A Deep Learning Framework for Skeleton-Based Human Recognition Using RGB-D Data

Vinoda Gopampallikar[1], Shashi Kant Gupta [2]

[1] Department of CSE (AI&ML) **CMR Technical Campus, Hyderabad, Telangana** 501401

Student, Department of Computer Science & Engineering **Lincoln University College, Malaysia**

email id : pdf.vinodaresearch@lincoln.edu.my, vinodaresearch@gmail.com, **vinodareddy.cse@cmrtc.ac.in**

[2] Department of Computer Science & Engineering, **Lincoln University College, Malaysia**

email id : **raj2008enator@gmail.com**, **shashigupta@lincoln.edu.my**

Abstract **-** Human Action Recognition (HAR) using 3D skeleton joint data has emerged as a key research area in Human–Computer Interaction (HCI) and visual surveillance applications. However, significant challenges arise due to viewpoint variations, which adversely affect the robustness of existing HAR systems. To address this limitation, a novel **Skeleton Joint Descriptor (SJD)** is proposed, which effectively captures and compensates for viewpoint changes. The proposed method leverages the stability of torso joints to transform all skeleton joints from the Cartesian coordinate system into a view-invariant coordinate framework. Furthermore, redundant joints are identified and eliminated by assigning weights based on their accumulated motion energy over an action sequence. The effectiveness of the proposed framework is validated through extensive experiments conducted on the benchmark NTU RGB+D dataset, comprising 60 distinct human actions. The proposed method achieves an effective recognition accuracy under the **cross-view** and **cross-subject** evaluation protocols, respectively. Comparative analysis with recent state-of-the-art methods clearly demonstrates the superiority of the proposed approach.

Keywords- **Human Action Recognition, Skeleton joints, Torso Matrix, Motion Energy, Accuracy.**

I INTRODUCTION

In recent years, Human Action Recognition (HAR) has attracted significant research attention due to its wide range of applications in video surveillance, gaming, and Human–Computer Interaction (HCI)**.** Despite this growing interest, reliable action recognition remains a challenging problem due to two major factors: **(i)** the inherent complexity of human behaviour as a spatio-temporal process [1], and **(ii)** variations in recording conditions and environmental factors such as viewpoint changes, occlusions, and illumination. Earlier HAR systems predominantly relied on RGB video data for action recognition. However, RGB-based approaches are highly sensitive to variations in human appearance and lighting conditions and often fail to capture reliable motion-related information. With the advent of affordable 3D depth sensors such as Microsoft Kinect, it has become possible to acquire rich three-dimensional information of the human body, significantly advancing research in HAR. Depth sensors provide two primary data modalities: depth images and 3D skeleton joints. Although depth images enable effective separation of the human body from the background, they are often affected by sensor noise and exhibit inconsistent representations when captured from different viewpoints. Consequently, skeleton-based HAR has emerged as a more robust and promising alternative. Moreover, Johansson [2] demonstrated that human actions can be effectively represented using skeletal motion patterns. Early skeleton-based HAR methods focused on handcrafted feature extraction techniques such as Euclidean distances between joint pairs [3], Histogram of Oriented Joint Gradients [4], Lie group [5] representations, and joint covariance matrices [6]. However, these handcrafted features often fail to capture complex motion patterns effectively in the presence of noise and occlusions present in skeleton data captured by Kinect-like sensors. Since skeleton joints are estimated from pixel-level features in individual frames, the extracted joint positions are frequently corrupted by noise, which negatively impacts recognition accuracy. In addition, modelling the spatio-temporal dynamics of human actions using handcrafted or traditional temporal models remains a challenging task. In recent years, the success of deep learning has led to the development of powerful

HAR frameworks based on Convolutional Neural Networks (CNNs) and Residual Neural Networks (ResNets) [7], [8]**.** Although these approaches have achieved notable improvements, skeleton data remain highly sensitive to viewpoint variations, especially when noise is present. As a result, a HAR model trained on one viewpoint often fails to generalize across unseen viewpoints, leading to increased misclassification rates. Furthermore, most existing skeleton-based methods utilize all available joints to represent an action, which introduces significant redundancy in the action descriptor.

To address these critical limitations, this paper presents a novel HAR framework that is robust to viewpoint variations and minimizes redundant joint information. The main contributions of this work are summarized as follows:

- **Viewpoint-invariant representation**: A novel Skeleton Joint Descriptor (SJD) is proposed to achieve robustness against viewpoint variations. The method utilizes stable torso joints as reference points and transforms all skeleton joints from the Cartesian coordinate system into a view-invariant coordinate space. A torso matrix is constructed using seven stable joints, namely hip center, mid-spine, base of spine, right hip, left hip, right shoulder, and left shoulder.
- **Redundancy reduction through motion energy**: To ensure compact yet discriminative action representation, the contribution of each joint is quantified using motion energy, which reflects its importance in action modelling. Based on these weights, only the most informative joints are retained, thereby reducing redundancy while preserving discriminative motion characteristics.

