IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 1, February, 2023, pp. 133-143
Journal homepage: http://iieta.org/journals/ts

Scale and View Invariant Informative Joint Descriptor (SVI^{2}JD) for Human Action Recognition from Skeleton Data

Dustakar Surendra Rao[1,2], Sudharsana Rao Potturu[1*], Vipparthi Bhagyaraju[3]

[1] Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram 500075, Andhra Pradesh, India
[2] Department of ECE, Guru Nanak Institutions Technical Campus, Hyderabad 501506, Telangana, India
[3] Department of ECE, Siddhartha Institute of Engineering and Technology, Hyderabad 501506, Telangana, India

Corresponding Author Email: potturusd54@klh.edu.in

https://doi.org/10.18280/ts.400112

Received: 26 July 2022
Accepted: 28 January 2023

Keywords:
action recognition, view invariance, informative joints, entropy, self-similarity, spherical system

ABSTRACT

One of the biggest challenges in the Human Action Recognition is View-point variations as the actions are captured under multiple views in real time. Furthermore, in the Skelton based action representation, for HAR, only few joints are informative and remaining joints constitutes redundancy. To sort out these problems, this paper proposes a new Action descriptor called as Scale and View Invariant Informative Joint Descriptor (SVI^{2}JD). SVI^{2}JD is a combination of three descriptors; they are namely Self-Similarity Joint Descriptor (SSJD), Informative Joint Descriptor (IJD) and Spherical Joint Descriptor (SJD). SSJD concentrates on the view invariance and employs a Self-Similarity Matrix (S^{3}M) which computes pair wise distance between joints in each frame of action sequence. Next, SJD aims at describing the action through restricted movements of joints because they can't move beyond particular angle and distance from origin of body. IJD removes the redundant joints those have less contribution towards the action. Further, a 2D Convolution Neural Network Model is proposed for feature extraction and classification. Different fusion rules are employed to fuse the individual results. The Effectiveness of proposed model is demonstrated through its simulation on two challenging datasets; NTU RGB+D dataset and Northwestern UCLA dataset.

1. INTRODUCTION

In recent years, the Human Action Recognition (HAR) has become an active research area due to its widespread applicability in different applications including Human-Machine Interaction, video understanding, Gaming, Virtual reality, Video Surveillance etc. [1-3]. However, the recognition of human actions is still a challenging task due to several reasons. (1) The complexity of spatio-temporal process of human behavior, and (2) environmental variations and changes in the settings at recordings including view points, occlusions and complex image backgrounds. Most of the earlier research on HAR has been done by considering the RGB videos an input. However, the action features extracted from RGB videos have so many problems like (1) lack of motion information, (2) human body appearance and (3) Illumination variations. If an action is observed form different viewpoints, it results in different intensity features. Further, Self-occlusion makes the recognition system to perform worse. Particularly, in the case of clutter background, the human body segmentation is very tough. Even though the detection of human body is done accurately, the challenge rises at their representation due to their complex movements. Furthermore, the human actions are also influenced by the emotion shift, personal character and different cultures. In such conditions, the extraction of inter class and intra class variations is very tough task.

To sort out all these problems, recently, the research on HAR has been diverted to other direction where the input data is of depth form [4-7]. With the invention of 3Dsensors like Microsoft Kinect, the 3D information of human body can be captured which provides an add-on feature about the motion information. Due to the provision of depth information, the research on HAR boosted up significantly. Depth images provide the segmented human body from background but it suffers from several problems like noisy data, varying movements in different directions etc. Hence the HAR based on skeleton data has gained an increased attention. As Johansson [8] mentioned that the skeleton is the most effective way for the human action representation.

Most of the earlier methods focused on the recognition of human action form single point of view. However, their performance is limited when the action is captured under different viewpoints. The action recognition under such conditions can be regarded as Cross View HAR and it is very challenging. The visual appearance of same action under different views looks like different [9], as shown in Figure 1. On the other hand, training the actions under multiple views is one possible solution which makes the HAR effective even under cross views. However, the multi-view training constitutes a huge computational complexity as well time complexity. Hence developing a view invariant HAR based on skeleton joints that can generalize better even under cross views is very much required.

In this paper, we develop an effective Skeleton Joint Descriptor called as SVI^{2}JD for action recognition from action sequence. SVI^{2}JD mainly concentrated over three problems namely scale, view variations and occlusions and proposed three different descriptors namely SSJD, SJD and IJD. Next, our method determines the features through a new 2D-CNN model and then fed to classification through fully connected layer. The 2D-CNN model is applied on three individual descriptor and the final results are fused in the third phase. At fusion, we employ different fusion rules to get the final action label. The major contributions of this research work are outlined as below:

- To reduce the features dimensionality, we propose a new Informative Joint Descriptor (IJD) which determines the most informative joints of an action and nullifies redundant joints. Here, the most informative joints have major contribution towards an action and it is determined through differential entropy
- To achieve view-invariance, we propose a new Skeleton Self-Similarity Joint Descriptor (SSJD) in which each action frame is represented through Skeleton Self-Similarity Matrix (S^{3}M). S^{3}M is determined through the computation of Euclidean distances between all joints in each frame.

The remaining paper is structured as follows; section II explores the details of literature survey. Section III explores the details of proposed method. Section IV explores the details of experimental emulations on several standard datasets and section V concludes the paper.

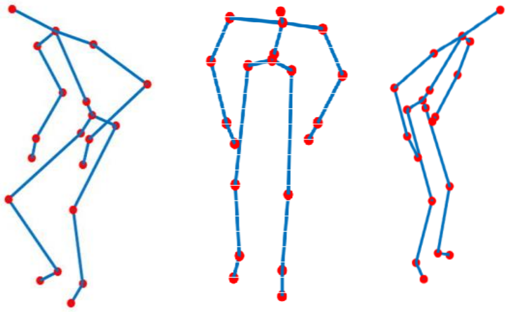


Figure 1. Action under different views

2. LITERATURE SURVEY

With the development of 3D sensor especially Microsoft Kinect, there is a sudden upsurge in the research on HAR. HAR through depth maps is categorized into two categories; they are the method used the depth maps directly and the methods used the skeleton from depth maps. An extensive research has been carried in both categories and some authors tried to integrate the both.

Liu et al. [10] proposed a new version of Long Short-Term Memory (LSTM) network called as Global Context Aware Attention LSTM (GCA-LSTM) for skeleton based action recognition. They considered the Global Memory Cell to select the informative joints from each skeleton frame. Further, they also introduced a recurrent attention mechanism to enhance the capability of their network and the training was done in a step-by-step process. Ke et al. [11] proposed to transform each channel of 3D location coordinates into a clip. The clip generated from each frame explores the temporal information of the overall skeleton sequence and one specific spatial relation between the skeleton joints. Further, they also proposed a Multi-task convolutional neural network (MTCNN) to learn the Spatio-temporal relationships of skeleton sequence.

