CHAPTER III

RESEARCH METHODOLOGY

Action detection in video is a popular topic of computer vision research. The incorporation of both visual and motion data has been the subject of extensive research for the purpose of action detection in videos. More accurate and compact than point trajectories, joint trajectories can be used to depict human motion. Human behavior encompasses a variety of bodily positions and engagements with the surrounding milieu. The configuration of the body's articulations determines a particular bodily arrangement, often referred to as a position, while the temporal evolution of these articulations characterizes the concept of motion. The majority of current skeleton-based action detection systems include manually-engineered characteristics to replicate the visual appearance and temporal behavior of human joints. The characteristics included in this study are the relative locations of body joints and the positions of angles formed between arms and legs, and the angles formed between limbs and the planes defined by body components.

From these facts, this study aims to deconstruct the issues plaguing existing CNN-based HAR models and present new approaches to generate HAR from massive video sequences. First, they offer a model for creating video descriptors that takes into consideration joint and trajectory points. While the third stage enhances feature aggregation in the neural network, the second stage proposes a model for extracting crucial spatiotemporal connections between various action types. In the last phase, a model is proposed to improve video descriptor production by extracting geometric information alongside the joints and trajectory points. By improving the learning of spatiotemporal features across different geometries, the fifth stage also works on a model to bolster the video descriptor for HAR. Finally, the models' performance in recognizing human actions is evaluated and compared to that of other models using the Penn Action dataset.

3.1 PROPOSED FRAMEWORK

An effective HAR model proposed by this research can involve five phases, as shown in Figure 3.1.