The remainder of this paper is organized as follows: Section II reviews the related literature on skeleton-based HAR. Section III presents the proposed SJD framework along with the motion-energy-based informative joint selection strategy. Section IV discusses the experimental setup and performance evaluation. Finally, Section V concludes the paper with key findings and future research directions.

II. RELATED WORK

J. Liu et al. [9] introduced a deep learning framework known as the Global Context–Aware Attention LSTM (GCA-LSTM) for skeleton-based Human Action Recognition. In this method, a global memory cell is designed to model the overall contextual information of an action sequence. Based on this global context, only the most informative joints are selected at each frame, which helps suppress the influence of less relevant joints and background noise. Additionally, a recurrent attention mechanism is integrated with the LSTM network to iteratively refine joint importance during training. This attention-guided learning enables the network to focus on discriminative body parts over time, thereby improving recognition accuracy. However, despite its strong temporal modelling capability, this approach remains sensitive to viewpoint changes. To further improve robustness against noisy skeleton data, J. Liu et al. [10] proposed an enhanced LSTM architecture with a gating mechanism. This model introduces specialized gates that adaptively regulate the information flow from noisy joint measurements to the memory cell. By selectively filtering unreliable joint inputs, the network becomes more tolerant to sensor-induced noise and occlusions. Although this gating-based LSTM improves recognition performance under noisy conditions, it still lacks inherent view-invariant modelling, which leads to degraded performance under cross-view evaluation settings.

To explicitly address the view-invariance problem, Qiang Nie et al. [11] proposed a dual-descriptor representation using Joint Euler Angles (JEAs) and the Joint Euclidean Distance Matrix (JEDM). The JEAs encode rotational motion information of joints in 3D space, while the JEDM captures pairwise distance relationships between joints to model body posture. By jointly exploiting angular and distance-based geometric features, the method attempts to reduce viewpoint dependency. However, the assumption of rotation consistency across subjects is often violated in real scenarios, since joint rotation angles vary significantly with factors such as body structure, age, muscle strength, and flexibility. As a result, complete viewpoint invariance cannot be guaranteed.

Motivated by the natural graph structure of human skeletons, several researchers have adopted Graph Convolutional Networks (GCNs) for skeleton-based HAR. C. H. Lin et al. [12] proposed a SlowFast-GCN framework by integrating a Spatio-Temporal Graph Convolutional Network (ST-GCN) with Slow FastNet. In this architecture, the ST-GCN branch captures fine-grained motion dynamics, while the SlowFast branch models

static semantic and long-term temporal information at different frame rates. By fusing fast and slow temporal representations, the method achieves improved multi-scale motion modelling. However, the complexity of dual temporal streams increases computational overhead.

D. Zhang et al. [13] developed a Graph Attention Convolutional Neural Network (GACNN) for harvesting discriminative spatial features from skeleton sequences. In this method, an attention mechanism is applied over graph nodes to dynamically assign importance to neighbouring joints. By emphasizing highly correlated joint relationships and suppressing irrelevant ones, the model effectively captures spatial dependencies within the human body structure. Nevertheless, its primary focus is on spatial attention, with limited emphasis on eliminating inter-view variation.

W. Zhang et al. [14] proposed an Attention-based Temporal Graph Convolutional Network (AT-GCN) that jointly learns spatial and temporal features from skeleton data. The model introduces an attention module to calculate adaptive weights for individual joints and temporal frames based on their contribution to action discrimination. This allows the network to automatically prioritize dominant joints and critical motion segments. Despite improved performance, the model still utilizes all joints and does not explicitly suppress redundant joints.

To further enhance relational modelling, Fang Liu et al. [15] proposed a Multi-Relational Graph Convolutional Network (MR-GCN) for HAR by establishing different types of joint relationships using graph knowledge modelling. Three types of relations are defined: global connection (long-range dependencies), symmetric connection (body symmetry), and natural connection (physical bone connectivity). Based on this formulation, a Two-Stream Multi-Relational GCN (2S-MRGCN) was introduced, where one stream processes the flow of body parts and the other processes the flow of skeleton joints. The dual-stream fusion enables comprehensive representation of local and global motion patterns. However, the reliance on full joint connectivity leads to redundant information encoding, which increases model complexity.