Rahmani et al. [12] proposed a Roust Non-Linear Knowledge Transfer Model (R-NKTM) for HAR form multiple views. R-NKTM is a fully connected layer that transfers the information of actions from any unknown view the high level virtual view after determining the non-linear relations between them. R-NKTM learned the knowledge from dense trajectories of synthetic 3D human models.

Zhang et al. [13] proposed a view invariant transfer dictionary and view invariance classifier. The dictionary projects the real time 2D video into a view invariant sparse representation and thus the classifier recognizes an action from any arbitrary view. They used synthetic data to determine the view invariance between 2D and 3D videos at the training phase. Further they employed dense trajectories for the effective encoding of action trajectory information.

Liu et al. [14] proposed a three stage view invariant HAR based on an enhanced skeleton visualization method. Initially, the developed a new transform that nullifies the view variations on spatio-temporal locations of skeleton joints. Next, the transformed images are viewed as color images that implicitly encode the skeleton joint's Spatio-temporal information. Finally, they employed CNN based model for the extraction of discriminative features from color images.

Liu et al. [15] proposed to extend the RNN into spatial domain as well as to temporal domain for the better analysis of hidden sources if action related information within the human skeleton sequence. With the help of pictorial structure of skeletal data, they proposed an effective tree structured based traversal framework. To handle the noise data, they proposed a new gating mechanism within LSTM module.

Wang and Wang [16] aimed to leverage the geometric relations between joints based on three primitive geometries such as Joints, Surfaces and Edges. They designed a new RNN framework by utilizing the temporal drop out layers and view point transformation layers to accommodate three inputs. For action detection, they employed a frame wise action classification followed by a multi-scale sliding window algorithm.

Shao et al. [17] build a Hierarchical Rotation and Relative Velocity (HRRV) descriptor to represent the action hierarchy at different scales of same action. They treated the action as a simultaneous movement of body arts and tried to group the bundles of body parts. Then the HRRV is encoded by the fisher vector and then properly arranged into the hierarchical model through mixed norm.

Nie et al. [18] 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 (JEDM). For learning the motion patterns they deployed two stream CNN models [19]. However, the accomplishment of Joint Euler Angles for skeleton recovery and estimation introduces an unnecessary complexity.

Li and Sun [20] proposed a CNN fusion model for skeletal HAR model. They represented the 3D skeletal sequence in three image formats and three image sequences through gray value encoding, referred to as Skeletal Trajectory Shape Images (STSIs) and Skeletal Pose Image (SPI) sequences. Further they build a CNN fusion model with three SPIs and three STSIs and the results are fused to get the final result.

Some authors concentered on the inclusion of Graphical Convolutional Networks (GCNs) to describe the action through skeleton joints. Yan et al. [21] proposed to learn the spatiotemporal features of an action through Spatial-Temporal Graph Convolutional Networks (ST-GCN). However, in GCN, the topology of network needs to set manually and it has fixed layers and input samples. Moreover, the nature attributes like bine lengths and directions are not much investigated in the GCN methods. Shi et al. [22] proposed an adaptive two-stream GCN in which the graph topology is learned uniformly and individually in an end-to-end manner. Zhang et al. [23] proposed to use the graph edges that regards to the bones in Human Skeleton. 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 edgeCNN with the help of shared intermediate layers [24].

Si et al. [25] 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. Zhang et al. [26] 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. Liu et al. [27] proposed to combine the GCN with Hidden Conditional Random Field (HCRF) to exploit the skeleton structure at the recognition of human actions. To capture the spatio-temporal information from action sequence, they proposed a multi-stream mechanism that considers the relative coordinates and bine directions as the features.

3. PROPOSED METHOD

3.1 Overview

As depicted in Figure 2, our method composed of three stages: skeleton joint descriptor, classification and fusion. In the first stage, to describe a skeleton action sequence, our method employed totally three Types of descriptors they are informative joint descriptor (IJD), self-similarity joint descriptor (SSJD), and Spherical Joint Descriptor (SJD).

Each descriptor is fed to the new deep learning model for feature extraction followed by classification. After classification of the input action sequence through individual models, they are fused to get the final action label. For Feature extraction and classification we employed a new to 2D-CNN model which is very simple and customized in nature. For fusion process, we employed totally two fusion rules such as max fusion and product fusion. In this section we initially explain the details of Skeleton descriptors and then classification model followed by fusion process.

3.2 Skeleton descriptor

For the recognition of Human actions, initially the action needs to be represented in such a way that the system can discriminate with other actions. Such kind of Representation can be regarded as descriptor and if the input data is the Skeleton joints, then it is called as skeleton joint description or simply skeleton descriptor. In this work, as a skeleton descriptor, we employed three methods based on three different statistics of skeletons to describe an action. The three methods are applied individually over the input Skeleton action sequence. The details of three individual descriptors are demonstrated in the following subsections.

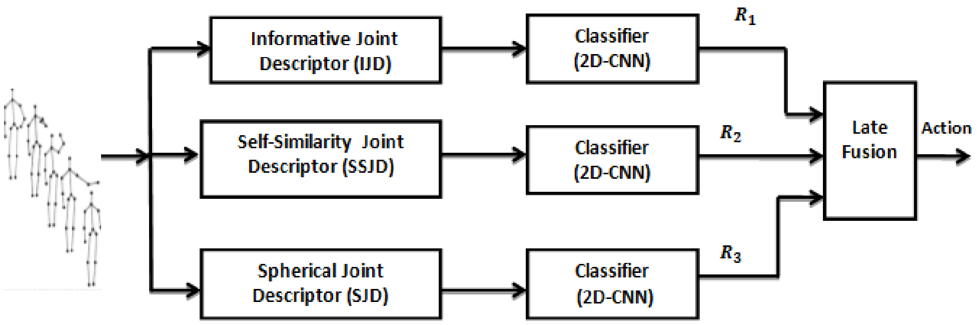

Informative Joint Descriptor (IJD)
Self-Similarity Joint Descriptor (SSJD)
Spherical Joint Descriptor (SJD)
Classifier (2D-CNN)
Classifier (2D-CNN)
Classifier (2D-CNN)
R_1
R_2
R_3
Late Fusion
Action

