INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Enhancement of Deep Positional Attention-Based Human Action Recognition by Using Geometric Positional Features

N Srilakshmi[1]*, **N Radha**[2]

1 Ph.D. Scholar, Department of Computer Science, PSGR Krishnammal College for Women, Tamilnadu, Coimbatore, India
2 Associate Professor, Department of Computer Science, PSGR Krishnammal College for Women, Tamilnadu, Coimbatore, India

OPEN ACCESS

Received: 20-02-2023
Accepted: 20-06-2023
Published: 03-08-2023

Citation: Srilakshmi N, Radha N (2023) An Enhancement of Deep Positional Attention-Based Human Action Recognition by Using Geometric Positional Features. Indian Journal of Science and Technology 16(29): 2190-2197. https://doi.org/10.17485/IJST/v16i29.379

* **Corresponding author**.

*srilakshmiphd123@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: To learn different geometric features of body joints from video frames, as well as trajectory point coordinates, for Human Activity Recognition (HAR). **Methods**: Joints and Trajectory-pooled 3D-Deep Geometric Positional Attention-based Hierarchical Bidirectional Recurrent convolutional Descriptors (JTDGPAHBRD)-based HAR framework is proposed. This framework considers the skeleton graph to extract geometric features such as joints, edges, and surfaces, along with the trajectory point coordinates. A new 3D-deep convolutional network with View Conversion (VC) and Temporal Dropout (TD) layers is designed that uses a Positional Attention-based Hierarchical Bidirectional Recurrent Neural Network (PAHBRNN) to learn more discriminatory high-level features. Then, a Fully Connected Layer (FCL) is applied to get the Video Descriptor (VD) of a particular frame. Moreover, the obtained VD is classified by the Support Vector Machine (SVM) classifier to recognize various kinds of human activities. **Findings**: The test findings show that the JTDGPAHBRD framework using the Penn Action database achieves a recognition rate of 99.7% compared to the existing HAR frameworks. **Novelty**: This framework has significantly improved the recognition of human activities. Thus, it represents a promising framework for the HAR.

Keywords: Human activity recognition; JTDPAHBRD; Geometric features; View conversion; Temporal dropout; SVM

1 Introduction

An efficient HAR can be difficult due to a variety of circumstances, views, and other factors. Over the past years, numerous HAR frameworks have been developed using deep learning algorithms. Deshpnande and Warhade [(1)] presented an improved model for HAR by integrated feature approaches such as Histogram of Gradient (HOG) local feature descriptor and Principal Component Analysis (PCA) as global features, as well as an optimized SVM classifier. But it cannot learn the local relationship among the image pixels and it needs a large number of input parameters. Weiyao et al. [(2)] developed

a multi-modal HAR framework based on Bilinear Pooling and Attention Network (BPAN). The RGB and skeleton information was pre-processed and a multimodal fusion network was devised to obtain fused characteristics. The FC 3-unit perceptron was used to make the final classification decision, but the training database was limited and the total accuracy was influenced by the weight value in the loss function. Muhammad et al.[3] designed a Bidirectional Long Short-Term Memory (BiLSTM)-based attention strategy with a dilated Convolutional Neural Network (CNN) to choose effective features in the input frame and recognize various human actions. Also, the center loss with softmax was used to minimize the loss function in video-based HAR. But it used a single-stream learning strategy, which was not suitable to learn more discriminative features from the video frames and recognize complex actions in large-scale datasets.

Khan et al.[4] developed a deep learning model, which comprises feature mapping, feature fusion, and feature selection. The feature mapping was conducted by DenseNet201 and InceptionV3. Then, deep features were extracted and fused by the serial-based extended model. The best features were chosen by the Kurtosis-controlled weighted K-Nearest Neighbor (KNN). Finally, those features were classified by many supervised learning algorithms. But it has a high computational time during the original deep feature extraction. Wang et al.[5] designed a novel HAR technique called Skeleton Edge Motion Networks (SEMN) to extract gesture data. The SEMN was designed by combining many spatiotemporal segments to obtain a deep interpretation of skeleton structures. A novel advanced rank error was applied to preserve sequential imperative data, but it was difficult to differentiate individual activities from granular skeleton images.

