Human Activity Recognition using Two-stream Attention Based Bi-LSTM Networks

by

Vaidehi Bhandarkar
202111064

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY
in
INFORMATION AND COMMUNICATION TECHNOLOGY
to

DHIRUBHAI AMBANI INSTITUTE OF INFORMATION AND COMMUNICATION TECHNOLOGY

May, 2023

Declaration

I hereby declare that

i) The thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology and has not been submitted elsewhere for a degree,

ii) Due acknowledgement has been made in the text to all the reference material used.

Vaidehi

Vaidehi Bhandarkar

Certificate

This is to certify that the thesis work entitled **Human Activity Recognition using Two-stream Attention Based Bi-LSTM Networks** has been carried out by Vaidehi Bhandarkar for the degree of Master of Technology in Information and Communication Technology at *Dhirubhai Ambani Institute of Information and Communication Technology* under my/our supervision.

M Khare 11/07/2023

Dr Manish Khare
Thesis Supervisor

Acknowledgments

The successful completion of my thesis work cannot be attributed solely to my own efforts. I am deeply grateful for the invaluable support and contributions of numerous individuals who have aided me along the way. I would like to take this moment to express my heartfelt appreciation to all those who have directly or indirectly assisted me in the fulfilment of my thesis.

I am grateful to my guide **Dr. Manish Khare** for sharing his knowledge, guidance and time during thesis. I am also thankful to **Dr. Avik Hati** for constant inspiration and great efforts to explain things clearly and simply.

I am thankful to **my Friends at Intelligence Surveillance Research Lab (ISRL)**, for backing me with their best knowledge and also participating in several discussions in order to achieve the defined objective of my thesis.

I owe a debt of gratitude to **my Friends and fellow Batch mates at DAIICT**, for constantly supporting me and always being there for me, giving me the feeling that I had a second family and a home away from home.

Lastly, and most importantly, I am deeply indebted to **my parents (Mr. Kiran Bhandarkar & Mrs. Manisha Bhandarkar)** and my brother **(Mr. Swapnil Bhandarkar)** for their moral support and continuous encouragement while carrying out this study. They raised me, supported me, taught me and loved me.

Contents

Abstract

Human Activity Recognition (HAR) is a challenging task that aims to identify the actions of humans from various data sources. Recently, deep learning methods have been applied to HAR using RGB (Red, Green and Blue) videos, which capture the spatial and temporal information of human actions. However, most of these methods rely on hand-crafted features or pre-trained models that may not be optimal for HAR. In this Thesis, we propose a novel method for HAR using Two-stream Attention Based Bi-LSTM Networks (TAB-BiLSTM) in RGB videos. Our method consists of two components: *a spatial stream and a temporal stream*. The spatial stream uses a convolutional neural network (CNN) to extract features from RGB frames, while the temporal stream uses an optical flow network to capture the motion information. Both streams are fed into an attention-based bidirectional long short-term memory (Bi-LSTM) network, which learns the long-term dependencies and focuses on the most relevant features for HAR. The attention mechanism is implemented by multiplying the outputs of the spatial and temporal streams, applying a softmax activation, and then multiplying the result with the temporal stream output again. This way, the attention mechanism can weigh the importance of each feature based on both streams. We evaluate our method on four benchmark datasets: UCF11, UCF50, UCF101, and NTU RGB. This method achieves state-of-the-art results on all datasets, with accuracies of 98.3%, 97.1%, 92.1%, and 89.5%, respectively, demonstrating its effectiveness and robustness for HAR in RGB videos.

List of Principal Symbols and Acronyms

$Bi - LSTM$	Bi directional LSTM
CNN	Convolutional Neural Network
GCN	Graph Convolutional Network
GRU	Gated Recurrent Unit
HAR	Human Activity / Action Recognition
HCI	Human Computer Interaction
$LSTM$	Long Short Term memory
RGB	Red, Green and Blue
RNN	Recurrent Neural Networks

List of Tables

List of Figures

CHAPTER 1

Introduction

Human activity recognition (HAR) is the task of identifying and classifying human actions in video data. It is an active area of Research in the field of Computer Vision for various contexts like security surveillance, human-computer interaction, etc. The goal of HAR is to automatically detect and analyze human activities from the nature of action from unknown video sequences.

HAR from videos is a rapidly evolving field of research that aims to develop intelligent systems capable of automatically analyzing and understanding human actions in video data. With the proliferation of video data from various sources such as surveillance cameras, smartphones, and social media platforms, the need for automated methods to recognize and interpret human activities has become increasingly important. HAR has numerous practical applications, including video surveillance, human-computer interaction, healthcare monitoring, and sports analysis, to name a few.

The primary objective of HAR is to enable machines to comprehend and interpret human behaviour in videos, similar to how humans effortlessly perceive and comprehend actions. This task involves not only detecting and recognizing the actions performed by individuals but also understanding the context, temporal dynamics, and spatial relationships within the video frames. Achieving accurate and reliable HAR is challenging due to the inherent complexity of human activities, variations in appearance, viewpoint, lighting conditions, occlusions, and the vast diversity of actions.

Over the years, significant advancements in computer vision, machine learning, and deep learning have revolutionized the field of HAR. These advancements have led to the development of sophisticated algorithms and techniques that can automatically extract discriminative features, model temporal dependencies, and leverage large-scale annotated datasets for training robust activity recognition models. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and their variants have demonstrated remarkable success in capturing spatial and temporal features from video data, enabling accurate recognition of human activities.

By exploring the advancements, challenges, and potential solutions in HAR, this thesis aims to contribute to the development of an intelligent system capable of accurately recognizing and understanding human activities from video data. The findings and insights presented in this work can have significant implications in various domains, including surveillance, healthcare, robotics, and human-computer interaction, ultimately leading to advancements in human-centric technologies.

