CHAPTER VII

AN ENHANCEMENT OF DEEP POSITIONAL ATTENTION-BASED HUMAN ACTION RECOGNITION BY USING GEOMETRIC POSITIONAL FEATURES

Due to a wide range of events, perspectives, and other elements, developing an effective HAR can be challenging. In recent years, various HAR frameworks have been built employing deep learning algorithms. Deshpnande and Warhade (2021) improved upon a previous model for HAR by fusing the Histogram of Gradient (HOG) local feature descriptor with the Principal Component Analysis (PCA) global features and an improved support vector machine (SVM) classifier. However, it requires a huge number of input parameters and cannot understand the local relationship between image pixels. Based on the Bilinear Pooling and Attention Network (BPAN), Weiyao et al. (2021) created a multi-modal HAR framework. A multimodal fusion network was developed to generate fused attributes from the pre-processed RGB and skeletal data. However, the FC 3-unit perceptron's overall accuracy was affected by the weight value in the loss function, and the training database was small. With the help of a Bidirectional Long Short-Term Memory (BiLSTM), a widened Convolutional Neural Network (CNN) was constructed by (Muhammad et al. 2021). This network is able to choose important features from the input frame in order to differentiate between various human activities. The loss function in video-based HAR was also minimized using the center loss with softmax. However, it only used a single learning stream, which was insufficient for recognizing complicated actions in videos and learning from video frames in large-scale datasets.

The deep learning model created by (Khan et al. 2021) includes the processes of feature mapping, feature fusion, and feature selection. DenseNet201 and InceptionV3 were employed to perform the feature mapping. The serialized augmented model then extracted deep characteristics and combined them. The best features were selected using a kurtosis-aware weighted K-Nearest Neighbor (KNN) algorithm. Finally, many supervised learning systems classified those features. However, it requires a lot of processing power to generate the initial deep feature extraction. Skeleton Edge Motion Networks (SEMN) is a unique HAR approach developed by (Wang et al. 2021) for extracting gesture information. To provide a comprehensive understanding of skeletal structures, the SEMN was developed by fusing many spatiotemporal segments together. It proved challenging to distinguish between individual activities and granular skeletal images despite the use of a novel advanced rank error to maintain sequential imperative data.

To categorize human movements, (Saleem et al. 2022) employed an SVM classifier and used pre-trained VGG-19 to extract body joints from a 2D body skeleton. However, it was less accurate since it could not learn the temporal and spatial correlations between individual pixels. Regarding skeletonized For HAR, Yadav et al. (2022) built a network consisting of Convolutional Long Short-Term Memory nodes. Skeleton coordinates were estimated using human identification and pose estimation, and these were then utilized in conjunction with geometric and kinematic features to generate reference traits. Although a categorizer head was used, it did not take into account edges and surface-related geometric features, which could have improved HAR performance. A Deep Neural Network (DNN) for human action classification was developed using transfer learning and shared-weight methods (Putra et al., 2022). This model consisted of pre-trained CNNs, attention layers, LSTMs trained with residual learning, and softmax layers. However, it did not meet the requirements of the online examples that were investigated, which required the classification of confusing action sequences.

A triboelectric gait sensor system was designed by (Li et al. 2023) for HAR. To improve HAR functionality, they extracted deep features from multichannel time-series gait data using long short-term memory (LSTM) and residual units. However, improved HAR functionality requires additional geometric details. To acquire additional discriminative characteristics from the video sequences and reliably recognize complicated behaviors, a two-stream learning technique has arisen in recent years, while the above-studied frameworks only used a single-stream learning strategy. (Nagarathinam et al. 2022) took this into account, and so they developed the Joints and Trajectory-pooled 3D-Deep Positional Attention-based Hierarchical Bidirectional Recurrent convolutional Descriptors (JTDPAHBRD) framework to improve the attribute concatenation task via the use of a Positional Attention-based Hierarchical Bidirectional Recurrent Neural Network (PAHBRNN). This PAHBRNN-based pooling separated the human skeleton-related attribute vectors across all clips into many subsets. In order to capture and combine the long-term spatio-temporal characteristics hierarchically, these pieces were fed to numerous PABRNNs. Additionally, the FCL was used to supply the absolute Video Descriptor (VD) that was used in the SVM's classification of HAR.

The geometric correlation between joints is not taken into account during feature extraction and concatenation. Instead, these frameworks simply fuse the joint and trajectory coordinates at each interval. Joint development in a fossil skeleton is a common occurrence. So, the relative joint geometries provide a useful description of actions. Joint trajectories only provide gesture information, not contour or geometrical information.

