Received: January 24, 2022. Revised: March 18, 2022.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Deep Positional Attention-Based Hierarchical Bidirectional RNN with CNN-Based Video Descriptors for Human Action Recognition

Srilakshmi Nagarathinam[1]* **Radha Narayanan[1]**

[1]*Department of Computer Science, PSGR Krishnammal College for Women, Coimbatore, India*
* Corresponding author's Email: srilakshmiphd123@gmail.com

Abstract: Human Action Recognition (HAR) is a highly notable area of study in contemporary computer vision. Many investigations focused on recognizing a person's actions from video streams based on extracting features regarding orientation and motion. This article presents a Joints and Trajectory-pooled 3D-Deep Positional Attention (PA)-based Hierarchical Bidirectional Recurrent Convolutional Descriptors (JTDPAHBRD) approach which uses a PA-based Hierarchical Bidirectional Recurrent Neural Network (PAHBRNN) for enhancing the feature aggregation process. First, the entire video is segregated into multiple blocks and they are provided to the 2-stream bilinear Convolutional 3D (C3D) model which applies the PAHBRNN as feature aggregation. In PAHBRNN, the feature vectors related to the different parts of a human skeleton in a certain clip are hierarchically aggregated using the position-aware guidance vector. Then, 2 different streams in the C3D network are fused and trained end-to-end using the softmax loss to get the final video descriptor for a particular video sequence. Further, the Support Vector Machine (SVM) classifier is applied to classify the resultant video descriptor to recognize the person's actions. At last, the investigational outcomes demonstrate the JTDPAHBRD achieves 99.6% better recognition accuracy than the classical state-of-the-art approaches.

Keywords: Human action recognition, Deep learning, JTDPABRD, Feature aggregation, Hierarchical BRNN, SVM.

1. Introduction

HAR is a technique used to identify videos that contain a particular task and retrieve relevant videos to distinguish the behavior of a person. It is often used in major domains such as object tracking, the development of human-computer interfaces, and hospital assistance. Thanks to surveillance systems, the internet, Livestream, etc., a vast quantity of videos is recorded every day. In computer vision, HAR is also extremely crucial in modern scenarios [1-3]. Automated detection of specific suspect activities in surveillance systems can also assist in understanding improper or irrelevant actions e.g., automated identification of a loitering person at places like aerodromes, subways, etc.

The motion recognition may allow different functionalities such as the automatic recognition of many gamers' movements. In the healthcare sector, patient rehabilitation can be supported by automated recognition of patients' actions [4-5]. Commonly, HAR is categorized into 3 levels: low, mid, and high. In low-level recognition, identification of edges, extraction of features, and recognition of actions are conducted. In mid-level recognition, identification of human-machine interaction and recognition of abnormal actions are conducted. Similarly, high-level recognitions are useful in different sophisticated applications.

Various findings have been reported in the previous decades to design different forms of HAR systems [6-8]. Alternatively, effective recognition of actions is also quite challenging to different situations, disparities in perception, and so on. In several latest approaches, video is captured under some conditions. However, those ideas have still not been applied in real-world applications.

Besides, a two-stage approach is implemented to learn and identify the features of original video streams by different classification models. The

DOI: 10.22266/ijies2022.0630.34

features that are important in many applications are hardly recognized since feature selection is highly problematic. Specifically, the orientation and trajectories of several scenes in the HAR can be completely distinct [9].

Thus, different deep learning approaches have been applied for training hierarchical characteristics by extracting low- and high-level features [10-13]. Such approaches are guided by either supervised or unsupervised classifiers to ensure an acceptable HAR performance. Unlike other deep learning approaches, Cao et al. [14, 15] designed Joints-pooled 3D-Deep convolutional Descriptors (JDD) to pool the convolutional activations of the 3D-deep Convolutional Neural Network (3DCNN) into the discriminated descriptors depending on the joint coordinates. Originally, a complete video sequence was partitioned into many blocks of fixed dimension and for every block, 3D convolutional attribute maps have been determined. Then, the stable joint coordinates were situated in the 3D attribute maps of a convolutional unit. Also, the activations of every joint coordinate in certain blocks were aggregated and resampled. Further, the mean pooling and l_2-norm were performed to pool these features into the video descriptors which were classified by the linear SVM.

Moreover, this approach was extended by the 2-stream C3D model to simultaneously train the reference from the joints and extract the spatiotemporal characteristics. In C3D, the joint coordinates were extracted using either preprocessing or skeleton extraction [16]. A max-min pooling was performed to pool the body joint-guided feature vector descriptions. Then, the feature and attention streams were multiplied with the bilinear product and given to the Fully Connected (FC) layers to form the resultant video descriptor. But, the time consumption for extracting the joint coordinates was high for complex datasets, and also extracting skeletons was a complicated process.

Thus, Joints and Trajectory-pooled 3D Descriptors (JTDD) have been designed to extract and concatenate the trajectory coordinates or optical flow between any video streams along with the joint coordinates in the C3D approach [17]. Then, the pooled feature descriptors are trained to get the resultant video descriptor which was applied to the SVM for categorizing the human actions. In contrast, the max-min pooling was performed to fuse the features which have more versatility to spatially perfect over the nearby filters. So, the required spatiotemporal disparities among classes were removed.

