Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7923019

# SPATIO-TEMPORAL JOINT DESCRIPTOR FOR ACTION RECOGNITION FROM SKELETON SEQUENCES

**Venkata Subbareddy K[1*], L. Nirmala Devi[2]**
[1]Research Scholar, ECE Department, Osmania University, Hyderabad, India.
[2]Professor, ECE Department, Osmania University, Hyderabad, India.
**Email:** subvish@gmail.com

**Abstract:** Human Action Recognition (HAR) gained huge research prospect due to its widespread applicability in different applications related to Human computer interaction, Human Robot Interactions and Visual surveillance etc. However, the traditional RGB videos assisted HAR has poor performance as they composed of different illuminations, viewpoints, clothing etc. The emergence of 3D skeleton sequences sort out these problems and shown a new direction for HAR. However, the skeleton sequences are sensitive to noises, and similar movements. Hence, this paper proposes a new Action Descriptor called as Spatio-Temporal Joint Descriptor (STJD) for action recognition from skeleton sequences. STJD encodes the both spatial and temporal movements of an action and ensures improved recognition accuracy especially for action with similar movements. Initially, STJD segments each frame of skeleton sequence into different local segments and each segment is encoded with Spatial Skeleton Joint Descriptor (SSJD). Further, the SSJDs of each frame are encoded with Temporal Skeleton Joint Descriptor (TSJD) and fed to a pre-trained deep learning model ImageNet for classification.

**Index Terms-** Human Action Recognition, Skeleton, Spatio-temporal information, Deep learning, F-Score.

## I. Introduction

In the recent years, HAR has gained a huge research interest in the computer vision field. Since the HAR is regarded as a fundamental task in many applications, so many researchers concentrated over it and suggested several methods. The major applicability of HAR includes visual surveillance [1], video streaming [2], health care [3], gaming entertainment [4], and detection of the complex object's movements [5]. Moreover, due to the continuous growing of robots and some automated services in the people's life; the HAR attained an ever growing importance. Enabling the robots and services to be aware of human action is very important because they can protect the people from so many danger situations, for instance from road accidents. In general, the HAR is accomplished with the help of RGB videos [6-10]. However, recognizing human actions from RGB videos is a challenging task as there exists several factors includes different illumination conditions, viewpoint variations, background noise, texture, clothing shape, colour, size and distance variations in the object and so on. Moreover, the humans also have different ways to perform an action. The traditional RGB videos show poor recognition performance if the action video has any clutters.

To overcome these shortcomings with RGB videos, recently, the HAR based on skeleton sequences has drawn an increasing research attention [11]. Compared with the traditional RGB videos, the skeleton data is more beneficial and ensure a better data analysis. The skeleton represents an action data in more compact way because it makes to disappear the background noises and illumination variations. Next, the skeleton is represented inherently with 3D joint positions those require very less memory. Alongside, with the invention of powerful depth sensors such as Microsoft Kinect sensor, accessing the skeleton joints position is very much easy and flexible. With the motivation of capturing the skeleton data through Kinect sensors in real time, so many works [12-14] are developed for action analysis through skeleton information. However, most of the earlier developed methods didn't utilize the hierarchical structure of human skeleton for action recognition. Hence, they had shown a limited recognition performance in recognizing actions, especially the actions with minor changes in their movements.

To solve these constraints in HAR, in this paper, we propose a new Action Descriptor which intends to suppress the ambiguity at the actions with similar movements. The major contributions of this paper are outlined follows;

1. First, a Spatio-Temporal Joint Descriptor (STJD) is proposed to ensure the robustness at the recognition of actions with similar movements. STJD is employed as a two-stage descriptor; in the first sage, the Spatial SJD (SSJD) is adopted and in the next stage Temporal SJD (TSJD) is adopted. SSJD measures the spatial changes within the frame through local segments and ensures a clear and reliable representation of each action. Next, TSJD measures the temporal changes between frames and provides much discriminative information for recognition system.
2. Secondly, to ensure simplicity, a new segmentation mechanism is developed in which the entire skeleton joints are segmented into several segments based on their location in human biological structure.
3. Lastly, this work adapts for a pre-trained CNN model, i.e., ImageNet for the purpose of classification.