Hence, the purpose of this research is to consider the relative geometries in the human body to improve HAR. In order to extract geometric features like joints, edges, and surfaces from the skeleton graph in addition to the coordinates of the points along the trajectory, a JTDGPAHBRD-based HAR framework is suggested. 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. To better learn discriminative high-level features, the PAHBRNN is implemented in a brand-new 3D-deep convolutional network with VC and TD layers. Next, to apply the FCL to a frame in order to obtain its VD. In addition, the SVM classifier is applied to the generated VD in order to identify distinct types of human behavior. Thus, this framework can increase the recognition rate of HAR systems.

7.1 PROPOSED METHODOLOGY

The JTDGPAHBRD framework for HAR is briefly described here. As shown in Fig. 7.1, the JTDGPAHBRD framework for HAR can be conceptualized as a generalized sketch. The primary objective is to provide a predicted activity label to a previously unseen video sequence. At first, a video is cut into individual frames. With the aid of a skeleton graph structure, elementary geometries like as joints, edges, and surfaces are defined for each frame. Each joint's position along the trajectory is also retrieved. Next, instead of using a max-min pooling technique, the PAHBRNN-based 2-stream C3D network is fed the geometry and trajectory coordinates to begin the pooling process. After that, the bilinear product combines the features from both streams (feature and attention, for example) and the FCL yields an absolute VD. In order to predict the action labels of test video sequences, the acquired VD is then trained by the SVM classifier.

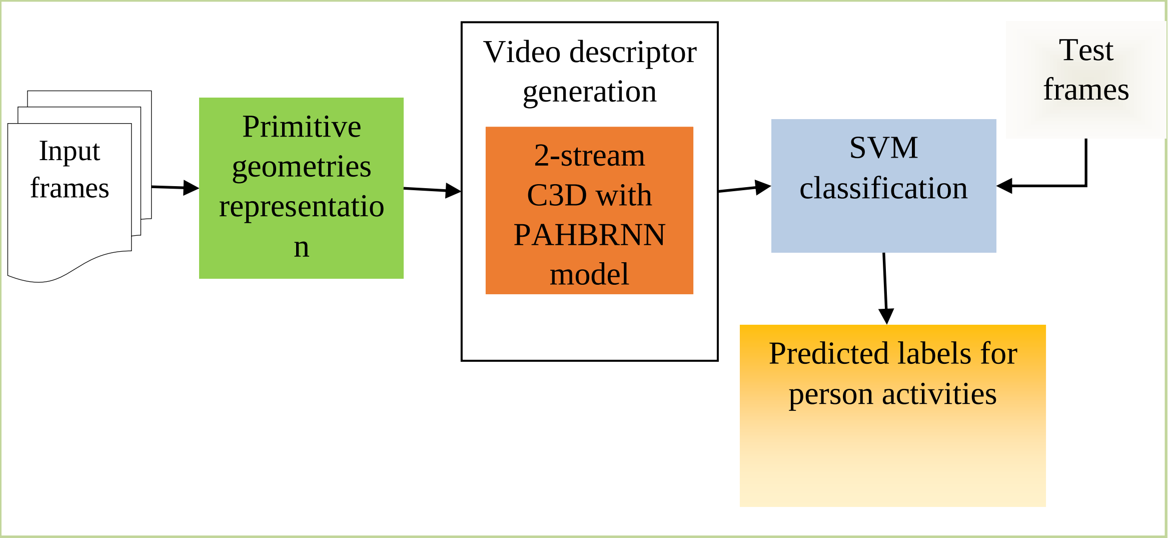
Input frames
Primitive geometries representatio n
Video descriptor generation
2-stream C3D with PAHBRNN model
Test frames
SVM classification
Predicted labels for person activities

Figure 7.1. Schematic representation of JTDGPAHBRD-based HAR.

7.1.1 Representation of Primitive Geometries from Skeleton Information

The skeleton data is a list of coordinates in three dimensions that represent the deformed shape of the body at certain spots. The body can be manipulated to create a variety of motions at the spots. Connecting these places follows the same structure as the joints in the human body. When the human body is mapped onto a graph, the bones become nodes and the joints become edges. The skeletal data for a given individual is limited by two geometric constraints: (1) The distance between any two sites along a connected bone fragment is always the same because of the bone's constant size, and (2) Any three points that generate two overlapping pieces are on the same 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.