To tackle this issue, JTDPABRD has been developed which exploits the PA-Bidirectional RNN (PABRNN) model rather than the max-min pooling-based feature aggregation in the two-stream bilinear C3D network [18]. By using PABRNN, the body joint and trajectory point coordinates extracted from two different streams were concatenated to get the final video descriptor for HAR. In contrast, the vanishing gradient problem was occurred due to the use of more parameters. Also, it needs to consider prior input sequences for extracting the long-term spatiotemporal features from long-range video sequences.

Therefore in this paper, a JTDPAHBRD approach is proposed which uses PAHBRNN for enhancing the feature aggregation process. First, the entire video pattern is segregated into multiple blocks and they are provided to the 2-stream C3D model. After extracting the joint and trajectory coordinates at the convolutional layer, the obtained feature vectors are passed to the PAHBRNN to perform feature aggregation rather than max-min pooling. In PAHBRNN, the feature vectors related to the different parts of a human skeleton in a certain clip are hierarchically aggregated using the position-aware guidance vector. Then, 2 different streams in the C3D network are multiplied by the bilinear product and trained end-to-end using the softmax loss to get the final video descriptor for a particular video sequence. Further, the created video descriptor is given to the SVM to recognize the person's actions. Thus, it can extract long-term spatiotemporal features and preserves sequence information over time. Also, it does not tend to vanish with back-propagation through time. As a result, the accuracy of recognizing human actions from video sequences is improved effectively.

The remaining sections of the article are prepared as follows: Section 2 studies the recent HAR systems using deep learning. Section 3 describes the methodology of JTDPAHBRD and Section 4 displays its performance. Section 5 concludes the research work and suggests future enhancements.

2. Literature survey

Spatio-Temporal Distilled Dense-Connectivity Network (STDDCN) [19] was designed to recognize the human actions in the video sequences. In this model, knowledge distillation and dense-connectivity were applied. The main goal of this blockwork was to find interaction policies among shape and movement points with various structures. The spatiotemporal interaction was enabled at the feature representation layer by using the block-level dense interactions among shape and movement pathways. Further, both points were interacted at the high-level layers based on the knowledge distillation between two streams and their final merging. Also, effective hierarchical spatiotemporal features were obtained. But, its accuracy was not effective for more complex datasets.

Cross-covariance [20] has been introduced to create Symmetric Positive Definite (SPD) matrix-based interpretations for recognizing 3D actions. The cross-covariance was created by the correlation data among the interval-elevated appearances to acquire highly informative attributes and interval-order configuration. Besides, a fashion of expression was devised to combine cross-covariance statics and covariance statistics as a greater SPD matrix obtained from the Riemannian geometry. After that, the symmetric cross-covariance was extended into the non-symmetric version in which the interval-order data was included in the associated matrix forms. However, it needs to extract highly complex non-linear correlations between factors, saliency weighting of appearances, spatiotemporal SPD representation, and so on to increase efficiency.

Timed-image-based CNN [21] has been suggested to recognize the actions in video sequences. Initially, intrinsic 3D attribute training was investigated from Hilbert-based meta-image representation of 3D information. After that, 2D+X representation was developed based on the duality among spatial 2D examples and an extra size X which may associate interval, frequency, or intensity. This description was used to acquire a better balance between spatial data and extra data provided by differences in X. But, it needs to combine interval-image characteristics and their respective spatial vs. temporal features.

A multiple-stream deep learning model [22] has been developed for characterizing global and local movement characteristics in a video sequence. At first, global movements were defined effectively using the intensity-based 3-layer movement record images. Then, the local spatiotemporal samples were mined from the skeleton. After that, the results of such levels were combined and the field information was considered for recognizing human actions. But, its efficiency was analyzed only for fixed background scenarios.

A Deep-Wide network (DWnet) [23] has been developed for human action recognition depending on 3D skeleton information. Initially, attributes were mined using the pruned deep model. After, these were synchronized to a large-dimensional attribute space and identified using the superficial configuration. But, its efficiency was less while dealing with more samples.

Correlation Network (CorNet) with a Shannon fusion [24] has been developed to learn a pre-trained CNN. The CorNet was used to capture a spatiotemporal correlation in a block-by-block manner without time correspondence. Also, Shannon fusion was applied to choose features depending on distribution entropy. The final layers of the pre-trained spatial and temporal networks were correlated for creating a 2D correlation tensor. After, this was fed to the FC layers to train the model. Further, predictions were made by combining the output of CorNet with that from the spatial and temporal stream's outcome. But, its performance was not effective when spatial and temporal streams were not balanced.

Integration of a new representation model [25] has been developed in which Multilayer Deep Features (MDF) of a person area and entire image region were integrated into an Extended Region-aware Multiple Kernel Learning (ER-MKL) scheme. First, the off-the-shelf semantic segmentation was applied to utilize the human cues. After, highly effective representations MDF were built via integrating activations at the final convolutional and FC layers. At last, ER-MKL was applied to train a strong classifier for combining person- and entire image-regions MDF. On the other hand, its classification accuracy was not effective and the computational cost was high.