Although existing deep learning and graph-based approaches demonstrate strong spatial–temporal modelling capabilities, most of them remain sensitive to viewpoint variations and rely on all skeleton joints, leading to view dependency and redundant feature representation. This critical gap directly motivates your proposed VS2JD with motion-energy-based joint selection.

III. PROPOSED HAR SYSTEM

The proposed Human Action Recognition (HAR) framework is designed to achieve robust view-invariant action recognition with a compact and discriminative skeleton representation. Initially, the 3D skeleton joint sequences acquired from the depth sensor are preprocessed and transformed from the Cartesian coordinate system into a view-invariant coordinate space by using torso joints as stable reference points. This transformation normalizes the skeletal structure across different camera viewpoints and significantly minimizes inter-view variations. Subsequently, an informative joint selection mechanism based on motion energy analysis is applied to quantify the contribution of each joint towards action representation. Joints with higher motion energy are assigned larger weights and selected to form a reduced yet discriminative skeleton descriptor, while less informative joints are discarded to eliminate redundancy. The resulting energy-weighted, view-normalized skeleton features are then encoded into an appropriate spatio-temporal representation and fed into a deep ResNet for action classification. This integrated view normalization, joint selection, and deep feature learning framework enables accurate and robust recognition of human actions under both cross-view and cross-subject settings.

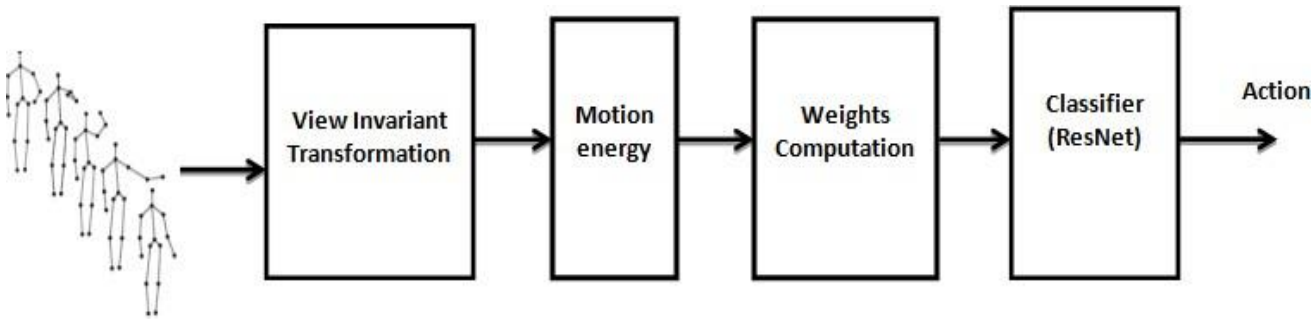


Figure.1 Block Diagram of proposed HAR system

Table.1 Notations

Notation	Meaning	Notation	Meaning
X	Input Action Sequence	N	Number of Frames
I	Number of Joints	**x**	X-axis of the New co-ordinate system
v	Mean distance between right and left hip joints	**z**	Z-axis of the New co-ordinate system
d	Translational Vector	**y**	Y-axis of the New co-ordinate system

A Skelton Joint Descriptor A

Towards the provision of view in variance in HAR, several methods in the past. However they eliminate the partial relative motions from the original skeleton sequence. For example for an instant consider a action called as waist rotating, the rotation of waist are eliminated by the existing methods because they transformed each skeleton into a standard pose as in the frontal view. So the this paper proposed a new view invariant transform called VS^{2}JD which transform all the skeleton in a synchronous manner thereby the relative motions are preserved.

Consider X be an action sequence with N number of frames, for nth frame , the ith joint is represented as $J^n_{i=(x^n_i, y^n_i, z^n_i)} n \in (1,2,\dots,N)$ and $i \in (1,2,\dots,I)$ where I denotes the total number of joints present in each skeleton frame. Figure2 shows the to ,most commonly used joints configurations. Here we used NTURGB+D dataset where every frames represented with 25 joints(shown in Table2), hence I=25. Each Joint is generally signified with five values among which three are local coordinates such as x-y- and z- axis, four this joint label and fifth is the time label n. Mathematically it can be expressed as 5D space vector as J=(x,y,z,i,n). This kind of skeleton representation is sensitive to viewpoint variations, and computed as especially the first three values such as x-y- and Z- AXIS. Hence we transform them into view invariant coordinators and after transformation, let they are denoted as $(\hat{x}, \hat{y}, \hat{z})$, and computed as

$$(\hat{x}, \hat{y}, \hat{z}) = [(R^{\theta}_{x}, 0)(R^{\theta}_{y}, 0)(R^{\theta}_{z}, d)][x, y, z] \quad (1)$$