A. Joints

Using M as the number of joints in the body, the $M \times 3$ matrix that results from the coordinates of the points along the interval is obtained. A tensor X of dimensions

$T \times M \times 3$ represents the skeletal data when the length of the video stream is T. Time-dependent shifts in joint coordinates reflect activity patterns across time. The rotation matrix maps one set of joint coordinates to the other set of coordinates in the scene. The define p_k as the coordinate vector of the joint at some instant in time, then obtain the subsequent coordinate vector by

$$\tilde{p}_k = Rp_k \tag{7.1}$$

The 3×3 revolution matrix is denoted by R in Eq. (7.1). Assume that in a given video, R holds true across a wide range of joints and distances. So, the unique $\tilde{X}$ identified from the other perspective for the joints tensor X is defined as

$$\tilde{X} = X \times_3 R^T \tag{7.2}$$

Multiplication by the 3-mode tensor $\times_3$ has equal magnitudes for $\tilde{X}$ and X in Eq. (7.2).

B. Edges

Bone movement, in addition to the temporal characteristics of joints, shapes a variety of activities. Joints' physical connections are specified using a graph. Nodes indicate joints and edges represent bones in this skeletal system.

There are $M - 1$ edges in a network with M nodes. The bone's orientation may be seen around the edge. Each node is represented by a vector of coordinates, and edges are named by subtracting the vectors of their respective starting and ending points. To clarify,

$$e_k = p_i - p_j \tag{7.3}$$

The edge, the endpoint, and the starting point are defined by the vectors e_k, p_i, p_j in Eq. (7.3). An edge of the skeleton is described by a tensor Y of size $T \times (M - 1) \times 3$. The edge vector whose terminal point is at the node serves as a symbol for the node itself. A zero-length edge-free vector represents a node that does not contain edge-ends. By doing so, Y is made one size larger, and X and Y are both made to be the same size.

A revolution matrix is used to transform the view's edge coordinate vectors into the other view's coordinate system. With the help of Eqns. (7.1) and (7.3), the transformation may be accomplished for the edge's vector of coordinates at a certain time interval.

$$\tilde{e}_k = \tilde{p}_i - \tilde{p}_j = Re_k \tag{7.4}$$

$\tilde{e}_k$ is the transformed vector of e_k in Eq. (7.4). Also, it is defined as:

$$\tilde{Y} = Y \times_3 R^T \tag{7.5}$$

The tensor of edges after conversion is denoted by $\tilde{Y}$ in Eq. (7.5). The joint and edge revolution matrices are found to be similar when comparing Eqs. (7.2) and (7.5).

C. Surfaces

The edges show how the joints are correlated with one another in pairs. It can't depict the case where two joints with adjacent edges are physically close together. The motions of neighboring bones are similarly advantageous to HAR. The standard vector is used to represent the surface since a plane is formed by two adjacent edges. To illustrate, let's say e_i, e_j are the vectors representing two adjacent edges, and s_k is the normative vector.

$$s_k = e_i \times e_j \tag{7.6}$$

The symbol × denotes the 3D cross product in Eq. (7.6). The magnitude of the vector represents the obliquity angle between the two edges, hence it is not normalized. The standard vector is multiplied by 100 to keep its size consistent with that of the coordinate vector. A body with M joints has $(M + 2)$ possible planes. Two surfaces with redundant information (the standard vector is determined by another standard vector) were removed so that a correct relationship could be established between joints and edges. This yields M surfaces, allowing a tensor Z of sizes $T \times M \times 3$ to determine a sequence's standard vectors.

The other perspectives are used to infer the standard vector of the first. The novel standard vector of a plane at a certain interval is defined by Eq. (7.6) and Eq. (7.4).

$$\tilde{s}_k = (Re_i) \times (Re_j) = Co(R)s_k \tag{7.7}$$

$Co(R)$ stands for the cofactor matrix of R, which is the adjoint matrix's transpose in Eq. (7.7). For a matrix R that can be inverted, then;

$$Co(R) = \big(det(R)\big)(R^{-1})^T \tag{7.8}$$

Transpose of the inverse R is denoted by T in Eq. (7.8) as $(R^{-1})^T$. A revolution matrix has a determinant of one and its transpose is R^{-1}. Hence, it may derive Eq. (7.8) as

$$Co(R) = (R^{-1})^T = R \tag{7.9}$$

This means that in the tensor interpretation of planes:

$$\tilde{Z} = Z \times_3 R^T \tag{7.10}$$

$\tilde{Z}$ in Eq. (7.10) represents the tensor of surface standard vectors after conversion. Joints, edges, and planes are all found to have the same revolution matrix by relating Equations (7.2), (7.5), and (7.10).