A novel hybrid deep learning blockwork [26] has been designed to recognize human actions depending on the motion tracking and extraction of spatial features in video sequences. In this blockwork, Gaussian Mixture Model (GMM) and Kalman Filter (KF) were used for identifying and mining the traveled people and Gated Recurrent Neural Networks (GRNN) for gathering the attributes in every block which helps to estimate the people activity. But, its accuracy was not effective for complex datasets and the time of video classification was high.

A Correlational Convolutional LSTM (C2LSTM) network [27] has been suggested which perceives motion data, spatial features, and temporal dependencies to recognize human actions. Initially, convolution and correlation functions were leveraged to credit both the spatial and motion data of the video sequence. Then, a deep network was designed by using the suggested units for recognizing human actions. But, it has less accuracy for more complicated datasets.

3. Proposed methodology

This section describes the JTDPAHBRD approach briefly. First, every video pattern is segregated into several clips and they are fed to the 2-stream C3D model. The 2 streams: attention and feature streams use the convolutional layer for mining the joint and trajectory coordinates, respectively along with the spatiotemporal features of different human skeleton parts in each clip. Then, the activations of every joint and trajectory coordinate with their related spatiotemporal features for specified human parts are aggregated from every channel. For this purpose, PAHBRNN is employed rather than the max-min pooling [8]. In PAHBRNN, the feature vectors related to the human skeleton are considered into 5 different categories as Left Arm (LR) and Left Leg (LL), Trunk (TK), Right Arm (RA) and Right Leg (RL). These are initially extracted from the convolutional layer of the C3D model along with the deep features. After the convolution layer, the extracted features are given to the five different PABRNNs as input. To form the motions from the adjacent skeleton features, the interpretation of the trunk feature is aggregated with those of another 4 types of features, accordingly. After, the occurrence of feature positions in all video clips associated with the particular sequence is obtained. The guidance of the extracted feature vectors to every other position is propagated using the position-aware guidance propagation method, which creates the position-aware guidance vector for each feature vector related to the entire human skeleton. Thus, the position-aware guidance vector gives separate feature vectors for different parts of the human skeleton. Further, these separate feature vectors are combined into their corresponding attention weight to get the resultant aggregated feature vectors. Thus, the CNN with PAHBRNN can extract and reduce the feature dimensionality efficiently.

Moreover, a two-stream bilinear C3D network is trained with the help of a position-aware guidance-based feature vector from the entire human skeleton automatically. The two streams are multiplied by the bilinear product. The entire network is trained end-to-end with softmax loss supervised by class labels. As a result, the feature descriptor for a particular video sequence is obtained and classified by the SVM to identify the person's actions. The schematic representation of JTDPAHBRD-based HAR and the 2-stream C3D using PAHBRNN is illustrated in Fig. 1 and 2, respectively.

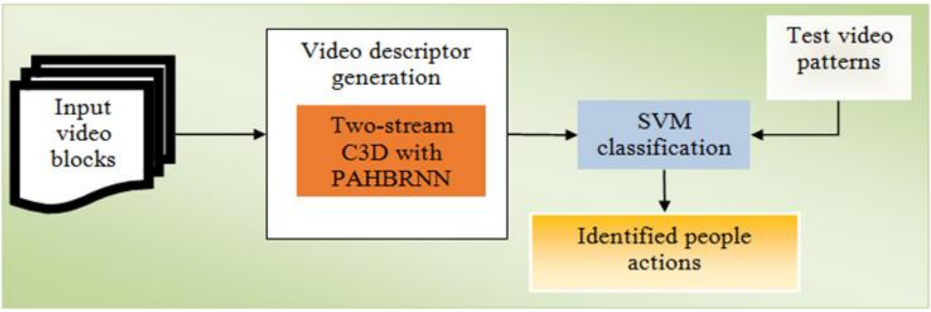

Input video blocks
Video descriptor generation
Two-stream C3D with PAHBRNN
SVM classification
Test video patterns
Identified people actions