Where R^{θ}_{x}, R^{θ}_{y} and R^{θ}_{z} are the rotation matrices along x-y-and z- directions respectively. Simply it can be represented as $R^{\theta}_{x,y,z} \in \mathbb{R}^{3\times3}$ is a 3x3 Matrix and 3 denotes the three coordinates x y and z . N ext d is the translational vector computed as

$$d = \frac{1}{N}\sum_{1=1}^{N} J^{n}_{1} \quad (2)$$

Eq(2) makes the origin to move from original point to hip center. The mathematically representation for R^{θ}_{x}, R^{θ}_{y} and $R^{\theta}_{z}$is given as

$$R^{\theta}_{x} = \begin{bmatrix} cos\theta & 0 & -sin\theta \\ 0 & 1 & 0 \\ sin\theta & 0 & cos\theta \\ 1 & 0 & 1 \end{bmatrix} \quad (3)$$

$$R^{\theta}_{y} = \begin{bmatrix} 0 & cos\theta & -sin\theta \\ 1 & sin\theta & cos\theta \end{bmatrix} \quad (4)$$

$$R^{\theta}_{z} = \begin{bmatrix} cos\theta & -sin\theta & 0 \\ sin\theta & cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

Now construct a matrix called torso matrix TM $\in \mathbb{R}^{7\times N}$ by concatenating the position of torso joints. The torso joints have very less positional deviation and here the joint {1,2,,5,9,13,17,21} are considered as torso joints . Figure 2 shows an example frame with only torso joints. Mathematically the TM is represented as

$$TM = \begin{bmatrix} J^{1}_{1} & J^{2}_{1} & J^{3}_{1} & \dots & J^{N}_{1} \\ J^{1}_{2} & J^{2}_{2} & J^{3}_{2} & \dots & J^{N}_{2} \\ J^{1}_{5} & J^{2}_{5} & J^{3}_{5} & \dots & J^{N}_{5} \\ J^{1}_{9} & J^{2}_{9} & J^{3}_{9} & \dots & J^{N}_{9} \\ J^{1}_{13} & J^{2}_{13} & J^{3}_{13} & \dots & J^{N}_{13} \\ J^{1}_{17} & J^{2}_{17} & J^{3}_{17} & \dots & J^{N}_{15} \\ J^{1}_{21} & J^{2}_{21} & J^{3}_{21} & \dots & J^{N}_{21} \end{bmatrix} \quad (6)$$

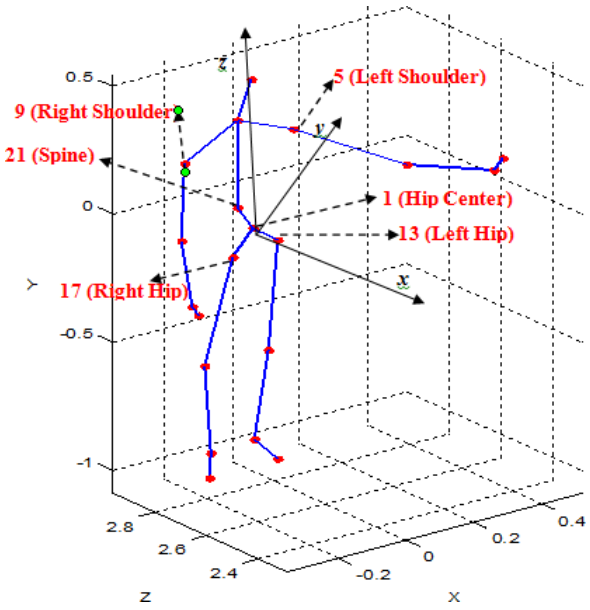


Figure 2. Action frame with only torso Joints

Here the torso matrix is a 7Xn matrix with 7 rows and n columns . then now apply Principal component Analysis (PCA)over Torso matrix and let z be the obtained first principal component which is always maintains the larger distance from the torso joints and is considered a Z-axis in the new view invariant coordinate system . A rough estimation of z can be done based on the orientation directing from hip center to spine. But it may cause an incorrect and inaccurate value due to Skelton joints which suffers from noise. Next the orientation of the second principle component obtained by PCA over TM is predicted to define both X- or Y-axis in the new view invariant coordinate system doing inference of orientation is not so easy . unlike this work consider x to signify the X-axis and it is obtained as

$$\mathbf{x}=\arg\min_{\mathbf{x}}\langle \mathbf{x}, \mathbf{v}\rangle \quad (7)$$

where v is a distance vector obtained based on the normalized accumulated distance of 17th and 13th mean distance . Mathematically it is expressed as

$$\mathbf{v}=\frac{\sum_{n=1}^{N}\sqrt{((x_{17}^{n}-x_{13}^{n})^{2}+(y_{17}^{n}-y_{13}^{n})^{2}+(z_{17}^{n}-z_{13}^{n})^{2})}}{N} \quad (8)$$

