Residual Neural Networks for Human Action Recognition from RGB-D Videos

K. Venkata Subbareddy [1,*], B. Pavani [2], G. Sowmya [3], and N. Ramadevi [4]

[1] Department of Electronics and Communication Engineering (ECE), Osmania University, Hyderabad, India
[2] Indian Institute of Technology (IIT) Bombay, Maharashtra, India
[3] Department of Electronics and Communication Engineering (ECE), Santhiram Engineering College, Andhra Pradesh, India
[4] Department of Computer Science Engineering (CSE), Santhiram Engineering College, Andhra Pradesh, India
*Correspondence: subvishk03@gmail.com (K.V.S.)

Abstract—**Recently, the RGB-D based Human Action Recognition (HAR) has gained significant research attention due to the provision of complimentary information by different data modalities. However, the current models have experienced still unsatisfactory results due to several problems including noises and view point variations between different actions. To sort out these problems, this paper proposes two new action descriptors namely Modified Depth Motion Map (MDMM) and Spherical Redundant Joint Descriptor (SRJD). MDMM eliminates the noises from depth maps and preserves only the action related information. Further SRJD ensures resilience against view point variations and reduces the misclassifications between different actions with similar view properties. Further, to maximize the recognition accuracy, standard deep learning algorithm called as Residual Neural Network (ResNet) is used to train the system through the features extracted from MDMM and SRJD. Simulation experiments prove that the multiple data modalities are better than single data modality. The proposed approach was tested on two public datasets namely NTURGB+D dataset and UTD-MHAD dataset. The testing results declare that the proposed approach is superior to the earlier HAR methods. On an average, the proposed system gained an accuracy of 90.0442% and 92.3850% at Cross-subject and Cross-view validations respectively.**

Keywords—**human action recognition, depth maps, Skeleton joints, view invariance, Residual Neural Network (ResNet), F-score**

I. Introduction

From the past few years, Human Action Recognition (HAR) has gained significant research interest in the field of computer vision. HAR is involved in numerous applications like visual surveillance [1], video streaming [2], Healthcare [3], gaming entertainment [4] and Complex objects movements' detection [5]. Generally, HAR is executed by considering the RGB videos as input [6, 7]. However, RGB videos are constrained to several challenges like different illuminations, viewpoints, sizes, colors, shapes, clothing texture and background noise. Moreover different persons perform even a single action in different ways and in such case the RGB videos based HAR has Limited performance.

To overcome these issues, recently, the HAR research is diverted in other direction where it considers the RGB-D videos as input [8]. Microsoft Kinect and other 3D-sensing devices have made it possible to record a person's 3D body shape and motion, adding a new dimension to the data (called as RGB-D data) on human movement. Unlike the conventional RGB videos, RGB-D videos comprises of additional depth information that ensure efficient motion analysis. The major data modalities of RGB-D are two; they are depth images and Body postures or Skeleton Joints. Depth images are robust for illumination variations and ensure uniformity in color and provide depth information that clears the ambiguity in motion. Under the body postures modality, action is represented through joints and each joint is represented with three positions. Due these advantages, most of current researchers focused on the RGB-D data based HAR. However, the existing methods have suffered from several problems. They are listed as follows;

- The existing skeleton based HAR methods mostly concentrated on the single view, i.e., the actions used for training and testing is acquired from only one view. In such case, the same action captured in different views may or may not get recognized. HAR in such instances is called as Cross view HAR which is tough task. Moreover, in skeleton based methods, the past researchers didn't concentrate on the redundancy of joints which constitutes computational burden on the recognition system [9].
- Depth images are composed of different types of noises like small body shaking movements, jumbled objects, cluttered backgrounds, ghost shadows etc. Due to these noises, the action descriptor consists of fake moving pixels which consequences to less recognition accuracy.

To overcome the above mentioned problems with individual modalities, this paper modeled a new HAR system that considers both Depth maps and skeleton

Manuscript received May 25, 2023; revised June 19, 2023; accepted July 21, 2023.

doi: 10.18178/joig.11.4.343-352

joints as inputs. Our system strengthens the weakness of the single data modality. In summary, the major contributions of this paper are outlined as follows:

- To remove the fake moving pixels from depth maps, this work proposes a new depth action descriptor named as Modified DMM (MDMM) which measures Weighted Motion Depth (WMD) for each pixel in the 2D depth image and discards if it found to have an undefined depth value.
- To ensure a least redundancy, this work proposes a new skeleton joint descriptor called as Redundant Joint Descriptor (RJD) which measures the redundancy of each joint through entropy measure. Based on the entropy values, the joints which contribute more redundancy are removed from each frame of input action sequence.
- To ensure the view invariance, this work proposes to transform the skeleton joints from Cartesian Coordinate System (CCS) to Spherical Coordinate System (SCS). SCS represent each joint with distance and angular deviation from other joints.

Rest of the article is organized as follows; the details of literature survey are explored in Section II. The details of proposed HAR system through newly proposed descriptors are explored in Section III. Section IV demonstrates the details of experimental analysis and the Section V concludes the paper.

II. Related Work

Depth maps and skeleton joints have additional depth coordinate which provides more information about the motion of an action. Hence, most of current researchers on HAR used either depth maps or skeleton joints or both data modalities. Hence, we surveyed both the depth map based HAR methods and skeleton joints based HAR methods.

A. Depth Maps

Wu *et al.* [10] proposed a Dynamic Image Sequence (DIS) which focused on Spatio-Temporal Attention Points and describes the action through local Spatio-temporal dynamics. Then they modeled Channel Attention (CA) model based CNNs for feature extraction followed by classification. Even though they employed deep learning for feature extraction, they didn't ensured a perfect discrimination between fake and original moving pixels.

Wang *et al.* [11] derived three compact representations from depth maps for action recognition; they are namely "Dynamic Depth Motion Normal Images (DDMNI)", "Dynamic Depth Normal Images (DDNI)" and "Dynamic Depth Images (DDI)". These dynamic images are generated from segmented sequence of depth maps through "Hierarchical Bi-Directional Rank Pooling (HBRP)". Over the obtained representations, they applied ConvNets for action prediction. This method had limited performance for the videos captured from real time which composed of several artifacts like Shadows and Jumbled objects.