Figure. 1 Schematic representation of JTDPAHBRD-based HAR

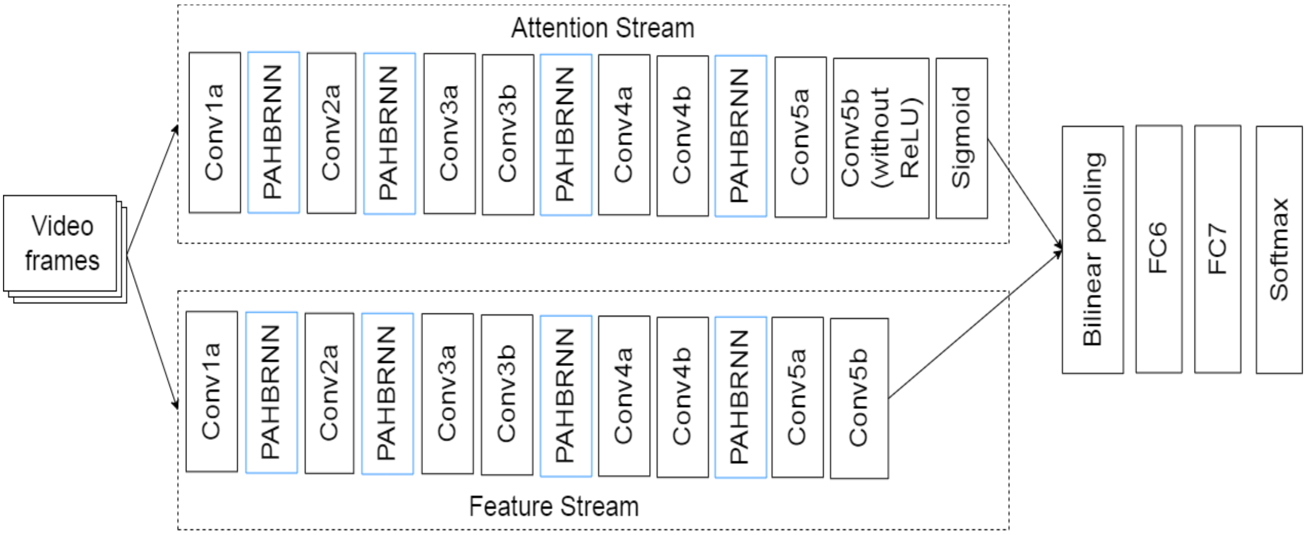


Figure. 2 Architecture of 2-stream bilinear C3D with PAHBRNN-based feature aggregation approach

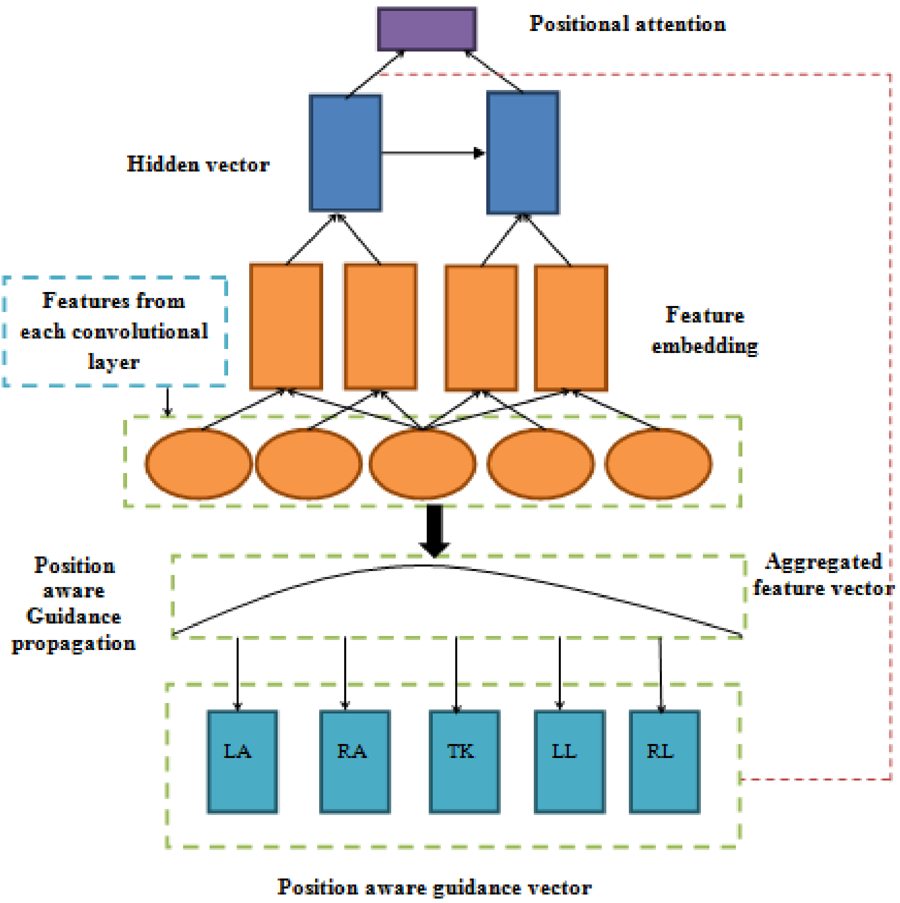

Positional attention
Hidden vector
Features from each convolutional layer
Feature embedding
Position aware Guidance propagation
Aggregated feature vector
LA
RA
TK
LL
RL
Position aware guidance vector

Figure. 3 Aggregated feature vector representation for entire human skeleton using PAHBRNN model

3.1 Positional attention-based hierarchical BRNN

Simple human actions are executed only by a particular segment of them e.g., hitting and kicking forwards are based on tilting the arms and legs, accordingly. Few activities are carried out by shifting the top or bottom body e.g., bowing down is primarily concerned with the top body. Also, highly complicated activities are created by the movements of such 5 segments e.g., jogging and sailing have to cooperate on the movement of the entire body.