1.1 Motivation

Human Activity Recognition is a challenging task that aims to identify the actions of humans from various data sources. It includes analyzing and understanding a person's behaviour and is fundamentally required for various applications such as video indexing, bio-metrics, surveillance and security. Due to this, video activity recognition has received great attention in the computer vision community. It aims at automatically recognizing human activity from video sequences. In HAR, various activities like walking, running, sitting, sleeping, standing, driving, opening the door, abnormal activities, etc., can be recognized.

Because of its use in a number of applications, such as healthcare, HCI, security, and surveillance, HAR has become a demanding topic in computer vision. The type of data produced by different sources, such as videos, photos, or signals, has a direct impact on HAR approaches. Video is used in HAR for security, monitoring, and recognising human actions and behaviours. Vision-based HAR has detected or predicted actions from video streams using a range of video sources, including CCTV, smartphone cameras, Kinect devices, and social media sites such as YouTube [3].

The percentage of each data source is depicted in Figure 1.1. CCTV cameras (52%) and cell phone sensors (26%) are the most popular data sources. Other data sources utilised less frequently include Kinect (1%), smartphone cameras (1%), camera photos (4%), social media images (3%), wearable body sensors (8%), and YouTube videos (5%).

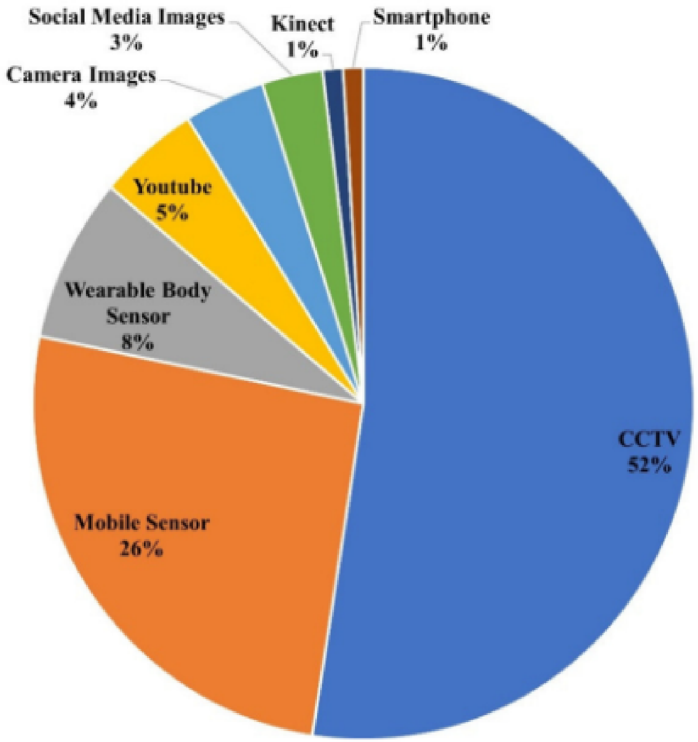


Figure 1.1: Data Sources for HAR [3]

Human Action Recognition (HAR) aims to automatically examine and recognize the nature of action from unknown video sequences. The motions of various human body components are frequently part of functional movements that do not reveal intents or thoughts. Human activities are classified into four categories based on the body parts involved and the intricacy of the action: Actions, gestures, interactions, and group activities [8].

- *Gesture:* It is a visible body activity that conveys a message. It is a movement made with the hands, faces, or other body parts, such as Okay gestures and thumbs up, rather than verbal or vocal communication.
- *Action:* It is a series of physical actions performed by a single person, such as walking and running.
- *Interactions:* It is a series of actions performed by no more than two actors. At least one subject is a person, and the other can be an individual or an item (handshakes, conversing, etc.).

- *Group activities:* It consists of a variety of motions, acts, or interactions. At least two performers are present, in addition to one or more interactive activities (such as volleyball, obstacle courses, etc.).

HAR is a rapidly growing field with the potential to revolutionize many different applications. Some of them are as follows :

- Human-computer interaction: HAR can be used to control devices by gestures or body movements.
- Video surveillance: HAR can be used to identify people and their activities in surveillance footage.
- Healthcare: HAR can be used to monitor the activities of elderly people or people with disabilities.
- Sports analytics: HAR can be used to analyze the performance of athletes.

1.2 Challenges

Recognizing human activity is challenging because it depends on the distance between the camera and the object and recognizing, with precision, the different types of activity performed. Activity Recognition using video data is challenging due to problems such as background clutter, partial occlusion, and changes in scale, viewpoint, lighting, and appearance. Moreover, most of the existing methods for HAR using video data rely on hand-crafted features or pre-trained models that may not be optimal for HAR. Therefore, there is a need to develop novel methods that can automatically learn features from video data and effectively model human activities from spatial and temporal dependencies.

Human activity recognition from videos is a challenging task that involves automatically identifying and categorizing human activities based on visual information captured in video sequences. This area of research finds applications in various domains, including surveillance, video analysis, healthcare, sports, and human-computer interaction. However, several inherent challenges need to be addressed for accurate and robust human activity recognition [17].

- **Variability in Human Actions:** Human actions exhibit significant variability due to variations in appearance, motion, viewpoint, scale, and occlusion.

Different individuals may perform the same action differently, and even a single individual may perform an action in various ways. This variability makes it difficult to design a single model that can effectively recognize all possible activity variations.

- **Temporal Dynamics:** Recognizing human activities from videos requires capturing the temporal dynamics of actions over time. Actions can have varying durations, and the speed at which they are performed may differ. Moreover, activities may consist of complex sequences of sub-actions, making it challenging to model and capture the temporal dependencies accurately.

- **Occlusions and Clutter:** Videos often suffer from occlusions, where parts of the human body or the entire person may be partially or completely hidden from view. Occlusions can occur due to objects, other people, or self-occlusions caused by body parts obstructing each other. Similarly, the presence of clutter in the background or foreground can further complicate the recognition task by introducing distractions and ambiguities.