Saleem et al.[6] utilized pre-trained VGG-19 for extracting the body joints from the 2D body skeleton and applied SVM classifier for classifying the human actions. But, its accuracy was less since it did not learn spatiotemporal relationships among different pixels.Yadav et al.[7] designed a Convolutional LSTM (ConvLSTM) network for skeletal-based HAR. Human identification and pose estimation were used to determine skeleton coordinates, which were combined with geometric and kinematic traits to create reference traits. A categorizer head was employed, but it did not consider edges and surface-related geometric traits to enhance HAR efficiency. Putra et al.[8] developed a Deep Neural Network (DNN) using transfer learning and shared-weight schemes for classifying human actions. This model consisted of pre-trained CNNs, attention layers, LSTM with residual learning, and softmax layers. But it did not satisfy outcomes analyzed for online cases, which need to classify sequences of ambiguous actions. Li et al.[9] developed a triboelectric gait sensor system for HAR. They applied LSTM and residual units to extract deep features from multichannel time-series gait data for improving HAR performance. But it needs more geometric features for effective HAR. The above-studied frameworks used only a single-stream learning strategy, whereas a two-stream learning strategy has emerged recently to learn more discriminative features from the video sequences and recognize complex actions accurately. From this perspective, the JTDPAHBRD framework has been developed, which employs PAHBRNN to improve the attribute concatenation task[10]. In this PAHBRNN-based pooling, the attribute vectors associated with the human skeleton in all clips were split into multiple parts according to the body structure. Such parts were fed to the multiple PABRNNs to hierarchically capture and concatenate the long-term spatiotemporal traits. Also, the FCL was utilized to provide the absolute VD that was classified by the SVM for HAR.

On the contrary, these frameworks merely fuse the joint and trajectory coordinates at every interval, while the geometric correlation among joints is ignored in the feature extraction and concatenation process. A typical activity is the formation of a fossil skeleton linked by joints. Therefore, a meaningful description of activities is provided by the relative geometries among joints. The trajectory of a specific joint only conveys gesture data and lacks contour or geometrical data.

Hence, the purpose of this research is to consider the relative geometries in the human body to improve HAR. The JTDGPAHBRD-based HAR framework is proposed, which considers the skeleton graph to extract geometric features such as joints, edges, and surfaces along with the trajectory point coordinates. The joints are separate points of the body. The edges are bones that link 2 nearby joints and are represented via the related joint's locations. The surfaces are the planes made through 2 nearby articulated bones. A new 3D-deep convolutional network with VC and TD layers is designed that uses the PAHBRNN to learn more discriminatory high-level features. Then, the FCL is applied to get the VD of a particular frame. Moreover, the obtained VD is classified by the SVM classifier to recognize various kinds of human activities. Thus, this framework can increase the recognition rate of HAR systems.

2 Methodology

This section briefly explains the JTDGPAHBRD framework for HAR. A general schematic representation of the JTDGPAHBRD framework for HAR is depicted in Figure 1. The major goal is to predict an activity label for an unknown video sequence. Initially, an entire video sequence is split into frames. For each frame, basic geometries like joints, edges, and surfaces are defined with the help of a skeleton graph structure. Also, the trajectory coordinates at each joint location are retrieved. Then, those geometry and trajectory coordinates are passed to the 2-stream C3D network, which comprises PAHBRNN for the pooling process rather than the max-min pooling strategy. Afterward, the features from both streams, such as feature and attention, are concatenated

by the bilinear product, and an absolute VD is obtained by the FCL. The obtained VD is further learned by the SVM classifier for predicting the action labels of test video sequences.