Figure 2. Block diagram of proposed system

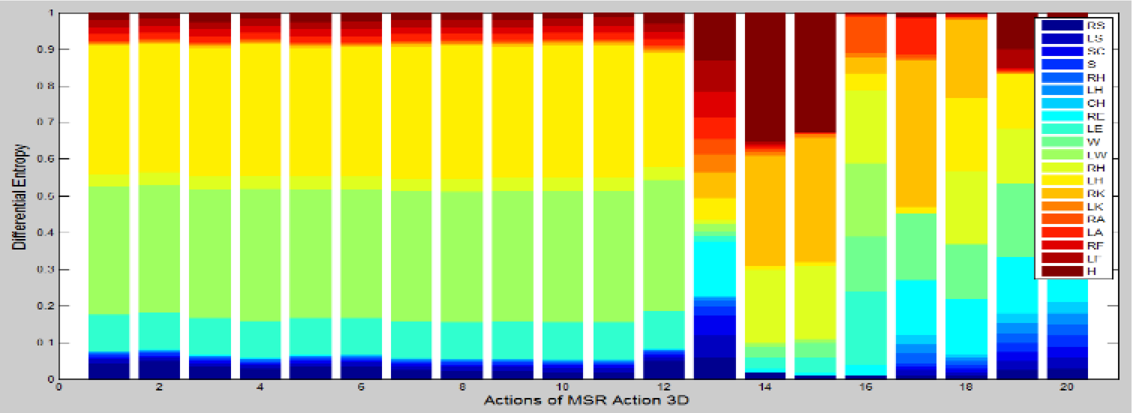


Figure 3. Joint's contribution of each action in MSR-Action 3D

3.2.1 Informative Joint Descriptor (IJD)

Generally, human beings pay more concentration over the moving targets and less concentrates on the static targets. This is the general human Psychology and it is the biggest secret of Magic. Inspired from the nature of human beings, our method tries to represent an action only with informative joints. During the execution of an action, not all joints participate and only few joint takes a major role. So the joints those have less contribution can be regarded as redundant joints. An example of such kind of joints is spine centre and hip centre etc. These joints are totally nonmoving in nature. Further, the redundant joints are also there which are very close for example wrist and hand. For each action, the contribution of joints is different and it is completely different for differ actions. To explore this kind of information, this work conducts an analysis with the help of moving distance variations of every joint during the execution of an action. Differential entropy is the most and common way to evaluate the information of a continuous two-dimensional signal. Differential entropy is an extended version of Shannon entropy hence we consider the concept of Differential entropy to evaluate the contribution of each joint for each action. Then we aggregate the differential entropy of all instances in each action j for each joint i. The following Figure 3 shows the contribution of joints for 20 different actions of MSR action 3D dataset.

The 20 joints are namely BEND (B), TWO HAND WAVE (THW), HANDCLAP (HC), JOGGING (J), SIDEKICK (SK), FORWARD KICK (FK), PICKUP & THROW (PT), GOLF SWING (GS), TENNIS SWING (TS), HIGH ARM WAVE (HW), HORIZONTAL ARM WAVE (HOW), FORWARD PUNCH (FP), HIGH THROW (HT), HAMMER (HA), HAND CATCH (HC), DRAW CROSS (DX), DRAW TICK (DT), DRAW CIRCLE (DC), and SIDE BOXING (SB). In the MSR action 3D dataset, each frame is represented with 20 joints namely "Hip Center (HC)", "Spine (S)", "Shoulder Center (SC)", "Head (H)", "Left Shoulder (LS)", "Left Elbow (LE)", "Left Wrist (LW)", "Left Hand (LH)", "Right Shoulder (RS)", "Right Elbow (RE)", "Right Wrist (RW)", "Right Hand (RH)", "Hip Left (HL)", "Left Knee (LK)", "Left Ankle (LA)", "Left Foot (LF)", "Hip Right (HR)", "Right Knee (RK)", "Right Ankle (RK)" and "Right Foot (RF)".

From Figure 3, we can see that for different actions the contributed joints are different. For High Arm wave action only three joint such as left elbow, left wrist and left hand have major contribution while for the forward kick action, the Skeleton joints of leg has major contribution.

Further, we can also understand that many joints contribute less for the recognition of actions. On the other hand, they introduce some extra noises. Motivated with these phenomena Ofli et al. [4] developed a new method called as Sequence of Most Informative Joints (SMIJ) to represent an action with most informative joints. They employed the calculation of variances of joint angle trajectory to calculate the joint's value and then represented an action as a sequence of most informative joints. But they had shown Limited performance at the actions like Draw X and High Arm Wave which have similar most informative joints. Furthermore, the experiments of Ofli et al. [4] on MSR action 3D had shown very poor performance (approximately 0%) at 8 actions those are performed by single arm. Unlike the most informative joint sequence, we opt to create a most suitable set of joints called as informative joints in which they contribute almost 80% of the entropy for an action. Upon the representation of each action with information joints the remaining joints are kept simply zero. Figure 4 shows the selected informative joints of the MSR action 3D. With this kind of Representation our method has resulted to a feature dimensionality reduction and also obtains good improvement in the recognition of actions.

Consider an action sequence **A** with **N** number of frames and let it represented as $A=\{A_1, A_2, ...A_N\}$ where A_i represents the i^{th} frame. For this action sequence initially our method compute the Euclidean distance between the same joints between the successive frames, thus the total number of distances are *N-1*. Let the distance between two successive frames for a joint k is represented as d_{ij}^{k} where i denotes the index of current frame j denotes the index of its next frame. Since we consider the successive frames for computation, the j is nothing but i+1. Next k denotes the index of joint and it varies as k=1, ...K where K is the total number of joints and it is different for different datasets. The K value for MSR action 3D is 20 while for NTU-RGB+D it is 25. Mathematically the distance $\mathrm{d}_{\mathrm{ij}}^{\mathrm{k}}$ is computed as

$$d_{ij}^{k} = \sqrt{\left(x_j^k - x_i^k\right)^2 + \left(y_j^k - y_i^k\right)^2 + \left(z_j^k - z_i^k\right)^2} \qquad (1)$$

	RS	LS	SC	S	RH	LH	CH	RE	LE	W	LW	RH	LH	RK	LK	RA	LA	RF	LF	H
HW									■		■									
HoW									■		■									
Ha									■		■									
HC									■		■									
FP									■		■									
HT									■		■									
DX									■		■									
DT									■		■									
DC									■		■									
HC								■	■	■	■									
THW								■	■	■	■									
SB								■	■	■	■									
B	■	■	■					■	■	■	■			■	■	■	■			■
FK															■		■			
SK											■				■		■			
J								■		■	■			■	■	■	■			
TS									■	■	■					■	■			■
TS1								■	■	■	■					■	■			■
GS	■	■	■					■	■	■	■									■
PT	■	■	■					■	■	■	■	■		■	■	■	■			■

Figure 4. Informative joints of MSR action 3D

where, $\left(x_i^k, y_i^k, z_i^k\right)$ and $\left(x_j^k, y_j^k, z_j^k\right)$ are the location coordinates of k^{th} joint in the frames at instances *i* and *j* respectively. Based on the obtained distances from successive frames, the differential entropy is calculated as

$$H(X_k) = -\sum_{i=1}^{N-1} p\left(d_{ij}^k\right) \log_b p\left(d_{ij}^k\right) \;\; \forall i, j \in N \qquad (2)$$

where, $H(X_k)$ is the differential entropy of k^{th} joint and $p\left(d_{ij}^k\right)$is the probability of occurrence of distance d_{ij}^k. Since the number of distances is *N-1*, Eq. (2) perform summation of all the probabilities. Eq. (2) is applied on every joint and the joints those have maximum entropy are considered as informative joints. In this manner, each action sequence is represented through only informative joints and the remaining joints are placed as 0 in every frame.