- **Scale and Viewpoint Variations:** Recognizing human activities across different scales and viewpoints is a challenging problem. The appearance of actions can vary significantly when observed from different distances or angles. Recognizing actions at different scales and viewpoints requires the models to be invariant to these variations while still capturing the discriminative information.

- **Contextual Understanding:** Human activity recognition often relies on contextual cues to distinguish between similar actions or to understand the intentions behind an activity. Contextual information, such as the scene, objects present, or the overall context of the video, can play a crucial role in accurately recognizing activities. However, effectively incorporating such contextual understanding into the recognition models is a non-trivial task.

- **Data Acquisition and Annotation:** Building accurate activity recognition models requires large-scale annotated datasets. Collecting and annotating such datasets is a labour-intensive and time-consuming process. Annotating activities in videos often requires manual labelling by human annotators, which can be subjective and prone to errors. Additionally, there is a need for diverse datasets that cover a wide range of activities and environmental conditions to ensure the generalizability of the models.

- **Computational Complexity:** Recognizing activities from videos involves analyzing and processing a large amount of visual data. This requires computationally expensive operations, such as feature extraction, motion estimation, and modelling temporal dependencies. Real-time or near real-time recognition in resource-constrained settings, such as embedded systems or mobile devices, poses additional challenges due to limited processing power and memory constraints.

- **Domain Adaptation and Generalization:** Activity recognition models trained on one dataset or in one environment may not generalize well to other datasets or real-world scenarios. Variations in camera setups, lighting conditions, background clutter, and human appearance can significantly impact the performance of recognition models. Domain adaptation techniques are required to ensure the models can generalize to unseen data and perform reliably in different environments.

Addressing these challenges requires interdisciplinary research involving computer vision, machine learning, and data annotation expertise. It involves developing robust algorithms to handle action variations, capture temporal dynamics, handle occlusions, incorporate contextual understanding, and efficiently process video data. Additionally, the availability of large-scale annotated datasets and the development of evaluation metrics are crucial to driving advancements in human activity recognition from videos.

Despite these challenges, there has been significant progress in HAR from videos in recent years. This is due to the development of new machine-learning techniques that can learn from large datasets of videos. As these techniques continue to improve, HAR from videos will likely become more accurate and reliable [16]. Here are some additional challenges that can be encountered in human activity recognition from videos:

- *Multi-person activity recognition:* In some cases, it may be necessary to recognize the activities of multiple people in a single video. This can be a challenging task, as the activities of different people can often overlap.

- *Unconstrained activities:* In many cases, people perform activities in unconstrained environments, where they may be moving around freely. This can make it difficult to track their movements and identify the activities they are performing.

- *Out-of-distribution activities:* When training a model for HAR, it is important to use a dataset that covers a wide range of activities. However, it is possible that the model will encounter activities that are not present in the training dataset. This is known as an out-of-distribution activity. When this happens, the model may not be able to recognize the activity accurately.

1.3 Contribution

The temporal motions, which are crucial for video-based human action recognition, are ignored by CNN deep models, which only learn the spatial feature. Two-stream convolutional networks are crucial for recognising human activity in videos as a potent feature extractor. Recent research has demonstrated the significance of two-stream feature extraction for the detection of human activity. This type of two-stream deep learning architecture has been demonstrated to be efficient for short-term temporal cubes capture in various sequence challenges and establishing the structural basis for the subsequently developed video-based human activity recognition.

In the thesis, we propose a novel method for HAR using Two-stream Attention Based Bi-LSTM Networks (TAB-BiLSTM) in RGB videos. This method is evaluated on four benchmark datasets: UCF11, UCF50, UCF101, and NTU RGB. Also, it achieves state-of-the-art results on all datasets, with accuracies of 98.3%, 97.1%, 92.1%, and 89.5%, respectively, demonstrating its effectiveness and robustness for HAR in RGB video.

1.4 Outline of the Thesis

The organization of the thesis is as follows:

- *Chapter 1* introduces the problem of Human Activity Recognition, the Motivation behind the research, challenges faced for HAR, its importance and the thesis's contribution.

- *Chapter 2* presents the General process for Human Activity Recognition and the detailed Literature Review of different types of Har by using Deep Learning methods and previous methods proposed for HAR.

- *Chapter 3* discusses the model architecture of the proposed methodology for Human Activity Recognition.
- *Chapter 4* discusses the state-of-the-art datasets on which the experiments are performed, experiments performed, the results and the comparison of results with previously proposed methods.
- *Chapter 5* gives the Conclusion and Future Scope of the Thesis.

Chapter 2
Literature Review

Human Activity Recognition is a field within computer vision that has gained significant attention due to its applications in diverse domains, including security surveillance, human-computer interaction, healthcare monitoring, sports analysis, and more. The primary objective of HAR is to develop algorithms and models that can detect and analyze human activities without human intervention. The goal is to extract meaningful information from video sequences, such as body movements, interactions, and gestures, and infer the corresponding actions being performed by individuals.

This literature review aims to provide a comprehensive understanding of human activity recognition from videos. The general process for activity recognition outlines the key steps involved, while the exploration of deep learning methods sheds light on the advancements and state-of-the-art approaches in the field.

2.1 General Process of Human Activity Recognition

Based on the type of data being processed, HAR can be separated into two main approaches: vision-based HAR and sensor-based HAR [1, 6], as shown in Figure 2.1. The former focuses on investigating raw data derived from wearable sensors and environmental sensors, while the latter focuses on analysing images or movies obtained from optical sensors. Based on the sort of data they collect, optical sensors can be distinguished from other types of sensors. Wearable sensors produce one-dimensional signal data, whereas optical sensors produce two-dimensional, three-dimensional, or video images [7].