The remaining structure of paper is structured as; section II briefly summarizes details of related work on HAR. Section III discusses the details of Segmentation and STJD. Section IV presents the simulation experiments details. Section V concludes the paper.

## II. Related Work

In earlier, several methods are developed to represent the action with an efficient set of features such that the HAR attains better recognition accuracy. Some researchers developed handcrafted methods while for researchers developed deep learning models to model the action through deep features.

Vemulapalli et al. [15] modeled an action descriptor from skeleton joints based on their 3D geometric relationships through different rotations and translations in 3D space. This kind of representation represents the actions as curves in the lie group. However, the classification is tough for curve classification. Hence, the curve is transformed into algebra and then fed to a combinational classifier for classification purpose. They employed totally three classifiers namely "Dynamic Time Warping (DTW)", "Fourier Temporal Pyramid (FTP)", and "Support Vector Machine (SVM)" algorithms.

G. Evangelidis et al. [16] proposed a compact 6D view invariant skeleton descriptor called as skeletal Squad which encodes the relative position of joint quadruples. Further, they used a Fisher Kernel to describe the Skeletal Quads in a sub-action. Gaussian Mixture Model is employed to train the data followed by classification. Y. Hsu et al. [17] applied Self Similarity Matrix (SSM) [18] which computes the similarities between frames through Euclidean distance and the obtained distances are formulated into a 2D symmetric matrix called as Spatio-Temporal matrix (STM). For the recognition of action, they employed SVM after describing the action with STM using the pyramid-structural bag-of-words (BoW-pyramid). With the availability of exact positions of skeleton joints, the skeletons can be made strictly view invariant after the transformation.

X. Diao et al. [19] suggested a new RNN model named as "Multi-Term Attention Networks (MTANs)" which can extract the temporal features at different scales. This network consists of MTA-RNN and ST-CNN. J. Liu et al. [20] proposed a new version of 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. A similar method is proposed by J. Liu et al. [21] by introducing a Gating Mechanism with LSTM to deal with noisy skeleton. However, the noise due to similar movements makes the recognition system more confused and results in larger number of false positives for large scale datasets.

**Problem Formulation:** Handcrafted methods uses basic statistical and computer vision methods to describe an action through its motion attributes. Most of the methods employed global methods which cannot reveal the inherent motion features. The global action descriptor hides the local variations (motion at individual part level) and hence, it can't provide sufficient discrimination between different actions with similar movements. For example draw cross and draw tick has only differ at the finger tips movements and the remaining body is same. In such situations, the HAR system with global descriptors shows larger false positives.

## III. Proposed Method

### 3.1. Method Overview

This section explores the full particulars of developed HAR model based on the skeleton joints information. Broadly speaking, the entire methodology of proposed approach is carried out in two phases; they are (1) Action representation through Spatio-Temporal Joint Descriptor (STSJD) and (2) ImageNet assisted action classification. Initially the action sequence is represented through the SJSD and then fed to ImageNet [23] for classification. Figure.1 depicts the overall schematic of developed HAR model.

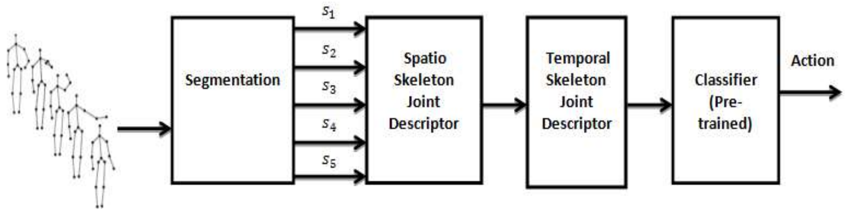


Figure.1 Block schematic of proposed HAR model

### 3.2. STJD

The main intention behind the development of STJD is to determine the Spatio-temporal changes in an action. The spatial changes are called as local changes while the temporal changes are called as global changes. For spatial changes evaluation, the local parts of the body are considered which lies within the frames. For temporal changes determination, entire frames are considered sequentially. Initially, every frame is subjected to a segmentation based on human bod anatomy. Under the segmentation process, each frame is segmented into several local segments based on the sides of human body. Then they are subjected to spatial changes determination and the descriptor obtained at this phase is termed as Spatial SJD (SSJD). Once the SSJD computation is done for every frame, then they are processed as an input for next phase descriptor, i.e., Temporal SJD (TSJD). So, in this section, initially we explore the details of segmentation process and then the proposed STJD.