To identify different person activities effectively, modeling the motions of such person's segments and their combinations is highly required. For this reason, the PAHBRNN is introduced to mine the long-term contextual data of spatiotemporal patterns. Fig. 3 shows the aggregated feature vector representation of an entire human skeleton using PAHBRNN.

The PAHBRNN adopts the HBRNN with positional attention method to represent the feature vectors by considering the different feature (body joints and trajectory points extracted from different parts of the human skeleton such as LA, RA, TK, LL and RL) embeddings as the input. Consider that the features contain position-aware guidance which is propagated to direct consecutive video clips based on the Gaussian kernel as:

$$Kernel(d) = e^{\left(-d^2/2\sigma^2\right)} \tag{1}$$

Where, d refers to the gap between the actual and aggregated feature vectors and σ represents the variable that limits the propagation scope. Then, the guidance base matrix G associated with the certain distance d and position i is defined as:

$$G(i,d) \sim N(Kernel(d), \sigma') \tag{2}$$

Where N defines the mean density with an estimated $Kernel(d)$ and standard deviation σ'. Moreover, the guidance vector for a feature at a particular position (i.e., LA, RA, TK, LL and RL) is obtained by aggregating the guidance of every feature extracted from the video clips:

$$A_j = Gc_j \tag{3}$$

Where

$$c_j(d) = \sum_{f\in F}[(j-d)\in pos(f)] + [(j+d)\in pos(f)] \quad (4)$$

In Eqs. (3) and (4), A_j is the aggregated guidance vector for the feature at the position j and c_j is the distance count vector which estimates the count of features with different distances. Also, F is the 3D feature maps containing multiple features, f is either a body joint location or a trajectory point feature in F, $pos(f)$ is the group of f's occurrence positions in different clips and $[\cdot]$ is an indicator function which equals to 1 if the criteria satisfy; or else, equals to 0.

Moreover, the position-aware guidance vector for a certain feature is combined into the aggregated feature's attentive weight (α_j) at the position j as:

$$F_a = \sum_{j=1}^{l} \alpha_j h_j \quad (5)$$

Where

$$\alpha_j = \frac{e^{\left(e(h_j,A_j)\right)}}{\sum_{k=1}^{l} e^{\left(e(h_k,A_k)\right)}} \quad (6)$$

$$e(h_j, A_j) = v^T tanh(W_H h_j + W_A A_j + b) \quad (7)$$

In Eqs. (5) and (6), F_a is the final aggregated feature vector for an entire human skeleton in a certain clip, h_j is the hidden vector at position j, A_j is the aggregated position-aware guidance vector obtained by Eq. (3), l is the video sequence length. In Eq. (7), $e(\cdot)$ is the score function which estimates the feature significance based on the hidden vector and the position-aware guidance vector, W_H and W_A are matrices, b is the bias vector, $tanh$ is the hyperbolic tangent function, v is the global vector and v^T is its transpose.

Thus, this PAHBRNN can generate the feature vectors based on the different parts of the entire human skeleton using five different PABRNNs efficiently. Moreover, the 2 streams in C3D are multiplied using the bilinear product and the aggregated feature vectors for all blocks are concatenated to get the final video descriptor [6] by training the C3D network end-to-end using the softmax loss. After obtaining the video descriptors, SVM is applied to learn these video descriptors and recognize human actions.

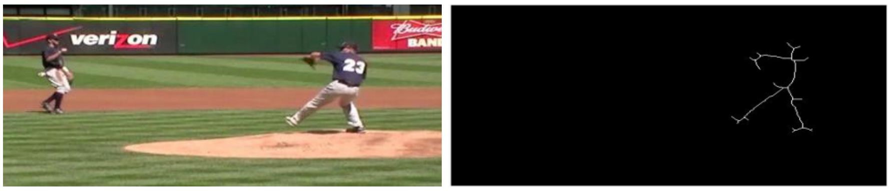
```
Algorithm:
Input: Training video patterns
Output: Human actions
Begin
Split video sequences into blocks;
$for (each\ frame)$
Set CNN variables for attention and feature streams;
Extract the features from different parts of the entire human skeleton such as LA, RA, TK, LL and RL at convolutional layers;
Concatenate the features extracted from each convolutional layer using PAHBRNN;
//PAHBRNN
Create the position-aware guidance propagation through Gaussian filter using Eq. (1);
Compute $G(i,d)$ by Eq. (2);
Aggregate the guidance of (i) RA and LA, (ii) RL and LL with TK features;
Aggregate the guidance of the upper and lower body to get the resultant aggregated position-aware guidance vector using Eqns. (3) (4);
Get the final combined feature vector belonging to a human skeleton in a single clip by calculating $\alpha_j$ and $e(h_j, A_j)$ using Eqns. (5), (6) & (7);
Fuse attention and feature streams in C3D network with the aid of bilinear product;
Train the two-stream C3D network end-to-end using softmax loss for a whole video sequence;
Obtain the final video descriptors;
Apply the SVM classification;
Identify the human actions from a specified video;
$end\ for$
End
```

4. Experimental results

This part analyzes the efficiency of the JTDPAHBRD approach on the Penn Action dataset by implementing it in MATLAB 2017b. This dataset comprises 2326 video sequences of 15 action labels. Each video is collected from different online video repositories and has 50-100 blocks including 13 body joints are annotated for every block. In this experiment, 1861 video sequences are used for training, and the remaining 465 video sequences are used for testing. The joint and trajectory coordinates, as well as C3D features, are considered as sources. So, the recognition accuracy of JTDPAHBRD with these features is analyzed by using different aggregation configurations.