3.2.2 SSJD

Self-similarity Matrix (SSM) has been introduced by Junejo et al. [28] and it has been proved as most stable and robust for cross views. For an action sequence, I. Junejo computed temporal self- similarities between frames at different time instances. For an action sequence $A=\{A_1, A_2, ...A_N\}$ discrete in space (x,y,t), the SSM is computed as a difference of pixels between all pairs of time frames. Such kind of computation ensures view invariance in HAR system. With this inspiration, our method generates skeleton self-similarities by evaluating pairwise differences between all skeleton joints in each time frame. The matrix is called as Skeleton Self-Similarity Matrix (S^{3}M). Unlike the traditional SSM which produces a single Matrix for an entire action sequence, our method produces effective skeleton SSMs for action sequence. Hence, our Method can provide more information about the movements than the temporal self-similarities. Since we evaluate S^{3}M for each time frame, our Method can be regarded as Spatio-temporal similarity information and it is invariant to multiple views. Figure 5 shows the SSJD of an action captured and different views. Even through the RGB and skeleton of same action captured from different views looks different their self-similarities exhibit similar properties.

At this phase, to deal with varying scales and occlusion problems we follow the decomposition rules and reform the action frame into three forms. The decomposition is done in a Coarse to fine manner and at finer scale; we include the entire joints in each frame. This kind of decomposition makes the HAR system robust against scale variations and occlusions. Due to this reason we get totally three SSJDs for each action frame as shown in Figure 6.

Consider an action **A** and we have three skeleton sequences formats with three different numbers of skeleton joints in each frame, such as $A=\{A_l\}$ where A_l represents the skeleton sequence at l^{th} scale. Henceforth, A_l is a set of 3D positions of skeleton joints and it is denoted as $A_l = \{X_l, Y_l, Z_l\} \in \Re^{T\times N_l \times 3}$, where $X_l \in \Re^{N\times J_l}$, $Y_l \in \Re^{N\times J_l}$ and $Z_l \in \Re^{N\times J_l}$. Here, *N* denotes the total number of frames, J_l denotes the number of joints considered at l^{th} scale with *x, y,* and *z* coordinates in 3D space.

Let's consider an action frame at l^{th} scale and n^{th} instance as $A_l(n)=\{X_l(n), Y_l(n), Z_l(n)\}$, then SSJD is represented as $S_l(t)$ and it is calculated as

$$S_l(n) = d_{ij} = \begin{bmatrix} 0 & d_{12} & \dots & d_{1N_l} \\ d_{21} & 0 & .. & d_{2N_l} \\ \dots & \dots & \dots & \dots \\ d_{N_l 1} & d_{N_l 2} & \dots & d_{NN_l} \end{bmatrix} \qquad (3)$$

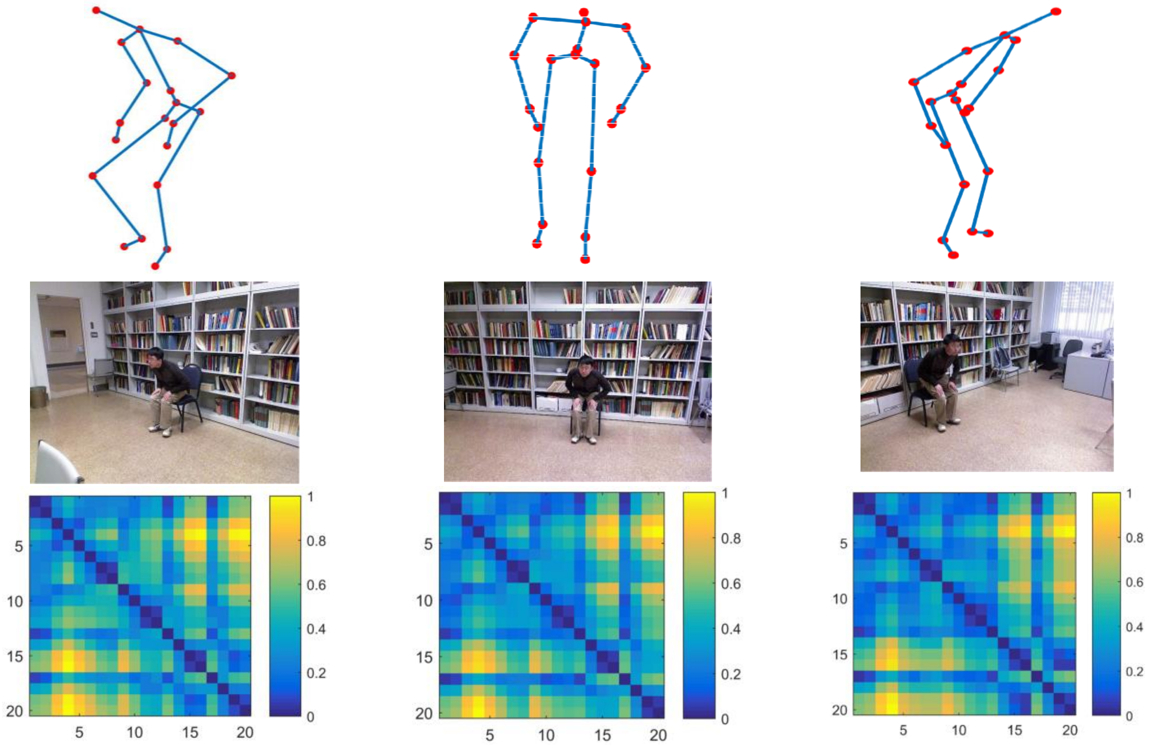


Figure 5. (a) Skeleton of an action under three views (b) RGB images in three poses and (c) SSJD in three poses

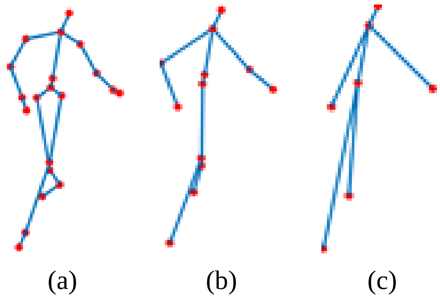


Figure 6. Coarse to fine decomposition of skeleton joints of MSR-action 3D (a) Scale (20 joints), (b) Scale 2 (12 Joints) and (c) Scale 3 (7 Joints)

where, d_{ij} is the distance between two joints i and j in the same frame. In the above Matrix we can see that the diagonal elements are zero because they are generated by the computation of distance between self-joints. Moreover, the Distance is a matrix $S_l(n)$ is a symmetric matrix and of size $J_l \times J_l$. This representation is most effective presentation as it provides in invariance against several transformations like scaling, rotating, translating, and even some affine transformations. At last the entire action sequence is represented with a set of SSJDs which are in equal number with the total number of frames in the action sequence.