Because wearables are worn by users to automatically identify and track a variety of activities, including sitting, jogging, running, and sleeping, they serve as repre-

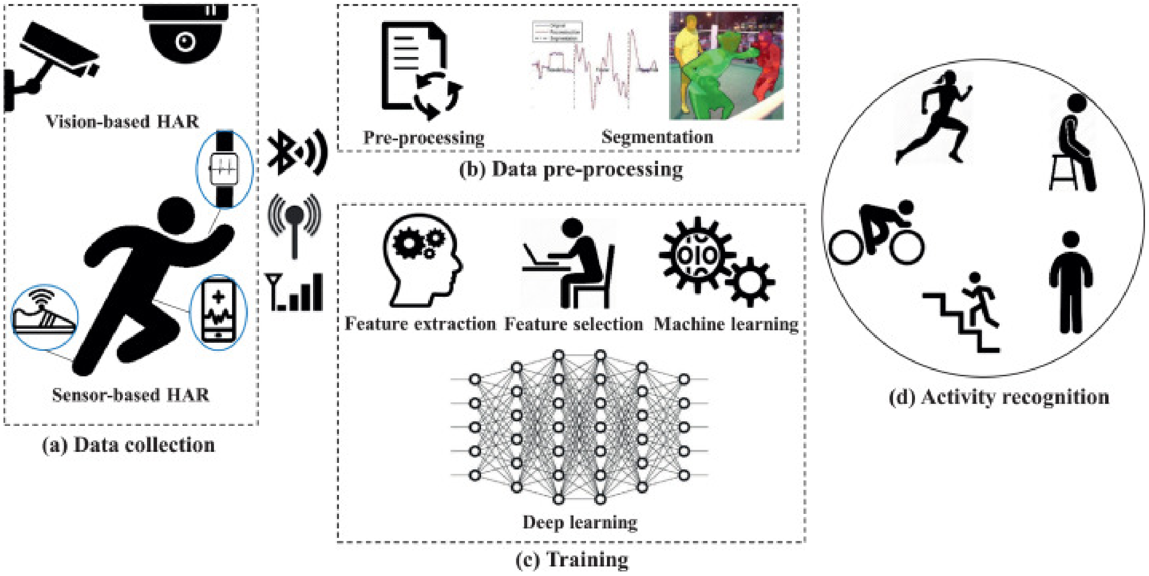


Figure 2.1: General Procedure for HAR [5]

sentative instances of sensor-based HAR. A sensor, however, is ineffective when a subject is either beyond of its detection range or engages in unknown behaviours.

CCTV systems, on the other hand, have long been used in vision-based HAR systems. Gesture and activity recognition systems based on video analysis have been thoroughly researched. Furthermore, this issue is particularly advantageous to security, surveillance, and interactive applications. In recent years, the vast majority of research has concentrated on vision-based HAR since vision-based data is less expensive and easier to obtain than sensor-based data. [5].

There are two main approaches to vision-based HAR:

- **Feature-based methods** extract features from the video, such as the position and velocity of body parts, and then use machine learning algorithms to classify the activities.

- **Deep learning methods** directly learn to classify activities from the video data without extracting features manually.

Deep learning has captured the computer vision community's interest in recent years owing to the excellent performance of deep learning-based research in a variety of study fields, including object detection and recognition, image classification, and natural language processing (NLP). While comparing with the traditional ML algorithm, Deep Learning methods significantly reduce the effort of selecting the right features by automatically extracting abstract features via several hidden layers. The deep learning structure has been demonstrated to work well with unsupervised learning, and reinforcement learning [15].

Deep learning methods have recently become the dominant approach for HAR due to their ability to learn complex patterns from large amounts of data. However, they require a large amount of training data, which can be expensive to collect. Deep Learning methods for human activity recognition can be briefly divided into three parts, i.e., HAR from 3D Skeleton data [12, 13, 28], spatio-temporal methods [9, 19, 24, 26, 27] and motion recognition [2, 14, 25, 26] which can be seen in Figure 2.2.

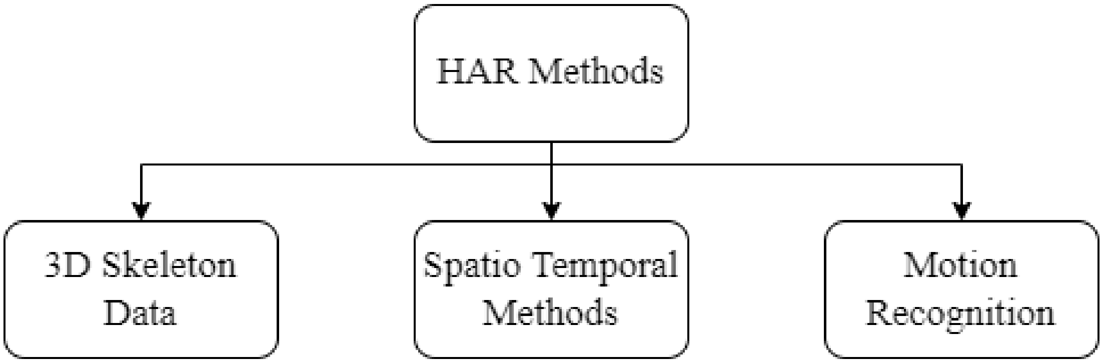


Figure 2.2: Human Activity Recognition: Deep Learning Methods

2.2 Human Activity Recognition from 3D Skeleton Data

Skeleton-based Action Recognition is a computer vision task that involves recognizing human actions from a sequence of 3D skeletal joint data. In recent years, skeleton-based action recognition has been attracting interest in the field of Computer Vision. Each joint of the human body is identifiable by a joint type, a frame index, and a 3D position in the skeleton, a type of well-structured data [28].

There are several advantages of using the skeleton for action recognition:

- First, the skeleton is a high-level depiction of the human body that abstracts human position and mobility. Biologically, humans can recognise the action category by witnessing simply joint motion, even in the absence of appearance information.