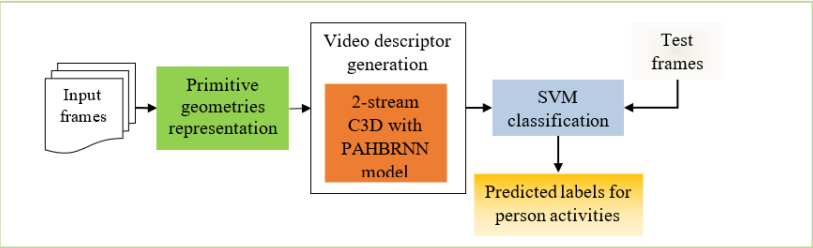

Fig 1. Schematic representation of JTDGPAHBRD-based HAR

2.1 Representation of Primitive Geometries from Skeleton Information

The skeleton information is an arrangement of 3D coordinates of points that create the distorted pattern of the body. Different motions of the points exist while the body changes deliberately. Such points are linked according to the physical pattern of body joints. The body's shape is represented as a graph, where joints are called points and bones are called edges. For a particular person, the skeleton information includes 2 geometric restraints: (1) since a bone's size is fixed, the gap between 2 nearby points along a linked fragment is constant, and (2) three points that create 2 overlapping fragments lie on a similar plane.

According to these interpretations, the skeleton information carries 3 kinds of data: the remote joints, the edges that represent the linked fragments, and the surfaces covered by overlapping fragments. These are explained below.

- Joints

Consider M joints for the body pattern, the coordinates of points at an interval create $M \times 3$ matrix. When the video sequence length is T, the skeleton information is represented by a tensor X with dimension $T \times M \times 3$. The joint coordinates that vary over the period reveal the temporal dynamics of activities. The joint coordinates of a given view are converted into the other view by the rotation matrix. Consider p_k is the joint's coordinate vector at a specific period, the new coordinate vector is attained by

$$\widetilde{p}_k = R p_k \tag{1}$$

In Eq. (1), R denotes the revolution matrix with a size of 3×3. For a given video, consider that R is equal for various joints and various intervals. So, for the joints tensor X, the novel $\widetilde{X}$ detected from the other view is defined by

$$\widetilde{X} = X \times_3 R^T \tag{2}$$

In Eq. (2), $\times_3$ is 3-mode tensor multiplication, $\widetilde{X}$ and X take equal magnitudes.

- Edges

In addition to the temporal features of joints, bone movement forms different activities. A graph is utilized to define the physical links of joints. The joints are represented by the nodes and the bones are represented by the edges.

For a graph of M nodes, there are $M - 1$ edges. The edge indicates the bone orientation. All nodes have a coordinate vector and all edges are denoted by subtracting the vector of the beginning point from that of the endpoint. That is,

$$e_k = p_i - p_j \tag{3}$$

In Eq. (3), e_k, p_i, p_j define the coordinate vectors of the edge, endpoint, and beginning point, correspondingly. The skeleton structure edges are defined by a tensor Y with sizes $T \times (M-1) \times 3$. The node is represented using the edge vector that terminates at that node. For a node that doesn't contain end points of edges, it is represented by a 0 vector. In this manner, the size of Y is incremented by 1 and the sizes of X and Y are created as equal.

The coordinate vectors of edges of a certain view are converted into the other view using a revolution matrix. For the coordinate vector of an edge at a specified interval, the conversion is realized depending on Eqns. (1) and (3):

$$\widetilde{e}_k = \widetilde{p}_i - \widetilde{p}_j = Re_k \tag{4}$$

In Eq. (4), $\widetilde{e}_k$ is the converted vector of e_k. Also, it is defined as:

$$\widetilde{Y} = Y \times_3 R^T \tag{5}$$

In Eq. (5), $\widetilde{Y}$ is the converted tensor of edges. Relating Eq. (2) and Eq. (5), it is observed that the revolution matrices of joints and edges are equal.

- Surfaces

The edges reflect the pairwise correlations among the joints. It is unable to represent the scenario wherein 2 joints with nearby edges are situated next to one another. Similarly, HAR also benefits from the relative motions of nearby bones. Because 2 nearby edges create a plane surface, the standard vector is utilized to represent the surface. Consider e_i, e_j are the vectors denoting 2 nearby edges, the standard vector s_k is:

$$s_k = e_i \times e_j \tag{6}$$

In Eq. (6), $\times$ is the cross product in the 3D area. The vector is not regularized because the magnitude represents the overlapping angle of the respective edges. To maintain the dimension of the standard vector similar to the coordinate vector, it is multiplied by 100. For the body with M joints, there are $(M+2)$ planes. To create a proper relation with joints and edges, 2 surfaces with duplicate data (the standard vector is defined by another standard vector) are omitted. This provides M surfaces so the standard vectors of a sequence are defined using a tensor Z with sizes $T \times M \times 3$.

