draw X (DX)	Sidekick (SK)
High throw (HT)	Draw tick (DT)	Jogging (JNG)
Hand clap (HC)	Draw circle (DC)	Tennis swing (TSN)
Bend (BND)	Two hand wave (THW)	Tennis serve (TS)
Tennis serve (TS)	Forward kick (FK)	Golf swing (GS)
Pickup throw (PT)	Side Boxing (SB)	Pickup throw (PT)

4.2 The sensitivity of parameters

A number of classifiers (Q) and feature sets (K) are important parameters for the RF. Various studies show the reliability of the value of K [37–39]. In [39], the value of K has been set to 2 when F has less than 10 attributes and for more than 10 attributes they set K =3. Several experiments has been conducted by Xia et al. [38] and the best results obtained by using K =12. Furthermore, he concludes that for K >10 the accuracy of the overall system slightly varies. For example; for K = 50 discussed in Kotsiantis et al. [39], where the author suggested that the value of K should be chosen based on of cost minimization criteria. Hence, there is no standard set to choose the value of K. In this paper, we have best possible results for all the experiments for K= 40. In addition, there is no literature review present to select the value of Q. Some of the previous works, have been using fixed number a of group of classifiers while some for variable size of Q [40]. Rodrguez et al. [34] have used the fixed value of Q =10. In [36], author claims that RF shows good classification ranging from 1 to 10 on average. However, Xia et al. [38] have suggested few numbers of base classifiers for the best performance of RF. In [41], they did experiment using different values of Q ranging from 5 to 100, where best accuracies have been achieved from Q =10 to 50. However, they also chose fixed value of Q for whole dataset. For our all experiments we used the fixed number of

classifier $Q = 3$ for classification of all the actions. Furthermore, we used k=3 for k-NN base classifiers.

For this study, after several experiments the best results have been obtained for extracting a complete BpoA using there parameters such as :$w_{min} = 8,\ w_{max} = 55,\ T_l = w_d^v(2w_{max} + 1)(2w_{max} + 1),\ w_d^f = 0.75, w_d^s = 1.25, w_d^t = 0.25, T_g = 0.06$. It is noticed that these parameters are very sensitive for detecting optimal window size. For our experiments, the pixel values are normalized between 0 and 1 for each BPoA_v.

4.3 Results and Discussion

For three subsets setting of MSR Action-3D dataset, the recognition accuracies of the proposed HAR system was compared with the accuracies of the conventional methods which are obtained from [7, 8, 10, 11, 13, 14, 25]. **Table 2** illustrates the comparison of proposed method with conventional methods using FTE. It is observed that performance of the proposed method for all tests in FTE is significantly better than conventional methods because BPoA considers only the maximum depth change while excludes the small variations in each action performed by the subjects.

Table 2. Performance Comparison of FTE for MSR-Action3D dataset

	Li et al. [7]	**Yang et al.**[8]	**Vieira et al.**[10]	**Yang et al.**[11]	**Chen et al.**[13]	**Chen et al.**[14]	**Lu et al.** [25]	**Our method**
Test I								
AS1	89.5	94.7	98.2	97.3	97.3	96.7	98.5	**99.2**
AS2	89	95.4	94.8	92.2	96.1	**100**	96.7	**100**
AS3	96.3	97.3	97.4	98	98.7	99.3	96.5	**100**
Average	91.6	95.8	96.8	95.8	97.4	98.7	96.2	**99.7**
Test II								
AS1	93.4	97.3	99.1	98.7	98.6	**100**	98.6	99.3
AS2	92.9	98.7	97	94.7	98.7	**100**	97.2	99.1
AS3	96.3	97.3	98.7	98.7	100	**100**	94.9	**100**
Average	94.2	97.8	98.3	97.4	99.1	**100**	97.2	99.4
Test III								
AS1	72.9	74.5	84.7	96.2	96.2	98.1	88	**98.2**
AS2	71.9	76.1	81.3	84.1	83.2	92	85.5	**94.6**
AS3	79.2	96.4	88.4	94.6	92	94.6	63.6	**96.8**
Average	74.7	82.3	84.8	91.6	90.5	94.9	79	**96.5**

Table 3 shows that the recognition rates of our proposed BPoA based feature extraction approach are much higher for all the tests as compare to the conventional DMM based feature extraction method [13] for RTE experiment. Confusion matrix of RTE for Test III is shown in **Fig. 8**, which illustrates that the recognition rates are improved as compared to a previous method [13].