- Second, the advancement of low-cost depth cameras and posture estimation technology makes skeleton access considerably easier.
- Third, when compared to RGB video, the skeletal representation is more resistant to changes in viewpoint and appearance.
- Fourth, due to its low dimensional representation, it is also computationally efficient [28].

Skeleton-based activity recognition complements RGB-based activity recognition in several ways: Pose Information, Viewpoint and Occlusion Independence, Data Efficiency, etc. Figure 2.3 depicts three ways for skeleton-based human activity recognition using deep learning.

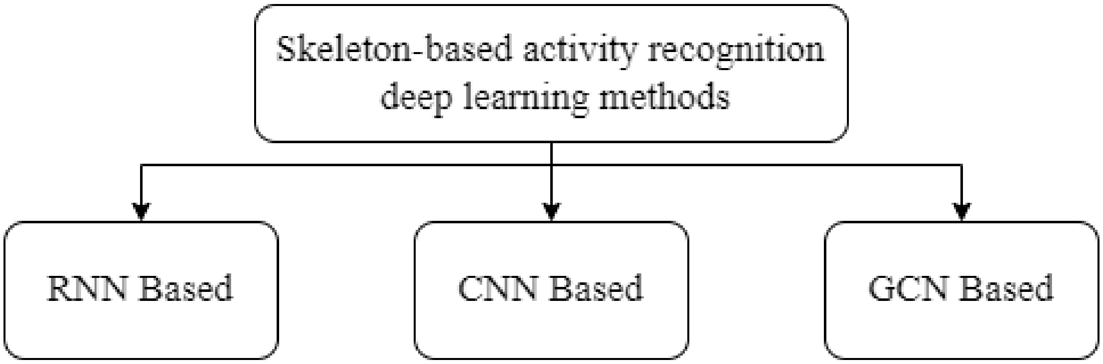
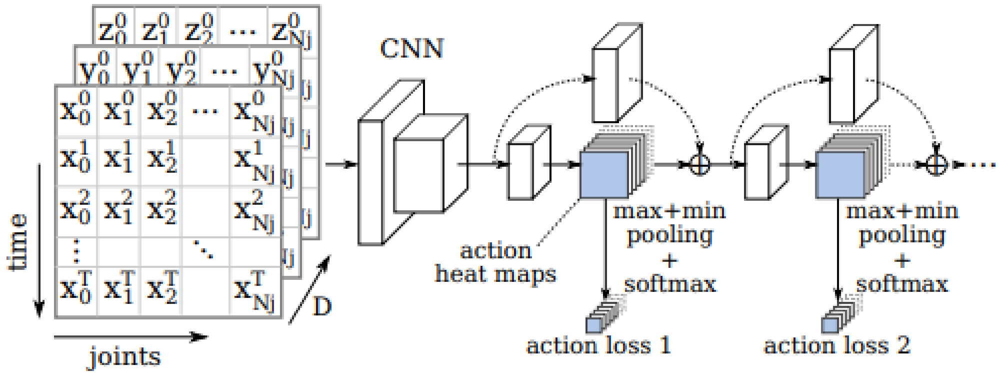


Figure 2.3: Skeleton-based human activity recognition using deep learning methods

- *Recurrent neural networks (RNN)*, such as LSTM and GRU are frequently used to describe the temporal dynamics of skeletal sequences. The input vector of a time slot is created by concatenating the 3D coordinates of all the joints in a frame in some order.
- A well-known network, such as ResNet, is used in the *The Convolutional Neural Network (CNN)*-based research to examine the spatial and temporal dynamics after transforming the skeleton sequence into a skeleton map of target size.
- Each joint is treated as a node in a graph by the *Graph Convolutional Network (GCN)*. Humans predefine the presence of the edge signifying the joint relationship based on prior knowledge. The edges for both physically separated and connected joint pairs were added to the predefined graph to improve its construction.

Figure 2.4 describes the general architecture for HAR from Skeleton data using CNN. As seen in figure 2.4, a fully convolutional neural network extracts feature from input poses and produce action heat maps. Human pose estimation forms the basis for human activity recognition from skeleton data.

A pooling operation on the action maps is carried out in order to generate the output probability of each activity for a video clip. The *max + min* pooling followed by a *Softmax activation* is employed in order to be more responsive to the strongest reactions for each action. Additionally, employing a stacked architecture with intermediate supervision in *K* prediction blocks, the predictions from the human pose regression approach are improved. Further, the action/activity recognition block is then injected with the action heat maps from each prediction block [12].



z_0^0 z_1^0 z_2^0 ⋯ z_{Nj}^0
y_0^0 y_1^0 y_2^0 ⋯ y_{Nj}^0
x_0^0 x_1^0 x_2^0 ⋯ x_{Nj}^0
x_0^1 x_1^1 x_2^1 x_{Nj}^1
x_0^2 x_1^2 x_2^2 x_{Nj}^2
x_0^T x_1^T x_2^T x_{Nj}^T
time
joints
D
CNN
action heat maps
max+min pooling + softmax
action loss 1
max+min pooling + softmax
action loss 2

Figure 2.4: General Architecture for HAR from Skeleton Data [11]

There are mainly two kinds of datasets for human activity recognition. Some are commonly used to evaluate video-based activity recognition algorithms, while others are commonly used to evaluate skeleton-based activity recognition algorithms. UCF101 and Kinetics, for example, are datasets used to evaluate video-based activity identification systems. Because no skeleton is provided in these datasets, the datasets can be more diverse because only RGB is required [22].

Due to the limitations of recording technology, datasets, including skeleton information, are often small-scale or in a controlled environment. Another significant distinction is that with RGB activity recognition datasets, the camera often does not record the entire body of a person, but skeleton-based datasets require full body capture for activity recognition. However, in the NTU RGB dataset, a popular skeleton-based action dataset, the whole body of the people is generally caught in the movies. PoseC3D dataset, NTU60-X dataset, NTU RGB 120 dataset, UWA3D dataset, N-UCLA dataset, and others are also available.

