Received: December 13, 2023. Revised: January 12, 2024. 251

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Multi-Modal Multi Feature Assisted Human Action Recognition through Deep Learning

Dustakar Surendra Rao[1, 2] L. Koteswara Rao[1]* Vipparthi Bhagyaraju[3]

[1] *Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Aziz Nagar, Hyderabad, 500075, Telangana, India*
[2]*Department of Electronics and Communication Engineering, Guru Nanak Institutions Technical Campus, Hyderabad, 501506, Telangana, India*
[3]*Department of Electronics and Communication Engineering, Siddhartha Institute of Engineering and Technology, Hyderabad, 501506, Telangana, India*
* Corresponding author's Email: koteswararao@klh.edu.in

Abstract: The discovery of Microsoft Kinect sensors opened a new research direction to human action recognition (HAR) from videos. However, the depth maps and body postures are of noisy and less reliable in their original form. Hence, they contribute limited contribution toward the actions recognition, especially in real time scenarios. Moreover, the HAR from single data modality have several limitations. Hence, this paper proposes a multi-modal HAR framework based on a simple and effective deep learning model. The proposed model considered depth maps and body postures as input and describes each action through two newly proposed descriptors namely improved motion history image (IMHI) and spatio-temporal posture descriptor (STPD). IMHI removes undefined motion regions and STPD ensure the provision of complete Spatio-temporal motion information to the training system. Alongside, a new temporal segmentation is also proposed to ensure robustness against speed variations. Finally, different fusion rules are adapted to determine the correct action based on different policies. We conduct extensive Simulations on three benchmark standard datasets namely MSRAction3D, MAD and PKU-MMD and the average accuracy obtained is observed as 95.2333%, 91.9945% and 93.3141% respectively. Experimental results prove that the proposed framework is discriminative for the actions with similar movements.

Keywords: Human action recognition, Depth maps, Skeleton, Motion history image, Segmentation, Accuracy.

1. Introduction

Recently, the vision based Human Action Recognition has gained a great research interest in computer vision field [1, 2]. The main aim of HAR is identify the activity of a human being performed in a video which was captured from uncontrolled background environment. The research on HAR has been widely accomplished due to its widespread applicability in different applications including computer games, virtual reality, security surveillance, human machine interaction, video understanding and assistive robots [3-6]. In the past, the initial research on HAR was carried out over the videos captured through RGB cameras. However, the RGB images are sensitive to light variations, view point variations, self-occlusions etc. For HAR from RGB videos, initially the human object needs to be segmented. In the case of RGB videos with cluttered background and occlusions, the human object detection is very tough. Human action is an aggregation of temporal and spatial information. Even though the segmentation of human object is done accurately, the recognition is bi challenge due to the complexity of action movements. Furthermore, the human actions are differentiated by several people, culture and emotion shifts.

To sort out the problems with traditional RGB videos, recently, depth maps have been introduced which are acquired through a special type of camera

called as microsoft kinect camera. Depth maps records the distance between the camera and the surface of the object. Depth information simplifies the segmentation of human objects from background thereby results in an improved recognition performance. Hence, the current research is diverted towards using depth data either from depth maps or body postures [7, 8]. The performance of HAR depends on a good action representation which can provide perfect and distinctive features of every action. However, depth maps suffer from some undefined regions introduced due to noise including jumbled objects, small body shaking movements and object boundaries. On the other hand, action representation through posture data is very sensitive to the joint's movements which can show a significant impact on the recognition of two actions with similar movements. A better solution is to use both data modalities for action recognition which can overcome the drawback of using single data model. Irrespective of the data model used, action representation, feature extraction and classification plays a vital role in the HAR. In the past HAR methods, most of the handcrafted feature extraction methods [9, 10] employed support vector machine (SVM) for classification. However, recent HAR directed towards the utilization of deep learning models, especially convolutional neural networks (CNNs) had shown its effectiveness compared to the traditional machine learning models as they gained a huge success in the image classification [11]. Hence, this work motivated to use two depth data modalities and deep learning algorithms to build an effective HAR.

This paper proposes a new mechanism for HAR from body postures and depth maps using a customized deep learning model using CNNs. The major contributions of this paper are outlined as follows;

1. To ensure an increased accuracy at HAR, we use both depth and body posture data modalities which can overcome the problem of using singe type. This paper proposes two new action representation methods namely improved motion history image (IMHI) and spatio-temporal posture descriptor (STPD) for depth maps and body postures respectively. IMHI nullifies the side effect due to the undefined regions and STPD provides perfect and distinctive features for each action through its local and global movements.
2. To extract perfect discriminative features from both action representations, this paper proposes a simple and effective CNN model. By keeping the complexity in mind, the proposed CNN model is designed based on network in network structure [12].
3. To enhance the accuracy of prediction, this paper proposes to use different fusion rules to fuse the outputs of two CNN models. Different fusion rules, both individually and combined are suggested to provide the flexibility.

Remainder of this paper is structured as follows; the particulars related to past HAR methods are discussed in section 2. The details related to the entire proposed methodology are discussed in section 3. Simulation results and analysis is explored in section 4 and section 5 provides the concluding remarks.

2. Related work

Since the depth maps and body postures are more advantageous than the traditional RGB videos, the earlier researchers used both of them in a proportionate way. Hence, we surveyed both the depth map based HAR methods and posture based HAR methods.

2.1 Depth maps

M. A. Faris et al. [13] applied deep learning algorithm over the fuzzy weighted multi-resolution depth motion maps (DMMs) [14] for HAR from depth action sequences. They proposed to provide larger weightage for the DMMs with significant motion information. The weightage analysis is carried out in three orientations; they are linear, reverse and central. A new CNN model is introduced to extract the features from Weighted DMMs followed by classification. However, the DMMs are susceptible to the noises due to the undefined motion regions.

Jiang Li et al. [16] applied CNNs on the depth action sequence after representing it with DMMs on three orthogonal planes. For three DMMs in three different views (front, side and top), they applied three CNNs models and the score obtained at each layer is processed softmax regression layer to determine the final action. Local binary pattern (LBP) is proved as best texture descriptor and its significance was proved even in HAR through several methods [15]. However, LBP suffers from huge information loss. Hence, an extended version of LBP called as local ternary pattern (LTP) is proposed by Li Z et al. [17] in their method for HAR. After the projection and DMM computation of input depth action sequence over three orthogonal planes, LTP is applied on each DMM to describe it's the texture. For feature extraction and classification, they employed

CNN model. But, the DMM on three orthogonal planes induces huge complexity and also not able to eliminate noises.

Xu Weiyao et al. [18] proposed a new model called as MSM which uses static history image (SHI) and motion history image (MHI) to describe an action through static and motion postures respectively. Besides MSM, they also proposed a multi-frame select sampling (MFSS) which captures key frames based on the motion energy. MSM is applied for all the three planes and encoded them with LBP followed by fisher kernels. For classification, they employed KELM algorithm. Even though MFSS reduces frames count to get processed, it results in the loss of temporal information.

Tianjin Yang et al. [19] proposed a 'Multi-label subspace learning (MLSL)' mechanism for action recognition from depth maps and named it as 'Depth sequential information entropy maps (DSIEM)'. DSIEM represented an action through Spatio-temporal features in which stitching and Entropy were employed to describe temporal and spatial features respectively. After representing the action in a single image, they computed HOG and passed to SVM for action prediction. Encoding the action only through temporal information results in the misclassification due to undefined motion regions. Such kind of regions exists due to the jumbled objects and small body shaking movements.