Where v also denotes the mean distance of action directing from right hip center to left hip center. Then the Y-axis of new invariant coordinate system is obtained as y =zXx coming a new positions with the time and joint labels the new representation is denoted as

$J=(\hat{x},\hat{y},\hat{z},i,n)$ which ensures reliance against view point variation

Table.2 Skeleton joints of NTURGB+D dataset

1 Base of spine	10 right shoulder	18 right knee
2 middle of spine	11 right elbow	19Right ankle
3 nee	12 right wrist	20 right foot
4 head	13 left knee	21 spine
5 left shoulder	14 left knee	22. tip of left mind
6 left elbow	15 left ankle	23 tip of right hand
7 left wrist	16 left foot	24 tip og right hand
8 left hand	17 right hip	25 right tumb
9 left hand		

B Motion descriptor

Once transformation of skeleton joints from Cartesian coordinate system to view invariant coordinate system, then they are processed for action describing based on motion attributes. Towards going we extract only informative joints which can contribute more in representing the action . In general, human beings pay more attention on moving objects than on the static objects , with this motivation we actually aimed to pick only informative joints to describe the action moreover the salient informative joints weights for each joints to choose them consider a joint i in the nth frame

$J_i^n=(\hat{x}_i^n,\hat{y}_i^n,\hat{z}_i^n)$, we compute the motion energy

$$ME_i=\sum_{n=2}^{N}\|J_i^n-J_i^{n-1}\| \quad (9)$$

Eq(9) is applied for all the joints and motion energy is calculated for every joint then the weight is calculated as

$$\omega = \rho . \|ME_i\| + 1 - \rho \quad (10)$$

Where $0 \leq \rho \leq 1$ is an arbitrary constant and is an is an arbitrary constant and $\| \ \|$ is a normalization function which normalization function which normalize the energy into the range of o and 1 Figure. 3 shows the accumulated motion energy of each joint

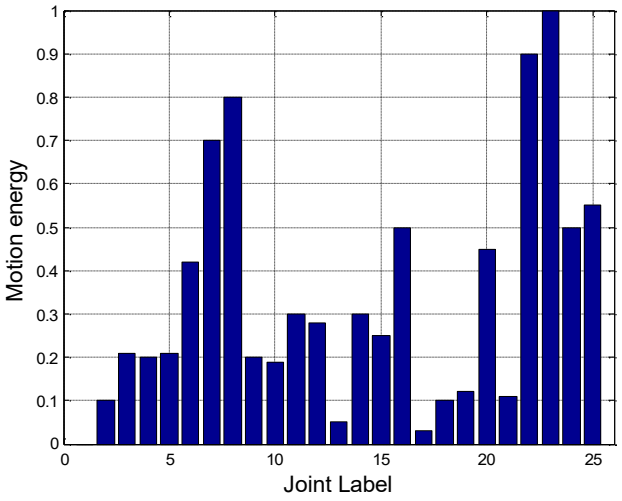
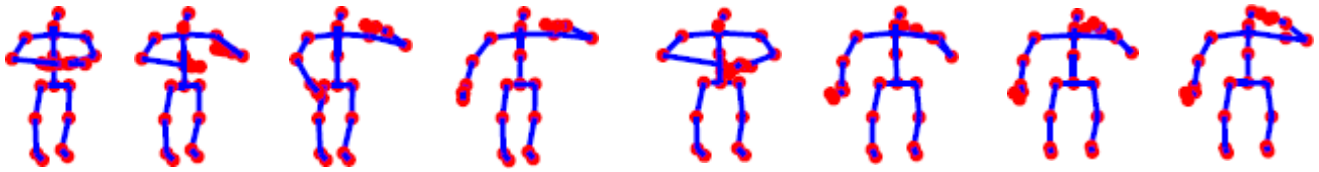


Figure.3 Accumulated motion energy of each joint

IV. EXPERIMENTAL ANALYSIS

The proposed Human Action Recognition (HAR) framework is extensively evaluated on the benchmark NTU RGB+D dataset [22], which is one of the largest and most widely used datasets for skeleton-based action recognition. This dataset comprises a total of 60 diverse human action classes, covering daily activities, mutual interactions, and health-related actions, thereby providing a comprehensive test bed for validating the robustness and generalization capability of HAR systems. The dataset was collected from 40 distinct subjects, and each action was recorded from 80 different viewpoint configurations using a Microsoft Kinect V2 depth camera. In total, the dataset contains 56,880 skeleton video sequences. Each frame in a sequence is represented by a 3D skeleton consisting of 25 anatomically defined joints, where each joint is encoded using three- dimensional spatial coordinates (x, y, z). This rich spatial and temporal representation enables accurate modelling of complex human motion dynamics.