The ratio of human activities which are correctly identified is called accuracy.

$$Accuracy = \frac{No.of\ recognized\ actions}{Total\ no.of\ actions\ tested} \times 100\% \quad (8)$$



(a) (b)

Figure. 4 (a) Sample input video block and (b) Outcomes of joint and trajectory coordinate extraction

Table 1. Recognition accuracy (%) of sources and JTDPAHBRD with different settings on Penn action dataset

	Aggregate all the activations	**JTDPAHBRD Ratio Scaling (1×1×1)**	**JTDPAHBRD Coordinate Mapping (1×1×1)**	**JTDPAHBRD Ratio Scaling (3×3×3)**	**JTDPAHBRD Coordinate Mapping (3×3×3)**
Joint + trajectory coordinates	0.6621	-	-	-	-
$FC7$	0.7758	-	-	-	-
$FC6$	0.7983	-	-	-	-
$conv5b$	0.7605	0.8533	0.9064	0.8542	0.8885
$conv5a$	0.6834	0.7956	0.8257	0.7961	0.8032
$conv4b$	0.5817	0.8134	0.8015	0.8385	0.8471
$conv3b$	0.4826	0.7517	0.7293	0.7554	0.7566

Table 2. Recognition accuracy (%) of aggregating JTDPAHBRDs from different units on penn action dataset

Aggregation Layers	**JDD [14]**	**STDDCN [19]**	**Dwnet [23]**	**CorNet [24]**	**JTDD [17]**	**JTDPABRD [18]**	**JTDPAHBRD**
$conv5b + FC6$	85.5	85.8	86.1	86.3	86.7	88.3	88.9
$conv5b + conv4b$	98.1	98.2	98.4	98.5	98.7	99.4	99.6
$conv5b + conv3b$	86.0	86.2	86.5	86.8	87.3	88.3	88.6

The experimental outcomes of extracting the joints and trajectory coordinates are illustrated in Fig. 4.

The recognition accuracy results of JTDPAHBRD on the Penn Action dataset are given in Table 1.

In Table 1, the 1st column indicates the accuracies of recognizing joint and trajectory coordinates including C3D features. It exhibits the accuracy of recognizing joint and trajectory coordinates directly as a feature is not sufficiently high. So, all the features in a specific layer should aggregate to achieve higher efficiency. The accuracy of $FC7$ is slightly less than the $FC6$. It is achievable since the real C3D cannot adjust $FC7$ that is well suitable to produce an effective video descriptor. For this reason, the results on PAHBRNN-based pooling at different 3D $conv$ units in JTDPAHBRD using more joints and trajectory coordinates are analyzed. The JTDPAHBRD attains greater performance than the JTDPABRD, JTDD, and JDD to concatenate the guided feature vectors of joint and trajectory coordinates in a video pattern based on 5 different segments.

Also, JTDPAHBRDs from different $conv$ units are concatenated to know if they can balance every other. Table 2 presents the outcomes of various mixtures using late fusion with the SVM scores on the Penn Action dataset. It compares the accuracy of JTDPAHBRD approach with the existing approaches such as JDD [14], STDDCN [19], DWnet [23], CorNet [24], JTDD [17] and JTDPABRD [18].

Fig. 5 demonstrates that the accuracy of fusing JTDPAHBRD from $conv5b + conv4b$ is higher than other combinations and it indicates that the features are interrelated. Thus, the accuracy of the JTDPAHBRD approach for identifying the human actions from the video sequences is effectively increased than all other existing approaches.

Similarly, Table 3 presents the outcomes of the effect of extracted joints + trajectory coordinates vs. Ground-Truth (GT) joints + trajectory coordinates for proposed and existing HAR approaches pooled from of $conv5b$ on the Penn Action dataset.

Fig. 6 proves that the JTDPAHBRD attains a very minimum variance between GT joints+ trajectory coordinates and extracted joints + trajectory coordinate. Hence, it achieves a high performance than the JDD, STDDCN, DWnet, CorNet, JTDD, and JTDPABRD approach on the Penn Action dataset.