2.2 Body postures

D. Warchol and K. Tomasz [20] proposed an action descriptor that encodes the angular correlation between the bone pairs. They employed totally five classifiers namely (1) "LogDet divergence based metric learning with triplet constraints (LDMLT)", (2) "Bidirectional long short-term memory network (BiLSTM)", (3) "Fully convolutional network (FCN)", (4) "DTW with city block distance (DTW-CBD)", and (5) "DTW with Euclidean distance (DTW-ED)". X. Diao et al. [21] suggested a new RNN model named as "Multi-term attention networks (MTANs)" for HAR from body postures. MTANs can extract the temporal features are different scales which consists of MTA-RNN and ST-CNN. However, they didn't encode the spatial information due to which misclassification is more at similar actions.

X. Zhang et al. [22] applied graph theory for HAR from skeleton sequences. The edge of graph is regarded as the combination of spatial as well as temporal relation between different bones. They described an edge by combining its spatial as well as temporal neighbor edges those explore the relation between different bones and consistency of the movements in an action sequence respectively. Further, they constructed graph node CNN and graph edge CNN with the help of shared intermediate layers. The accomplishment of graph theory makes the HAR sensitive to locations errors of skeleton joint positions.

P. Zhang et al. [23] proposed a simple semantics guided neural network that explicitly includes the high level semantics of skeleton joins such as frame index and joint type. Further, they exploited the hierarchical relationship between joints through two modules; they are correlation between joints in same frame and dependencies between frames. The correlation between frames ensures spatial relationship between joints but makes the system susceptible to speed of actions.

Z. Shao et al. [24] proposed a new action descriptor called as hierarchical rotation and relative velocity (HRRV) to describe the action hierarchy at different scales. They treated the action as a simultaneous movement of body parts and grouped the all body parts into some bundles. Then the HRRV is encoded by the fisher vector and then properly arranged into the hierarchical model through mixed norm. HRRV ensures robust against action with different speeds but not reduces misclassification at inter-class similar actions.

Q. Nie et al. [25] proposed a view invariant HAR mechanism by recovering the corrupted skeletons based on a 3D bio-constrained model. The bio-constrained model is formulated based on joint's motion limit and constant bone length. They described an action through two motion features; they are joint Euler angles and Euclidean distance matrix between joints. For learning the motion patterns they deployed two stream CNN models [26]. Euler angles won't consider the restricted movement of joints and hence it is invariant to view points.

L. Shi et al. [27] proposed an adaptive two-stream graphical convolutional network GCN) in which the graph topology is learned uniformly and individually in an end-to-end manner. C. Si et al. [28] proposed attention Enhanced GC-LSTM network for HAR from skeleton information. AGC-LSTM represents an action through Spatio-temporal features but also explores their co-occurrence relationship. They represent the skeleton in a hierarchical structure thereby the learning ability will get boost up and also reduces the computation cost significantly. GCN are very much sensitive to location errors at joint positions and not able to analyze the motion at fine body parts like fingers.

Some authors [43, 44] considered the skeleton joints as 3D data and applied for HAR after

processing them though recurrent neural networks (RNNs). However, they didn't ensure the robustness against view point variations and inter class similarities.

2.3 Hybrid methods

Even though the individual data modalities have gained an efficient recognition performance in HAR, a single modality has its own problems. Hence, some researchers considered two and more data modalities for HAR. A. Kamel *et al*. [29] considered depth maps and body postures as input data modalities and proposed a new CNN model with three channels. Further, they proposed two new action descriptors namely depth motion image (DMI) for depth maps and motion joint descriptor (MJD) for body postures. For final action prediction, they suggested several fusion rules. Even though the MJD provide view invariance, the DMI is not able to reduce the ambiguity between intentional unintentional motion pixels.

X. Ji *et al.* [30] proposed a temporal invariant action descriptor which was adequate for complex backgrounds in HAR. They embedded the skeleton information into the depth maps and partitioned the human body into several set of motion parts. For feature representation, they applied a Spatio-temporal scaled pyramid (STSP) and then encoded through fisher vectors to accumulate local coarse features. However, the Spatio-temporal features extracted from an entire video could not ensure speed invariance. Y. Fan *et al.* [31] proposed a cross attention module for action recognition which is an integration of cross attention and self-attention branches. This approach can extract informative joints which are highly correlated with the context of scenario. However, cross attention induces more redundancy if the noises are not removed before features extraction.

Qin Cheng *et al.* [32] considered RGB and Depth data modalities and proposed a Spatio-temporal information aggregation module (SITAM) which utilizes CNNs to acquire Spatio-temporal information from input data. Further, they introduced a cross modality interactive module (CMIM) to aggregate the multi-modal complementary information completely. Finally, an integrated model called as Multi-modal interactive network (MIMINet) by fusing the SITAM and CMIM. But, they didn't revealed the inherent characteristics of actions with different speeds and time periods

C. Liang *et al.* [33] proposed a multi-modal HAR by considering the depth maps and skeleton as input data models. They proposed a segmental architecture to explore the relations between sub-actions, jointly with the fusion of heterogeneous information and class privacy preserved collaborative representation (CPPCR). Then the sub-actions based depth motion and skeleton features are fused. Next, CPPCR is applied to address the issues in sub-action sharing scenario, learning from high-level discriminative representation. However, the CPPCR didn't ensure robustness against view point variations.

H. Wei *et al.* [34] utilized inertial and video sensing to perform action detection followed by action recognition. The Inertial data and video data are transformed into 2D and 3D images and then fed to 2D and 3D CNN model and the decisions are fused. M. Ehatisham-Ul-Haq *et al*. [35] proposed a multi-modal features level fusion approach for robust HAR which uses the data of RGB camera, Wearable inertial sensor and depth sensor. They applied histogram of oriented gradients (HOG) on RGB/Depth videos and statistical features on wearable sensor data. For classification, they used K-nearest neighbour and SVM algorithms. RGB videos are sensitive to cluttered backgrounds, clothing colors and illuminations.

X. Wang *et al*. [36] proposed a fusion based HAR framework that uses inertia signals and skeleton data as input models. For skeleton data transformation, they proposed a weighted front-view skeleton motion map (WF-SMM), a weighted multi-view skeleton motion map (WM-SMM), and a 3D weighted skeleton motion map (3DW-SMM). Next the inertial data is transformed into 2D images and then fed to dilated CNN. They applied two types fusion such as decision level and fusion level. J. Cheng *et al*. [37] proposed cross modality compensation CNN called as cross modality compensation block (CMCB) which can learn the compensation features jointly from depth and RGB data modalities. CMCB is incorporated into VGG and ResNet to verify the performance. However, the inertia data is not resilient to disturbances like jumbled objects and small body shaking movements.

X. Weiyao *et al.* [38] proposed a multi-modal HAR model based on Bi-linear pooling and attention network (BPAN) on skeleton and RGB videos. Initially, they pre-processed the both data models and then processed through a multi-modal fusion network. Finally, the BPAN effectively compress the features and project them into a latent space and then fed to a 3-layer perceptron for classification. RGB is more susceptible to noises due to illumination variations, background clutters and scales.

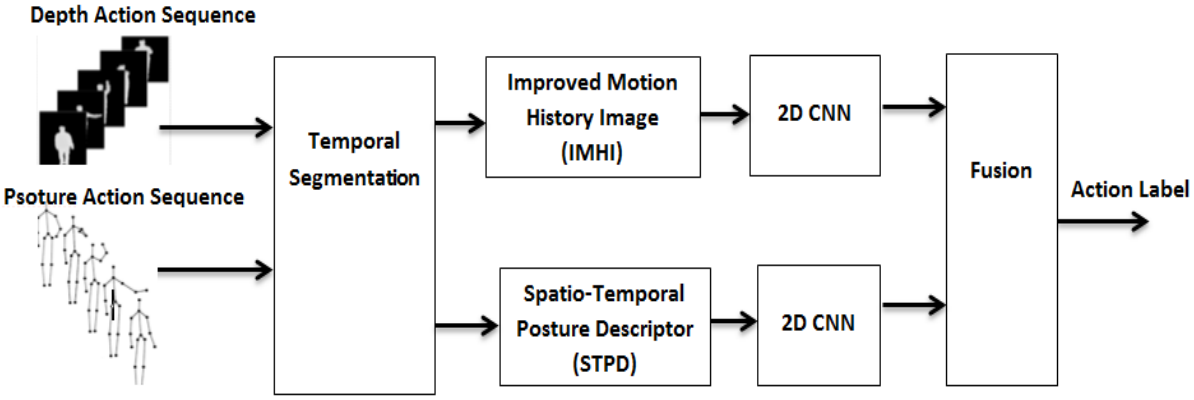

Figure. 1 the overall architecture of the proposed HAR framework

Summary of problem: In the case of depth maps based HAR methods, the action descriptors didn't focus on the removal of undefined motion regions due to the unintentional movements like Small body shaking movements, jumbled objects etc. On the other hand, the body posture based HAR methods didn't ensure the robustness against view point variations.