Even though DMM can capture Spatio-temporal depth cues related to motion, it neglects static information. By including static formation, Xu *et al.* [12] 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 over all the three planes and encoded them with Local Binary Pattern (LBP) LBP followed by Fisher Kernels. For classification, they employed Kernalized Extreme Learning Machine (KELM) algorithm. However, MHI and SHI are very sensitive to noises due to fake moving pixels which have less interconnectivity. The motion region has larger pixel connectivity while the non-motion region has less pixel connectivity.

Yang *et al.* [13] 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 through SVM for action prediction. However, entropy based motion computation could not ensure a better discrimination between actions with similar movements, for example actions like *'draw cross'* and "*draw tick*".

Sanchez-Caballero *et al.* [14] proposed two deep learning architectures for HAR from raw depth videos. They are namely Stateless Conv_LSTM and State full Conv_LSTM. The later model allows the HAR system to accumulate the discriminative features from previous frames without showing impact on the memory of computer.

B. Skeleton Joints

Shao *et al.* [15] proposed a hierarchal model which simultaneously selects discriminated body parts at same scale and groups the bundles of body parts at different scales. At preprocessing, they decomposed the entire skeleton joints into hierarchical body parts with different scales. Then, a descriptor called as "Hierarchical Rotation and Relative Velocity (HRRV)" is proposed to describe the hierarchy of body parts and then the encoded through Fisher vectors. However, the HRV is sensitive to viewpoints variations.

To handle view point variations and noisy skeleton joints, Nie *et al.* [16] proposed a view invariant mechanism which recovers the damaged skeleton joints based on 3D bio-constrained model and visualizes the motion features at body level at recovering process. Two constraints namely joint's motion limit and fixed length of bones are defined under 3D bio-constrained model. Two motion features namely Joint Euler Angles (JEAs) and Euclidean Distance Matrix between Joints (JEDM) are derived for representing human action. Further, two stream deep learning models [17] are employed to train the system. Even though it is view invariant, it could not encodes the spatial-temporal relations between Skeleton joints.

Recently, Warchol and Tomasz [18] proposed a new Bone pair Descriptor (BPD) and five different time series classification models for recognizing human actions from skeleton joints. Under BPD, initially the bone pairs are subjected to compute their angular correlation and then they are combined with distance descriptor which explores the spatial relationship between skeleton joints. Five classifiers are employed for classification; they are 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)". However, BPD introduces larger storage burden on the recognition system because it constructs a symmetrical matrix with equal dimensions.

Nguyen *et al.* [19] proposed an improved version of Double-Feature Double-Motion Network (DD-Net) [20] called as Double-Feature Double-Motion Network (DD-Net) which solves the problem of weak connections with global trajectories. Alongside, TD-Net is added as an additional branch which takes the Normalized Coordinates of Joints (NCJ) to highlight the spatial information. However, DD-Net can't ensure resilience against multiple views.

Xie *et al.* [21] proposed a Spatio-Temporal Mixing of Global and Local Self attention Graph Convolutional Network (STGL-GCN) for HAR from Skeleton Joints data. The global self-attention matrix can acquire the dependencies between non-physical correlations between joints and while the local self-attention matrix captures the connection strength of local edges between the joints. But, the GCN's are unable to handle the long-distance between joints. Hence, Rahevar *et al.* [22] proposed a Spatio-Temporal Dynamic Graph Attention Network (ST-DGAT) which integrates the self-attention mechanism with graph convolutions to extract significant joints information. ST-DGAT also leans the Spatio-temporal patterns of skeleton frames.

C. Hybrid Methods

Even through the individual modality based HAR has gained significant accuracy, they can't solve the problems with other models. Hence, some of the recent researchers considered multiple data modalities for HAR. Kamel *et al.* [23] considered both body postures and depth maps as inputs and proposed a new CNN model with three channels. Further, they proposed two new action descriptors such as "Depth Motion Image (DMI)" for depth maps and "Motion Joint Descriptor (MJD)" for body postures. For final action prediction, they suggested several fusion rules. DMJ is a basic motion descriptor which doesn't have any additional capabilities to label the non-motion regions. Due to this reason, it has less performance at Cross subjects validation.

Fan *et al.* [24] proposed a cross attention module for HAR based on the integration of several self-attention branches. Cheng *et al.* [25] considered RGB and Depth data modalities as inputs for HAR and proposed a "Spatio-temporal Information Aggregation Module (SITAM)" model which utilizes CNNs to acquire Spatio-temporal information. Further, they introduced a "Cross Modality Interactive Module (CMIM)" to aggregate the multi-modal complementary information. Finally, an integrated model called as "Multi-modal interactive network (MIMINet)" is proposed by fusing the SITAM and CMIM.

Hafeez *et al.* [26] proposed a hybrid descriptor and Logistic Regression (LR) based HAR. Two data modalities namely inertial sensor and RGB silhouettes are considered as input. The action is described through various features set like Geometric, Skewness, Entropy and Temporal Movement. Further the features are optimized with Zero Order Optimization method and then fed to LR for classification.

Compared to the single data modality, multiple data modalities ensure an increased recognition performance in HAR. However, there are several constraints: 1) inappropriate data modalities creates additional storage burden; 2) Pure deep learning algorithms can't ensure the perfect discrimination between actions. 3) Fake moving pixels are not nullified. The major novelty of this work is of two fold: 1) In the case of depth maps, the past methods didn't focus on the nullification of fake moving pixels. Towards such problem, this work proposed MDMM which monitors each pixel and discards if it found as a fake moving pixel; 2) Next, view invariance and joints redundancy is not concentrated much in earlier skeleton based HAR. Towards such task, this work proposed SRJD which ensures less storage and view invariant action recognition.