### A. Segmentation

In general, the human body consists of 20 major joints which can be segmented into five segments based on their location side. For example, the joints located on the left side of human body can be segmented into one segment. Based on this strategy, the entire joints are segmented into five Segments. They are namely, Axial, Left-Top, Right-Top, Left-Bottom and Right –Bottom. For example, if we consider MSR Action 3D dataset, each skeleton frame has 20 joints. They are namely “HIP CENTER (HC), SPINE (S), SHOULDER CENTER (SC), HEAD (H), RIGHT SHOULDER (RS), LEFT SHOULDER (LS), LEFT ELBOW(LE), RIGHT ELBOW (RE), LEFT WRIST (LW), RIGHT WRIST (RW), LEFT HAND (LH), RIGHT HAND(RH),LEFT KNEE (LK), RIGHT KNEE (RK), HIP LEFT (HL), HIP RIGHT (HR), LEFT ANKLE (LA), RIGHT ANKLE (RA), LEFT FOOT (LF) AND RIGHT FOOT (RF)”. After segmentation, all these 25 joints are represented as $\mathrm{Segment\ 1(Axial)} \rightarrow \{\mathrm{HC, S, SC, H}\}$, $\mathrm{Segment\ 2(Left - Top)} \rightarrow \{\mathrm{LS, LE, LW, LH}\}$, $\mathrm{Segment\ 3(Right - Top)} \rightarrow \{\mathrm{RS, RE, RW, RH}\}$, $\mathrm{Segment\ 4(Left - Bottom)} \rightarrow \{\mathrm{HL, LK, LA, LF}\}$ and $\mathrm{Segment\ 5(Right - Bottom)} \rightarrow \{\mathrm{HR, RK, RA, RF}\}$.

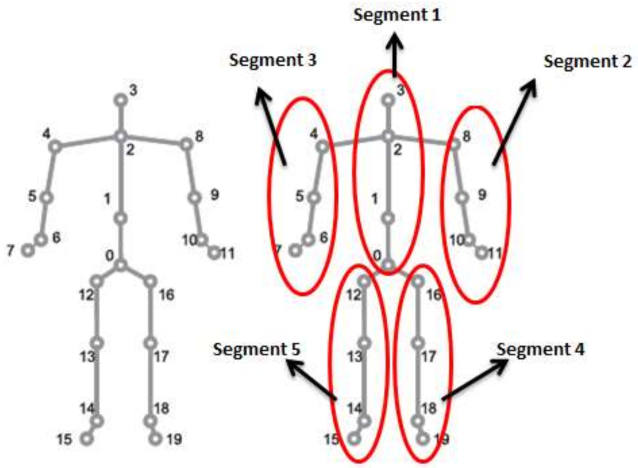


Figure.2 Segmented skeleton frame

Figure.2 shows the segmented skeleton frame of an action in MSR action 3D dataset [24]. Since it composed of 20 joints, the size of all segments is equal. This kind of segmentation is called as uniform segmentation which is possible only in the presence of uniform joints.

**B. SSJD**

To measure the spatial changes in the frame, the local segments are considered one-by-one. The spatial changes are measured with respect to one origin point which was defined for every segment. Initially, for every segment, one origin point is defined which has much restrictions on its movement and less deviation in their position.. For segment 1, the base of spine (SB) is considered as origin point, while for segment 2, 3, 4 and 5, the origin points are LS, RS, HL and HR respectively. All the above mentioned origin points have very less deviations in their positions even for complex actions. Hence they are chosen as origin points. In SSJD, the spatial changes are measured as a distance between origin point and remaining joints in each segment. Let a skeleton action sequence composed of N frames and each frame consists of 25 joints, each joint position on three axes is represented as $J_n^p = \left(x_n^p, y_n^p, z_n^p\right)^T$ where $p \in P$ denotes the joint and $n \in N$ denotes the frame number. For example consider the Hip left joint in $n^{th}$ frame, it can be represented as $J_n^{HL} = (x_n^{HL}, y_n^{HL}, z_n^{HL})^T$. Based on this kind of representation, the SSJD is measured along three axes, as

$$D_i^n(x,y,z) = \left(\frac{1}{l(s_i)-1}\sum_{k=1}^{K} J_{ik}^n(x,y,z)\right) - (J_{i1}^n(x,y,z)) \qquad (1)$$