2.3 Spatio-Temporal Methods for Human Activity Recognition

Spatio-Temporal means relating to both space and time. Understanding the spatial and temporal interaction between video frames is at the heart of human activity recognition. When it comes to ensuring the security and safety of residents, activity recognition faces many challenges, including industrial monitoring, violence detection, person identification, virtual reality, and cloudy environments due to significant improvements in camera movements, occlusions, complex backgrounds, and variations in illumination. The spatial and temporal information is critical in recognising various human actions in videos.

The spatio-temporal characteristic has long been used to describe action recognition from intensity video. The video is viewed as a 3D volume with space x-, y-, and temporal t-axes. Two-stream convolutional networks play an essential role as a powerful feature extractor in human activity recognition in videos. Recent studies have shown the importance of two-stream feature extraction for human activity recognition. The spatial stream processing of RGB images and the temporal stream processing of successive optical flow fields make up the two components of the dual-stream network. The dual-stream network's primary goal is to separately simulate the mobility stream using optic flow and the appearance stream using RGB data.

Figure 2.5 shows the Spatio-temporal pyramid network, one of the efficient ways for HAR from Spatio-Temporal data [25]. Each block of spatio-temporal convolution (STC) consists of two 2D convolutions that follow each other, then a pseudo-1D convolution. The temporal dimension is subjected to the temporal pooling, which shortens it to length 1.

Widely used spatio-temporal datasets for human activity recognition in videos are NTU RGB, UCF11, UCF50, MSR Daily Activity3D, Florence 3D action, UCF101, Kinetics400, kinetics700, HMDB51, etc.

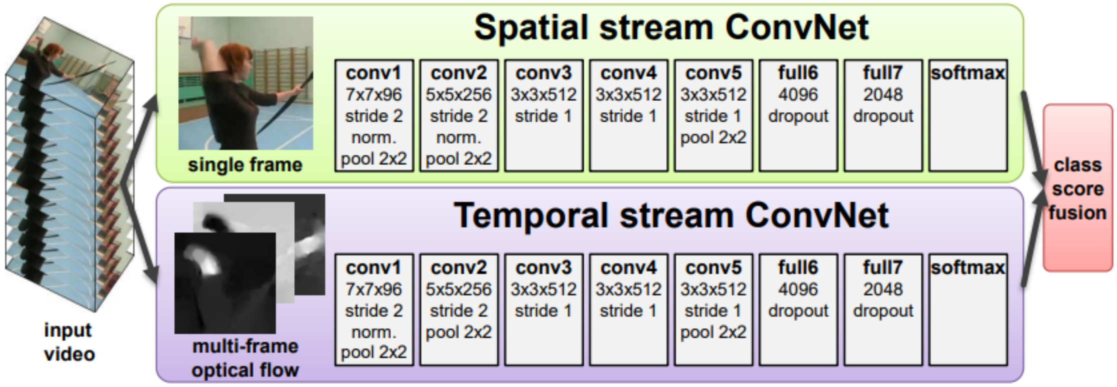


Figure 2.5: Architecture for HAR from Spatio-Temporal data [21]

2.4 Human Activity Recognition from Motion

The technique for getting 3D information on humans directly uses a Motion Capture system (Mocap) [2]. It is a crucial method for recording and studying human articulations. Mocap has been extensively utilised to animate computer-generated characters in films and video games. Additionally, it is utilised to examine and improve the sequencing mechanics of top athletes and track the success of physical therapy in helping patients recover. Only the 3D locations of the chosen points are recorded in a motion capture system; therefore, action recognition algorithms using this type of data frequently construct features based on joint positions or joint angles.

As seen in Figure 2.6, in human movement sciences, a motion capture system must be sufficiently accurate to evaluate human activities. To test the accuracy of OpenPose, numerous synchronised video cameras are used. To reliably recognise human activity, participants conducted three motor tasks (walking, countermovement jumping, and ball throwing), and these movements were captured using both marker-based optical motion capture and OpenPose-based markerless motion capture.

Examples of motion capture datasets with sizable collections of activities include the CMU Motion Capture Database, the MPI HDM05 Motion Capture Database, the CMU Kitchen Dataset, the LACE Indoor Activity Benchmark Data Set, the TUM Kitchen dataset, etc.