Table 3. Performance Comparison of Random Test for MSR-Action3D dataset

	Chen et al. [13]	**Our method**
Test I AS1	97.4	**99.1**
AS2	96.1	**98.4**
AS3	97.7	**98.5**
Average	97.1	**98.6**
Test II		
AS1	98.5	**99.2**
AS2	97.8	**98.6**
AS3	98.9	**99.7**
Average	98.4	**99.1**
Test III		
AS1	84.8	**92.7**
AS2	67.8	**82.2**
AS3	87.1	**88.7**
Average	79.9	**87.8**

Table 4 provides experiment results of DMM features with RF classifier to verify the performance improvement of this classifier. The results illustrated in **Table 4** (only RF) and **Table 3** (new feature BPoA and RF) show that robust feature representation of an action has an equal importance with selecting a suitable classifier. BPoA covers the most important features of each action that further supported by the RF classifier.

Table 4. Performance Comparison of Random Test for MSR-Action3D dataset

	Chen et al. [13]	Chen et al. [13] using RF classifier	Our method
Test I AS1	97.4	98.1	**99.1**
AS2	96.1	96.9	**98.4**
AS3	97.7	98.2	**98.5**
Average	97.1	97.73	**98.6**

Test II			
AS1	98.5	98.7	**99.2**
AS2	97.8	98.3	**98.6**
AS3	98.9	99.3	**99.7**
Average	98.4	98.76	**99.1**
Test III			
AS1	84.8	85.3	**92.7**
AS2	67.8	68.5	**82.2**
AS3	87.1	87.9	**88.7**
Average	79.9	80.6	**87.8**

We further conducted cross-validation experiments to show that our method does not depend on any specific training data. By considering all the 252 combinations, we choose 5 out of 10 subjects for training and remaining subjects for testing. As a result, **Table 5** shows that our method achieves an accuracy of 93.1 %, which is higher than the previous methods.

Table 5. The performance of our method on MSR Hand Action 3D dataset, compared to previous methods using cross-validation

Method	**Mean accuracy % ± STD**
HON4D [22]	82.2 ± 4.2
Rahmani et al. [23]	82.7 ± 3.3
Tran et al.[24]	84.5 ± 3.8
Chen et al.[14]	87.9 ± 2.9
Our proposed method	**93.1 ± 2.5**

	HTW	**HAM**	**FP**	**HT**	**HNC**	**BND**	**TS**	**PT**
HTW	**100**	0	0	0	0	0	0	0
HAM	0	**85.6**	14.4	0	0	0	0	0
FP	0	3.2	**84.3**	5.7	0	0	6.8	0
HT	0	0	0	**100**	0	0	0	0
HNC	0	0	0	0	**100**	0	0	0
BND	0	0	0	0	0	**100**	0	0
TS	0	0	0	0	6.1	0	**86.3**	7.6
PT	0	9.5	0	0	0	0	5.3	**85.2**

(a)

	HW	HC	DX	DT	DC	THW	FK	SB
HW	**93.2**	0	0	0	6.8	0	0	0
HC	7.5	**71.1**	11.5	9.9	0	0	0	0
DX	0	5.6	**70.4**	8.6	8.7	6.7	0	0
DT	0	13.2	7.2	**66.9**	12.7	0	0	0
DC	0	0	13.4	11.5	**75.1**	0	0	0
THW	8.5	10.8	0	0	0	**80.7**	0	0
FK	0	0	0	0	0	0	**100**	0
SB	0	0	0	0	0	0	0	**100**