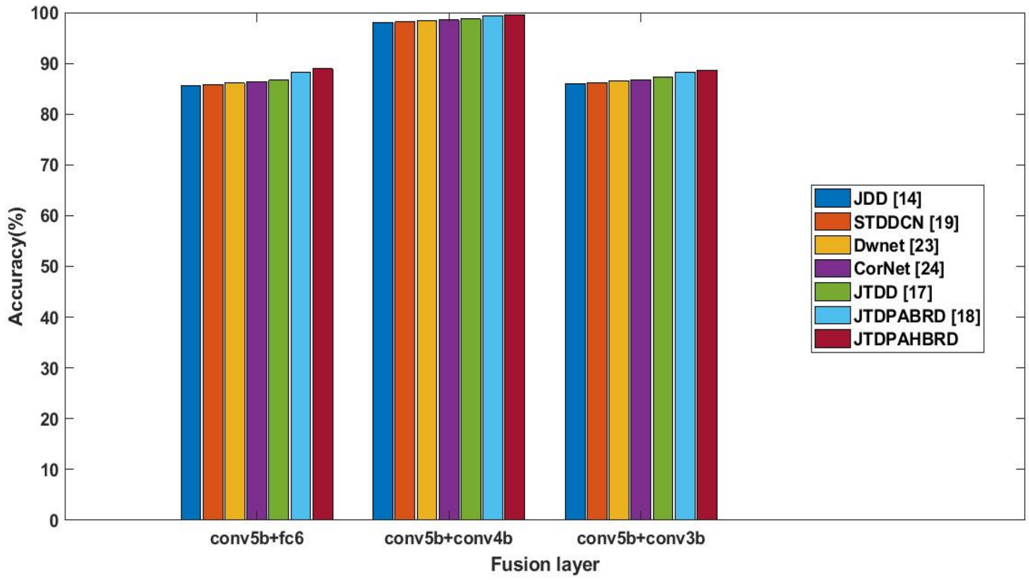


Figure. 5 Accuracy of aggregating JTDPAHBRD from different units on Penn action dataset

Table 3. Effect of extracted joints + trajectories vs. GT joints + trajectories for proposed and existing approaches on Penn action dataset

Approaches Pooled from $conv5b$	**GT**	**Extracted**	**Variance**
JDD [14]	0.819	0.777	0.042
STDDCN [19]	0.823	0.782	0.041
DWnet [23]	0.826	0.786	0.040
CorNet [24]	0.829	0.791	0.038
JTDD [17]	0.835	0.810	0.025
JTDPABRD [18]	0.847	0.828	0.019
JTDPAHBRD	0.860	0.849	0.011

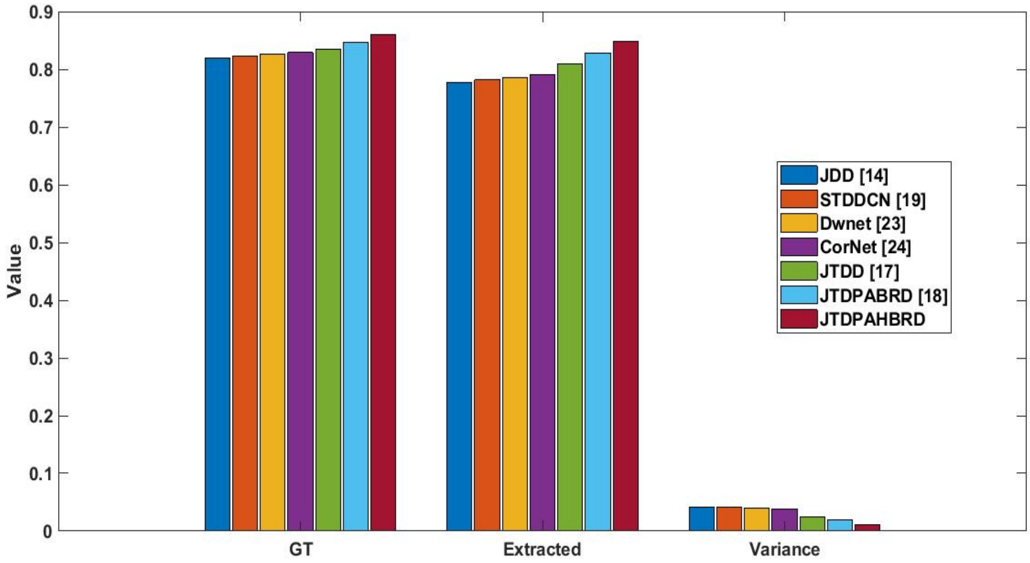


Figure. 6 Influence of identified joints + trajectories vs. GT joints + trajectories for various approaches on Penn action dataset

5. Conclusion

This study proposes the JTDPAHBRD approach in which PAHBRNN is employed to develop the aggregation of features from each video sequence. Initially, each block is fed to the 2-stream C3D model to capture the different features from the different parts of a human skeleton. After that, these characteristics are given to the PAHBRNN which aggregates them hierarchically into a single feature vector. Also, two streams in the C3D network are combined and trained end-to-end using the softmax loss to acquire the resulting video descriptor. Further, the SVM is trained on the obtained video descriptor and used to recognize the person's actions. To conclude, the investigational outcomes proved that JTDPAHBRD on Penn Action dataset has an accuracy of 99.6% when concatenating it from $conv5b$ and $conv4b$, which is 1.07% greater than all other approaches. While concatenating $conv5b + FC6$ layers, the JTDPAHBRD on Penn Action dataset has an accuracy of 88.9%, which is 2.83% greater than all other approaches. Similarly, concatenating $conv5b + conv3b$ layers, the JTDPAHBRD on Penn Action dataset has an accuracy of 88.6%, which is 2.01% greater than all other approaches.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Y. Zhang, W. Qu, and D. Wang, "Action-scene model for human action recognition from videos", *AASRI Procedia*, Vol. 6, pp. 111-117, 2014.