7.1.2 Recognition of Human Activities

The 2-stream C3D network is supplied the three types of skeleton data necessary for HAR: joint, edge, and surface coordinates, as well as the trajectory coordinates. Fig. 7.2 depicts the entire network architecture of the HAR. The VC and TD layers are included in this architecture to better facilitate attribute mining and VD creation.

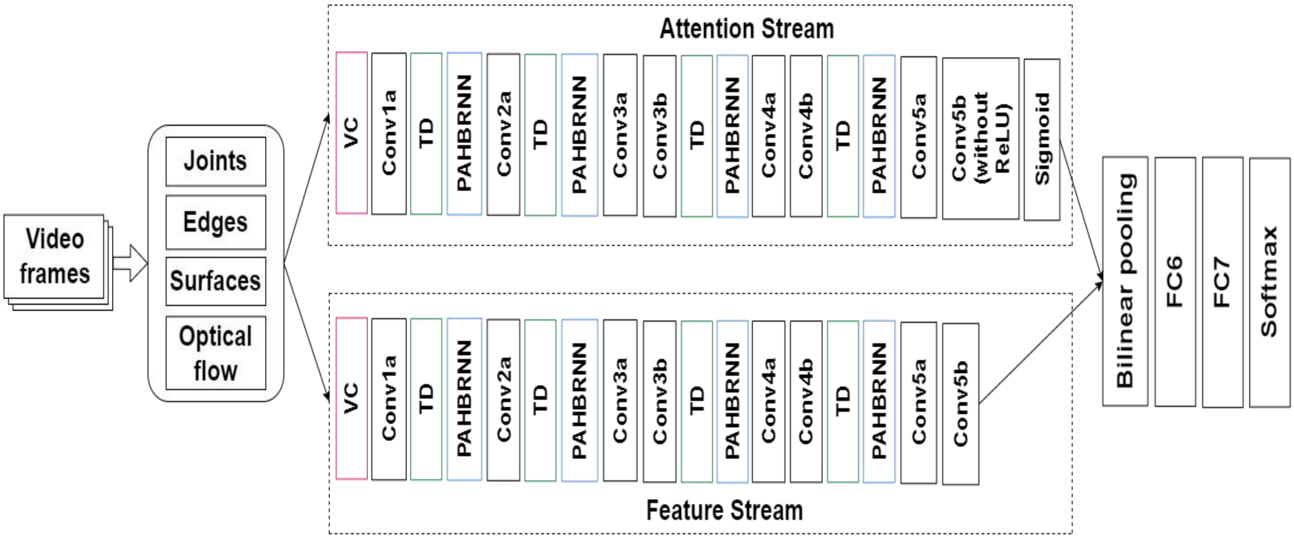

Figure 7.2. Structure of proposed 2-stream bilinear C3D network for HAR.

A. **View conversion:** In real life, human skeletons can be photographed from a random camera perspective. This framework plans on using the VC layer to translate the skeleton information into 3D space by collecting the joints, edges, and planes in order to generate view-invariant interpretations.

The video's X, Y, Z coordinates are transformed using a comparable transformation matrix, R. R is a combination of revolutions about the x, y, z axes, as described by Euler's rotation theory.

$$R = R_x(\alpha)R_y(\beta)R_z(\gamma) \tag{7.11}$$

The x, y, z rotation angles are denoted by α, β, γ in Eq. (7.11). R is computed from three independent orientation variables given a skeletal structure by developing a small number of important hypotheses. In the learning task, R is calculated to transform the inputs once the orientations in a given value have been arbitrarily selected. Given that the surface is almost perpendicular to the z-axis, the coordinates and α, β are drawn from $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and the γ is set to 0. 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. The spatial dropout is the inspiration for this technique used to investigate the 4D tensor convolution feature. In this study, it is altered for 3D tensor and applied for attribute training from frames. As illustrated in Fig. 2, the TD is conducted before the PAHBRNN.

Consequently, the 2-stream C3D network is educated to generate the absolute VD of a specified sequence. In order to categorize the behaviors of persons inside a video sequence, the resulting VD is put into a support vector machine algorithm (SVM).

7.2 Results and Discussion

The JTDGPAHBRD framework is evaluated based on its performance in MATLAB 2017b. This analysis makes use of the 2326 video sequences from the Penn Action Corpus, each of which is tagged with 15 different types of activities. Each clip contains between 50 and 100 blocks, and all of the annotated body joints total 13. This dataset includes 1861 video sequences used for training and 465 video sequences used for evaluation. Some examples of data sources are C3D features, primitive geometry coordinates, and trajectory data. Several aggregation configurations are utilized to evaluate JTDGPAHBRD's recognition accuracy in light of these features.