3.2.3 Spherical Joint Descriptor (SJD)

Generally, the skeleton joints in the Cartesian co-ordinate system are represented in the form of (x,y,z). However, such kind of representation makes the HAR system sensitive and it can make the system to recognize the two similar actions as different actions. Next, the moment of joints has several restrictions as they can't move beyond certain distance and angle. The joints can't move farther than a limited distance from hip center joint and also restricted by certain angle. Such kind of restriction can be used to describe an action and spherical co-ordinate system is found as the best solution. In Spherical co-ordinate system, each joint is represented with three attributes such as r, θ and ϕ where r is the distance of corresponding joint from hip centre, and θ and ϕ are the directions of movements in two different angles. For the transformation purpose, we have chosen the hip centre as reference point or origin. Consider the joints in Cartesian co-ordinate system as J={0, J_1, ...J_N}, in the spherical coordinate system, it can be represented as [29, 30].

$$J_s = (r,\theta,\phi) \tag{4}$$

where

$$r = \sqrt{(x_{HC} - x_k)^2 + (y_{HC} - y_k)^2 + (z_{HC} - z_k)^2} \tag{5}$$

$$\theta = \arccos\left(\frac{z}{r}\right) \tag{6}$$

$$\phi = \arctan\left(\frac{y}{x}\right) \tag{7}$$

where, (x_{HC}, y_{HC}, z_{HC}) and (x_k, y_k, z_k) are the Cartesian coordinates of Hip Center and k^{th} skeleton joint respectively. The angle θ denotes the vertical angle of joint with z-axis and the angle ϕ denotes the horizontal angle with x-axis. The above mentioned Eqns. (5)-(7) are applied on every joint in every frame. For an action sequence with N number of frames, the obtained SJD can be represented as

$$SJD = \begin{bmatrix} J_{s_1 1} & J_{s_1 2} & \cdots & J_{s_1 19} \\ J_{s_2 1} & 0 & .. & J_{s_2 19} \\ \cdots & \cdots & \cdots & \cdots \\ J_{s_N 1} & J_{s_N 2} & \cdots & J_{s_N 19} \end{bmatrix} \tag{8}$$

where, $J_{s_i j}$ represents the SJD of j^{th} joint in the i^{th} frame.

3.3 CNN model

After the motion representation of an action video through EDMM, the EDMM (Eq. (5)) 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. Since we apply on depth data, the pixel relations are either maximum or minimum. Hence, the max pooling is most adaptable pooling operation for dimensionality reduction. Our proposed CNN model has one fully connected layer of size 1×n where n denotes the total number of actions. Figure 7 shows the architecture of proposed CNN model.

Table 1. CNN model attributes

Layer	Filter count	Size	Filter size	Stride
Conv₁	64	112x112	7x7	2x2
Pool₁	-	-	-	2x2
Conv₂	128	56x56	5x5	1x1
Conv₃	256	56x56	5x5	1x1
Conv₄	512	56x56	5x5	1x1
Pool₂	-	-	-	2x2
Conv₅	1024	28x28	3x3	2x2
Pool₃	-	-	-	2x2

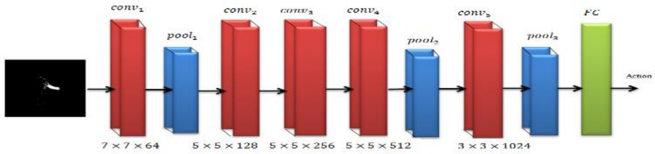


Figure 7. CNN model

As shown in Table 1, the *Conv₁* has applied 64 convolutional filters and the size of each filter is defined as7x7. The *Conv₂*, *Conv₃ and Conv₄* applies 128, 256 and 512 filters respectively and the size of each filter at these layers is 5x5. At the *Conv₅*, the size of each convolutional filter is defined as 3x3 and the total number of convolutional filters applied is 1024. Since the filter size is small, we have applied more number of convolutional filters at this layer. The texture of two different actions through EDMM makes the system challenging to extract distinct features when the size of convolutional filters is small. For instance, the size of 3x3 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 7x7 at the staring convolutional layer. Next, the size of filter at max-pooling layer is fixed as 2x2 and its main intention is to reduce feature map size. In this work, we used a total of 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 112x112 to 56x56. Next, due to the accomplishment of max-pooling layer after fourth convolutional layer, the size of feature map is reduced from 56x56 to 28x28. 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.

3.4 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 three descriptors, the results obtained at the softmax layers are of three 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 three 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 than the remaining two methods. Consider R_1 be the output of softmax layer of phase 1, R_2 be the output of softmax layer of phase 2 and R_3 be the output of softmax layer of phase 3, they are fused as

$$F_1 = Max(R_1, R_2, R_3) \tag{9}$$

$$F_1 = \mathrm{Product}(R_1, R_2, R_3) \tag{10}$$

From the two values such as F_1 and F_2 the final action is obtained as

$$Action = Max(F_1, F_2) \tag{11}$$

where, Action is the name of an action which has highest score and it represents the final action class prediction.

4. SIMULATION EXPERIMENTS

This section describes the details of simulation experiments of the proposed approach. For the experimental validation, we applied our method on two multi-view datasets namely NTURGB+D dataset [31] and Northwestern – UCLA dataset [32]. Since these two datasets are multi-view datasets, we process them for Cross view as well as cross subject validation. Under Cross view, the training and test views are different while under cross subject, the subjects used for training and testing are different.