3. Proposed HAR framework

3.1 Overview

Here we proposed a new and effective HAR Framework which uses two types of data for action representation such as depth maps and body posters. Each input data is processed through an efficient action descriptor and then fed to a deep learning model for feature extraction followed by classification. For death maps based action representation, we proposed an improved version of MHI called as IMHI. Next for body postures, we proposed Spatio-temporal posture descriptor (STPD). As a preprocessor, we proposed a simple temporal segmentation mechanism and applied on both data formats. IMHI captures the variations of motion during the body motion and STPD captured spatial and temporal variations of motion. For feature extraction and classification, we proposed a new customized Deep learning model which is simple and effective. After getting the classified results from individual data, they are fused through different fusion rules to determine the final action label. The overall working schematic of proposed human action recognition Framework is shown in Fig. 1.

3.2 Temporal segmentation

Here, the main intention of temporal Segmentation is to divide the entire action sequence into different segments thereby we can acquire the detailed motion information. Generally the earlier Depth Map descriptors like DMM and MHI are applied over the entire data sequence. Hence they may not be able to capture detailed motion information. Hence, to acquire the detailed motion information, here we suggest applying the IMHI and STPD on temporal segments instead of on entire action sequence. The temporal segments are derived after the division of the input action sequence into a set of 3D overlapped segments. The segmentation is done in such a way the number of frames in each segment must be same. The major advantage with temporal segmentation based action representation is that the recognition system can cope up with the speed variations. The actions are generally performed by different people have different speeds. Hence temporal segmentation is performed at multiple resolutions. Here we consider totally three different resolutions with frame count of all, 5 and 10. For entire frame count, we consider it as level 0 (l_0) for frame count of five and 10 we consider as level 1 (l_1) and level 2 (l_2) respectively.

Note that the frame count at l_1 and l_2 is variable and it varies based on the size of input action sequence. Even though the temporal segmentation helps in the improvisation of recognition accuracy by providing detailed motion cues, it consequences to an increased computational complexity. Hence we consider only three levels at segmentation which includes the default level also, i.e., l_0 which considers entire frames. Further, the overlap size (OS) is mentioned as 3. The overlap size is defined as the number of frames present between two neighboring temporal segments. Simply it is the total number of common frames between two successive temporal segments. To ensure a simple segmentation here we use the same overlap size for both levels such as l_1 and l_2. Here we apply temporal segmentation

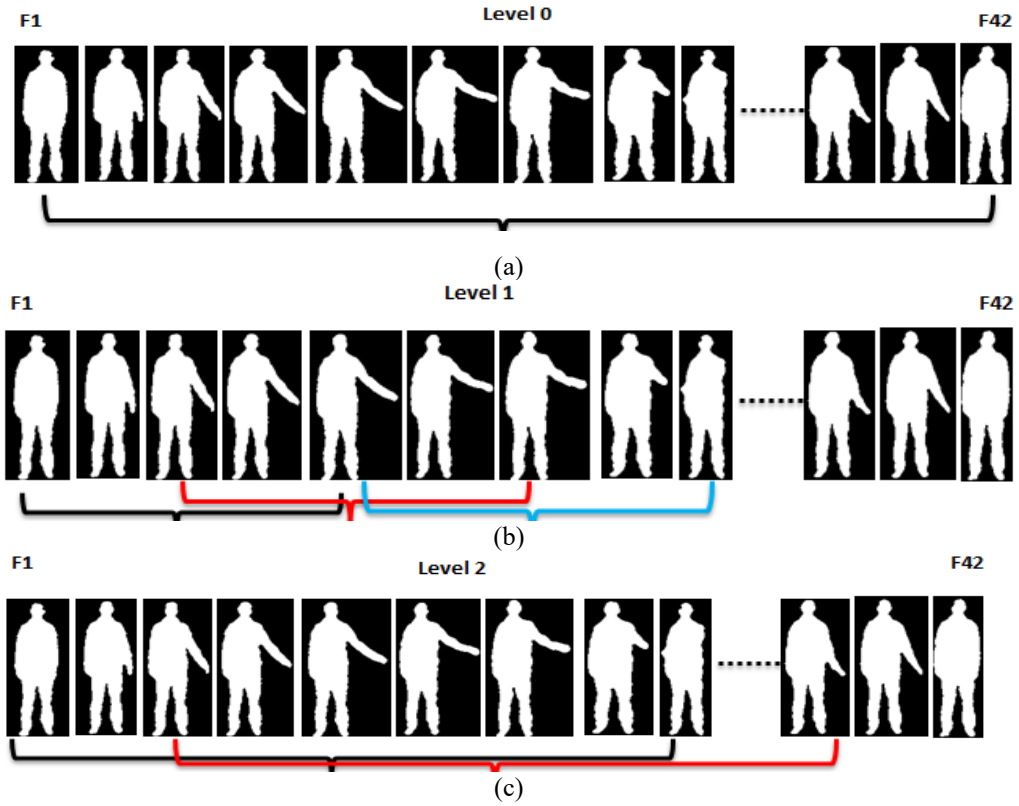

Figure. 2 Temporal segmentation at three different levels: (a) Level 0, (b) Level 1, and (c) Level 2

on both data formats such as death maps and body postures. On segmented depth maps we applied MHI to represent motion information. Since we get three MHIs for three levels we apply a motion based fusion mechanism to provide a single MHI which intern called as IMHI. For the temporal segmented body postures, we compute the spatial distances between successive frames and final action is represented with three statistical features such as mean, variance and range. Fig. 2 shows the process of temporal segmentation at three different levels.

3.3 IMHI

Motion representation is an important task in the HAR. Generally motion representation attempts to encode the motion information through the entire videos sequence into a single 2D representation. Here we consider one of the popular motion representation technique called as MHI which was introduced by Bobick and Davis [39]. Generally the MHI is a visual method which records the history of template through their temporal variations at every pixel in the whole video sequence. MHI effectively encodes the action moment's spatial distributions based on the pixel intensity in a temporal manner. Here we measure the MHI on the depth map modality. The main advantage with MHI is its less sensitivity nature towards shioullete noises and illumination changes. For MHI computation initially the depth sequence is converted into binary sequence which compactly reserves the motion information. Here we compute MHI for multiple temporal segments and finally IMHI is obtained after their fusion based on motion threshold. Consider a depth action sequence as $D(x, y, t), t = 1,2, \dots, T$ Where T denotes the number of frames. The first step of MHI is the conversion of $D(x, y, t)$ into binary format and let it be represented as $B_D(x, y, t)$ and it is obtained based on the distance between successive frames as

$$B_D(x, y, t) = \begin{cases} 1, & if\ |D(x, y, t) - D(x, y, t+1) \geq \delta| \\ 0, & Otherwise \end{cases} \quad (1)$$

Where $D(x, y, t)$ and $D(x, y, t+1)$ are the pixels at position *(x,y)* in t and $t+1$ frames respectively. δ is the motion threshold and based on the experimental analysis we found $\delta = 40$ provides better results. Based on the obtained $B_D(x, y, t)$ the MHI is computed as

$$MHI = \sum_{t=0}^{N-2} w_t . B_D(x, y, t) \quad (2)$$

Where w_t is the weight of the frame at t^{th} instants and its value is varied between 0 and