Recognition accuracy refers to the degree to which a person's behaviors are accurately interpreted. It can be calculated with Eq. (7.12).

$$Accuracy = \frac{Number\ of\ recognized\ actions}{Total\ number\ of\ actions\ tested} \times 100\% \tag{7.12}$$

Input video frame and skeleton picture used to represent primitive geometry coordinates are shown in Fig. 7.3.

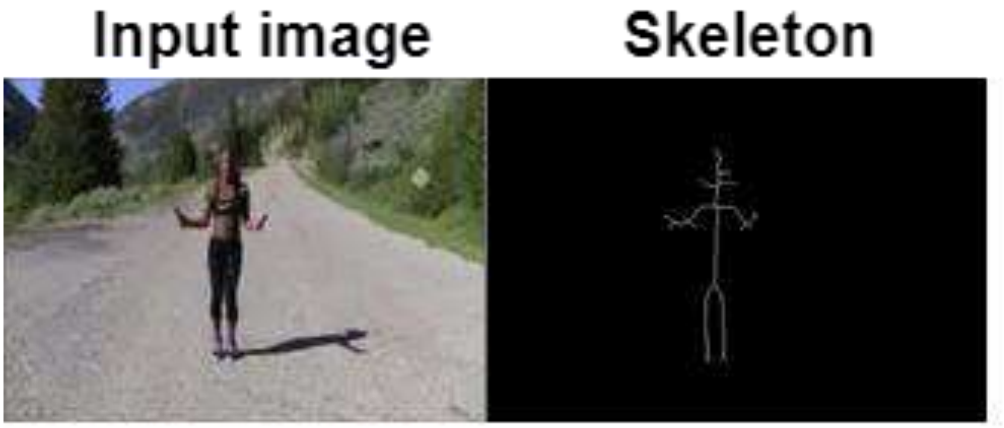

Figure 7.3. Input image and its corresponding skeleton image for primitive geometry coordinates representation.

The Penn Action dataset's JTDGPAHBRD identification rates are shown in Table 7.1.

Table 7.1. Recognition Rate (%) of Sources and JTDGPAHBRD with Distinct Settings on 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	-	-	-	-
$conv5b$	0.7931	0.8763	0.9231	0.8718	0.9042
$conv5a$	0.7084	0.8251	0.8518	0.8146	0.8275
$conv4b$	0.6102	0.8406	0.8227	0.8555	0.8633
$conv3b$	0.5095	0.7794	0.7504	0.7791	0.7727

The first column of Table 7.1 displays the accuracy with which coordinates for primitive geometries and trajectories can be identified. This research proves that it is not possible to reliably identify primitive geometries and trajectories by direct recognition alone. Therefore, to achieve higher accuracy, it is essential to concatenate all attributes inside a specific layer. The accuracy of $fc7$ is somewhat lower than that of $fc6$. The real C3D can't alter $fc7$, therefore this is a viable option for making a functional VD. Due to the incorporation of more primitive geometries and trajectory data, the outcomes of PAHBRNN-based pooling in JTDGPAHBRD are studied at a variety of 3D $conv$ units. It is evident that the JTDGPAHBRD outperforms the other HAR systems when video patterns are combined with primitive geometries and trajectory coordinates according to different sections of the human body (such as the right leg, right arm, trunk, left leg, and left arm).

Furthermore, JTDGPAHBRDs from many $conv$ units are integrated to ascertain their compatibility with one another in terms of maintaining equilibrium. Figure 7.4

displays the results of implementing late merging in a variety of settings and the SVM scores on the Penn Action database. Existing frameworks like BPAN, SEMN, VGG19-SVM, ConvLSTM, JTDD, JTDPABRD and JTDPAHBRD are compared to JTDPAHBRD in terms of accuracy.