4.1 Datasets

4.1.1 NTURGB+ D dataset

NTURGB+Dis a large scale dataset consists of totally 56,880 samples of 60 actions. All these samples are collected with the help of 60 subjects. The entire actions are classified into three categories; they are daily actions – 40 (ex. Reading, eating, drinking etc.), health related actions – 9 (ex. Falling down, staggering, sneezing etc.) and mutual actions – 11 (hugging, kicking, punching etc.). Every action is captured under 17 different scene environments and hence there are 17 video sequences labeled as S001 to S017. All the actions are acquired under three viewing points with an angular deviation of −45°,0°, and +45°. Figure 8 shows a samples frames of handshaking action in Four different models. This dataset provides the action sequences in multiple models including Infrared frames, RGB frames, 3D skeleton joints and depth maps. This is quite challenging dataset for recognition as it composed of similar actions, and larger noises in the dataset. At the validation, w employed both cross view validation as well as cross subject validation. At the cross view validation, we used totally 18960 sample videos captured through Camera1 (at +45°) for training and the remaining 37920 sample videos captured through camera 2 and camera 3 are used for testing. Next, under the cross subject validation, the sample videos acquired through 20 subjects (1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38) are used for training and the sample videos of remaining 20 subjects are used for testing. At this phase, we conduct a fivefold cross subject validation by changing the subjects used for training and testing, both cross view validation as well as cross subject validation. At the cross view validation, we used totally 18960 sample videos captured through Camera1 (at +45^0) for training and the remaining 37920 sample videos captured through camera 2 and camera 3 are used for testing. Next, under the cross subject validation, the sample videos acquired through 20 subjects (1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38) are used for training and the sample videos of remaining 20 subjects are used for testing. At this phase, we conduct a fivefold cross subject validation by changing the subjects used for training and testing.

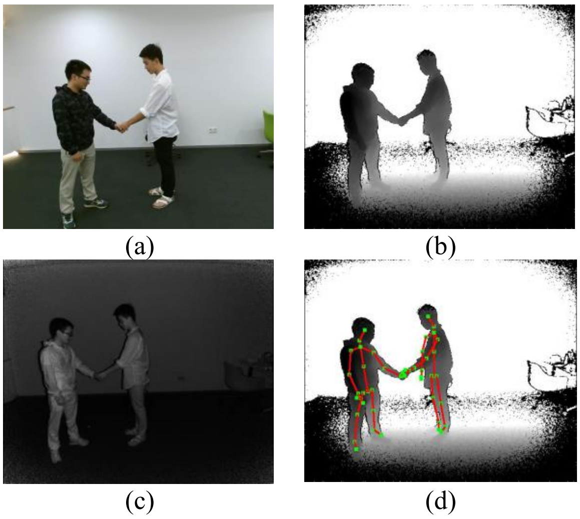


Figure 8. Handshaking action in different models (a) RGB, (b) Depth, (c) Infrared and (d) Depth + skeleton joints

4.1.2 Northwestern – UCLA dataset

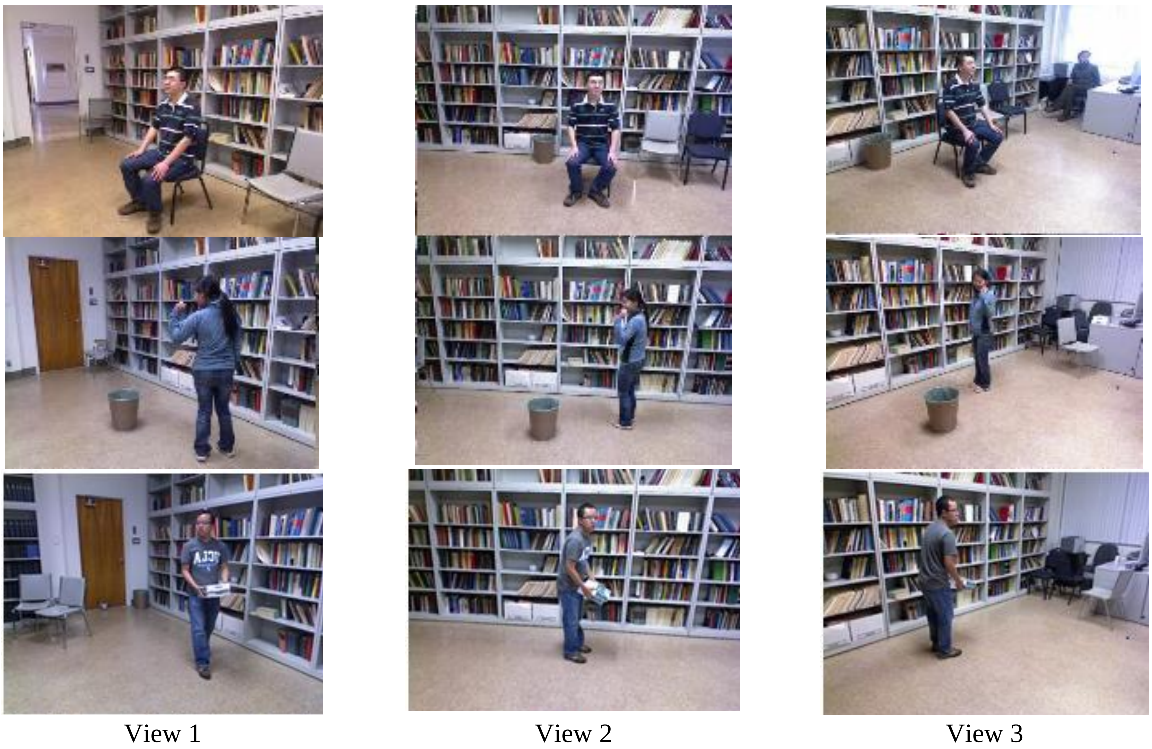
Northwestern – UCLA (NUCLA) is a multi-view action dataset that was captured with the help of three Kinect cameras placed a three different views. This dataset composed of totally1494 action videos and the total number of actions present is 10. Each action is acquired with the help of 10 subjects after repeating it from 1 to 6 times. Since this dataset composed of the actions with similar movements, it is very much challenging in nature. Furthermore, during the acquirement of action videos through multiple cameras, the skeletons have brought several self-occlusion problems. Moreover, the human actions involve objects thereby they have introduced different occlusions in the action videos. At this dataset, we have done cross view validation by considering the action video samples from two cameras for training and one camera for testing. Some samples of this dataset are shown in Figure 9.

4.2 Results

At the simulation of both datasets we have conducted a numerous cross validation with respect to both subject and viewpoints. For NTU RGB+D dataset, totally there are 40 subjects. Hence we conduct a fivefold cross subject validation by interchanging the subjects used for training and testing. For the selection of subjects, a random selection process is accomplished. At every validation, the subjects used for training and testing are shown in Table 2 along with the obtained accuracy.