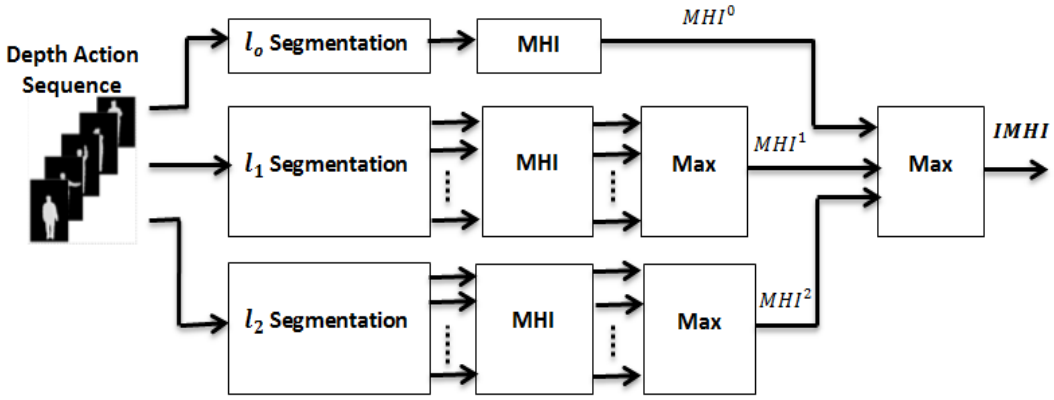

Figure. 3 IMHI Computation from depth action sequence

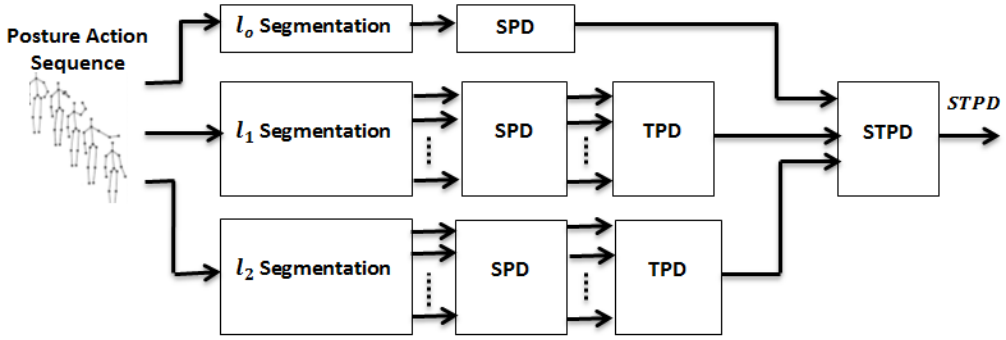

Figure. 4 STPD computation from posture action sequence

255. Mathematically w_t computation is done as follows;

$$w_t = \frac{255}{N}.(t+1) \tag{3}$$

Based on the Eq. (3), we can understand that w_t is a linear early rising function with the time. Since the motion attributes presence rises with time, more weight is assigned to the frames with later time instances than the frames at starting time instances.

In our work the MHI is applied on temporal segments at multiple resolutions. Consider a depth action sequence of size N and its entire frames are segmented into P segments. Then the MHI of each segment is represented as $MHI_{p=1,2,\ldots,P}$. For level 1 it can be represented as $MHI^1_{p=1,2,..p}$ and for level 2 it can be represented as $MHI^2_{p=1,2,..P}$. Since for level 0 we have only one MHA it can be simply represented as MHI^0 . Consider $MHI^l_{p=1,2,..P}, l = 0,1,2$ which denotes the MHA at l^{th} level, it is obtained by capturing the the maximum values at each pixel. Since MHI is obtained as a difference between frames the existence of motion enlarges the value at corresponding pictures. The accumulated MHI at l^{th} level is obtained as

$$MHI^l(x,y,t) = max\left(MHI^l_p(x,y,t)\right)\forall p = 1,2 \tag{4}$$

Where $MHI^l_p(x,y,t)$ is the MHI at p^{th} segment at l^{th} level. Here P is different for l_1 and l_2 while for l_0 not there is no existence of P. Now the IMHI is obtained based on the average of MHI's at l_0, l_1 and l_2. Mathematically the IMHI is calculated as

$$IMHI(x,y,t) = \frac{1}{3}\sum_{l=0,1,2} MHI^l(x,y,t) \tag{5}$$

The final IMHI is a 2D representation of depth map action sequence which consists of only motion information and it can cope up with speed and length variations. Fig. 3 shows the process of IMHI computation over a depth action sequence.

3.4 STPD

STPD mainly aims at the action representation from skeleton sequences with respect to the moments of joints at different spatial and temporal distributions. For a given input skeleton sequence with N number of frames, initially it was divided into several segments and then each segment is subject to spatial posture descriptor (SPD) followed by temporal posture descriptor (TPD). Since the segment has less number of frames, they can explore

the spatial moments of joints. Next TPD considers the SPDs as a Input and represent each action with three statistical features. Here, the SPD shows resilience towards misclassifications for the actions with similar moments and TPD ensure global motion of an action. At the TPD, we use three statistical measures namely mean, variance and range to represent each action. Fig. 4 shows the process of STPD computation

Consider a skeleton action sequence with T number of frames and each frame is represented with Q number of joints, then each joint is represented as $J_q^t = [x_q^t, y_q^t, z_q^t]$. Means J_q^t indicate the position of a joint on three Axes such as X- Y- and Z-axis. After processing the input skeleton sequence through temporal segmentation we get several segments. The segmentation is done in such a way each segment must have same frame count. Consider l^{th} level segmentation in which the entire sequence is divided into P segments, and then SPD computes the Euclidean distance between two same joints located at two different and successive frames. Let $J_{q_t}^{l_i} = [x_{q_t}^{l_i}, y_{q_t}^{l_i}, z_{q_t}^{l_i}]$ be the q^{th} joint in t^{th} frame of i^{th} segment obtained at l^{th} level segmentation and $J_{q_{t+1}}^{l_i} = [x_{q_{t+1}}^{l_i}, y_{q_{t+1}}^{l_i}, z_{q_{t+1}}^{l_i}]$ be the q^{th} joint in $t+1^{th}$ frame of i^{th} segment obtained at l^{th} level segmentation, then the Euclidean distance is computed between these two joints is computed as

$$d_{q_t}^{l_i} = \sqrt{\left(x_{q_{t+1}}^{l_i} - x_{q_t}^{l_i}\right)^2 + \left(y_{q_{t+1}}^{l_i} - y_{q_t}^{l_i}\right)^2 + \left(z_{q_{t+1}}^{l_i} - z_{q_t}^{l_i}\right)^2} \quad (6)$$

With the help of Eq. (6), the Euclidean distance is computed for all joints. Hence, for every fame, we get Q number of distances and if we consider there is L number of frames in each segment then we get a distance matrix of size Q*(N-1) and it is represented as

$$d_{q\in 1,2,\dots,Q}^{l_i} = \begin{bmatrix} d_1^{12} & d_1^{23} & \dots & d_1^{(N-1)N} \\ d_2^{12} & d_2^{23} & \dots & d_2^{(N-1)N} \\ \vdots & \vdots & \ddots & \vdots \\ d_Q^{12} & d_Q^{23} & \dots & d_Q^{(N-1)N} \end{bmatrix} \quad (7)$$