To ensure fair and standardized performance evaluation, two widely accepted evaluation protocols are adopted, namely the Cross-Subject (CS) and Cross-View (CV) protocols: For experimental validation, a total of 40,320 samples are used for training, and 16,560 samples are reserved for testing, in accordance with the standard dataset split defined by the NTU RGB+D benchmark. This large-scale training setup enables robust learning of spatio-temporal motion patterns, while the diverse testing set rigorously evaluates the system's recognition performance under both subject and viewpoint variations. Furthermore, Figure 5 illustrates representative skeleton samples of two action categories, namely "brushing teeth" and "drinking water", highlighting the structural and motion differences captured by the 3D joint representations.



(a) (b)

Figure.5 Sample Actions from NTURGB+D dataset (a) Drinking Water and (b) Brushing Teeth

Figure.6 Detection rates of several actions in the NTURGB+D dataset

Figure.6 illustrates the class-wise detection (recognition) accuracy (%) of the proposed HAR system on the NTU RGB+D dataset, considering all 60 action classes. The horizontal axis represents the individual action categories (1–60), while the vertical axis denotes the detection rate in percentage. Each bar corresponds to the recognition performance of a specific action. From the figure, it can be observed that the proposed method achieves consistently high recognition accuracy across the majority of action classes, with most detection rates lying above 80%, and several classes reaching above 95% accuracy. A few actions exhibit comparatively lower performance, which can be attributed to high inter-class similarity, subtle motion patterns, or occlusion effects present in those activities. Nevertheless, the overall trend demonstrates the robustness and discriminative capability of the proposed view-invariant HAR framework across diverse human actions

. Table.3 Comparison with existing methods on accuracy (%)

Approach	**Cross View (CV)**	**Cross Subject (CS)**
LARP [5]	52.7600	50.0800
Deep RNN [16]	64.0900	56.2900
Deep LSTM [16]	67.2900	60.6900
Part-Aware LSTM [16]	70.2700	62.9300
SlowFast-GCN [12]	90.0000	83.8000
Trust Gate + ST-LSTM [18]	77.7000	69.2000
HSGAC [13]	90.3000	81.8300
ATGCN [14]	89.7000	83.3000
ST-GCN [17]	88.3000	81.5000
Clips + CNN + MTLN [19]	84.8300	79.5700
Clips + CNN + Pooling [19]	80.4600	76.3700
3Scale ResNet152 [20]	-	85.0000
Proposed	**90.3490**	**85.9080**

Table 3 presents a comprehensive comparison of the proposed approach with several state-of-the-art Human Action Recognition (HAR) methods under two standard evaluation protocols: Cross-View (CV) and Cross-Subject (CS). From the quantitative results, it is evident that the proposed method achieves superior performance with accuracies of 90.3490% (CV) and 85.9080% (CS), outperforming most existing methods. The earlier deep learning–based approaches such as Deep RNN [16], Deep LSTM [16], and Part-Aware LSTM [16] demonstrate moderate performance, achieving CV accuracies of 64.09%, 67.29%, and 70.27%, respectively. These methods primarily focus on learning temporal dependencies in skeleton sequences but do not explicitly address viewpoint invariance. As a result, their performance degrades

significantly under cross-view conditions. This limitation is particularly evident for actions involving high inter-class similarity, such as *"brushing teeth"* and *"drinking water"*, where subtle motion differences must be distinguished across varying viewpoints.

The LARP method [5], which represents human motion using Lie group modeling, records the lowest performance with only 52.76% (CV) and 50.08% (CS) accuracy. Although Lie group representation captures motion geometry, it fails to ensure robustness against viewpoint variations and inter-class similarity, leading to poor recognition for actions such as *"pat on back"*, *"point finger"*, *"point to something"*, and *"pushing"*. Recent Graph Convolutional Network (GCN)-based approaches, including SlowFast-GCN [12], HSGAC [13], ATGCN [14], and ST-GCN [17], show significantly improved performance due to their ability to model spatio-temporal joint relationships using graph structures. These methods achieve CV accuracies in the range of 88.30% to 90.30% and CS accuracies of 81.50% to 83.80%. However, despite their strong modeling capability, GCN-based methods remain sensitive to noisy skeleton data and generally utilize all skeleton joints, which introduces redundant information and increases computational complexity.