III. Proposed HAR Framework

Fig. 1 shows the overall block diagram of proposed HAR system which considers both depth maps and skeleton joints data as inputs and recognizes the action based on the fusion of individual output probability scores. For both modalities, we introduced two new descriptors; they are MDMM and Spherical RJD (SRJD). Further at training, we adapted to a standard pre-trained deep learning model called as a ResNet50. The major reason behind the consideration of ReNet50 is its faster training capability at each layer. The ResNet50 model trains the HAR system individually with two different action descriptors and then the obtained results are fused based on different fusion rules. To get the action prediction results, here we apply two types of fusion rules namely maximum fusion and product fusion. For each action, the softmax layer of ResNet50 produces a vector of N probability scores. Due to the consideration two models, each action gets two probability score and they are fed as input to fusion process.

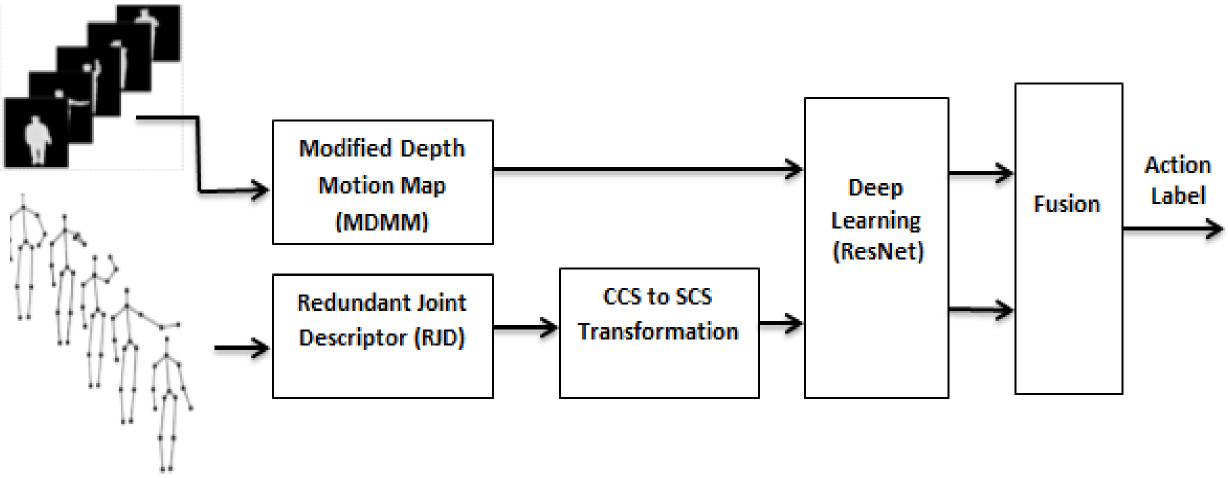

Modified Depth Motion Map (MDMM)
Redundant Joint Descriptor (RJD)
CCS to SCS Transformation
Deep Learning (ResNet)
Fusion
Action Label

Figure 1. Block diagram of proposed HAR framework.

A. Action Descriptor

Under this section, we explore the details of proposed new action descriptors such as MDMM and SRJD. MDMM is derived from depth maps and SRJD is derived from skeleton joints.

1) MDMM

MDMM is an extended version of Depth Motion Map (DMM) introduced by Yang *et al.* [27]. For a given action video sequence, DMM is computed as an accumulated energy of motion information that was thresholded against a particular threshold. Later, Chen *et al.* [28, 29] also introduced one more version of DMM by aggregating the absolute difference between successive frames of a depth action sequence. Chen's DMM is preferred over Yang's DMM, because it has better capability in preserving the spatial motion information. Hence, we consider the same in our work. For a depth action sequence with N frames $V(x,y,t)_{t=0,\ldots,N-1}$ the DMM is computed as:

$$DMM = \sum_{t=0}^{N-2} |V(x,y,t+1) - V(x,y,t)| \qquad (1)$$

DMM is more advantageous because it captures the shape and motion cues of an action. For an input action sequence, DMM computes a spatial energy distribution map which ensures a perfect discrimination between the actions. But, DMM is susceptible to the presence of some undefined regions incurred due to camera's unstable reflection and video's low resolution. In addition, the depth videos comprised of jumbled objects, Ghost Shadows and pixels with undefined depth values. Moreover, the small movement due to body shakings also produces some fake moving edges at the boundaries of subject. These edges are called as false edges which don't have any significance in the provision of additional motion information. Hence, this work developed a modified version of DMM called as MDMM; the block diagram is shown in Fig. 2. At first, MDMM build a binary action video sequence $D_B(x,y,t)_{t=0,\ldots,N-2}$ by comparing the successive frames as

$$D_B(x,y,t) = \begin{cases} 1, & if\ V(x,y,t) \neq V(x,y,t+1) \\ 0, & Otherwise \end{cases} \qquad (2)$$

Next, to find out whether the motion at the corresponding pixel is significant or not, MDMM computes WMD (w_m). For this purpose, MDMM locates as a spatial window over every pixel and compute WMD as:

$$w_m(x,y,t) = \frac{1}{(P+1)(P+1)} \sum_{x-\left(\frac{P}{2}\right)}^{x+\left(\frac{P}{2}\right)} \sum_{y-\left(\frac{P}{2}\right)}^{y+\left(\frac{P}{2}\right)} D_B(x,y,t) \qquad (3)$$

where P is the size of spatial window, and $w_m(x,y,t)$ denotes WDM of a pixel located at (x,y,t). Based on $w_m(x,y,t)$, and a depth motion threshold D_T , a difference map $D_m(x,y,t)$ is generated between successive frames as:

$$D_m(x,y,t) = \begin{cases} V(x,y,t) \neq V(x,y,t+1), if S_m(x,y,t) \geq D_T \\ \quad 0, \qquad\qquad Otherwise \end{cases} \qquad (4)$$

Based on our simulation experiments we found that our method has shown optimal performance for the values of P = *8* and $D_T = 0.6$. Finally the MDMM is constructed based on the accumulation of difference maps as

$$MDMM = \sum_{t=0}^{N-2} D_m(x,y,t) \qquad (5)$$

Fig. 3 shows some examples of the DMM and MDMM. These figures shows a clear spatial energy distribution maps which has no fake moving regions and consists of only original motion information.