Eq. (7) represents the distance matrix of i^{th} segment in l^{th} level. Based on the obtained $d_{q\in 1,2,\dots,Q}^{l_i}$, we compute the mean displacement of each joint in i^{th} segment by measuring the average of distances in each row and it is mathematically expressed as

$$d_q^{l_i} = \frac{1}{N-1}\sum_{n=2}^{N} d_q^{(N-1)N} \quad (8)$$

Here $d_q^{l_i}$ is a one dimensional Matrix consists of a Q number of rows and one column where each value represents mean displacement of each joint in the local segment. To find the global displacement of all joints in the local segment, we compute the statistical measures such as mean, variance and range. Consider μ^{l_i}, v^{l_i} and r^{l_i} the mean, variance and range of i^{th} segment in l^{th} level they are computed as

$$\mu^{l_i} = mean\left(d_q^{l_i}\right), q \in 1,2,\dots,Q \quad (9)$$

$$v^{l_i} = variance\left(d_q^{l_i}\right), q \in 1,2,\dots,Q \quad (10)$$

$$r^{l_i} = max\left(d_q^{l_i}\right) - \min\left(d_q^{l_i}\right), q \in 1,2,\dots,Q \quad (11)$$

The mean, variance and range shown in Eq. (9), Eq. (10) and Eq. (11) respectively are belongs the statistical features of i^{th} segment. These features measure the global motion of an action in local segment. These three features are computed for all segments in l^{th} level. Finally the mean, variance and range of l^{th} level is computed as

$$\mu^{l} = mean\left(\mu^{l_i}\right), i \in 1,2,\dots,P, \quad (12)$$

$$v^{l} = variance\left(v^{l_i}\right), i \in 1,2,\dots,P \quad (13)$$

$$r^{l} = max\left(r^{l_i}\right) - \min\left(r^{l_i}\right), i \in 1,2,\dots,P \quad (14)$$

Here μ^l, v^l and r^l denotes the statistical features of SPD for one level. The same process is applied for all the three levels and we get totally a final action descriptor of size 3*3 where the first three represents the statistical measures such as mean, variance and range and the next 3 represent the three levels such as l_0, l_1 and l_2. The final action descriptor is called as STPD which can cover both spatial and temporal motion variations of an action. The major advantage with STPD is its effective motion information provision towards HAR. Especially the SPD is more helpful in the reduction of misclassification of similar actions. For example consider three motions namely draw cross mark, draw circle and the draw tick of MSR action 3D which have similar motion at the fingers. Such kinds of actions are needed to be analyses deeply to recognize them accurately and SPD provides such kind of flexibility. Since SPD is able to identify the displacements of actions at finer level, it can provide perfect discrimination between the actions with similar moments.

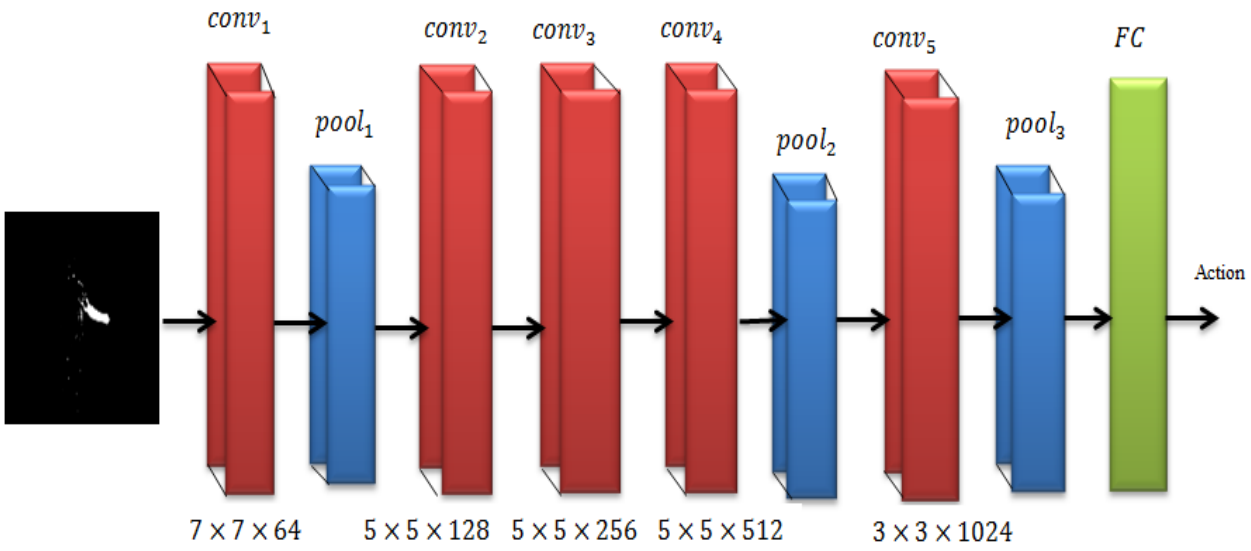

Figure. 5 CNN model

3.5 CNN model

After the motion representation of an action video through IMHI or STPD, it is resized into 112×112 before feeding as an input to CNN model. The proposed CNN model is composed of five convolutional (*conv*) layers, two pooling layers (PL) and one fully connected layer (FCL). Here, the *conv* layers are used for the extraction of features and pooling layers are used for the reduction of features dimensionality. Here, we applied max pooling operation which can find the maximum values for a given array or matrix. Our proposed CNN model has one fully connected layer of size $1 \times n$ where n denotes the total number of actions. Figure.5 shows the architecture of proposed CNN model. As shown in Fig. 5, the *Conv1* has applied 64 convolutional filters and the size of each filter is 7×7. The *Conv2, Conv3 and Conv4* applies 128, 256 and 512 filters respectively and the size of each filter is 5×5. At the *Conv5*, the size of convolutional filter is 3×3 and the total number of convolutional filters applied is 1024. The texture of two different actions through proposed action descriptors makes the system challenging to extract distinct features when the size of convolutional filters is small. For instance, the size of 3×3 accomplishment on image at starting is not efficient because two action images may have similar characteristics in the small-sized region. Therefore, we decided to apply convolutional filter with size 7×7 at the staring convolutional layer. Next, the size of filter at max-pooling layer is fixed as 2×2 and its main intention is to reduce feature map size. In this work, we used two max-pooling layers, where one is used after *conv1* and second max-pooling layer is used after *conv4*. Due to the accomplishment of max-pooling layer after the *conv1*, the feature map size is reduced from 112×112 to 56×56. Next, due to the accomplishment of max-pooling layer after *conv4*, the size of feature map is reduced from 56×56 to 28×28. Finally, the features maps are processed through FCL and its size of equal to total actions to test. At testing phase, we used softmax regression layer to produce a score for each action with the help on the trained weights. The action with highest score is treated as the action present in the input video.

3.6 Fusion

After the action is described through the proposed descriptors, then they are subjected to classification through 2D-CNN model. The model is applied three times each time the descriptor is different. Due to the consideration of two descriptors, the results obtained at the softmax layers are of two values. The output of softmax layer is a vector which has the length equal to the number of actions trained to the system. The values of softmax layer output are posterior probabilities those denotes that the probability of input action to be the trained action. However, we have two probabilities for every action. Hence we applied fusion mechanism to derive the final results. For fusion, we consider two strategies namely product and maximum since they have better fusion capability. Consider R_1 be the output of IMHI+2DCNN, R_2 be the output of STPD+2DCNN,

they are fused as $F_1 = Max(R_1, R_2)$ and $F_2 = Product(R_1, R_2)$. From the two values such as F_1 and F_2 the final action is obtained as $Action = Max(F_1, F_2)$. Where Action is the name of an action which has highest score and it represents the final action class prediction.

4. Simulation experiments

4.1 Datasets

This section explores the details of simulation experiments of proposed HAR mechanism on three standard datasets such as MSRAction3D [40], multi-modal action dataset (MAD) [41] and PKU-MMD [42].