Where $D_i^n(x,y,z)$ are the SSJD's, $J_{i1}^n(x,y,z)$ are the positions of origin points, and $J_{ik}^n(x,y,z)$ are the positions of remaining joints along three axes x-, y-, and z- axis respectively. Further, $n$ signifies the frame number and $i$ signifies the segment number. In Eq.(1), $k$ varies from 2 to K (total joints present in each segment). After excluding the first joint (origin point), in every segment, there exists only length of segment minus one joints. Hence the summation is divided by $l(s_i) - 1$ in which $l(s_i)$ denotes the length of $i^{th}$ segment. The $D_i^n(x,y,z)$ is a vector of size $N \times S \times 3$ where N signifies frame count, S signifies segments count and 3 represents three distances along three axes. The main advantage with SSJD is its ability to ensure a perfect discrimination between different actions with similar movements. For example, consider three actions namely Draw circle, Draw tick ad Draw cross. In these three actions, the movements are differed only at finger tips. In such kind of actions, the global motion descriptors have limited discrimination capability. Hence the local motion analysis is required which can capture the local changes and provide sufficient discrimination capability to the recognition system. Moreover, as the number of joints increases in the Skeleton frame, the recognition accuracy will be more if SSJD is used as an action descriptor.

### C. TSJD

TSJD mainly intended towards the provision of sufficient discrimination between different actions with larger deviations in their movements. This can be called as a generalized SJD or global SJD. Since the entire frames are subjected to describe the movements of an action it is called as global SJD. The evaluation of TSJD is done between the reference frame and remaining frames in the action sequence. Here, we consider the first frame as a reference frame and it was discriminated from remaining frames to derive the TSJD. TSJD is measured with the help of SSJDs. For an input action sequence having *N* frames, there are totally $N \times S$ SSJDs and they are employed as an inputs for TSJD. The TSJD is measured segment wise between reference frame and current frames. Consider $D_i^n(x,y,z)$ be the SSJD of $i^{th}$ segment in *n*th frame, then the TSJD is obtained by the subtraction of $D_i^n(x,y,z)$ from $D_i^1(x,y,z)$. Here $D_i^1(x,y,z)$ is termed as the SSJD of $i^{th}$ segment in the reference frame. Let $T_i^{n-1}(x,y,z)$ be the TSJD of $i^{th}$ segment in *n*th frame, it is calculated as

$$T_i^{n-1}(x,y,z) = \left\| D_i^n(x,y,z) - D_i^1(x,y,z) \right\| \qquad (2)$$

Where $T_i^{n-1}(x,y,z)$ are the TSJDs along three axes such as x-, y-, and z-axis respectively. In the above three expressions, the N value is varied from 2 to N and at every *n*th instant, the *n*th frame is considered as current frame. After the completion of TSJDs of every frame, they are concatenated and represented as $T_i(x,y,z) = \{T_i^1(x,y,z), T_i^2(x,y,z), \ldots, T_i^{n-1}(x,y,z)\}$. Based on these TSJDs, the action is described globally through three statistical measures; they are namely mean, range and variance. The mathematical representation of these three metrics is shown below;

$$\tau\big(T_i(x,y,z)\big) = \max\big(T_i(x,y,z)\big) - \min\,(T_i(x,y,z)) \qquad (3)$$

Where $\tau\big(T_i(x,y,z)\big)$ is the range of the movements of an action along 3 axes such as x-, y- and z- axis respectively.

$$\mu\big(T_i(x,y,z)\big) = \frac{1}{length(T_i(x))} \sum_{n=1}^{N} T_i^n(x,y,z) \qquad (4)$$