The standard vector of a particular view is detected from the other views. According to the Eq. (6) and Eq.(4), the novel standard vector of a plane at a certain interval is defined by

$$\widetilde{s}_k = (Re_i) \times (Re_j) = Co(R)\, s_k \tag{7}$$

In Eq. (7), $Co(R)$ denotes the cofactor matrix of R, which is the transpose of the adjoint matrix. For an invertible matrix R, get:

$$Co(R) = (det(R)) \left(R^{-1}\right)^T \tag{8}$$

In Eq. (8), $\left(R^{-1}\right)^T$ denotes the transpose of the inverse R. It must be observed that the determinant of a revolution matrix is one and R^{-1} is its transpose. Eq. (8) is deduced as:

$$Co(R) = \left(R^{-1}\right)^T = R \tag{9}$$

Thus, for the tensor interpretation of planes, it pursues that:

$$\widetilde{Z} = Z \times_3 R^T \tag{10}$$

In Eq. (10), $\widetilde{Z}$ denotes the converted tensor of surface standard vectors. Relating Eq. (2), Eq.(5), and Eq. (10), it is determined that joints, edges, and planes possess an equal revolution matrix.

2.2 Recognition of Human Activities

For HAR, the 3 categories of skeleton information such as the coordinates of joints, edges, and surfaces along with the trajectory coordinates are fed to the 2-stream C3D network. The entire network structure for HAR is depicted in Figure 2. In this structure, the VC layer and TD layer are added to enhance attribute mining and VD generation.

A. View conversion: Human skeletons can be captured from a random camera perception in a real-time circumstance. To create view-invariant interpretations, this framework intends to utilize the VC layer to convert the skeleton information into 3D space by capturing the joints, edges, and planes.

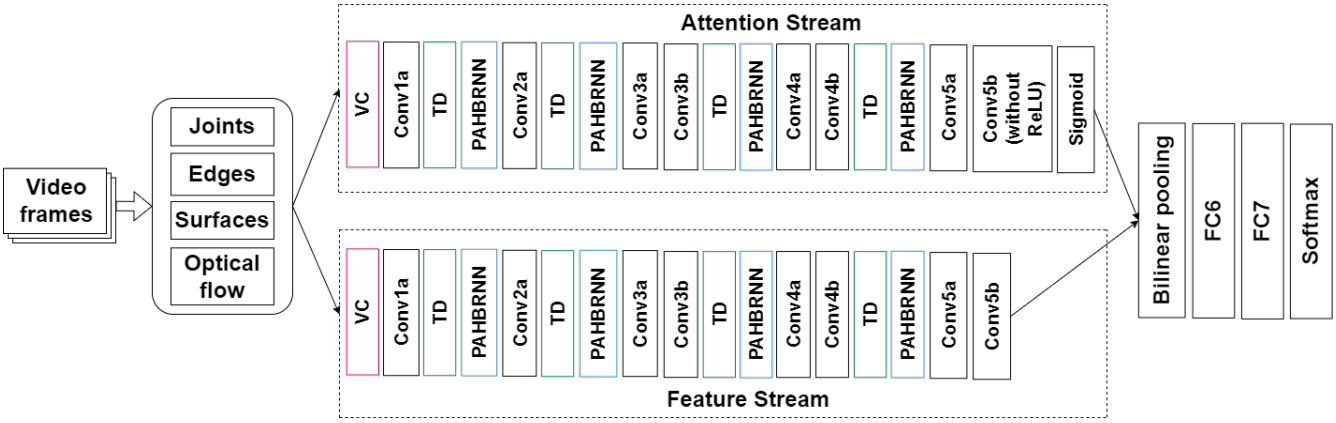

Fig 2. Structure ofproposed 2-stream bilinear C3D network for HAR.