A. MSR Action 3D dataset: The actions in MSRA3D are captured with the help of a depth camera, and the pose of the subject is located in the front view. This dataset is acquired with the help of ten subjects, and every subject performs each action two to three times. This dataset is very puzzling due to the speed variations on each topic. Also, this dataset consists of 20 activities: BEND, TWO HAND WAVE, HANDCLAP, JOGGING, SIDEKICK, FORWARD KICK, PICKUP & THROW, GOLF SWING, TENNIS SWING, HIGH ARM WAVE, HORIZONTAL ARM WAVE, FORWARD PUNCH, HIGH THROW, HAMMER, HAND CATCH, DRAW CROSS, DRAW TICK, DRAW A CIRCLE, and SIDE BOXING.

B. MAD Dataset: It is one of the largest depth datasets which consists of totally 35actions and they are performed with the help of 20 subjects. Every subjected is asked to perform each action for two times. The 35 actions are namely crouching, running, walking, jumping, left arm & left swipe, jump & side-kick, left arm & right swipe, left arum punch, left arm wave, left arm pointing to ceiling, left arm dribble, swing from left, left arm throw, left arm back receive, left arm receive, left leg left kick, left leg front kick, right arm left swipe, right arm right swipe, right arm wave, right arm dribble, right arm punch, right arm pointing to ceiling, swing from right, right arm throw, right arm back receive, right arm receive, right leg front kick, right leg right kick, cross arms in chest, basketball shooting, both arms pointing to screen, both arms pointing to both sides, both arms pointing to left side and both arms pointing to left side.

C. PKU-MMD dataset: This dataset consists of totally 51 actions and are acquired with the help of 66 subjects using three Microsoft Kinect V2 cameras from right, middle and front views simultaneously. It contains totally 1076 long video sequences in 51 action categories and 2000 short video sequences with 49 action categories. It consists of approximately 20,000 action instances with 5.4 million frames. Each video contains approximately 20 action instances and the total scale of this dataset is 5,312,580 frames of 3000 minutes with 21,545 temporally localized actions. Among the 51 actions, 41 are daily actions (ex. Waving hand, drinking etc.) and 10 are interactive actions (ex. Hand shaking, hugging etc.). This dataset provides totally five types of data modalities such as RGB videos, infrared, skeleton, depth images and RGB images. The resolution of depth maps is 512*424 and the number of joints present in each skeleton frame is 25.

4.2 Results

Two metrics namely Recall and Accuracy are used to assess the performance of developed HAR model. They are defined as follows;

$$Recall = \frac{TP}{TP+FN} \tag{15}$$

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \tag{16}$$

Where TP denotes correctly classified action videos and FN denotes falsely classified action videos.

A set of experiments have been conducted on the three datasets through different descriptors with same CNN model. As an initial set of experiments, we simulated the proposed HAR framework with three different combinations, such as IMHI + 2DCNN, STPD + 2DCNN and IMHI + STPD + 2DCNN. From the results shown in Fig. 6, it can be seen that the integrated descriptor had shown better performance than the individual descriptors. Further, among the remaining two combinations, the maximum performance is achieved for STPD. Compared to the IMHI, STPD features are different which helps in the recognition of action even with similar movements. From the results, we can see that, for three similar actions namely, High arm Wave, Hammer and Horizontal arm wave, the recall is observed as 100%, 100% and 96% respectively. The average recall of IMHI, STPD and Fused descriptor with common 2DCNN is observed as 91.2345%, 92.7785% and 94.6633% respectively.

In the case of simulation over MAD dataset, it needs an additional background removal process which was done by IMHI. As the IMHI considers the binary frames as input and thresholding through a motion threshold, the background is completely removed and the remaining frames consist of only motion of human objects. Fig. 7 shows the recall of MAD dataset fort three combinations IMHI +

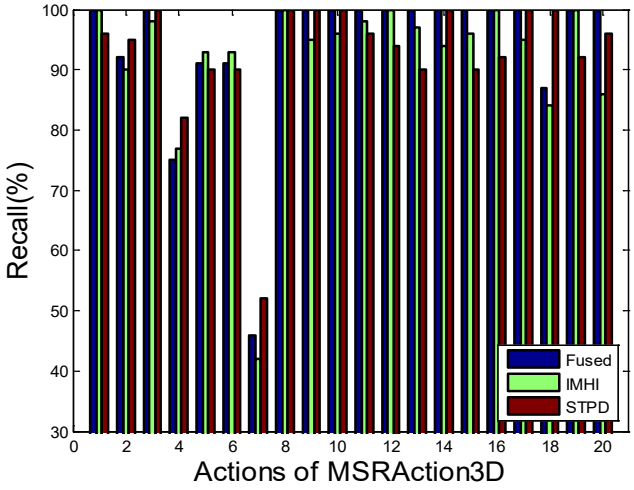

Figure. 6 Recall (%) for different actions of MSR Action 3D dataset

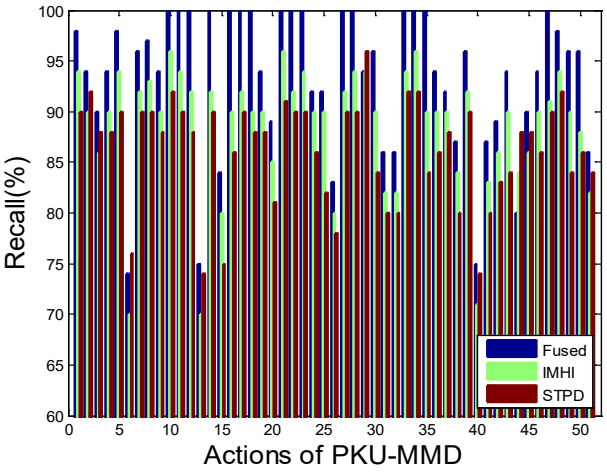

Figure. 8 Recall (%) for different actions of PKU-MMD dataset

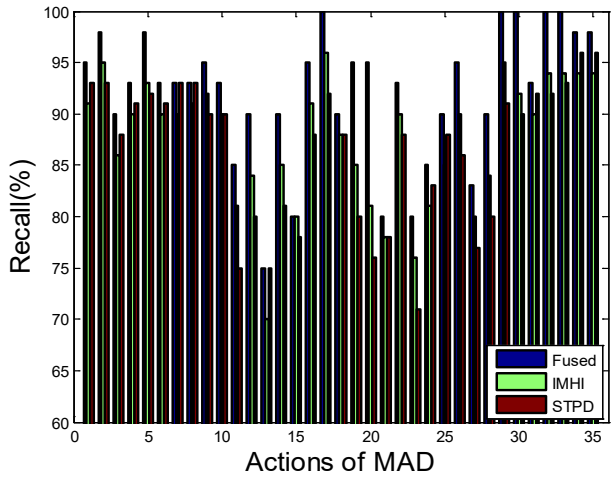

Figure. 7 Recall (%) for different actions of MAD dataset

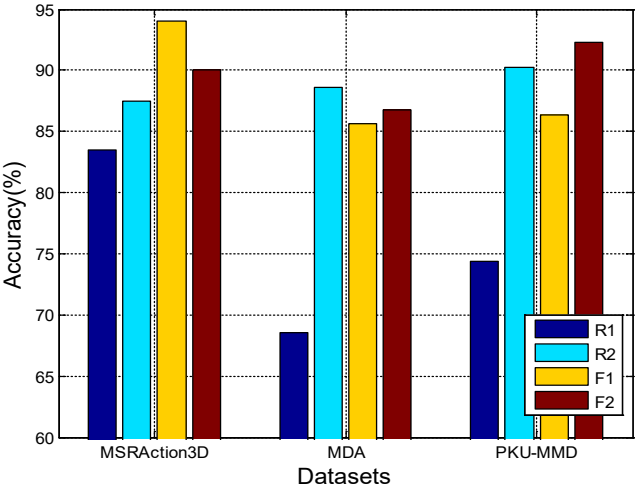

Figure. 9 Accuracy at different fusion rules for different datasets

2DCNN, STPD + 2DCNN and IMHI + STPD + 2DCNN. From the results, it was noticed that the four actions namely left arm throw, left arm punch, left arm pointing to ceiling and left arm wave had gained similar recall rate and it is approximately noticed as 90.2314%, 95.2233%, 85.4120% and 93.2214% respectively. On the other hand, the IMHI had shown its significance on four actions those were performed by left hand. Even though IMHI had shown better performance, the STPD had shown relatively great performance because the features spatiotemporal features extracted are distinctive and ensure more discrimination between the actions with similar movements. The average recall of IMHI, STPD and Fused descriptor with common 2DCNN is observed as 86.5413%, 88.9633% and 92.4712% respectively.