(b)

	HT	FK	SK	JNG	TSN	TS	GS	PT
HT	**90.6**	0	0	0	0	0	0	9.4
FK	0	**91.2**	8.8	0	0	0	0	0
SK	0	0	**100**	0	0	0	0	0
JNG	0	0	0	**100**	0	0	0	0
TSN	0	0	0	0	**83.3**	16.7	0	0
TS	0	0	0	0	12.5	**78.2**	0	9.3
GS	0	0	0	0	16.6	0	**83.4**	0
PT	4.9	0	0	0	6.3	5.9	0	**82.9**

(c)

Fig. 8. Confusion matrix for RTE a) AS1. b)AS2. c) AS3
HW= High wave, **HTW**= Horizontal wave, **HAM**= Hammer, **HC**= Hand catch, **FP**= Forward punch, **HT**= High throw, **DX**= Draw X, **DT**= Draw tick, **DC**= Draw circle, **HNC**= Hand clap, **THW**= Two hand wave, **SB**= Side boxing, **BND**= Bend, **FK**= Forward kick, **SK**= Sidekick, **JNG**= Jogging, **TNS**= Tennis swing, **TS**= Tennis serve, **GS**= Golf swing, **PT**= Pickup throw.

Fig. 9 and **Table 6** show the confusion matrix and performance comparison, respectively. Similar with three subsets setting of more challenging setting, which employed 20 actions of MSRActions3D, results show that the proposed human action recognition architecture outperformed several well-known methods.

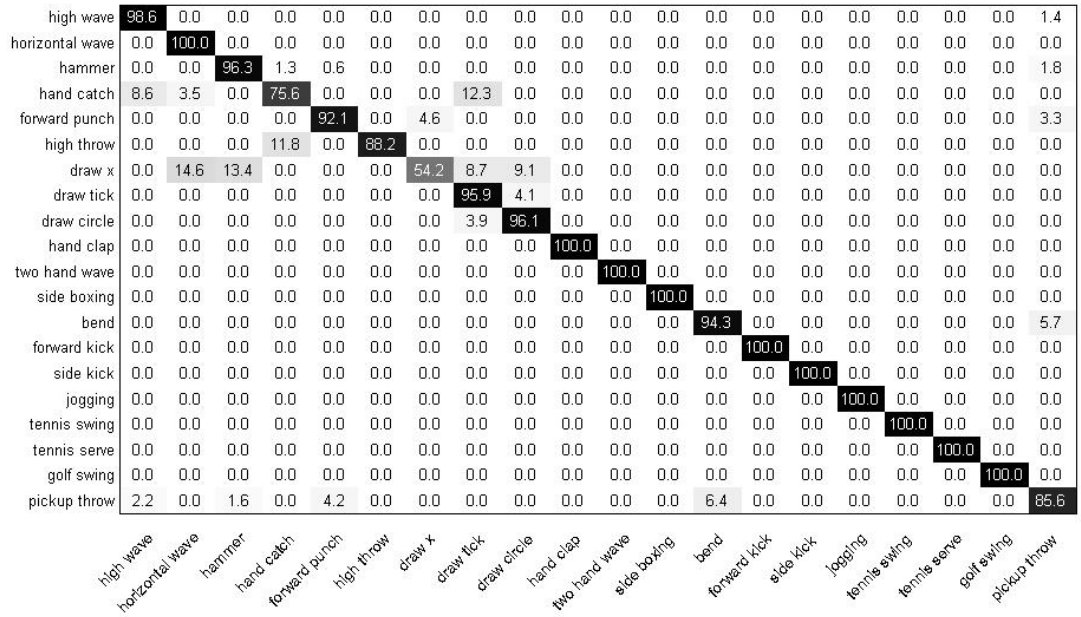

Fig. 9. Confusion matrix using the protocols [37] for MSR-Action3D dataset

Table 6. Performance Comparison of protocols [28] for MSR-Action3D dataset

Method	Recognition rate %
Actionlets [28]	88.20
Points in a lie group [29]	89.48
LM^3 TL [30]	90.53
Our proposed method	**93.84**