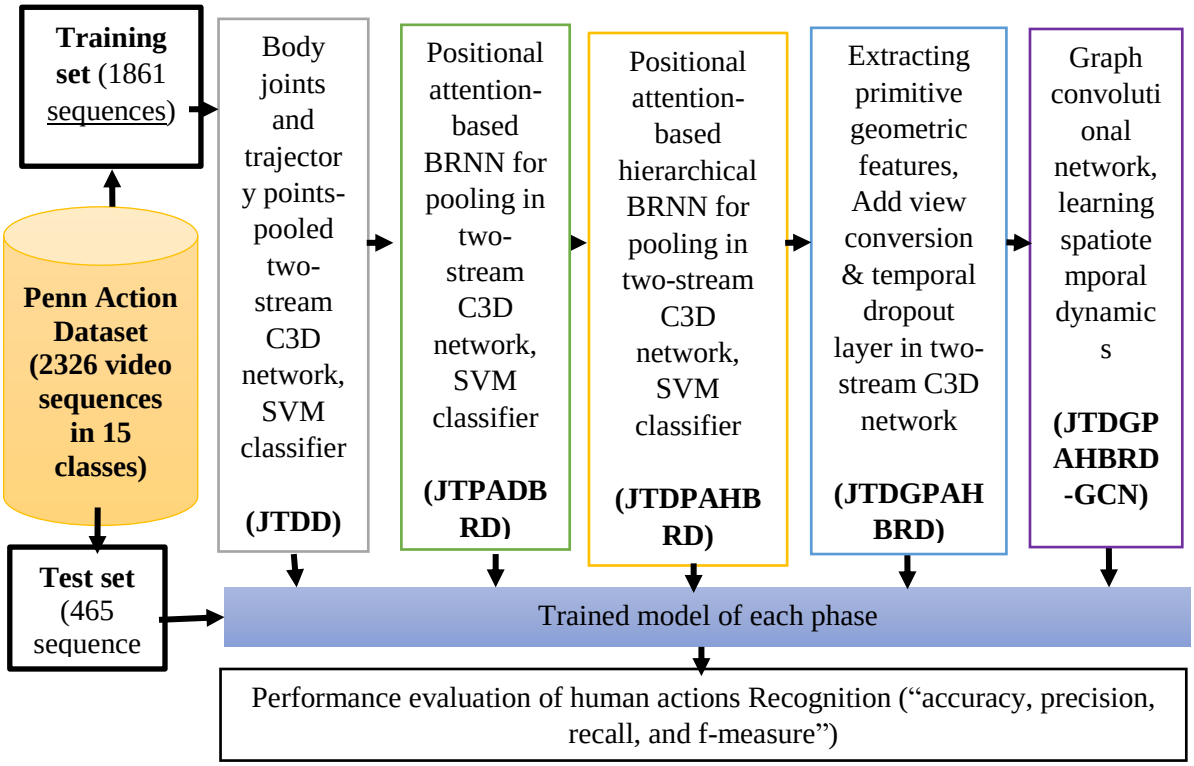

Figure 3.1. Structure of the Proposed Research

3.1.1 Body Joints and Trajectory Guided 3D Deep Convolutional Descriptors for Human Activity Identification

In order to enhance recognition precision, they propose a Joints and Trajectory-pooled 3D-Deep Convolutional Descriptor (JTDD) that blends an optical flow with a two-stream bilinear model to improve recognition accuracy. In this paradigm, the JTDD receives not only the body's joint locations but also an optical flow, or a set of trajectory points between video frames. Final video descriptors are calculated using a two-stream C3D multiplied by the bilinear product once the entire network has been trained. Linear support vector machines are used to classify these video features in order to identify human actions.

3.1.2 Deep Positional Attention-based Bidirectional RNN with 3D Convolutional Video Descriptors for Human Action Recognition

In the study's second half, Using a two-stream Convolutional 3D network and the PABRNN model to extract critical spatiotemporal information, a method for activity recognition is proposed in the second half of the study. Joints and Trajectory-

pooled 3D-Deep Positional Attention-based Bidirectional Convolutional Descriptor (JTPADBRD) is the name of this technique. In order to feed the video into the C3D program, it must first be cut up into smaller pieces. The PABRNN is then used to construct an aggregate clip descriptor from the convolutional feature vector representations of each clip in the video. Bilinear multiplication of these two streams is used to train the entire pipeline with class labels. The final visual description is likewise a sum of the changes in activation across spatial and temporal layers that are all fully interconnected. This video description is given into a support vector machine trained to detect events within videos.

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

Improve feature aggregation with the Joints and Trajectory-pooled 3D-Deep Positional Attention (PA)-based Hierarchical Recurrent Convolutional Descriptors (JTDPAHBRD), which was developed in the third stage of this research. Before feeding the C3D network, which uses the PAHBRNN to aggregate features, the full video is split up into smaller pieces. PAHBRNN builds a hierarchical representation of the human skeleton using the feature vectors corresponding to its various parts using the position-aware guiding vector. When training a C3D network end-to-end with the softmax loss, fusing two streams together yields the final video descriptor for a given video sequence. To determine who is performing each action in the videos, the support vector machine classifier is applied to the generated video descriptor.

3.1.4 An Enhancement of Deep Positional Attention-based Human Action Recognition by using Geometric Positional Features

In the final stage of the investigation, a JTD-Geometric and PAHBRD (JTDGPAHBRD) is proposed to investigate the joint, edge, and surface geometries of the graph structure for HAR by seeing skeletons as a chain of graph joints. It is argued that the trajectory coordinates and other forms of primitive geometric information can be used to learn discriminative interpretations of activities. In addition, the PAHBRNN now includes a View Conversion (VC) layer and a Temporal Dropout (TD) layer for learning more secure interpretations of HAR data. Human activities are detected by a

support vector machine classifier once the absolute VD is derived from a fully connected layer.