For a given video, X, Y, Z are converted by a similar conversion matrix R. According to Euler's rotation theory, R is denoted as a mixture of revolutions regarding x, y, z axes:

$$R = R_x(\alpha) R_y(\beta) R_z(\gamma) \tag{11}$$

In Eq. (11), α, β, γ denote rotate angles of x, y, z, correspondingly. Given a skeleton structure, R is calculated using 3 separate orientation variables, which are predicted from the skeletons by creating a few significant hypotheses. In the learning task, the orientations in a specific value are arbitrarily chosen and R is computed to convert the inputs. Here, α, β are sampled from $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and γ is set as 0 since the surface is nearly perpendicular to the z-axis. In the test phase, α, β, γ are set as 0, and the actual tensors of joints, edges, and planes are utilized.

B. Temporal dropout: The skeletons gathered might not often be accurate because of noise and pose variations. To solve this issue, a method is adopted depending on dropout, which enhances the framework's robustness. For a typical dropout, all hidden units are arbitrarily neglected from the model with a chance of p_{drop} in the learning. For the test stage, each activation is utilized and $1 - p_{drop}$ is multiplied to consider the rise in the estimated bias. TD is marginally varied from the typical dropout. For $T \times d$ matrix interpretation of a frame, where T denotes the frame size and d denotes the feature size, merely T dropout tests are executed, and the dropout range is extended among the feature size. This method is motivated by the spatial dropout to analyze the convolution feature 4D tensor. In this study, it is altered for 3D tensor and applied for attribute training from frames. As illustrated in Figure 2, the TD is conducted before the PAHBRNN.

Thus, the 2-stream C3D network is trained to create the absolute VD of a given sequence. The obtained VD is provided to the SVM algorithm to categorize the activities of subjects in specified videos.

3 Results and Discussion

The effectiveness of the JTDGPAHBRD framework is assessed by executing it in MATLAB 2017b. The Penn Action Corpus is used in this scrutiny, which comprises 2326 video sequences that each include 15 activity tags. All clips are assembled from several web video libraries and involve 50–100 blocks, each of which has 13 body joints annotated. With this dataset, 1861 video sequences are utilized for learning, while 465 video sequences are utilized for testing. Sources include C3D features, coordinates of primitive geometries, and trajectory coordinates. To assess the recognition accuracy of JTDGPAHBRD using these characteristics, several aggregation setups are used.

The proportion of a person's actions that are correctly recognized is referred to as recognition accuracy. It is computed by using Eq. (9).

$$Accuracy = \frac{Number\ of\ recognized\ actions}{Total\ number\ of\ actions\ tested} \times 100\%$$

The sample input video frame and its corresponding skeleton image for representing the coordinates of primitive geometries are displayed in Figure 3.

Table 1 displays the JTDGPAHBRD recognition rate values for the Penn Action dataset.

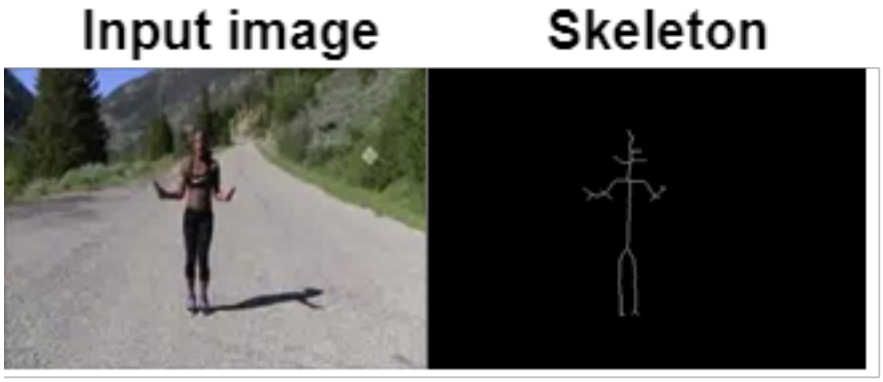

Fig 3. Inputimage and its corresponding skeleton image for primitive geometry coordinates representation.