From the results, the maximum accuracy is attained at 3rd cross subject validation. At this validation, the sample videos of first 20 subjects are used for training and the sample videos of remaining 20 subjects are used for testing. The least accuracy is observed at second validation where the sample videos of odd subjects are used for training and the sample videos of even subjects are used for testing. From these validations, the average accuracy is observed as 84%.

For the assessment of fusion rules effect on the recognition accuracy, the simulation is done with respect to different fusion rules and the obtained accuracy at both cross view and cross subjects are shown in Table 3. At this phase, two types of fusion are considered; they are early fusion and late fusion. In the early fusion, the feature vectors are fused before processing them to the CNN model. The all action descriptors are fused by concatenated them horizontally and each action is represented as a composite feature vector as the composition of IJD, SSJD and SJD. Over the composite vector, we applied the 2D-CNN model for feature extraction followed by classification.



View 1 View 2 View 3

Figure 9. Sit down, throw, and pick up with two hands actions under multiple views

Table 2. Five-fold cross subject validation on NTU RGB+D dataset

CS No.	Training subjects	Testing subjects	Accuracy (%)
CS1	1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38	3, 6, 7, 10, 11, 12, 20, 21, 22, 23, 24, 26, 29, 30, 32, 33, 36, 37, 39, 40	**85.4000**
CS2	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39	2, 4, 6, 8, 10, 12, 14, 16, 18, 20,22 24, 26, 28, 30, 32, 34, 36, 38, 40	**82.2000**
CS3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	21, 22, 23, 24, 25, 26, 27, 28, 29, 30 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	**86.3000**
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	**83.6000**
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	**82.9000**

Table 3. Recognition accuracy (%) of different fusion methods over different datasets

Fusion scheme	Fusion function	NTU RGB+D		N-UCLA
		CV	CS	CV
Early	Concatenation	**88.2345**	**81.4578**	88.3147
Late	Maximum	83.2455	77.6589	86.4785
	Product	82.4444	78.6637	87.8647
	Average	87.9678	79.7845	**89.5674**

From the results, the maximum accuracy is observed at early fusion process for NTU RGB+D while for N-UCLA, the maximum accuracy is gained at late fusion through the average fusion rule. Since the NTU RGB+D is a very big and challenging dataset, the early fusion ensures better results than the late fusion. The confusion matrix of N-UCLA dataset is shown in Table 4 where it has totally ten actions namely 'pick up with 1 hand (P1H)', 'pick up with two hands (P2H)', 'Drop trash (DT)', 'Walk Around (WA)', 'Sit down (SD)', 'Stand Up (SU)', 'Donning (DN)', 'Doffing (DF)', 'Throw (TH)' and 'Carry (CR)'. From the results, the maximum detection rate is observed for Stand Up action while the minimum detection rate is observed for Pick up with one hand action. Further the first two actions such as pick up with one hand and pick up with two hands is observed to have more confusion. Since these two have similar movements and appearance they have observed a higher confusion at recognition. The further two actions those have larger confusion is throw and carry actions.

Table 4. Confusion matrix of CV on the N-UCLA dataset

	P1H	P2H	DT	WA	SD	SU	DN	DF	TH	CR
P1H	0.75	0.20	0.03	0.02	0	0	0	0	0	0
P2H	0.05	0.95	0	0	0	0	0	0	0	0
DT	0	0	0.95	0.05	0	0	0	0	0	0
WA	0	0	0.06	0.88	0	0	0	0.04	0	0.02
SD	0	0.02	0.02	0	0.94	0.02	0	0	0	0
SU	0	0.02	0	0	0.02	0.96	0	0	0	0
DN	0	0	0	0	0	0	0.94	0.06	0	0
DF	0	0	0	0	0	0	0.04	0.92	0.04	0
TH	0	0	0	0	0	0	0	0.06	0.88	0.06
CR	0	0	0.04	0	0	0	0	0	0.12	0.84

In the Table 5, we compare our method with several standard methods those employed handcrafted features and deep learning methods like CNN and RNN for HAR with skeleton data [10, 21, 31]. All these methods used skeleton data as input for recognizing human actions.

Table 5. Comparison of recognition accuracy (%) on NTU RGB+D dataset

Method		NTU RGB+D		N-UCLA
		CV	CS	CV
Deep-LSTM [31]		67.3000	60.7000	-
Res-TCN [24]		83.1000	74.3000	-
Spatio-Temporal LSTM [15]		77.7000	69.2000	-
P-LSTM [31]		70.3000	62.9000	-
SK-CNN [14]		87.2000	80.0000	92.6000
GCA-LSTM [10]		84.0000	76.1000	-
HBPL [17]		82.0000	74.9000	-
Beyond joints [16]		87.6000	79.5000	-
Bio-Constrained [18]		91.8000	**86.9000**	94.4000
ST-GCN [21]		88.3000	81.5000	-
Clips + MTCNN [11]		87.4000	81.1000	93.4000
Proposed	**SSJD**	**91.2300**	**86.3020**	**90.2300**
	SJD	**89.4510**	**84.5350**	**88.2140**
	IJD	**85.6600**	**80.2250**	**86.9330**
	SVI^{2}JD	**93.2000**	**84.0000**	**93.6600**

From the results, we can see that our method has achieved great recognition accuracy at cross views. As our method employed a self-similarity based joint descriptor that can provide view invariance for all the 80 views in NTU RGB+D dataset. The proposed method shown outstanding performance than the RNN based method [31] and also the attention based methods [10, 15]. CNN based approaches [11, 14] employed skeleton visualization strategy for the improvisation of recognition performance in HAR. However, the proposed method can achieve nearer results with CNN methods which needs a sequence of intensive transformations from skeleton to color images. Recently, the GCN which is a general form of CNN have gained a great interest which can represent the skeleton as a set of nodes and edges. They formulated the action with respect to the length of edges and strength of edges. ST-GCN [21] gained a better accuracy on NTU RGB+D dataset. Due to the accomplishment of SJD and SSJD, our method outperformed all the earlier methods in the cross views validation. The SSJD provides great view invariance such that the developed system can recognize the action at any view. Our method has gained good performance for all actions in NTU RGB+D except for two actions; they are reading and writing which are very much difficult to identify.

5. CONCLUSION

In this paper, we propose a composite action descriptor for recognizing human actions from skeleton action sequences. The proposed descriptor is a composition of three descriptors such as IJS, SSJD and SJD. IJD ensures a less complexity by describing an action only with informative joints. SSJD and SJD ensure view invariance and motion restricted action describing such that the multi-view actions and similar actions can also be recognized much accurately. The effectiveness is demonstrated by the simulation of proposed model on two challenging datasets such as NTU RGB+D and N-UCLA. Experimental results demonstrate that our method can handle problem of cross view action recognition through SSJD description with SJD. Particularly, the proposed method exhibited robust performance at large view changes.