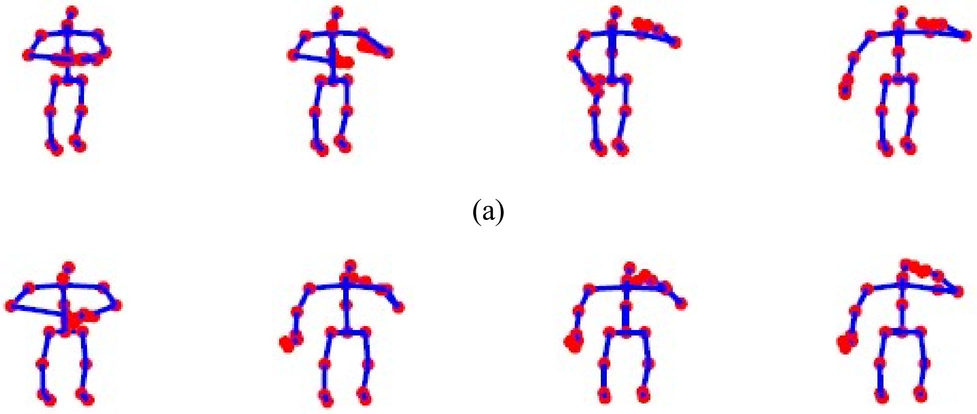
Where $\mu\big(T_i(x,y,z)\big)$ is the mean movements of an action along 3 axes such as x-, y- and z- axis respectively.

$$\sigma^2\big(T_i(x,y,z)\big) = \frac{\sum_{n=1}^{N}\Big(T_i^n(x,y,z)-\mu\big(T_i(x,y,z)\big)\Big)^2}{N-1} \qquad (5)$$

Where $\sigma^2\big(T_i(x,y,z)\big)$ is the movement's variance of an action along three axes x-axis, y-axis and z- axis respectively. To lessen the burden at training phase, we referred these three statistical measures. Among these three measures, the first measure, i.e., range signifies the minimum and maximum possible distances up to which the skeleton joint can move. Sine each action has its restrictions over the movement, this metric is more helpful in discriminating the actions. The next two measures explore the mean and variance distance that was incurred due to the movements of joints in an action. These can help in the determination of overall distance changes due to the joints movements. For an input action sequence having N frames, the size of TSJD descriptor is $(N-1)\times S\times 3$. Along with the TSJD features, the three statistical measures are concatenated at the end of each row.

## IV. Experiments and Discussion

The developed action recognition model is simulated on NTURGB+D dataset [22]. It composed of noisy skeleton data hence we had chosen them for experimental validation. We conduct a vast set of experiments with different view and different subjects. This dataset totally consist of 60 actions acquired through 40 subjects under different views. The total number of skeleton sequences present is 56,880. The video samples are captured with the help of Microsoft Kinect V2 cameras. The 3D skeletal data is represented by 3D locations of 25 major joints of body. The action sequences are approximately having 80 different views. They used totally three cameras to acquire three different horizontal views for the same action. For every camera setup, the cameras are located at the same height at different angles $-45^0$, $0^0$, and $+45^0$. Each person is asked to perform an action twice, once for right camera and another for left camera. In this manner, they captured one right side $45^0$ view, one left side $45^0$ view, one right side view, one left side view and two frontal views.  Figure.3 shows some action skeleton samples.



(a)

(b)

Figure.3 sample skeleton frames of different actions of NTURGB+D dataset (a) Brushing Teeth and (b) Drinking Water