The accuracy evaluation of the proposed method using MSRDailyActivity3D dataset [28] is compared with only joints position features [28], moving pose features [42] which worked on body pose information as well as differential quantities of the human body joints and HON4D [22]. **Table 7** shows that recognition rates of our proposesd method compared to conventional methods . The results suggest that proposed method has significantly outperformed the conventional methods. However, it is also noticed that recognition rate of HON4D is slight higher than the proposed method. This is because the proposed BPoA method does not perform well with the activities where there is the interaction of human activity with different things such as read book, write on paper, and play guitar in which it is very difficult to clear the background objects.

Table 7. Accuracy evaluation of proposed and conventional methods using MSRDailyActivity3D dataset

Method	Recognition rate %
Only Joints position features [28]	68 ± 5%
Moving pose [42]	73.8 ± 2%
Our proposed method	76.3 ± 3%
HON4D[22]	80 ± 2%

In spite of these limitations, the overall performance of our proposed method shows that BPoA produces more prominent features as compared to state-of-the-art methods

5. Conclusion

In this paper, we present a computationally efficient BPoA based feature extraction method for action recognition system using RF classifier. An optimal size window encapsulates BPOA from each DMM and discards the remaining regions (i.e., BPoNA). Experimental results on MSR Action 3D dataset demonstrates that the performance of proposed HAR system achieves the mean recognition rate of 98% for Fixed Test, 88% for Random Test, and 93.1% with cross-validation experiment, which outperform the existing state-of-the-art methods. Furthermore, we observe that generating feature vector using BPoA is more compact and separable as compare to the feature extraction methods in previous systems. The processing time of the proposed HAR system is slightly longer, but the performance increases to 5% as compared to the existing systems. Thus, it is possible to use the proposed system for many real-time applications including healthcare systems, automatic video surveillance, and smart homes.

References

[1] O. Masoud, N.Papanikolopoulos, "A method for human action recognition," *Image Vis. Comput*., vol. 12, no. 8, pp. 729–743, 2003. Article (CrossRef Link)

[2] Z. Gao, J.-M. Song, H. Zhang, A.-A.Liu, Y.-B.Xue, G.-P. Xu, "Human action recognition via Mmulti-modality Information," *J. Electr. Eng. Technol*., vol 9 no. 2, pp. 739–748, 2014. Article (CrossRef Link)

[3] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, J. Gall, "A Survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications'(Springer Berlin Heidelberg)*, pp. 149–187, 2013. Article (CrossRef Link)

[4] J.W. Davis, A.F.Bobick, "The representation and recognition of human movement using temporal templates," in *Proc. of Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 928–934, 1997.Article (CrossRef Link)

[5] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, 2005. Article (CrossRef Link)

[6] R. Gupta, A. Y.-S. Chia, R. Rajan, "Human activities recognition using depth images," in *Proceedings of the 21st ACM International Conference on Multimedia*' , pp. 283–292, 2013. Article (CrossRef Link)

[7] W. Li, Z. Zhang, Z. Liu "Action recognition based on a bag of 3D points," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 9–14, 2010. Article (CrossRef Link)

[8] X. Yang, Y.L. Tian, "EigenJoints-based action recognition using Na #x00EF;ve-Bayes-Nearest-Neighbor," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 14–19, 2012. Article (CrossRef Link)

[9] J. Sung, C. Ponce, B. Selman, A. Saxena, "Unstructured human activity detection from RGBD image," in *Proc. of IEEE International Conference on Robotics and Automation*, pp. 842–849, 2012. Article (CrossRef Link)