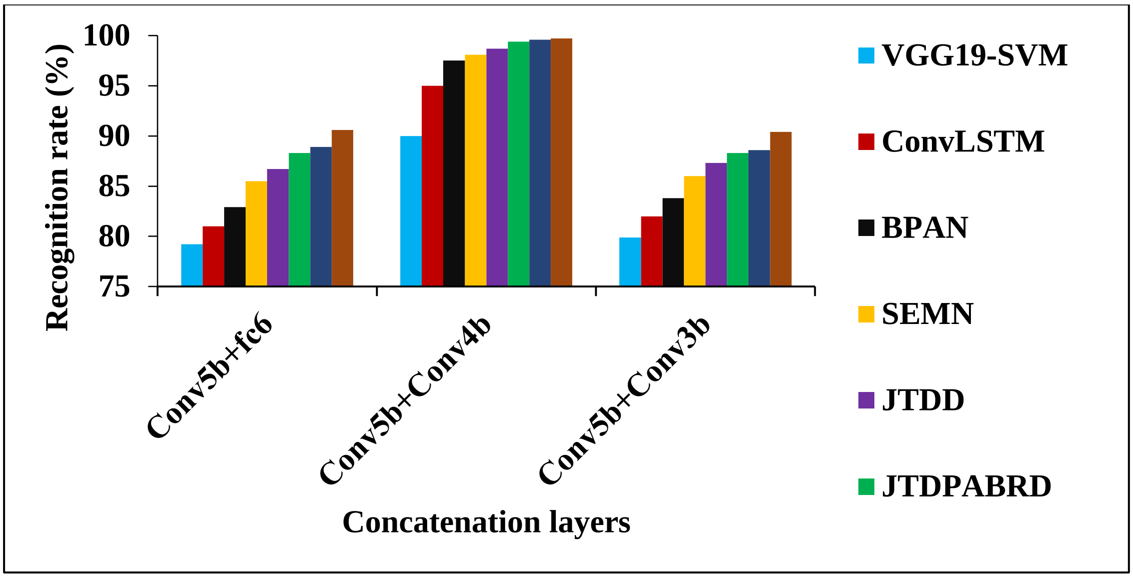

Figure 7.4. Recognition rate of JTDGPAHBRD by concatenating different layers for penn action dataset.

Concatenating $conv5b + conv4b$ in the JTDGPAHBRD has a higher recognition rate than other groupings, as seen in Fig. 7.4. This shows that the qualities are related. Therefore, it is argued that the JTDGPAHBRD framework is superior to the other current frameworks in its ability to reliably distinguish human activities across a variety of video sequences.

Precision, recall, and f-measure for fusing multiple layers on the Penn action dataset are shown in Table 7.2.

Table 7.2. Precision, Recall, and F-measure of Fusing Multiple Layers Together on Penn Action Dataset

Performance Metrics	Fusion Layers					
	$conv5b + fc6$		$conv5b + conv4b$		$conv5b + conv3b$	
	JTD-PAHBRD	JTD-GPAHBRD	JTD-PAHBRD	JTD-GPAHBRD	JTD-PAHBRD	JTD-GPAHBRD
Precision	0.881	0.885	0.989	0.992	0.879	0.883
Recall	0.886	0.890	0.994	0.996	0.885	0.890
F-measure	0.884	0.888	0.992	0.994	0.882	0.887

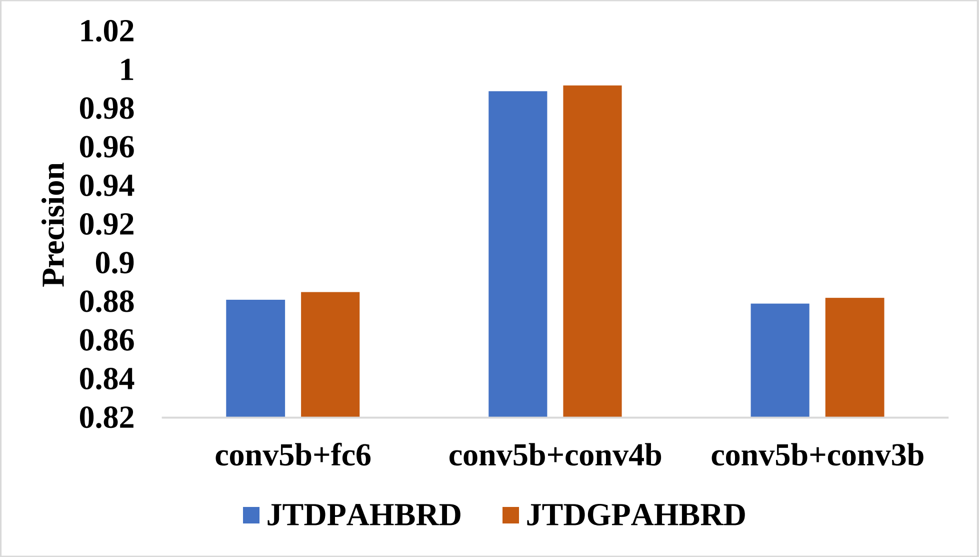

Figure 7.5. Recognition rate of Precision of JTDGPAHBRD by concatenating different layers for penn action dataset.