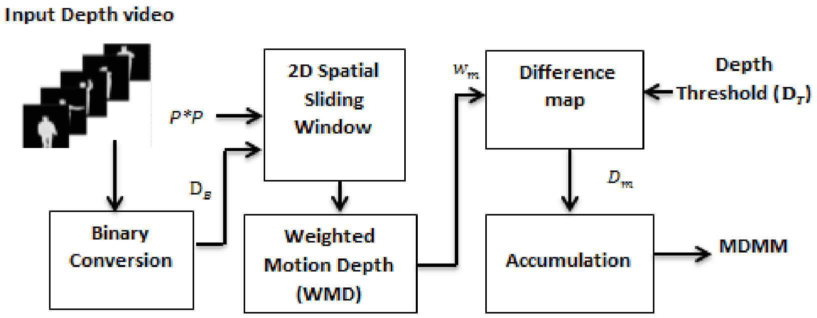
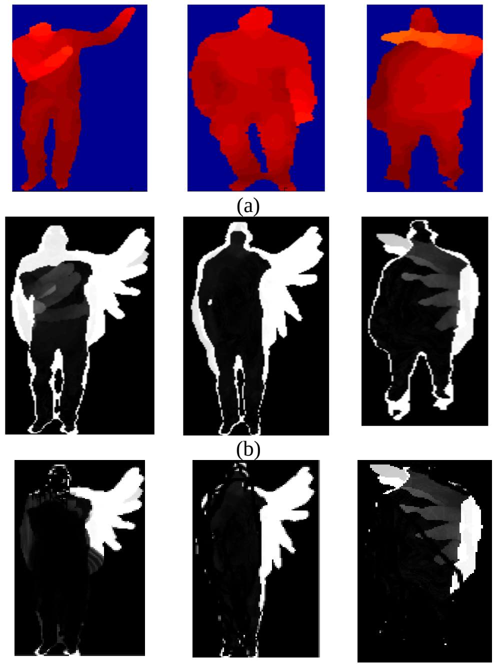


Figure 2. Block diagram of MDMM.


(a)

(b)

(c)

Figure 3. (a) Original depth frame, (b) DMM output and (c) MDMM output.

2) SRJD

SRJD aimed to describe an action with only few and view invariant joints. For a given action frame with N number of joints, only few joints has significant contribution towards the action representation. Hence, the joints which contribute less motion information can be regarded as redundant joints. Examples of such joints are *'hip centre'* and *'spine centre'* which are non-moving in nature. For example consider an action called *'High Arm wave'* from MSR-Action 3D dataset, only three joints namely *'left hand', 'left wrist' and 'left elbow'* give more information about motion. Similarly, for an action called as '*Forward Kick'*, only the joints at left leg such as *'left hip', 'left elbow', 'left ankle'* and *'left foot'* has maximum motion information. The joints which have less contribution towards an action are called as redundant joints and are needs to be removed.

Considering the above problems as a serious issue, Ofli *et al.* [30] aimed to describe an action with only informative joints and proposed a new descriptor called as "Sequence of Most Informative Joints (SMIJ)". SMIJ computes the variance of each joint from its angular trajectories and selected only few joints those have maximum variance. However, they experienced limited recognition accuracy for the actions (like *'High Arm Wave' and 'Draw X')* those are described with common informative joints. Approximately eight actions of MSR-Action 3D dataset [31] have experienced 0% accuracy because they are executed with only one arm. Unlike SMIJ, we propose to select the informative joints based on the Differential Entropy [32]. The joints are selected which can preserve approximately 80% entropy of an action. Once the informative joints are chosen, then the redundant joints are simply kept as zero.

Consider B be the action video having N frames, it is representation through N fames as $B = \{b_1, b_2, \dots, b_N\}$, where b_i denotes i^{th} frame. Consider K joints are there in each frame, initially; this method compute Euclidean distance between successive frames at same joint positions. Let's assume the spatial coordinates of k^{th} joint in i^{th} and j^{th} frames are denoted as (x_j, y_j, z_j) and (x_i, y_i, z_i) and $d_i^j(k)$ be the k^{th} joint's Euclidean distance, it is obtained as:

$$d_i^j(k) = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}, k \in K \quad (6)$$

Here $d_i^j(k), k = 1, 2, \dots, K$ is used to compute the differential entropy $(H(X_k))$ of k^{th} joint as

$$H(X_k) = -\sum_{i=1}^{N-1} p(d_{ij}^k) \log_b p(d_{ij}^k)\ \forall i, j \in N \quad (7)$$

Eq. (7) is computed for every joint and the informative joints are chosen those have maximum differential entropy. In this manner, each action frame is represented with only selected joints and the redundant joints are simply kept as zero.

Once the redundant joints are removed from each frame, then they are transformed from CCS to SCS. CCS makes the HAR system sensitive to view point variations, i.e., the system trained with actions in one view can't recognize the same action under different views. Hence, we proposed to transform them from CCS to SCS. Generally, the joints of human body are restricted to move beyond certain distance and angel from the center of body (i.e., '*hip center'*). These restricted movements can be taken as an advantage and are used here to describe an action such that the HAR system becomes resilient to view point variations. For this purpose, we considered SCS where the restricted movements of joins are modeled through three coordinates such as r, θ and ϕ. Here r measures the Euclidean distance between *hip center* and corresponding joint. Next, ϕ and θ measures the angular distances with respect to horizontal angle and vertical angle respectively. Let's the joints in CCS are represented as $J = \{O, J_1, J_2, \dots, J_N\}$, in the SCS, they are represented as $J_S = (r, \theta, \phi)$ [21] where

$$r = \sqrt{(x_{HC} - x_i)^2 + (y_{HC} - y_i)^2 + (z_{HC} - z_i)^2} \quad (8)$$

$$\theta = \arccos \frac{z}{r} \quad (9)$$

$$\phi = \arctan \frac{y}{x} \quad (10)$$

where (x_{HC}, y_{HC}, z_{HC}) and (x_i, y_i, z_i) are the Hip Center's and i^{th} skeleton joint's positions respectively.