[10] A. W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, M.F.M. Campos, "STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences," in *Proc. of Iberoamerican Congress on Pattern Recognition Springer Berlin Heidelberg*, pp. 252–259, 2012. Article (CrossRef Link)

[11] X. Yang, C. Zhang, Y. Tian, "Recognizing actions using depth Motion maps-based histograms of oriented gradients," in *Proc. of 'Proceedings of the 20th ACM International Conference on Multimedi*, pp. 1057–1060, 2012. Article (CrossRef Link)
[12] N. Dalal, B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, 2005. Article (CrossRef Link)
[13] C. Chen, K. Liu, N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps," *J. Real-Time Image Process*., vol 12, no. 1, pp. 155–163, 2016. Article (CrossRef Link)
[14] C. Chen, R. Jafari, N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns," in *Proc. of IEEE Winter Conference on Applications of Computer Vision*, pp. 1092–1099, 2015. Article (CrossRef Link)
[15] T. Ojala, M. Pietikainen, T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002. Article (CrossRef Link)
[16] P. Turaga, R. Chellappa, V.S. Subrahmanian, O. Udrea, "Machine recognition of human activities a survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 11, pp. 1473–1488, 2008. Article (CrossRef Link)
[17] A.V. Le, S.-W. Jung, C.S. Won, "Nonuniform video size reduction for moving objects," *Sci. World J*., p. e832871, 2014. Article (CrossRef Link)
[18] J.C. Niebles, H. Wang, L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis*, vo. 79, no. 3, pp. 299–318, 2008. Article (CrossRef Link)
[19] J. Sun, X. Wu, S. Yan, L.F. Cheong, T.S. Chua, J. Li, "Hierarchical spatio-temporal context modeling for action recognition," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2004–2011,2009. Article (CrossRef Link)
[20] M. Raptis, S. Soatto, "Tracklet descriptors for action modeling and video analysis," in *Proc. of European Conference on Computer Vision, (Springer Berlin Heidelber)*, pp. 577–590, 2010. Article (CrossRef Link)
[21] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, "Learning realistic human actions from movies,", in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008. Article (CrossRef Link)
[22] O. Oreifej, Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proc. of Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 716–723, 2013. Article (CrossRef Link)
[23] H. Rahmani, A. Mahmood, A, D. Q. Huynh, A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *Proc. of IEEE Winter Conference on Applications of Computer Vision*, pp. 626–633, 2014. Article (CrossRef Link)
[24] Q.D. Tran, N.Q. Ly, "Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences," in *Proc. of 'The 2013 RIVF International Conference on Computing Communication Technologies - Research, Innovation, and Vision for Future (RIVF)*, pp. 253–258, 2013. Article (CrossRef Link)
[25] L. Xia, C.C. Chen, J.K. Aggarwal, "View invariant human action recognition using histograms of 3D joints, " in *Proc. of '2 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, 2012. Article (CrossRef Link)
[26] Kamal, Shaharyar, Ahmad Jalal, and Daijin Kim. "Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM," *J. Electr. Eng. Technol*, vol. 11, pp. 1857-1862, 2016. Article (CrossRef Link)
[27] Jalal, Ahmad, Yeonho Kim, Shaharyar Kamal, Adnan Farooq, and Daijin Kim. "Human daily activity recognition with joints plus body features representation using Kinect sensor," in *Proc. of Informatics, Electronics & Vision (ICIEV), 2015 International Conference on*, pp. 1-6, 2015. Article (CrossRef Link)

[28] Wang, J., Liu, Z., Wu, Y. and Yuan, J. "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. of Computer Vision and Pattern Recognition (CVPR) IEEE Conference on* pp. 1290-1297, 2012. Article (CrossRef Link)

[29] Vemulapalli, R., Arrate, F. and Chellappa, R., "Human action recognition by representing 3d skeletons as points in a lie group," in *Proc. of Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 588-595, 2014. Article (CrossRef Link)