3.1.5 An Improvements of Deep Learner-based Human Activity Recognition with the Aid of Graph Convolution Features

In the fifth and final phase of the study, a Graph Convolutional Network (GCN) trained on the JTDGPAHBRD (JTDGPAHBRD-GCN) was used to create a video descriptor for HAR. For better end-to-end learning, the GCN can acquire supplementary information between frames, like more advanced spatial-temporal properties. Also, an adaptive graph-based search space is built to enhance feature representation capability. Then, this space is probed using an evolution technique that is both computationally and statistically efficient. In addition, the temporal dynamics of the skeleton pattern are provided by the resulting GCN, which is combined with geometric properties of the skeleton body joints and trajectory coordinates from the JTDGPAHBRD to produce a more accurate video descriptor for HAR. The SVM system then categorizes these collected descriptors to recognize a wide range of human behaviors.

3.2 DATASET DESCRIPTION

Penn Action dataset are taken into account for the experimental analysis which consists of 2326 video sequences from the 640x480. The training set consists of 1861 video sequences, whereas the testing set consists of 465 sequences. Some examples of such data are C3D features, primitive geometry coordinates, trajectory coordinates, and spatiotemporal correlations. The provided models are evaluated using MATLAB 2017b.

3.2.1 Penn Action Dataset

It has 2326 video sequences annotated for human joints across 15 movement classes. Each frame is tagged with information on 13 separate joints. Baseball pitches, baseball swings, bench presses, bowling, clean and jerks, golf swings, jumping jacks, jumping rope, pull-ups, push-ups, sit-ups, squats, strumming a guitar, tennis forehands, and tennis serves are only some of the sports and workouts represented in the collection. Each clip contains fifty to a hundred blocks, and the various body parts are

represented by names such as "head," "left shoulder," "left elbow," "left wrist," "left hip," "left knee," "left ankle," "right shoulder," "right elbow," "right hip," "right knee," and "right ankle." The dataset is available at http://dreamdragon.github.io/PennAction/. Figure 3.2 shows the sample image of penn action dataset.

Figure 3.2 Examples from the Penn Action dataset for a sequence of frames

3.3 PERFORMANCE METRICS

3.3.1 Accuracy

Correct categorization as a percentage of all examined instances constitutes the accuracy metric. Percentage of HAR in each class that were correctly categorised as opposed to the total number of HAR.

$$Accuracy = \frac{True\ Positive\ (TP)+True\ Negative\ (TN)}{TP+TN+False\ Positive\ (FP)+False\ Negative\ (FN)} \quad (3.1)$$

The TP (True Positive) result in Eq. (3.1) above indicates that the Human Activity, in this case, bench pressing, has been accurately identified. The model correctly identifies other acts as "others," such as "other actions except bench press," as indicated by the TN (True Negative) outcome. If the model returns an FP (False Positive) result, it means it misclassified some genuine instances of an action, such a bench press action, as something else. The model incorrectly identifies the other

activities as real, such as other acts being categorized as bench press, which would result in a FN (False Negative) outcome. Accurate results are obtained by taking both positive and negative actions.

3.3.2 Precision

Precision is the degree to which individual measurements are consistent with one another. It's dependent on how many TP and FP actions were correctly identified. Accurate measurements are easy to replicate, even if they are only slightly off.

$$Precision = \frac{TP}{TP+FP} \tag{3.2}$$

3.3.3 Recall

Most recall measures concentrate on how well a model can pick out correct information. Recall, also known as true positive rate, is the percentage of instances in which a forecast was accurate relative to the total number of times one was attempted.

$$Recall = \frac{TP}{TP+FN} \tag{3.3}$$

3.3.4 F-measure

The F1 Score is a combined recall and precision metric. Since there is a trade-off between accuracy and recall, F1 can be used to evaluate how well our models strike this balance. When both the precision and recall halves of the F1 score are 0, the result is also 0. As a result, it gives extremely negative values in both variables.

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3.4}$$

3.4 CHAPTER SUMMARY

This section outlines the general methodology used in the present investigation. The suggested methodologies and their operational mechanisms are presented in a comprehensive manner. The stages of this study's approach are also explained. In addition, a general overview of the methods employed in the study described here is provided. The datasets and performance indicators that were utilized to evaluate the new and old models are briefly described.