B. Classification and Fusion

Once the action is described through MDMM and SRJD, then they fed to 2D ResNet50 [33] for feature extraction followed by action prediction. Here we consider 2D ResNet50 which learn high level features from action descriptors. The 2D ResNet50 is separately trained for both modalities individually after fixing the size of each action descriptor to 224×224. Here, we consider ResNet50 for two channels, one for depth maps and another for skeleton joints. The 1st channel considers MDMM and 2nd channel considers SRJD as inputs. After classification at each channel by softmax layer, the output is a vector of actions probabilities. The size of each vector is equal to the length of number of actions trained to the system. Since, we employed two channels, each action has two probabilities and hence we applied fusion mechanism to determine the final action. Let's the output vectors of first and second channel's softmax layers are represented with R_1 and R_2 respectively, the fusion is done as follows:

$$F_1 = Max(R_1, R_2) \quad (11)$$

$$F_2 = Product(R_1, R_2) \quad (12)$$

Based on the results obtained at two fusion strategies, the final action prediction is done as:

$$Action = Max(F_1, F_2) \quad (13)$$

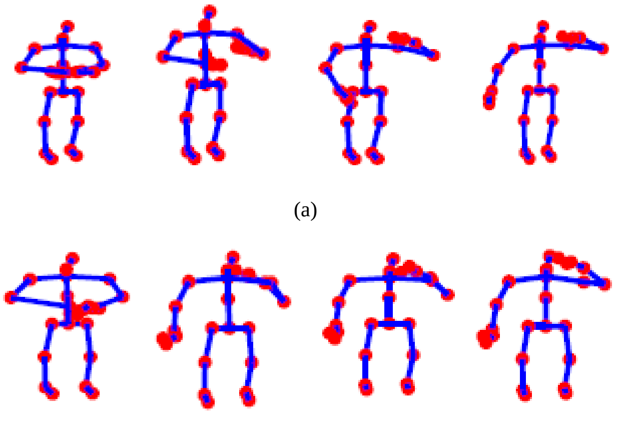
where *Action* signifies the label of action that has largest probability score.

IV. Experimental Investigations

This section explores the details of experimental investigations carried out on the proposed HAR mechanism with two standard datasets namely NTU RGB+D [34] and UTD-MHAD [35].

A. Datasets

1) NTU RGB+D: NTURGB+D is a very large sized dataset, it consists larger number of action samples and classes with rich inter- and intra-class variations. NTURGB+D 60 composed of totally 56,880 depth action sequences and are acquired from 40 subjects. Totally, 60 different actions are acquired with the help of three Microsoft Kinect V2 cameras. NTURGB+D consist of totally four different data modalities; they are depth maps, postures, Infrared videos and RGB videos. For depth maps and IR videos, each frame has a resolution of 512×424 and 1920×1080 respectively. The posture data model is represented with different frames and each frame is described through 25 joints and each joint is represented with three positions *x*, *y* and *z*. Fig. 4 shows example actions of NTURGB+D dataset.



(a)

(b)

Figure 4. Sample actions NTURGB+D dataset (a) Drinking Water and (b) Brushing Teeth.

2) UTD-MHAD: This dataset is captured through eight subjects among which four subjects are male and remaining four subjects are female. This dataset consists of totally 27 actions. Each subject performed the actions repeatedly for four times and resulted in totally 846 sequences. After removing three sequences, the total action sequences count becomes 861.

B. Evaluation Results

Under the evaluation, we conduct extensive experiments on the both datasets and the performance is measured through Detection rate, F-score and Accuracy. Mathematically, these metrics are defined as follows:

$$Detection\ Rate = \frac{TP}{TP+FN} \quad (14)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (15)$$

And

$$F - score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (16)$$

where *TP* stands for True Positive, *TN*, *FN* and *FP* stands for True Negative, False Negative and False Positive Respectively. For a given action, if the proposed HAR system recognizes correctly, then it is considered as *TP* otherwise *FP* or *FN*.

Majorly, we conduct two types of cross validations; they are Cross Subject (CS) validation and Cross View (CV) validation. Since our one of the objective is to make the HAR view invariant, we conduct CV validation by considering different views for training and testing. Similarly, at CS validation, we considered different subjects for training and testing. A one more case study s performed by training the system with individual and combined descriptors. At individual descriptors, the HAR system uses MDMM and SRJD individually whereas at combined descriptors, the system uses both descriptors and the obtained results are fused.

Fig. 5 shows the F-score for different actions of NTURGB+D dataset. From this figure, we observed that the maximum F-score is attained by combined descriptor for an action ‘*wear jacket’* and minimum F-score is attained for an action ‘*wipe face*’. Further it was noticed that the action with inter class similarities experiences lesser F-score due to the similar movements at fingers. For instance, 10% of ‘*drinking water’* action is mistakenly recognized as ‘*brushing teeth*’ because they have similar movement at fingers. A one more action pair for such example is *‘walk towards each other’* and *‘walk apart from each other’*. Such kind of misclassifications is reduced by proposed combined action descriptor. Similar observations are observed based on the F-score shown in Fig. 6 for UTD-MHAD dataset. In UTD-MHAD dataset, approximately 18 actions have gained 100% F-score. Since UTD-MHAD has diversified actions, the proposed approach was succeeded in recognizing almost all actions accurately. On an average, the combined descriptor gained an F-score of 86.3320% while the individual descriptors such as MDMM and RJMD have gained 83.1120% and 82.5600%, respectively.