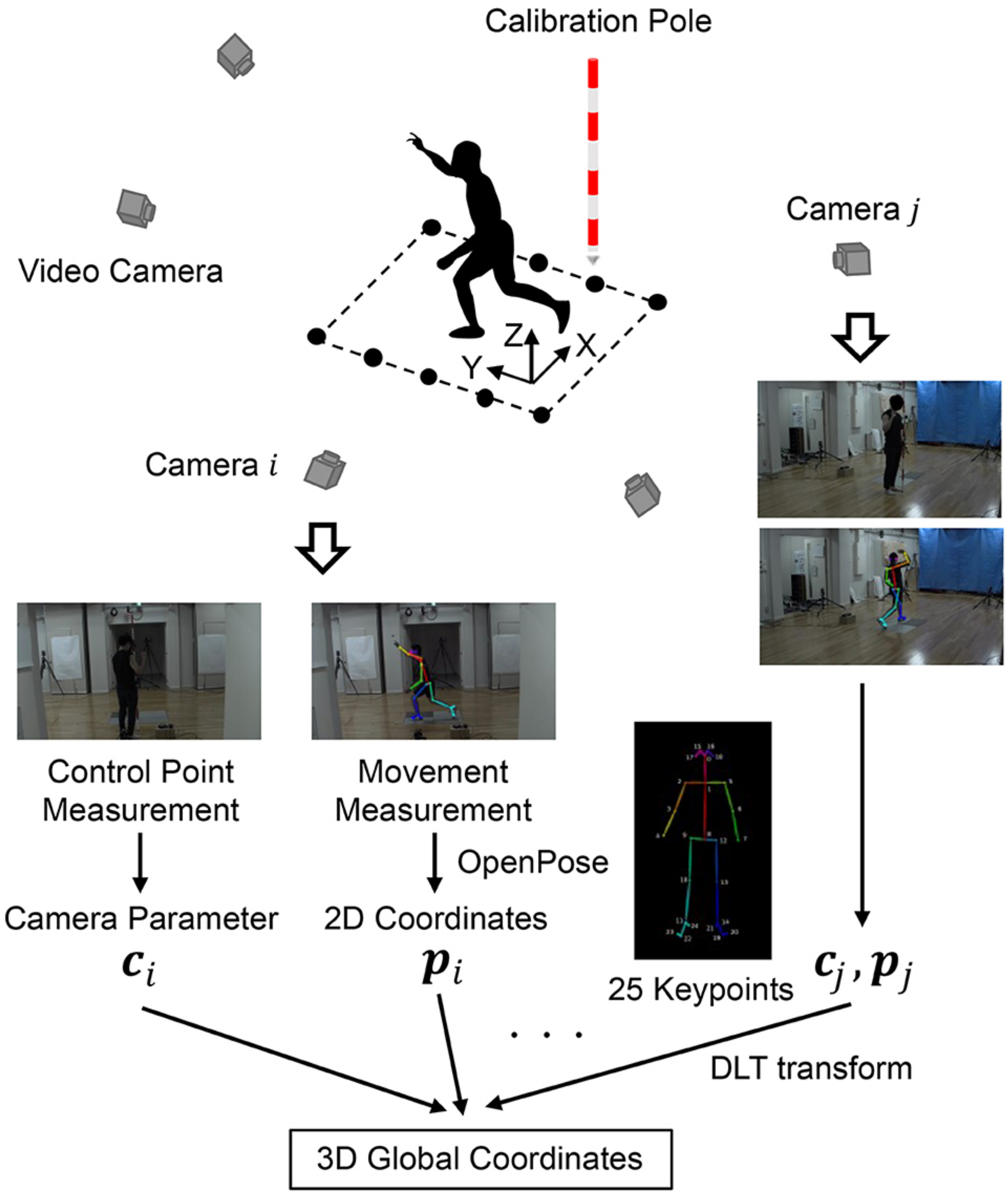


Figure 2.6: Human Activity Recognition from Motion [2]

Skeleton-based methods rely on extracting and analyzing human joint positions or poses obtained from depth sensors or pose estimation algorithms. While skeleton-based methods offer advantages such as viewpoint invariance and reduced sensitivity to appearance variations, they suffer from several limitations. Firstly, accurately estimating human poses from video data can be challenging, particularly in complex scenes or when dealing with occlusions. Errors in pose estimation can adversely affect the performance of skeleton-based methods, leading to decreased accuracy. Secondly, the representation of human actions solely based on joint positions may not capture the full spatial and temporal dynamics of the activity, as it neglects other important cues such as appearance and motion information.

Motion-based methods, on the other hand, leverage motion information extracted from video sequences to recognize human activities. These methods often use optical flow or other motion descriptors to capture temporal dynamics and movement patterns. While motion-based methods are effective in certain scenarios and can capture dynamic aspects of actions, they have their limitations as well. One limitation is their vulnerability to camera motion or background clutter, which can introduce noise in the motion cues and degrade the recognition accuracy. Additionally, purely motion-based methods may struggle with activities that do not exhibit significant motion or rely heavily on subtle spatial cues.

In contrast to skeleton-based and motion-based methods, the spatio-temporal methods combine both spatial and temporal information to achieve superior human activity recognition performance. By considering both the appearance and motion cues in a unified framework, the spatio-temporal method can capture the complete dynamics of actions, leading to enhanced discriminative power and robustness.

The proposed method exploits the spatial stream, which processes the video frames to extract spatial features using convolutional layers. This stream captures fine-grained appearance information and spatial relationships among different body parts, enabling the recognition of activities based on visual cues. Simultaneously, the temporal stream processes the same video frames but with permuted channels, emphasizing the temporal relationships and capturing motion dynamics. This approach allows the model to effectively encode both spatial and temporal features, providing a more comprehensive representation of human activities.

The fusion of spatial and temporal streams in the proposed method facilitates the extraction of discriminative features that capture both appearance and motion cues. By integrating the strengths of both skeleton-based and motion-based approaches, the spatio-temporal method overcomes the limitations of relying solely on joint positions or motion cues. It achieves improved accuracy in recognizing human activities, even in challenging scenarios involving occlusions, background clutter, or subtle spatial cues.

The proposed spatio-temporal method for human activity recognition offers significant advantages over skeleton-based and motion-based methods. By effectively combining spatial and temporal cues, it captures the complete dynamics of actions, resulting in improved accuracy and robustness.

CHAPTER 3
Proposed Methodology

This Chapter describes the proposed method for Human Activity Recognition using Two-stream Attention Based Bi-LSTM Networks (TAB-BiLSTM) in RGB videos. The overall architecture of our method is shown in Figure 3.1. It consists of two major components: a spatial stream known as Spatial Feature Extraction Block and a temporal stream known as Temporal Feature Extraction Block.

The spatial stream takes in the twenty frames of the video and processes them sequentially. Each frame has a resolution of 64x64 pixels in RGB format. To extract relevant spatial features, the spatial stream consists of three convolutional layers, each with 64 filters. Convolutional layers are commonly used in deep learning models to extract visual features by applying a set of filters to the input data. By utilizing multiple convolutional layers, the model can capture increasingly complex spatial patterns.

After the convolutional layers, max-pooling layers are applied with a pool size of (2, 2) and a stride of (2, 2). Max-pooling reduces the spatial dimension of the data while retaining the most important information. It achieves this by dividing the input into non-overlapping regions and only keeping the maximum value within each region. This pooling operation helps to down-sample the data and focus on the most salient features.