REFERENCES

[1] Poppe, R. (2010). A survey on vision-based human action recognition. Image and Vision Computing, 28(6):

976-990. https://doi.org/10.1016/j.imavis.2009.11.014

[2] Aggarwal, J.K., Xia, L. (2014). Human activity recognition from 3d data: A review. Pattern Recognition Letters, 48: 70-80. https://doi.org/10.1016/j.patrec.2014.04.011

[3] Aggarwal, J.K., Ryoo, M.S. (2011). Human activity analysis: A review. Acm Computing Surveys (Csur), 43(3): 1-43. https://doi.org/10.1145/1922649.1922653

[4] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R. (2014). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. Journal of Visual Communication and Image Representation, 25(1): 24-38. https://doi.org/10.1109/CVPRW.2012.6239231

[5] Evangelidis, G., Singh, G., Horaud, R. (2014). Skeletal quads: Human action recognition using joint quadruples. In 2014 22nd International Conference on Pattern Recognition, pp. 4513-4518. https://doi.org/10.1109/CVPR.2016.213

[6] Feichtenhofer, C., Pinz, A., Zisserman, A. (2016). Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1933-1941. https://doi.org/10.1109/CVPR.2016.213

[7] Wang, Z., Liu, S., Zhang, J., Chen, S., Guan, Q. (2016). A spatio-temporal CRF for human interaction understanding. IEEE Transactions on Circuits and Systems for Video Technology, 27(8): 1647-1660. https://doi.org/10.1109/TCSVT.2016.2539699

[8] Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. Perception & Psychophysics, 14(2): 201-211. https://doi.org/10.3758/BF03212378

[9] Rahmani, H., Mahmood, A., Huynh, D., Mian, A. (2016). Histogram of oriented principal components for cross-view action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(12): 2430-2443. https://doi.org/10.1109/TPAMI.2016.2533389

[10] Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C. (2017). Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Transactions on Image Processing, 27(4): 1586-1599. https://doi.org/10.1109/TPAMI.2016.2533389

[11] Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F. (2018). Learning clip representations for skeleton-based 3d action recognition. IEEE Transactions on Image Processing, 27(6): 2842-2855. https://doi.org/10.1109/TIP.2018.2812099

[12] Rahmani, H., Mian, A., Shah, M. (2017). Learning a deep model for human action recognition from novel viewpoints. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(3): 667-681. https://doi.org/10.1103/PhysRevD.94.065007

[13] Zhang, J., Shum, H.P., Han, J., Shao, L. (2018). Action recognition from arbitrary views using transferable dictionary learning. IEEE Transactions on Image Processing, 27(10): 4709-4723. https://doi.org/10.1109/TIP.2018.2836323

[14] Liu, M., Liu, H., Chen, C. (2017). Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition, 68: 346-362. https://doi.org/10.1016/j.patcog.2017.02.030

[15] Liu, J., Shahroudy, A., Xu, D., Kot, A.C., Wang, G. (2017). Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12): 3007-3021. https://doi.org/10.48550/arXiv.1706.08276

[16] Wang, H., Wang, L. (2018). Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection. IEEE Transactions on Image Processing, 27(9): 4382-4394. https://doi.org/10.1109/TIP.2018.2837386

[17] Shao, Z., Li, Y., Guo, Y., Zhou, X., Chen, S. (2018). A hierarchical model for human action recognition from body-parts. IEEE Transactions on Circuits and Systems for Video Technology, 29(10): 2986-3000. https://doi.org/10.1109/TCSVT.2018.2871660

[18] Nie, Q., Wang, J., Wang, X., Liu, Y. (2019). View-invariant human actionrecognition based on a 3D bio-constrained skeleton model. IEEE Transactions on Image Processing, 28(8): 3959-3972. https://doi.org/10.1109/TIP.2019.2907048

[19] Li, C., Xie, C., Zhang, B., Han, J., Zhen, X., Chen, J. (2021). Memory attention networks for skeleton-based action recognition. IEEE Transactions on Neural Networks and Learning Systems, 33(9): 4800-4814. https://doi.org/10.1109/TNNLS.2021.3061115

[20] Li, M., Sun, Q. (2021). 3D skeletal human action recognition using a CNN fusion model. Mathematical Problems in Engineering, 2021: 1-21. https://doi.org/10.1155/2021/6650632

[21] Yan, S., Xiong, Y., Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Thirty-Second AAAI Conference on Artificial Intelligence, pp. 7444–7452. https://doi.org/10.1609/aaai.v32i1.12328

[22] Shi, L., Zhang, Y., Cheng, J., Lu, H. (2019). Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12026-12035. https://doi.org/10.48550/arXiv.1805.07694

[23] Zhang, X., Xu, C., Tian, X., Tao, D. (2019). Graph edge convolutional neural networks for skeleton-based action recognition. IEEE Transactions on Neural Networks and Learning Systems, 31(8): 3047-3060. https://doi.org/10.48550/arXiv.1805.06184

[24] Soo Kim, T., Reiter, A. (2017). Interpretable 3d human action analysis with temporal convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1-9. https://doi.org/10.1109/CVPRW.2017.207

[25] Si, C., Chen, W., Wang, W., Wang, L., Tan, T. (2019). An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1227-1236. https://doi.org/10.48550/arXiv.1902.09130

[26] Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N. (2020). Semantics-guided neural networks for efficient skeleton-based human action recognition. In proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1-10. https://doi.org/10.1109/CVPR42600.2020.00119

[27] Liu, K., Gao, L., Khan, N.M., Qi, L., Guan, L. (2020). A multi-stream graph convolutional networks-hidden conditional random field model for skeleton-based action

recognition. IEEE Transactions on Multimedia, 23: 64-76. https://doi.org/10.1109/TMM.2020.2974323

[28] Junejo, I.N., Dexter, E., Laptev, I., Perez, P. (2010). View-independent action recognition from temporal self-similarities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(1): 172-185. https://doi.org/10.1109/TPAMI.2010.68

[29] Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D.D. (2018). Deep convolutional neural networks for human action recognition using depth maps and postures. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49(9): 1806-1819. https://doi.org/10.1109/TSMC.2018.2850149

[30] Rani, S.S., Naidu, G.A., Shree, V.U. (2021). Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition. Materials Today: Proceedings, 37: 3164-3173. https://doi.org/10.1016/j.matpr.2020.09.052

[31] Shahroudy, A., Liu, J., Ng, T. T., and Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and pattern Recognition, pp. 1010-1019. https://doi.org/10.48550/arXiv.1604.02808

[32] Wang, J., Nie, X., Xia, Y., Wu, Y., Zhu, S.C. (2014). Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2649-2656. https://doi.org/10.48550/arXiv.1405.2941