[2] J. Brownlee, "A gentle introduction to a standard human activity recognition problem", *Deep Learning for Time Series*, Vol. 2019, 2019.

[3] C. Jobanputra, J. Bavishi, and N. Doshi, "Human activity recognition: a survey", *Procedia Computer Science*, Vol. 155, pp. 698-703, 2019.

[4] S. Ranasinghe, F. A. Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation", *International Journal of Distributed Sensor Networks*, Vol. 12, No. 8, pp. 1-22, 2016.

[5] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method", *Journal of Healthcare Engineering*, Vol. 2017, pp. 1-31, 2017.

[6] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods", *Frontiers in Robotics and AI*, Vol. 2, pp. 1-28, 2015.

[7] A. L. H. Ps and U. Eranna, "A simplified machine learning approach for recognizing human activity", *International Journal of Electrical & Computer Engineering*, Vol. 9, No. 5, pp. 3465-3473, 2019.

[8] K. P. Reddy, G. A. Naidu, and B. V. Vardhan, "View-invariant feature representation for action recognition under multiple views", *International Journal of Intelligent Engineering and Systems*, Vol. 12, No. 6, pp. 1-13, 2019, doi: 10.22266/ijies2019.1231.01.

[9] J. Basavaiah, C. Patil, and C. Patil, "Robust feature extraction and classification based automated human action recognition system for multiple datasets", *International Journal of Intelligent Engineering and Systems*, Vol. 13, No. 1, pp. 13-24, 2020, doi: 10.22266/ijies2020.0229.02.

[10] H. Kim, S. Lee, and H. Jung, "Human activity recognition by using convolutional neural network", *International Journal of Electrical and Computer Engineering*, Vol. 9, No. 6, pp. 5270-5276, 2019.

[11] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep learning models for real-time human activity recognition with smartphones", *Mobile Networks and Applications*, pp. 1-13, 2019.

[12] I. R. Moreno, J. M. M. Otzeta, B. Sierra, I. Rodriguez, and E. Jauregi, "Video activity recognition state-of-the-art", *Sensors*, Vol. 19, No. 14, pp. 1-25, 2019.

[13] S. Ding, S. Qu, Y. Xi, A. K. Sangaiah, and S. Wan, "Image caption generation with high-level image features", *Pattern Recognition Letters*, Vol. 123, pp. 89-95, 2019.

[14] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Action recognition with joints-pooled 3D deep convolutional descriptors", In: *Proc. of Twenty-Fifth International Joint Conf. on Artificial Intelligence*, pp. 3324-3330, 2016.

[15] C. Cao, Y. Zhang, C. Zhang, and H. Lu, "Body joint guided 3-d deep convolutional descriptors for action recognition", *IEEE Transactions on Cybernetics*, Vol. 48, No. 3, pp. 1095-1108, 2017.

[16] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221-231, 2012.

[17] N. Srilakshmi and N. Radha, "Body joints and trajectory guided 3D deep convolutional descriptors for human activity identification", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8, No. 12, pp. 1016-1021, 2019.

[18] N. Srilakshmi and N. Radha, "Deep positional attention-based bidirectional RNN with 3D

Convolutional video descriptors for human action recognition", *IOP Conf. Series: Materials Science and Engineering*, pp. 1-10, 2021.

[19] W. Hao and Z. Zhang, "Spatiotemporal distilled dense-connectivity network for video action recognition", *Pattern Recognition*, Vol. 92, pp. 13-24, 2019.

[20] J. Zhang, Z. Feng, Y. Su, and M. Xing, "Cross-covariance matrix: time-shifted correlations for 3D action recognition", *Signal Processing*, Vol. 171, pp. 1-13, 2020.

[21] A. M. Atto, A. Benoit, and P. Lambert, "Timed-image based deep learning for action recognition in video sequences", *Pattern Recognition*, Vol. 104, pp. 1-13, 2020.

[22] Y. Gu, X. Ye, W. Sheng, Y. Ou, and Y. Li, "Multiple stream deep learning model for human action recognition", *Image and Vision Computing*, Vol. 93, pp. 1-10, 2020.

[23] Y. Dang, F. Yang, and J. Yin, "DWnet: deep-wide network for 3D action recognition", *Robotics and Autonomous Systems*, Vol. 126, pp. 1-8, 2020.

[24] N. Yudistira and T. Kurita, "Correlation net: spatiotemporal multimodal deep learning for action recognition", *Signal Processing: Image Communication*, Vol. 82, pp. 1-9, 2020.

[25] B. Sheng, J. Li, F. Xiao, and W. Yang, "Multilayer deep features with multiple kernel learning for action recognition", *Neurocomputing*, Vol. 399, pp. 65-74, 2020.

[26] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition", *Journal of King Saud University-Computer and Information Sciences*, Vol. 32, No. 4, pp. 447-453, 2020.

[27] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition", *Neurocomputing*, Vol. 396, pp. 224-229, 2020.