[30] Y. Yang et al., "Latent max-margin multitask learning with skelets for 3D action recognition," *IEEE Trans. Cybern.*,vol. 47, no2, pp.439-448, 2017. Article (CrossRef Link)

[31] C. Chen R. Jafari, N. Kehtarnavaz,, "Improving human action recognition using fusion of depth camera and inertial sensors," *IEEE Trans. Hum.-Mach. Syst.*, vo. 45, no. 1. , pp. 51–61, 2015. Article (CrossRef Link)

[32] G. Stiglic, P. Kokol, "Effectiveness of rotation forest in meta-learning based gene Expression classification," in *Proc. of Twentieth IEEE International Symposium on Computer-Based Medical Systems*, pp. 243–250, 2007. Article (CrossRef Link)

[33] K.-H. Liu, D.-S. Huang, "Cancer classification using rotation forest," *Comput. Biol. Med.*, vo. 38, no. 5, pp. 601–610, 2008. Article (CrossRef Link)

[34] J.J Rodriguez,L.I. Kuncheva, C.J Alonso, "Rotation Forest: A new classifier ensemble method," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1619–1630, 2006. Article (CrossRef Link)

[35] S. Wold, K. Esbensen, P. Geladi, "Principal component analysis," *Chemom. Intell. Lab. Syst.*, vo. 2, no. 1, pp. 37–52, 1987. Article (CrossRef Link)

[36] L.I. Kuncheva, J.J. Rodríguez, "An experimental study on rotation forest ensembles," in *Proc. of International Workshop on Multiple Classifier Systems, (Springer Berlin Heidelberg)*, pp. 459–468, 2007. Article (CrossRef Link)

[37] C.-X. Zhang, J.-S. Zhang, "RotBoost: A technique for combining rotation forest and AdaBoost," *Pattern Recognit. Lett.*, vol. 29, no. 10, pp. 1524–1536, 2008. Article (CrossRef Link)

[38] J.-F. Xia, K. Han, D.-S. Huang, "Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor," *Protein Pept. Lett.*, vo. 17, no. 1, pp. 137–145, 2010. Article (CrossRef Link)

[39] S.B. Kotsiantis, P.E. Pintelas, "Local rotation forest of decision stumps for regression problems," in *Proc. of 2nd IEEE International Conference on Computer Science and Information Technology*, pp. 170–174, 2009. Article (CrossRef Link)

[40] M. Shaheryar, M. Khalid, A.M. Qamar, "Rot-SiLA: A novel ensemble classification approach based on rotation forest and similarity learning using nearest neighbor algorithm," in *Proc. of 12th International Conference on Machine Learning and Applications*, pp. 46–51, 2013. Article (CrossRef Link)

[41] J. Xia, P. Du., X. He, J. Chanussot, "Hyperspectral remote sensing image classification based on rotation forest," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 239–243, 2014. Article (CrossRef Link)

[42] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proc. of Proceedings of the IEEE International Conference on Computer Vision*, pp. 2752-2759, 2013. Article (CrossRef Link)

Adnan Farooq received his B.S degree in Computer Engineering from COMSATS Institute of Science and Technology, Abbottabad, Pakistan and M.S. degree in Biomedical Engineering from Kyung Hee University, Republic of Korea. His research interest includes Image Processing, Computer vision.

Faisal Farooq received his B.S degree in Electronics Engineering from COMSATS Institute of Science and Technology, Abbottabad, Pakistan and M.S. degree in Biomedical Engineering from Kyung Hee University, Republic of Korea. His research interest includes Image Processing, Signal Processing, and Pattern Recognition.

Anh Vu Le is at Optoelectronics Research Group, Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. He received his BS in Electronics and Telecommunications from Ha Noi University of Technology, Mater and PHD degrees in Electronics and Electrical from the Dongguk University in 2007 and 2012, respectively. His current research interests include Robotics vision, human detection, action recogntion, feature matching, 3D video processing.