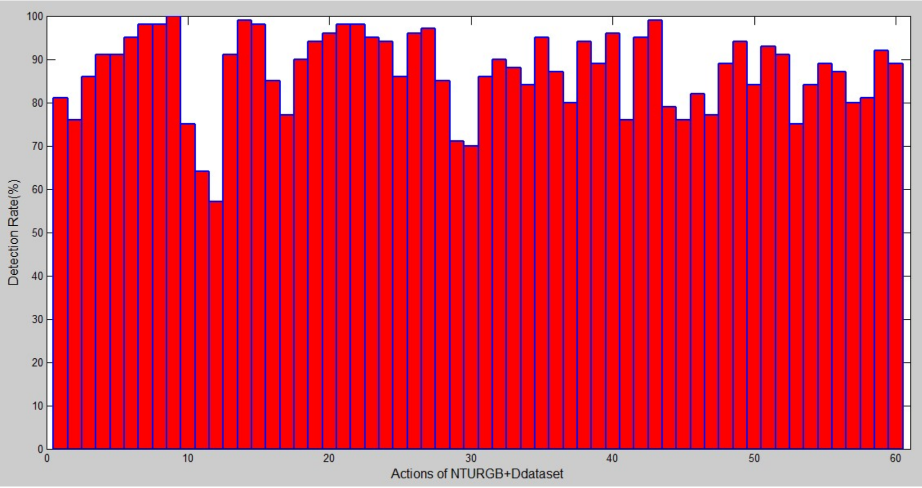


Figure.4 Detection rates of several actions in the NTURGB+D dataset

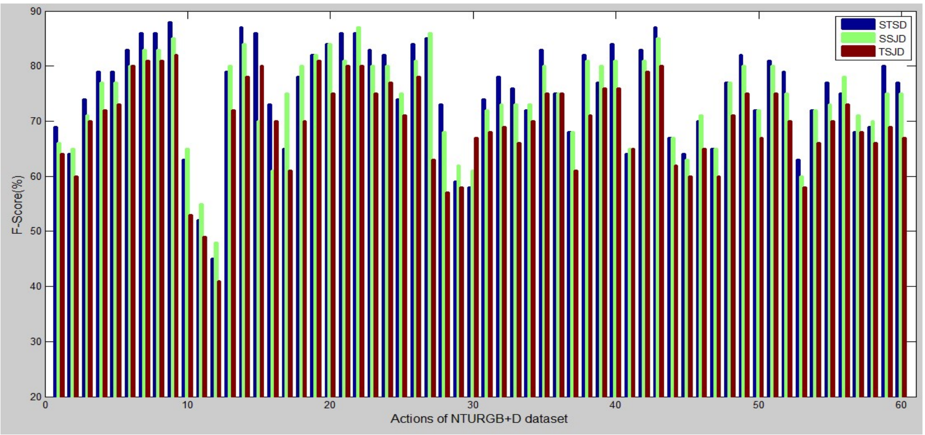


Figure.5 F-score of different actions of NTURGB+D dataset

Table.1 Accuracy (%) Comparison on NTURGB+D Dataset

| **Method** | **Cross View** | **Cross Subject** |
|---|---|---|
| LARP [15] | 52.7600 | 50.0800 |
| Skeletal Squads [16] | 41.3600 | 38.6200 |
| Deep-LSTM [22] | 67.3000 | 60.7000 |
| P-LSTM [22] | 70.3000 | 62.9000 |
| GCA-LSTM [20] | **84.0000** | 76.1000 |
| ST-LSTM + Trust Gate [21] | 77.7000 | 69.2000 |

| | | |
|---|---|---|
| TS-LSTM + MTAN [19] | 80.7000 | 74.6800 |
| STSD + ImageNet | 81.2200 | **78.9670** |

Figure.4 depicts the Detection rate of different actions of NTURGB+D dataset. Among the all tested actions, the maximum DR is observed at an action ‘standing up (from sitting position)’ and it is approximately 98.7854%. Next the minimum DR is observed at an action ‘put something in pocket/take something from packet’ and it is approximately 57.3361%. The overall DR is observed as 87.2167% which is shows that the proposed method can recognize individual actions perfectly. Even though the NTURGB+D dataset consists of a huge number of actions and most of the actions have similar movements, the proposed STSJD can provide describe them in such a way, the discrimination is perfect. Next, the results shown in Figure.5 demonstrate the comparison of F-scores of proposed descriptors. This comparison is done to analyze the individual impact on the actions recognition. As it can be observed from the results, the maximum F-score is observed for the combined descriptor and minimum F-score is observed for TSJD. The average F-score of combined descriptor is observed as 88.1241% while for SSJD and TSJD, it is observed as 85.1324% and 80.2212% respectively. Since the combined action descriptor have more information regarding the movements of an action, it can helps to the recognition system in the accurate recognition. Next, Table.1 shows the comparison of accuracies of different methods. Among the methods employed only skeleton data for action recognition; the GCA-LSTM [20] has a noticeable accuracy. However, it is not focused on the skeleton noise removal which rises due to the similarity of movements in actions. Hence, it has gained only an accuracy of 76.1000% at cross subjects. Due to the SSJD, the proposed has gained a remarkable accuracy of approximately 78.9670% at cross subject simulation.  Next, the one more method TS-LSTM + MTAN [19] also gained a more accuracy than the proposed method at cross views. However, it had shown poor performance at cross subjects.