Table 1. Recognition Rate (%) of Sources and JTDGPAHBRD with Distinct Settingson Penn Action Database

	Cumulative all the activations	**JTDGPAHBRD Ratio Scaling (1×1×1)**	**JTDGPAHBRD Coordinate Mapping (1×1×1)**	**JTDGPAHBRD Ratio Scaling (3×3×3)**	**JTDGPAHBRD Coordinate Mapping (3×3×3)**
Primitive geometries + trajectory coordinates	0.7018	-	-	-	-
$fc7$	0.8045	-	-	-	-
$fc6$	0.8298	-	-	-	-
*conv*5*b*	0.7931	0.8763	0.9231	0.8718	0.9042
*conv*5*a*	0.7084	0.8251	0.8518	0.8146	0.8275
*conv*4*b*	0.6102	0.8406	0.8227	0.8555	0.8633
*conv*3*b*	0.5095	0.7794	0.7504	0.7791	0.7727

The accuracy of identifying coordinates for primitive geometries and trajectories is shown in Table 1's first column. It demonstrates that the direct recognition of primitive geometries and trajectories as a trait is insufficiently accurate. Therefore, to increase accuracy, each trait in a given layer must be concatenated. $fc7$'s accuracy is somewhat worse than $fc6$'s accuracy. It is possible because the genuine C3D can't change $fc7$, which is necessary to create a meaningful VD. Since additional primitive geometries and trajectory coordinates were used, the outcomes of PAHBRNN-based pooling at various 3D *conv* units in JTDGPAHBRD are examined. It is clear that when concatenating the primitive geometries and trajectory coordinates in video patterns according to separate parts of the human body (e.g., right leg, right arm, trunk, left leg, and left arm), the JTDGPAHBRD performs better than the other HAR systems.

Additionally, JTDGPAHBRDs from several *conv* units are combined to determine whether they can balance one another. The outcomes of various configurations applying late merging and the SVM grades on the Penn Action database are shown in Figure 4. It compares the accuracy of JTDPAHBRD framework with the existing frameworks: BPAN [2], SEMN [5], VGG19-SVM [6], ConvLSTM [7], JTDD [11], JTDPABRD [12], and JTDPAHBRD [10].

Figure 4 displays that concatenating *conv*5*b* + *conv*4*b* in the JTDGPAHBRD has a greater recognition rate than other groupings and that the traits are interconnected. Thus, it is concluded that the JTDGPAHBRD framework can accurately recognize human activities in different video sequences compared to the other existing frameworks. For various HAR frameworks on the Penn Action dataset, Table 2 provides the performance outcomes of the obtained coordinates of the primitive geometries and trajectories vs. Ground-Truth (GT) geometries + trajectory coordinates. From these analyses, it is clear that the proposed JTDGPAHBRD framework outperformed the other HAR frameworks applied to the Penn Action database.

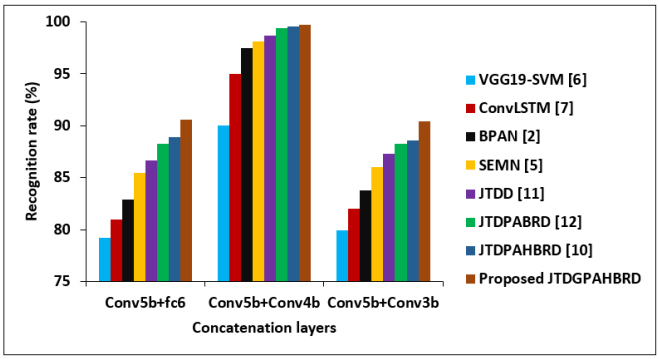

Fig 4. Recognitionrate of JTDGPAHBRD by concatenating different layers for penn action dataset

Table 2. Effect of Obtained Primitive Geometries + Trajectories vs. GT Geometries + Trajectoriesfrom *conv5b* for Various HAR Frameworks on Penn Action Database

Frameworks	GT	Obtained	Variation
VGG19-SVM [6]	0.733	0.671	0.062
ConvLSTM [7]	0.751	0.694	0.057
BPAN [2]	0.784	0.735	0.049
SEMN [5]	0.819	0.777	0.042
JTDD [11]	0.835	0.810	0.025
JTDPABRD [12]	0.847	0.828	0.019
JTDPAHBRD [10]	0.860	0.849	0.011
JTDGPAHBRD	0.893	0.886	0.007