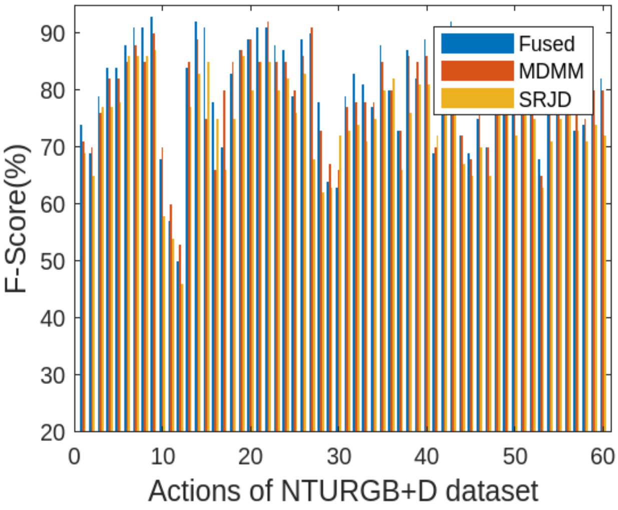


Figure 5. F-Score Comparison at individual and combined descriptors over action in NTURGB+D dataset.

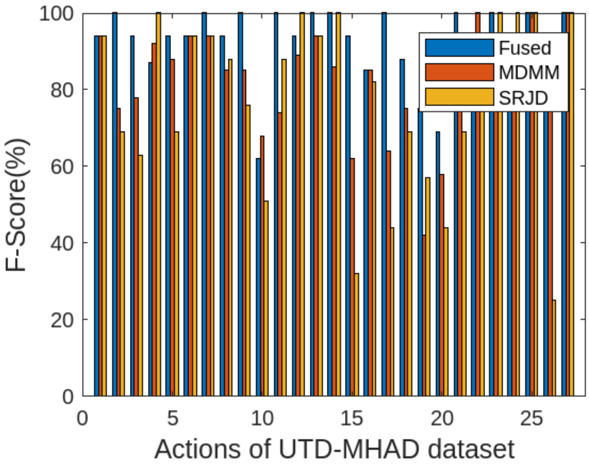


Figure 6. F-Score Comparison at individual and combined descriptors over action in UTD-MHAD dataset.

Table I shows the results of five-fold CS validation of proposed method over NTURGB+D dataset. The maximum accuracy (92.3520%) is observed at successive subjects where the first 20 subjects are used for training and reaming 20 subjects are used for testing. Similarly, Table II shows the results of three-fold CS validation of proposed method over UTD-MHAD dataset. Since the number of subjects of UTD-MHAD is 8, we conduct only three-fold validation. At this case, the proposed method had shown superior performance by achieving an average accuracy of 93.0173%.

TABLE I. FIVE-FOLD VALIDATION ON CROSS SUBJECTS (CS) OF NTU RGB+D DATASET

CS No.	Training Subjects	Testing Subjects	Accuracy (%)
CS1	Even subjects (2,4,…,40)	Odd Subjects (1,3,…,39)	90.3320
CS2	Odd Subjects (1,3,…,39)	Even Subjects (2,4,…,40)	89.3410
CS3	1–20 Subjects	21–40 Subjects	92.3520
CS4	1, 2, 3, 4, 5, 11, 12, 13, 14, 15, 21, 22, 23, 24, 25, 31, 32, 33, 34, 35	6, 7, 8, 9, 10, 16, 17, 18, 19, 20 26, 27, 28, 29, 30, 36, 37, 38, 39, 40	90.9960
CS5	1, 4, 7, 8, 12, 13, 16, 19, 20, 23, 24, 28, 29, 32, 35, 36, 39, 37, 38, 40,	3, 5, 6, 9, 11, 15, 17, 18, 21, 25, 27, 30, 2, 10, 14, 22, 26, 31, 33, 34	87.2000
	Average		**90.0442**

TABLE II. THREE-FOLD VALIDATION ON CROSS SUBJECTS OF UTD-MHAD DATASET

CS No.	Training Subjects	Testing Subjects	Accuracy (%)
CS1	2,4,6,8	1,3,5,7	95.4120
CS2	1,2,3,4	5,6,7,8	93.2200
CS3	1,4,5,8	2,3,6,7	90.4200
	Average		**93.0173**

Tables III and IV show the accuracies of proposed method on NTU RGB+D and UTD-MHAD datasets respectively. At this case study, we can see that the maximum accuracy is achieved only when the system has trained with both descriptors. Because, the combined descriptor provides more knowledge to HAR system about the actions. Furthermore, it can also ensure a perfect discrimination between actions even at finger movements. Such advantage is not available with individual descriptors; hence, they had shown limited accuracy. Particularly, the MDMM has shown its superiority at CS validation while the SRJD had shown its superiority at CV validation. Hence the combined descriptor is regarded as best descriptor which can ensure more accurate recognition even for similar actions.

TABLE III. COMPARATIVE ACCURACY OF PROPOSED METHOD UNDER DIFFERENT COMBINATIONS ON NTU RGB+D DATASET

Method	Cross Subject	Cross View
MDMM + ResNet50	85.2350%	80.2100%
SRJD + ResNet50	82.3300%	87.4520%
SRJD + MDMM + ResNet50	**90.0442%**	**92.3850%**

TABLE IV. COMPARATIVE ACCURACY OF PROPOSED METHOD UNDER DIFFERENT COMBINATIONS ON UTD-MHAD DATASET

Method	Cross Subject	Cross View
MDMM + ResNet50	88.5230%	86.7020%
SRJD + ResNet50	84.5620%	92.3800%
SRJD + MDMM + ResNet50	**93.0173%**	**95.6300%**

Next, to analyze the impact of fusion mechanisms on the action recognition, we conduct a case study by applying different fusion rules on the output probabilities of individual methods. Fig. 7 shows the effect of fusion on recognition accuracy at different cross fold validations. From the results, we can see that the accuracy gained at fusion of scores is more than the individual results such as R1 and R2. Among the two fusion mechanisms, the maximum accuracy is observed at Product fusion, approximately 95.7110%. Even though the fusion takes additional time to derive final scores, its necessity is there when the multiple data modalities are used. Further, the overall detection rates of proposed combined system are explored for all the action of NTURGB+D dataset, results are shown in Fig. 8. From the results, the maximum detection rate is obtained at the action at *'Jump up'* action and minimum detection rate is observed at *'Eat Meal/Snack'*.