Fig. 8 shows the recall of PKU-MMD dataset fort three combinations IMHI + 2DCNN, STPD + 2DCNN and IMHI + STPD + 2DCNN. From the results, for more number of actions, the proposed fused descriptor had gained 100% recall. In this case also, the STPR had shown its significance in the detection of actions with similar movements. Further, we found that the actions namely 'typing on keyboard', is misclassified as 'tear up paper', 'taking a selfie' is misclassified as 'hand waving', 'rub two hands' and 'clapping' are misclassified each other. Since these actions are similar in nature, they are misclassified. The average recall of IMHI, STPD and Fused descriptor with common 2DCNN is observed as 86.3137%, 88.5686% and 93.3333% respectively.

As our method proposed to apply multi-fusion mechanism, we conduct a simulation study over three action datasets with different fusion rules. The performance is described through accuracy, as shown in Fig. 9. From Here the R1 denotes the output of softmax layer of layer 1 it is a vector with N number of elements where N is equal to the number of classes. Each value is a probability of an action to be the

Table. 1 classification accuracy at different validations on different datasets

Fusion	MSRAction3D	MAD	PKU-MMD
R1	83.6522	69.7845	74.4555
R2	88.9645	88.4471	90.2352
F1	95.2333	90.3692	93.3141
F2	91.2345	91.9945	89.9963
Max	**95.2333**	**91.9945**	**93.3141**

correct class in the layer 1. Similarly, the R2 denotes the output of softmax layer of layer 2 and each value denotes a probability of an action to be the correct class in the layer 2. Next, F1 is a *Max* based fusion operation which determines the maximum value by comparing each value in R1 and R2 element wise. Finally F2 is a product (*Prod*) based fusion operation that calculates the dot product between the elements of R1 and R2. At last we apply base fusion on the results of R1, R2, F1 and F2 to determine the final action label. From the results, we can see that the F1, R2 and F2 had shown effective performance for MSRAction3D, MAD and PKU-MMD datasets respectively. The average accuracy is observed as 94.8964%, 88.2231% and 92.34152% respectively.

To further alleviate the performance of proposed approach, we conduct a case study with cross subjects, i.e., the subjects used for training and testing are different. For MSRAtion3D and MAD, we consider the actions of first 10 subjects for training and the actions of remaining 10 subjects for testing. Next, for PKU-MMD dataset, we used the actions of first 33 subjects for training the actions of remaining 33 subjects for testing. At this case study also, we applied different fusion rules and the observed accuracies are shown in Table 1. From the results, we observed that the F1, F2 and F1 had shown better accuracy for MSRAction3D, MAD and PKU-MMD datasets respectively. The maximum values of accuracies obtained at fusion operations F1, F2 and F1 for MSRAction3D, MAD and PKU-MMD datasets respectively are 95.2333%, 91.9945% and 93.3141% respectively.

4.3 Comparison

Compared to the single data modality, the HAR with multiple data modalities has more capability in HAR. As there exists some features which cannot be revealed by one data model, they may get revealed by other data model. Hence, recent researchers on HAR have concentrated on the integration of multiple data modalities. The earlier depth map based methods [40, 14] had shown poor performance in HAR due to their inefficiency at the removal of obstacles like undefined motion regions due to small body shaking movements, jumbled objects etc. Compared to the depth maps based HAR methods, the skeleton based methods had shown an improved performance because; the joints distinctively represent an action with perfect movements. Among the skeleton based methods, [43] and [44] applied long short-term memory in different variations but they had shown limited performance than the proposed method on PKU-MMD dataset. They didn't concentrate on the provision of view invariance because they considered raw skeleton joints information directly.

Among the hybrid models, X. Ji et al. [30] integrated depth maps in skeleton model and trained the system to recognize actions effectively. But, the Spatio-temporal features are extracted from the complete video which could not ensure better performance at different actions with different speeds. A. Kamel et al. [29] adapted depth maps and skeleton joints and proposed a new CNN model with three layers. They also proposed two new descriptors called as DMI and MJD for depth maps and joints respectively. However, the external effects in depth maps like jumbled objects and small body shaking movements are not eliminated before processing video for feature extraction. Moreover, they were not provided local motion attributes to the HAR system which results in more misclassification at the action with similar movements. Even though Q. Cheng *et al.* [32] adapted for temporal attention model along with cross model learning, they didn't reveal the inherent characteristics of actions with different speeds and time periods. Further, we can see that the single data model based HARs have gained very less recognition performance. Even though CPPCR [33] provisioned collaborative learning, no pre-processing is applied over input data modalities. Finally, on an average, the proposed method gained an improvement of 14% from depth maps based methods, 7% from skeleton based methods and 3% from hybrid methods.

5. Conclusion

This paper proposed an integrated HAR framework by integrating two data modalities such as depth maps and skeleton. In each model, every action is newly described through two new descriptors such as IMHI and STPD of Depth maps and skeleton joint respectively. IMHI concentrated on the nullification of several undefined motion regions present due to small body shaking movements and jumbled objects. Next STPD concentered on the provision of both local and global motion information such that the similar actions can get recognized effectively.

Table. 2 accuracy comparison of the proposed method with existing method on different datasets

Author	Data Modalities	Method	Dataset	Accuracy (%)
W. Li et al. [40]	Depth maps	Bag of 3D Points	MSRAction3D	74.7077
C. Chen et al. [14]	Depth maps	DMM, LBP and Fisher kernel	MSRAction3D	89.5214
S. Song *et al.* [43]	Skeleton	LSTM	PKU-MMD	83.7000
		SA-LSTM	PKU-MMD	86.3000
		TA-LSTM	PKU-MMD	86.6000
		STA-LSTM	PKU-MMD	86.9000
P. Elias *et al.* [44]	Skeleton	Bi-LSTM	PKU-MMD	86.5000
A. Kamel *et al*. [29]	Depth maps and Skeletons	DMI, MJD and CNN	MSRAction3D	94.5100
			MAD	91.8600
X. Ji et al. [30]	Depth maps and Skeletons	Segmentation followed by STSP	MSRAction3D	90.8000
Q. Cheng *et al.* [32]	Depth maps and Skeletons	SITAM and CMIM	PKU-MMD	93.1000
C. Liang *et al*. [33]	Depth maps and Skeletons	Temporal Segmentation and CPPCR	MSRAction3D	94.1800
Proposed	Depth maps and Skeletons	Temporal Segmentation, IMHI, STPD and 2DCNN	MAD	**91.9945**
			PKU-MMD	**93.3141**
			MSRAction3D	**95.2333**

Alongside, we also proposed a new temporal segmentation mechanism which ensures an invariance to speed variations. The proposed new 2D CNN model is simple and less complex and used different fusion rules to determine the action. Simulation on three standard datasets proves the effectiveness of proposed approach in terms of recognition accuracy. The average improvement in accuracy is observed as 7.93%, 6.13% and 0.1345% for MSRAction3D, PKU-MMD and MAD datasets respectively.

References

[1] T. K. Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection", *IEEE Trans. Pattern Anal. Mach. Intell*., Vol. 31, No. 8, pp. 1415-1428, 2009.

[2] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context", In: *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR*), Miami, FA, USA, pp. 2929-2936, 2009.

[3] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response", *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 38, No. 1, pp. 14-29, 2016.