In contrast, the proposed approach explicitly addresses two major limitations that are not jointly handled by existing methods:

1. Viewpoint invariance, achieved through transformation of skeleton joints into a new view-invariant coordinate system using stable torso joints, and
2. Redundancy elimination, achieved through motion-energy-based informative joint selection, which discards less contributivel joints before classification.

Because of these two complementary mechanisms, the proposed method demonstrates consistent superiority over GCN-based methods, achieving an average performance improvement of 2.05% under CV and 4.336% under CS protocols. This clearly indicates that the proposed framework not only provides robust view-invariant representation but also enhances noise tolerance and compactness of the action descriptor. The method Clips + CNN + MTLN [19] and Clips + CNN + Pooling [19] achieve reasonable performance but are limited by the use of clip-level representations, which do not fully capture long-range temporal dynamics. The 3Scale ResNet152 [20], although achieving 85.00% under CS, lacks reported CV performance and primarily focuses on deep spatial feature learning without explicit viewpoint normalization.

Computational Efficiency: The computational efficiency of the proposed method is also evaluated. All experiments were conducted on a system with 8 GB RAM, a 2.5 GHz processor, using MATLAB 2014. The proposed system required 79,575.5 seconds for training, while the testing time was only 0.39 seconds per sequence, which includes both feature extraction and classification. This demonstrates that, despite its strong performance, the proposed system remains suitable for real-time and near-real-time HAR applications.

V. CONCLUSION

This paper presented a novel view-independent Human Action Recognition (HAR) framework that effectively addresses two critical challenges in skeleton-based HAR, namely viewpoint variation and redundant joint representation. To overcome viewpoint dependency, a View-Stabilized Skeleton Joint Descriptor (VS2JD) was introduced, which transforms 3D skeleton joints from the Cartesian coordinate system into a new view-invariant coordinate space using stable torso joints as reference points. Further, a motion-energy-based informative joint selection strategy was adopted to assign contribution-based weights to each joint, enabling the elimination of redundant joints while preserving discriminative motion characteristics. The resulting compact and view-normalized skeleton representation was then classified using a deep Residual Network (ResNet).

Extensive experiments conducted on the benchmark NTU RGB+D dataset using both Cross-Subject (CS) and Cross-View (CV) evaluation protocols clearly demonstrate the effectiveness and robustness of the proposed approach. The proposed method achieved 90.3490% accuracy under the CV protocol and 85.9080% under the CS protocol, outperforming several state-of-the-art LSTM-based, CNN-based, and GCN-based HAR methods. Unlike many existing approaches that either lack explicit view-invariant modelling or suffer from redundancy due to full-joint utilization, the proposed framework jointly ensures view invariance, redundancy reduction, and noise robustness, leading to consistent performance

improvements. Furthermore, the system achieved a low testing time of 0.39 seconds per sequence, making it suitable for real-time HAR applications such as surveillance, healthcare monitoring, and human–computer interaction systems.

In future work, the proposed framework can be extended by integrating attention-based temporal modelling, multi-person interaction analysis, and lightweight deep architectures for deployment on edge and embedded platforms. Additionally, fusion with RGB and depth modalities can further enhance recognition performance in complex real-world environments.

REFERENCES

1 Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. DOI:10.1016/j.imavis.2009.11.0142

2 G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception, Psychophys.*, vol. 14, no. 2, pp. 201–211, Jun. 1973. https://link.springer.com/article/10.3758/BF03212378

3 X. Yang and Y. L. Tian, "Eigen Joints-based action recognition using Naïve-Bayes-nearest-neighbor," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 14–19. https://doi.org/10.22937/IJCSNS.2022.22.6.16

4 L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20– 27. https://doi.org/10.1109/TCSVT.2017.2715045

5 R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595. https://openaccess.thecvf.com/content_cvpr_2014/papers/Vemulapalli_Human_Action_Recognition_2014_CVPR_paper.pdf

6 M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. IJCAI*, vol. 13, 2013, pp. 2466–2472. https://dl.acm.org/doi/10.5555/2540128.2540483

7 Nguyen, Hung-Cuong, Thi-Hao Nguyen, Rafał Scherer, and Van-Hung Le. 2023. "Deep Learning for Human Activity Recognition on 3D uman Skeleton: Survey and Comparative Study" *Sensors* 23, no. 11: 5121. https://www.mdpi.com/1424-8220/23/11/5121