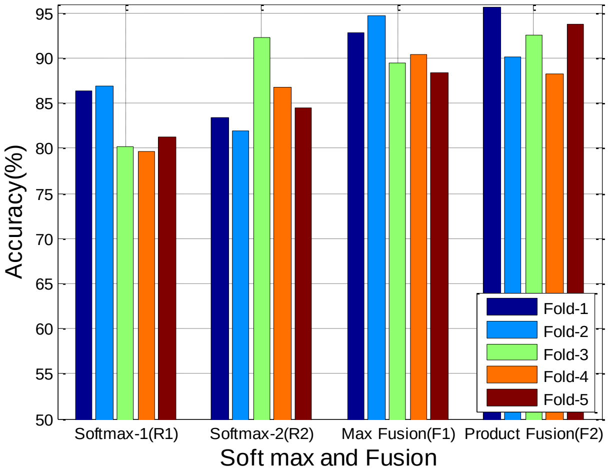


Figure 7. Accuracy comparison between fusion rules at different folds on NTURGB+D dataset.

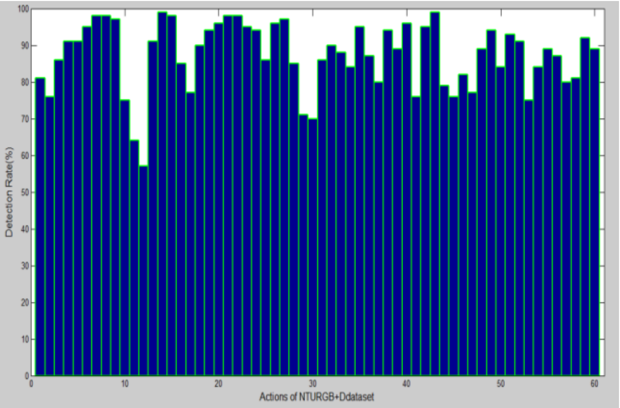


Figure 8. Detection rates of different actions from NTURGB+D dataset

C. Comparison and Discussion

Table V shows the accuracy comparison between proposed and existing methods. This comparison is done between the methods those used a common dataset for validation, i.e., NTURGB+D. As we have mainly aimed at lessening the misclassification rate, the large sized dataset is required to validate because it contains more number of actions. NTURGB+D is one such dataset and hence most of the past methods have considered it for validation. The accuracy values mentioned in this table are all referred from the corresponding articles only. From the comparison, we can see that the proposed approach has attained maximum accuracy in both CS validation and CV validation. Compared with all the existing methods, the proposed approach achieved more accuracy as 90.0442% and 92.3850% at CS and CV validations respectively. These values declare a huge improvement in accuracy from the single data modality based HAR method like [11, 14–16, 34]. Next, the multiple data modality based methods such as [24, 26, 28, 36, 37] has gained a better improvement in accuracy, especially they boost up the performance at CV validations. For instance, Deep Bilinear [22] proposed a view invariant descriptor based on the angular restrictions of human joints and hence they gained approximately 90.70% accuracy at CV. Next, the pure deep learning based Action recognition methods such as Deep LSTM (D-LSTM) and Part aware LSTM (P-LSTM) [34] has achieved an accuracy of 60-70%. These methods are the basic methods applied by the creators of NTURGB+D dataset. The main reason behind less accuracy is that they didn't focus on the effective action descriptor which can nullify the external effects like noises, undefined regions, shadows etc. Even though the combinational methods have gained an improvement in recognition accuracy, they had shown limited performance because most of them adapted only deep learning methods for feature extraction and classification. Directly applying deep learning algorithms over a noisy RGB-D data cannot ensure sufficient discrimination between original moving pixel and fake moving pixel.

TABLE V. COMPARISON BETWEEN PROPOSED AND EXISTING APPROACHES; S: SKELETON, D: DEPTH

Method/Dataset	**Modality**	**Accuracy (%)**	
		CS	**CV**
DDI + DDNI + DDMNI [11]	D	87.1000	84.2000
Stateless Conv-LSTM [14]	D	75.2600	75.4500
State full Conv-LSTM [14]	D	80.4300	79.9100
Deep LSTM [34]	S	60.7000	67.3000
HRRV [15]	S	74.9000	82.0000
STGL-GCN [21]	S	87.1600	88.3000
ST-DGAT [22]	S	90.9000	88.9000
Bio-Constrained [16]	S	86.9000	91.8000
P-LSTM [34]	S	62.9000	70.3000
TD-net [19]	S	38.8000	47.6000
Hybrid Descriptor with LR [26]	Silhouettes + RGB	90.2300	-
Cross Attention [24]	RGB + S	84.2000	89.3000
SC- ConvNets [36]	RGB + D	86.9000	87.7000
Deep Bilinear [28]	RGB + D + S	85.4000	90.7000
TSN [37]	RGB + D	78.9000	79.9000
Proposed	**D + S**	**90.0442**	**92.3850**

V. CONCLUSION

This paper aims at the improvisation of HAR accuracy from RGB-D (Depth maps and Skeleton joints) videos. Towards such aim, two innovative action descriptors are designed namely MDMM and SRJD for depth maps and skeleton joints respectively. MDMM nullifies the external noises like background clusters, ghost shadows from depth maps while SRJD ensure resilience against view point variations. The proposed HAR system is trained with both descriptors through ResNet50. At classification, the obtained results are fused to predict the action. Extensive simulations are carried out over the proposed system through two benchmark RGB-D datasets and gained an improvement of 2.9442% and 4.5230% at NTURGB+D and UTD-MHAD datasets respectively.