Concatenating $conv5b + conv4b$ in the JTDGPAHBRD has a higher recognition rate of precision, as seen in Fig. 7.5. This shows that the qualities are related. Therefore, it is argued that the JTDGPAHBRD framework is superior to the other current frameworks in its ability to reliably distinguish human activities across a variety of video sequences.

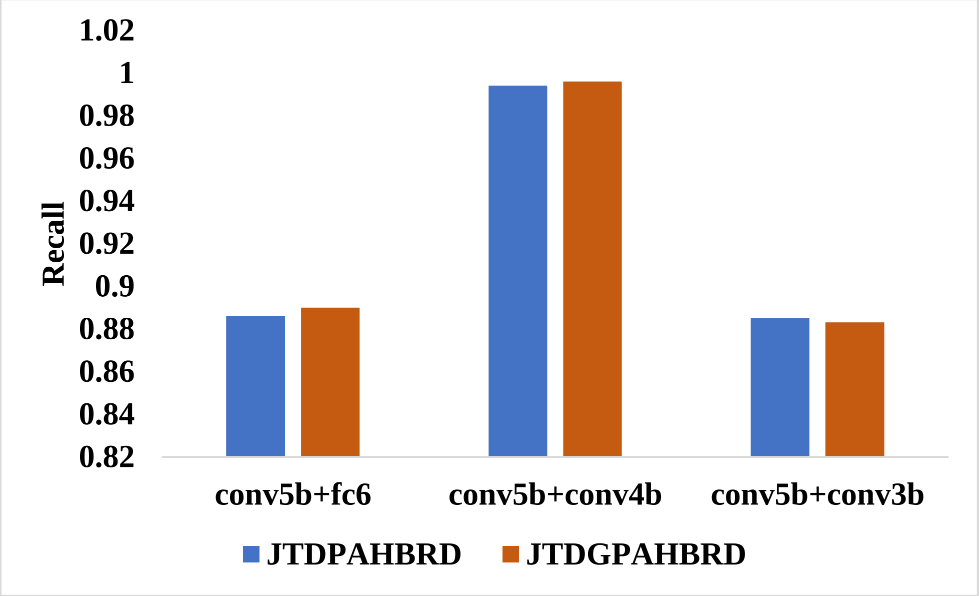

Figure 7.6. Recognition rate of Recall of JTDGPAHBRD by concatenating different layers for penn action dataset.

Concatenating $conv5b + conv4b$ in the JTDGPAHBRD has a higher recognition rate of recall, as seen in Fig. 7.6. This shows that the qualities are related. Therefore, it is argued that the JTDGPAHBRD framework is superior to the other current frameworks in its ability to reliably distinguish human activities across a variety of video sequences.

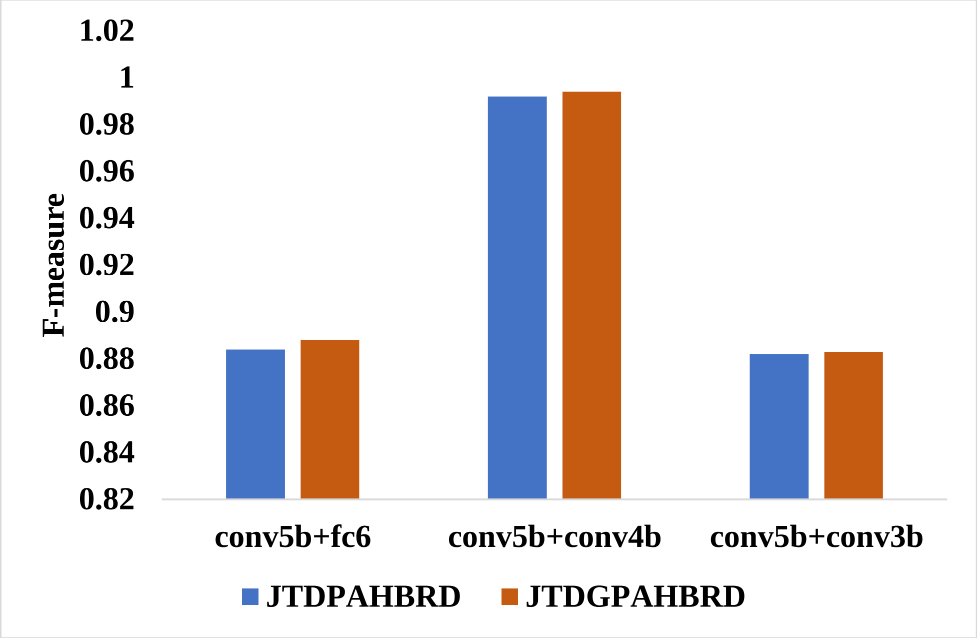

Figure 7.7. Recognition rate of F-measure of JTDGPAHBRD by concatenating different layers for penn action dataset.