[4] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition", In: *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NA, USA, pp. 1933-1941, 2016.

[5] Z. Wang, S. Liu, J. Zhang, S. Chen, and Q. Guan, "A Spatio-temporal CRF for human interaction understanding", *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 27, No. 8, pp. 1647-1660, 2017.

[6] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton based human action recognition with global context-aware attention LSTM networks", *IEEE Trans. Image Process.*, Vol. 27, No. 4, pp. 1586-1599, 2018.

[7] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey", *Pattern Recognition*, Vol. 60, pp. 86-105, 2016.

[8] Shaikh, M. Bilal, and D. Chai, "RGB-D Data-Based Action Recognition: A Review", *Sensors*, Vol. 21, No. 12, p. 4246, 2021.

[9] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d Normals for activity recognition from depth sequences", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, USA, pp. 716-723, 2013.

[10] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences", In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 804-811, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks",

Communications of the ACM, Vol. 60, No. 6, pp. 84-90, 2017.

[12] M. Lin, Q. Chen, and S. Yan, "Network in network", *arXiv preprint arXiv:1312.4400*, 2013.

[13] M. A. Faris, J. Chiverton, Y. Yang, and D. Ndzi, "Deep Learning of Fuzzy Weighted Multi-Resolution Depth Motion Maps with Spatial Feature Fusion for Action Recognition", *J. Imaging*, Vol. 5, No. 82, pp. 1-25, 2019.

[14] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps", *Journal of Real-time Image Processing*, Vol. 12, No. 1, pp. 155-163, 2016.

[15] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action recognition from depth sequences using depth motion maps-based local binary patterns", In: *Proc. of IEEE Winter Conf., on Applications of Computer Vision (WACV)*, Waikola, HI, USA, pp. 1092-1099, 2015.

[16] J. Li, X. Ban, G. Yang, Y. Li, and Y. Wang, "Real-time human action recognition using de6th motion maps and convolutional neural networks", *International Journal of High Performance Computing and Networking*, Vol. 13, No. 3, pp. 312-320, 2019.

[17] Z. Li, Z. Zheng, and F. Lin, et al. "Action Recognition from depth sequence using depth motion maps based local ternary patterns", *Multimedia Tools Appl.*, Vol. 78, pp. 19587-19601, 2019.

[18] X. Weiyao, W. Muqing, Z. Min, L. Yifeng, L. Bo, and X. Ting, "Human Action Recognition Using Multilevel Depth Motion Maps", *IEEE Access*, Vol. 7, pp. 41811-41822, 2019.

[19] T. Yang, Z. Hou, J. Liang, Y. Gu, and X. Chao, "Depth Sequential Information Entropy Maps and Multi-Label Subspace Learning for Human Action Recognition", *IEEE Access,* Vol. 8, pp. 135118-135130, 2020.

[20] D. Warchol and T. Kapuscinski, "Human Action Recognition Using Bone Pair Descriptor and Distance Descriptor", *Symmetry*, Vol. 12, No. 1580, pp. 1-12, 2020.

[21] X. Diao, X. Li and C. Huang, "Multi-Term Attention Networks for Skeleton-Based Action Recognition", *Appl. Sci.*, Vol. 10, No. 5326, pp. 1-19, 2020.

[22] X. Zhang, C. Xu, X. Tian, and D. Tao, "Graph edge convolutional neural networks for skeleton-based action recognition", *IEEE Trans. Neural Netw. Learn. Syst.*, Vol. 31, No. 8, pp. 3047-3060, 2020.

[23] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics guided neural networks for efficient skeleton-based human action recognition", In: *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, pp. 1112-1121, 2020.

[24] Z. Shao, Y. Li, Y. Guo, X. Zhou, and S. Chen, "A hierarchical model for human action recognition from body-parts", *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 29, No. 10, pp. 2986-3000, 2019.

[25] Q. Nie, J. Wang, X. Wang, and Y. Liu, "View-invariant human action recognition based on a 3D bio-constrained skeleton model", *IEEE Trans. Image Process.*, Vol. 28, No. 8, pp. 3959-3972, 2019.

[26] C. Li, C. Xie, B. Zhang, J. Han, X. Zhen and J. Chen, "Memory attention networks for skeleton-based action recognition", *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 33, No. 9, pp. 4800-4814, 2022.

[27] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition", In: *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, pp. 12026-12035, 2019.

[28] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition", In: *Proc. of IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, pp. 1227-1236, 2019.

[29] A. Kamel, B. Sheng, Y. Po, P. Li, and R. Shen, "Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 49, No. 9, pp. 1806-1819, 2019.

[30] X. Ji, J. Cheng, W. Feng, and D. Tao, "Skeleton embedded motion body partition for human action recognition using depth sequences", *Signal Process*, Vol. 143, pp. 56-68, 2018.

[31] Y. Fan, S. Weng, Y. Zhang, B. Shi and Y. Zhang, "Context-Aware Cross-Attention for Skeleton-Based Human Action Recognition", *IEEE Access*, Vol. 8, pp. 15280-15290, 2020.

[32] Q. Cheng, Z. Liu, Z. Ren, J. Cheng, and J. Liu, "Spatial-Temporal Information Aggregation and Cross-Modality Interactive Learning for RGB-D-Based Human Action Recognition", *IEEE Access*, Vol. 10, pp. 104190-104201, 2022.

[33] C. Liang, D. Liu, L. Qi, and L. Guan, "Multi-Modal Human Action Recognition with Sub-Action Exploiting and Class-Privacy Preserved Collaborative Representation Learning", *IEEE Access*, Vol. 8, pp. 39920-39933, 2020.

[34] H. Wei and N. Kehtarnavaz, "Simultaneous Utilization of Inertial and Video Sensing for Action Detection and Recognition in Continuous Action Streams", *IEEE Sensors Journal*, Vol. 20, No. 11, pp. 6055-6063, 2020.

[35] M. E. U. Haq, A. Javed, M. A. Azam, H. M. A. Malik, I. Aun, I. H. Lee, and M. T. Mahmood, "Robust Human Activity Recognition Using Multimodal Feature-Level Fusion," *IEEE Access*, Vol. 7, pp. 60736-60751, 2019.

[36] X. Wang, T. Lv, Z. Gan, M. He, and L. Jin, "Fusion of Skeleton and Inertial Data for Human Action Recognition Based on Skeleton Motion Maps and Dilated Convolution", *IEEE Sensors Journal*, Vol. 21, No. 21, pp. 24653-24664, 2021.

[37] J. Cheng, Z. Ren, Q. Zhang, X. Gao, and F. Hao, "Cross-Modality Compensation Convolutional Neural Networks for RGB-D Action Recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 32, No. 3, pp. 1498-1509, 2022.

[38] X. Weiyao, W. Muqing, Z. Min, and X. Ting, "Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition", *IEEE Sensors Journal*, Vol. 21, No. 17, pp. 19157-19164, 2021.

[39] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257-267, 2001.

[40] W. Li, Z. Zhang, and 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*, San Francisco, CA, USA, pp. 9-14, 2010.

[41] D. Huang, S. Yao, Y. Wang, and F. D. L. Torre, "Sequential max-margin event detectors", In: *Proc. of European Conference on Computer Vision*, Zurich, Switzerlan, pp. 410-424, 2014.

[42] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding", In: *Proc. of the Workshop on Visual Analysis in Smart and Connected Communities*, Mountain View, CA, USA, pp. 1-8, 2017.

[43] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention based LSTM networks for 3D action recognition and detection", *IEEE Trans. Image Process.*, Vol. 27, No. 7, pp. 3459-3471, Jul. 2018.

[44] P. Elias, J. Sedmidubsky, and P. Zezula, "Understanding the gap between 2D and 3D skeleton-based action recognition", In: *Proc. of IEEE Int. Symp. Multimedia (ISM)*, San Diego, CA, USA, pp. 192-1923, 2019.