## V. Conclusion

The main objective of this article is the recognition of human actions through skeleton sequences at an action with complex movements. For this purpose, a new STSD is proposed which encodes the both spatial and temporal motion attributes of an action and lessens the ambiguity. The STSJD preserves the Spatio-temporal information thereby the action with similar length and movements are also recognized effectively. Extensive simulations conducted on NTURGB+D action datasets prove the robustness of developed system against similar movements and noisy skeleton. Our approach gained approximately 3% improvement in the recognition accuracy on the NTURGB+D dataset. This result explores the effectiveness of proposed approach compared to the existing state-of-the-art approaches.

From the results, it is observed that the proposed method had not shown much significance at the detection of actions under multiple views. Hence, this work can be extended towards the development of a new view invariant action descriptor.

## 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, Oct. 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," S*ensors*, vol. 14, no. 7, pp. 11735-11759, Jul. 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. of 6th Int. Symp. Sustain. Healthy Buildings*, Seoul, South Korea, vol. 27, 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, Jun. 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. of 11th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Seoul, South Korea, Aug. 2014, pp. 74-80.

[6] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1-16:43, Apr. 2011.

[7] Ronald Poppe, "A survey on vision-based human action recognition", *Image and Vision Computing*, vol.28, no.6, pp.976-990, June 2010.

[8] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen, "A Comprehensive Survey of Vision-Based Human Action Recognition Methods", *Sensors,* vol.19, no.1005, pp.1-20, 2019.

[9] Yu, K.; Yun, F., "Human Action Recognition and Prediction: A Survey," arXiv 2018, arXiv:1806.11230.

[10] Dawn, D.D.; Shaikh, S.H., "A comprehensive survey of human action recognition with Spatio-temporal interest point (STIP) detector", *Vis. Comput.*, vol.32, pp.289–306, 2016.

[11] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 2, pp. 224-241, Feb. 2011.

[12] X. Yang, Y. Tian, "Eigen joints-based action recognition using Naive–Bayes-nearest-neighbor", in *Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, 2012, pp. 14–19.

[13] Liu, A.A., Nie, W. Z., Su, Y. T., Ma, L., Hao, T., Yang, Z. X., "Coupled hidden conditional random fields for RGB-D human action recognition," *Signal Process.*, vol.112, pp.74–82, 2015.

[14] M. Ding, and G. Fan, "Multilayer joint gait-pose manifolds for human gait motion modeling," *IEEE Trans. Cybern.,* vol.45, no.11, pp.2413–2424, 2015.

[15] R. Vemulapalli, F. Arrate , R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in: P*roc. of CVPR*, 2014, pp. 588–595.

[16] G. Evangelidis, G. Singh, R. Horaud, "Skeletal quads: human action recognition using joint quadruples," in *Proc. of the Int. Conf. on Pattern Recognition*, Stockholm, Sweden 2014, pp. 4513–4518.

[17] Y. Hsu, C. Liu , T. Chen, and L. Fu, "Online view-invariant human action recognition using RGB-D Spatio-temporal matrix," *Pattern Recognit*., vol.60, pp.215–226, 2016.

[18] I.N. Junejo, E. Dexter, I. Laptev, P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell*., vol.33, no.1, pp.172–185, 2011.

[19] Xiaolei Diao, Xiaoqiang Li and Chen Huang, "Multi-Term Attention Networks for Skeleton-Based Action Recognition", *Appl. Sci.*, vol.10, no.5326, pp.1-19, 2020.

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

[21] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton based action recognition using Spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

[22] A. Shahroudy, J. Liu, T.T. Ng, G. Wang, "NTU RGB+D: a large scale dataset for 3D human activity analysis", in *Proc. CVPR*, Las Vegas, NV, USA, pp. 1010–1019, 2016.

[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of ACM,* Vol.6, no.6, pp.84-90, 2017.

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