Concatenating $conv5b + conv4b$ in the JTDGPAHBRD has a higher recognition rate of F-measure, as seen in Fig. 7.7. This shows that the qualities are related. Therefore, it is argued that the JTDGPAHBRD framework is superior to the other current frameworks in its ability to reliably distinguish human activities across a variety of video sequences.

Table 7.3 compares the computed coordinates of the primitive geometries and trajectories to the Ground-Truth (GT) geometries + trajectory coordinates for a number of HAR frameworks on the Penn Action dataset. The suggested JTDGPAHBRD framework outperforms previously tested HAR frameworks, according to an analysis of the Penn Action database.

Table 7.3. Effect of Obtained Primitive Geometries + Trajectories vs. GT Geometries + Trajectories from $conv5b$ for Various HAR Frameworks on Penn Action Database

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

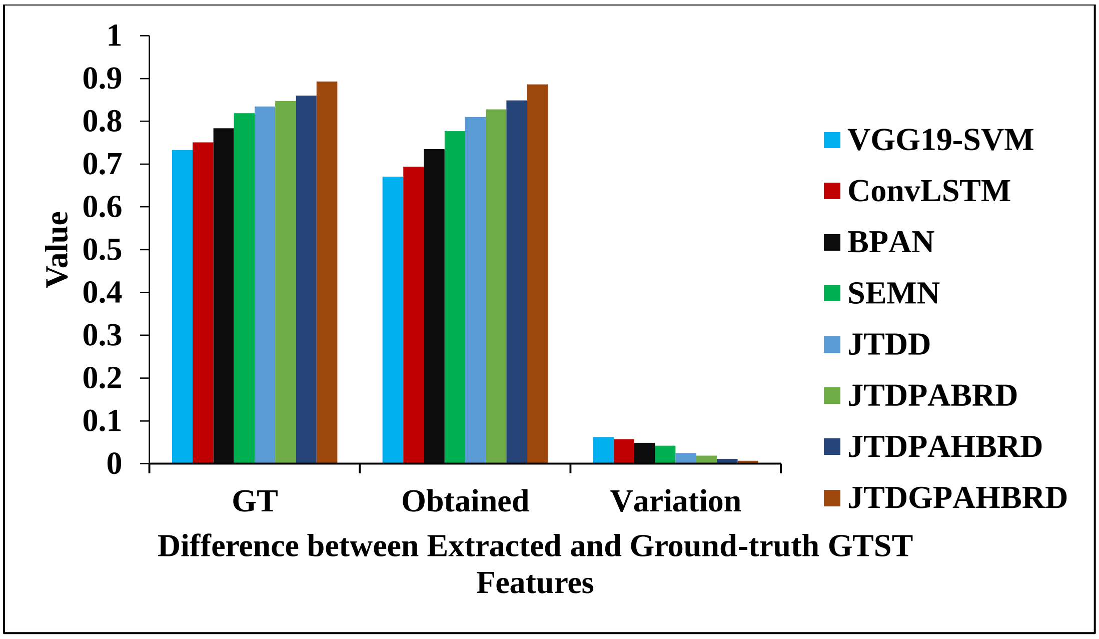

Figure 7.8 Effect of Extracted GTST vs. Ground-truth GTST for Different HAR Models on PAD

As can be shown in Figure 7.8, the JTDGPAHBRD minimizes the gap between the extracted GTST and the true GTST. After conducting these tests, the PAD found that the proposed JTDGPAHBRD model had the best recognition performance of all of the HAR models it had tried.

7.3 CHAPTER SUMMARY

In this research, the JTDGPAHBRD framework was created to learn HAR trajectories and body joint coordinates from many video frames. An impressive recognition rate of 99.7% was attained using this approach. This framework has potential applications in video surveillance systems, sports, military, etc., where accurate action recognition is required. Improved HAR functionality in large-scale video sequences was achieved through the extraction of joint geometry and trajectories. Despite its ability to identify a wide range of human activities, it has been unable to effectively learn the spatiotemporal correlations between distinct geometries, and manual extraction of geometries from long-range video sequences remains a challenging task. In order to develop more reliable video descriptors, researchers plan to implement a graph-based neural network for automatically learning spatiotemporal correlations among geometric data.