4 Conclusion

The JTDGPAHBRD framework was developed in this study that learned both the coordinates of primitive geometries of body joints and trajectories from multiple video frames for HAR. This framework achieved promising results with a recognition rate of 99.7%. This framework could also be used in video surveillance systems, sports, defense, etc., for recognizing a person's actions precisely. The extraction of different geometries among body joints along with the trajectories enhanced the performance of HAR when using large-scale video sequences. Though it recognizes different human actions, it did not efficiently learn the spatiotemporal relationships among various geometries and a manual extraction of geometries from long-range video sequences was complex. So, future work will focus on introducing a graph-based neural network for automatically learning spatiotemporal relationships among geometric features to create more robust video descriptors.

References

1) Deshpnande A, Warhade KK. An Improved Model for Human Activity Recognition by Integrated feature Approach and Optimized SVM. In: 2021 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE. 2021;p. 571–576. Available from: https://doi.10.1109/ESCI50559.2021.9396914.
2) Weiyao X, Muqing W, Min Z, Ting XZ. Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition. *IEEE Sensors Journal*. 2021;21(17):19157–19164. Available from: https://doi.10.1109/JSEN.2021.3089705.
3) Muhammad K, Mustaqeem, Ullah A, Imran AS, Sajjad M, Kiran MS, et al. Human action recognition using attention based LSTM network with dilated CNN features. *Future Generation Computer Systems*. 2021;125:820–830. Available from: https://doi.org/10.1016/j.future.2021.06.045.
4) Khan S, Khan MA, Alhaisoni M, Tariq U, Yong HSS, Armghan A, et al. Human Action Recognition: A Paradigm of Best Deep Learning Features Selection and Serial Based Extended Fusion. *Sensors*. 2021;21(23):7941–7941. Available from: https://doi.org/10.3390/s21237941.
5) Wang H, Yu B, Xia K, Li J, Zuo X. Skeleton edge motion networks for human action recognition. *Neurocomputing*. 2021;423:1–12. Available from: https://doi.org/10.1016/j.neucom.2020.10.037.
6) Saleem R, Ahmad T, Aslam M, Martinez-Enriquez AM. An Intelligent Human Activity Recognizer for Visually Impaired People Using VGG-SVM Model. In: Advances in Computational Intelligence;vol. 13613. Springer Nature Switzerland. 2022;p. 356–368. Available from: https://doi.org/10.1007/978-3-031-19496-2_28.
7) Yadav SK, Tiwari K, Pandey HM, Akbar SA. Skeleton-based human activity recognition using ConvLSTM and guided feature learning. *Soft Computing*. 2022;26(2):877–890. Available from: https://doi.10.1007/s00500-021-06238-7.

8) Putra PU, Shima K, Shimatani K. A deep neural network model for multi-view human activity recognition. *PLOS ONE*. 2022;17(1):e0262181. Available from: https://doi.org/10.1371/journal.pone.0262181.
9) Li J, Xie Z, Wang Z, Lin Z, Lu C, Zhao Z, et al. A triboelectric gait sensor system for human activity recognition and user identification. *Nano Energy*. 2023;112. Available from: https://doi.org/10.1016/j.nanoen.2023.108473.
10) Nagarathinam S, Narayanan R. Deep Positional Attention-Based Hierarchical Bidirectional RNN with CNN-Based Video Descriptors for Human Action Recognition. *International Journal of Intelligent Engineering & Systems*. 2022;15(3):406–415. Available from: https://doi:10.22266/ijies2022.0630.34.
11) Srilakshmi N, Radha N. Body Joints and Trajectory Guided 3D Deep Convolutional Descriptors for Human Activity Identification. *International Journal of Innovative Technology and Exploring Engineering*. 2019;8(12):1016–1021. Available from: https://doi.10.35940/ijitee.K1985.1081219.
12) Srilakshmi N, Radha N. Deep Positional Attention-based Bidirectional RNN with 3D Convolutional Video Descriptors for Human Action Recognition. *IOP Conference Series: Materials Science and Engineering*. 2021;1022(1):1–10. Available from: https://doi.10.1088/1757-899X/1022/1/012017.