8 C. Wang and J. Yan, "A Comprehensive Survey of RGB-Based and Skeleton-Based Human Action Recognition," in *IEEE Access*, vol. 11, pp. 53880-53898, 2023. Digital Object Identifier 10.1109/ACCESS.2023.328

9 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. DOI:10.1109/TIP.2017.2785279

10 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. DOI:10.1109/CVPR.2017.391

11 Qiang Nie, Jiangliu Wang, Xin Wang, Yunhui Liu, "View-Invariant Human Action Recognition Based on a 3D Bio-Constrained Skeleton Model", *IEEE Transactions on Image Processing*, Vol. 28, No. 8, August 2019, pp.3959-3972. https://ieeexplore.ieee.org/document/8672922.

12 C. -H. Lin, P. -Y. Chou, C. -H. Lin and M. -Y. Tsai, "SlowFast-GCN: A Novel Skeleton-Based Action Recognition Framework," *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, Taipei, Taiwan, 2020, pp. 170-174. December 2020DOI:10.1109/ICPAI51961.2020.00039

13 D. Zhang, H. Gao, H. Dai and X. Shi, "Human Skeleton Graph Attention Convolutional for Video Action Recognition," *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, Shenyang, China, 2020, pp. 183-187. https://www.computer.org/csdl/proceedings/isctt/2020/1rHeKX6WcSc

14 W. Zhang, L. Zhou and X. Qian, "Skeleton-based Action Recognition with Attention and Temporal Graph Convolutional Network," *2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP)*, Nanjing, China, 2021, pp. 19-23.

15 F. Liu, Q. Dai, S. Wang, L. Zhao, X. Shi and J. Qiao, "Multi-Relational Graph Convolutional Networks for Skeleton-Based Action Recognition," *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing&Networking (ISPA/BDCloud/SocialCom/SustainCom)*, Exeter, United Kingdom, 2020, pp. 474-480. DOI:10.1088/1742-6596/2030/1/012091

16 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. DOI:10.1109/CVPR.2016.115

17 Yan S, Xiong Y, Lin D. “Spatial temporal graph convolutional networks for skeleton-based action recognition”, *Proceedings of the AAAI conference on artificial intelligence*. 2018, 32(1). DOI:10.1609/aaai.v32i1.1232

18 Jun Liu, Amir Shahroudy, Dong Xu, Alex C. Kot, and Gang Wang, “Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates”, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 40, Dec. 2018, pp. 3007-3021.
https://ieeexplore.ieee.org/document/8101019

19 Q. Ke, M. Bennamoun, A. Senjian, F. Sohel, B. Faird, “A New representation of Skeleton Sequences for 3D action representation”, *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition,* pp.4570-4579, 2017. DOI:10.1109/CVPR.2017.486

20 Bo Li, Yuchao Dai, Xuelian Cheng, Huahui Chen, Yi Lin and Mingyi He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," *Proceedings of IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Hong Kong, pp. 601-604, 2017. DOI:10.48550/arXiv.1704.05645

21 Vinoda Reddy, P Suresh Varma and A. Govardhan, “Recurrent Feature Grouping and Classificationfor action model prediction in CBMR”, International Journal of Daa Management and Knowledge process Vol.7,No.5/6, November 2017, http://dx.doi.org/10.5121/ijdkp.2017.7605

22. Gopampallikar Vinoda Reddy,; Kongara Deepika; Lakshmanan Malliga; Duraivelu Hemanand; Chinnadurai Senthilkumar, “Human Action Recognition Using Difference of Gaussian and Difference of Wavelet” BIG DATA MINING AND ANALYTICS ISSN 2096-0654 07/10 pp336 –346 Volume6,Number3,
DOI:10.26599/BDMA.2022.9020040

23 Vinoda Reddy, P Suresh Varma and A.Govardhan, “MultiLinear Kernel Mapping for FeatureDimension Reduction in Content Based Multimedia Retrieval System”, The International Journal of Multimedia & Its Applications (IJMA), Vol. 8, No. 2, April 2016, pp. 1-16.http://dx.doi.org/10.5121/ijma.2016.8201.
https://aircconline.com/ijma/V8N2/8216ijma01.pdf

24 Vinoda Gopampallikar1 , Shashi Kant Gupta 2,”A Survey on Human Action Recognition Using Depth Maps and Skeleton Postures SGS Engineering &Sciences,Vol.3no.1, 2025 LGPRSGS Engineering &Sciencehttps://spast.org/techrep/article/view/5605