Most the current action datasets consist of simple actions and single actions which has common motion patterns. Such kind of actions can be recognized easily with effective action descriptors. However, there exists complex actions which consist of multiple and short actions, for example cooking and car repairing etc. Recognition of these kinds of actions needs very effective recognition system and it is suggested as one more possible future work. Further, the interactions are also considered as complex activities which again classified into three classes; they are Group Activities, Human to Human interaction and Human to Object Interactions.

CONFLICT OF INTEREST:

The authors declare no conflict of interest.

AUTHORS CONTRIBUTIONS:

Conceptualization, design, development, drafting, investigation and experimental validation by K. Venkata Subbareddy; Formal Analysis, and editing by Dr. B. Pavani; Dataset Organization by G Sowmya and Proof reading and Editing by N. Ramadevi; all authors had approved the final version.

REFERENCES

[1] H. Y. Cheng and J. N. Hwang, "Integrated video object tracking with applications in trajectory-based event detection," *J. Vis. Commun. Image Represent.*, vol. 22, no. 7, pp. 673–685, 2011.

[2] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14, no. 7, pp. 11735–11759, 2014.

[3] A. Jalal, J. T. Kim, and T. S. Kim, "Human activity recognition using the labeled depth body parts information of depth silhouettes," in *Proc. 6th Int. Symp. Sustain Healthy Buildings*, Seoul, South Korea, 2012, pp. 1–8.

[4] A. Y. Yang, S. Iyengar, P. Kuryloski, and R. Jafari, "Distributed segmentation and classification of human actions using a wearable motion sensor network," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Workshops*, Anchorage, UK, 2008, pp. 1–8.

[5] A. Jalal and S. Kamal, "Real-time life logging via a depth silhouette-based human activity recognition system for smart home services," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal Based Surveill., (AVSS)*, Seoul, South Korea, 2014, pp. 74–80.

[6] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 1005, pp.1–20, 2019.

[7] K. Yu and F. Yun, "Human action recognition and prediction: A survey," *Int J. Comput. Vis*, vol. 130, pp. 1366–1401, 2022.

[8] P. Khaire and P. Kumar, "Deep learning and RGB-D based human action, human—human and human—object interaction recognition: A survey," *Journal of Visual Communication and Image Representation*, vol. 86, pp. 103531–103556, 2021.

[9] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 12, pp. 2430–2443, 2016.

[10] H. Wu, X. Ma, and Y. Li, "Convolutional networks with channel and STIPs attention model for action recognition in videos," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2293–2306, 2020.

[11] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based large-scale 3-D action recognition with convolutional neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1051–1061, 2018.

[12] W. Xu, M. Wu, M. Zhao, Y. Liu, B. Lv, and T. Xia, "Human Action recognition using multilevel depth motion maps," *IEEE Access*, vol. 7, pp. 41811–41822, 2019.

[13] 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.

[14] A. Sanchez-Caballero, D. Fuentes-Jimenez, and C. Losada-Gutierrez, "Real time human action recognition using raw depth video based recurrent neural networks," *Multimedia Tools Appl*, vol. 82, pp. 16213–16235, 2023.

[15] 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.

[16] 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.

[17] 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, 2021.

[18] D. Warchol and T. Kapuscinski, "Human action recognition using bone pair descriptor and distance descriptor," *Symmetry*, vol. 12, no. 1580, pp. 1–12, 2020.

[19] T. T. Nguyen, D. T. Pham, Ha Vu, and T. L. Le, "A robust and efficient method for skeleton-based human action recognition and its application for cross-dataset evaluation," *IET Computer Vision*, vol. 16, pp. 709–726, 2022.

[20] Y. Fan, S. Sakthi, Y. Wu, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proc. the ACM Multimedia Asia*, Beijing, China, 2019, pp. 1–6.

[21] Z. Xie, G. Zheng, L. Miao, and W. Huang, "STGL-GCN: Spatial–temporal mixing of global and local self-attention graph convolutional networks for human action recognition," *IEEE Access*, vol. 11, pp. 16526–16532, 2023.

[22] M. Rahevar, A. Ganatra, T. Saba, A. Rehman, and S. A. Bahaj, "Spatial–temporal dynamic graph attention network for skeleton-based action recognition," *IEEE Access*, vol. 11, pp. 21546–21553, 2023.

[23] A. Kamel, Bin Sheng, Yang Po, Ping Li, and Ruimin 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.

[24] 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.

[25] 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.

[26] S. Hafeez, S. S. Alotaibi, A. Alazeb, N. A. Mudawi and W. Kim, "Multi-sensor-based action monitoring and recognition via hybrid descriptors and logistic regression," *IEEE Access*, vol. 11, pp. 48145–48157, 2023.

[27] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion map- s-based histograms of oriented gradients," in *Proc. the 20th ACM International Conference on Multimedia*, New York, NA, USA, 2012, pp. 1057–1060.

[28] J. F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang, "Deep bilinear learning for RGB-D action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 335–351.

[29] C. Chen, M. Liu, B. Zhang, J. Han, J. Jiang, and H. Liu, "3D action recognition using multi-temporal depth motion maps and fisher vector," in *Proc. Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, NA, USA, 2016, pp. 3331–3337.

[30] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition," *J. Vis.* Commun. *Image Represent*, vol. 25, no. 1, pp. 24–38, 2014.

[31] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW*), San Francisco, CA, 2010, pp. 9–14.

[32] S. Y. Park and A. K. Bera, "Maximum entropy autoregressive conditional heteroskedasticity model," *Journal of Econometrics*, vol. 150, no. 2, pp. 219–230, 2009.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the IEEE conference on computer vision and pattern recognition,* Lasvegas, NV, USA, 2016, pp. 770–778.

[34] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "NTURGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, US, 2016, pp. 1010–1019.

[35] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, Canada, 2015, pp. 168–172.

[36] Z. Ren, Q. Zhang, J. Cheng, F. Hao, and X. Gao, "Segment spatial temporal representation and cooperative learning of convolution neural networks for multimodal-based action recognition," *Neurocomputing*, vol. 433, pp. 142–153, 2021.

[37] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, 2019.