The output of the pooling layer is then flattened, converting the multi-dimensional tensor into a vector and passed through two fully connected layers. The first fully connected layer has 512 units, and the second has 256 units. Fully connected layers, also known as dense layers, connect each neuron to every neuron in the previous and subsequent layers, allowing for complex relationships to be learned.

To prevent overfitting and improve generalization, dropout layers with a dropout

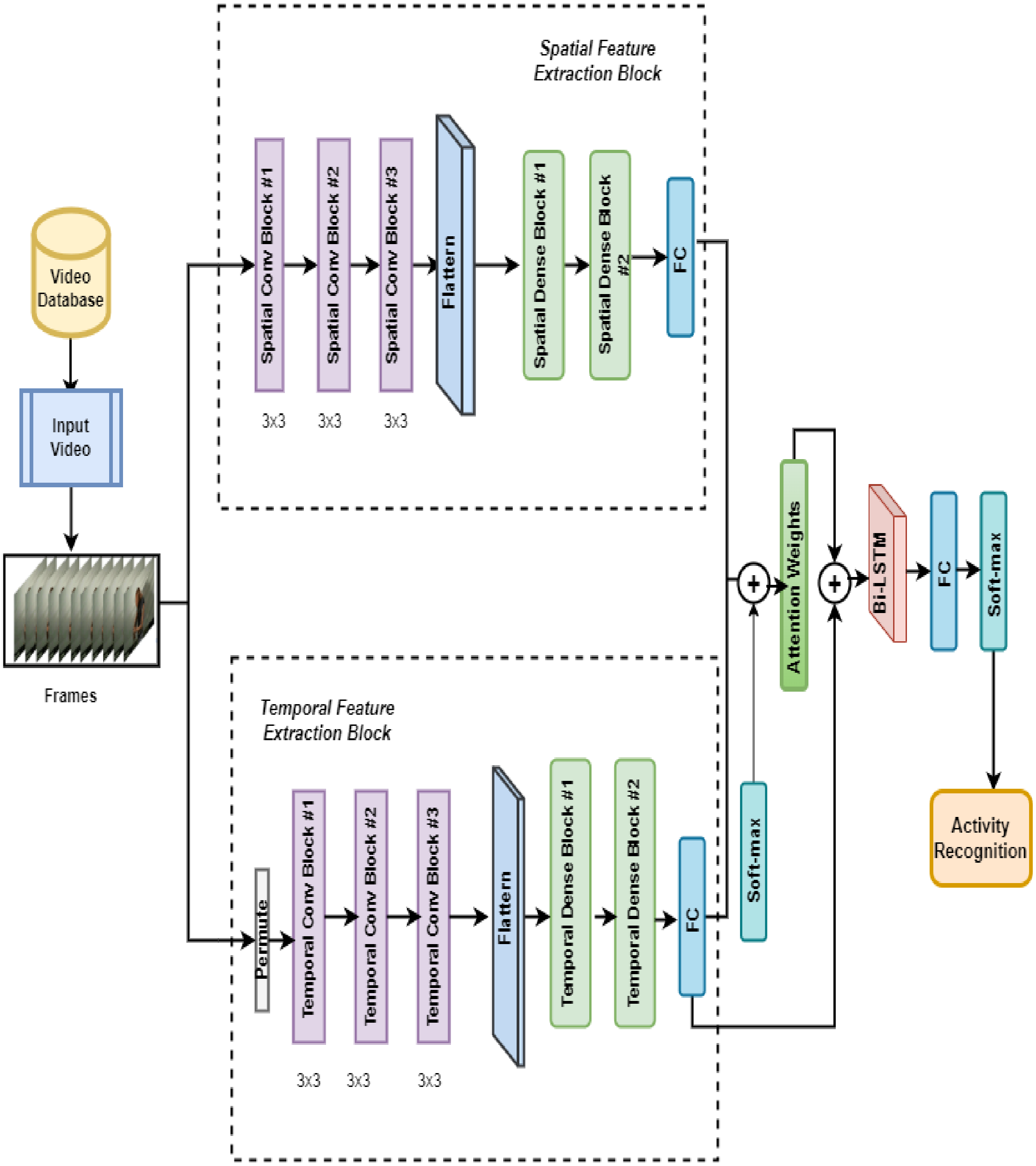


Figure 3.1: **Proposed Architecture for Human Activity Recognition**

rate of 0.5 are applied after the fully connected layers. Dropout randomly deactivates a fraction of the neurons during training, forcing the network to learn more robust and generalized representations.

The temporal stream processes the same twenty frames as the spatial stream but with their channels permuted. By permuting the channels, the temporal relationships between consecutive frames are emphasized, allowing the model to capture motion and temporal dynamics. The permuted frames go through the same three convolutional layers, max-pooling layers, flattening operation, and fully connected layers as the spatial stream.

After passing through both streams, the output of the final fully connected layer from each stream is multiplied element-wise. This multiplication operation combines the information from both streams and creates an attention map. The attention map highlights the most relevant features that were captured by the spatial and temporal streams, emphasizing the regions and moments in the video that contribute the most to the overall understanding and analysis. This can be better understood from Figure 3.2.

The model uses two streams, spatial and temporal, to process video data. Each stream consists of three convolutional layers, max-pooling layers, fully connected layers, and dropout layers. By combining the outputs of the final fully connected layers from both streams, an attention map is generated, which focuses on the most important features from both the spatial and temporal aspects of the video.

3.1 Conv Block

The Conv (Convolution) Block layers are applied to both Spatial and Temporal streams before flattening. It applies a convolutional and a pooling layer to reduce the spatial and increase the receptive field of the features. The output of the pooling layer is assigned to the spatial and temporal stream variables for the next iteration. The loop output is a sequence of feature maps, one for each RGB frame and optical flow image for spatial and temporal streams, respectively. It can be seen in Figure 3.3.

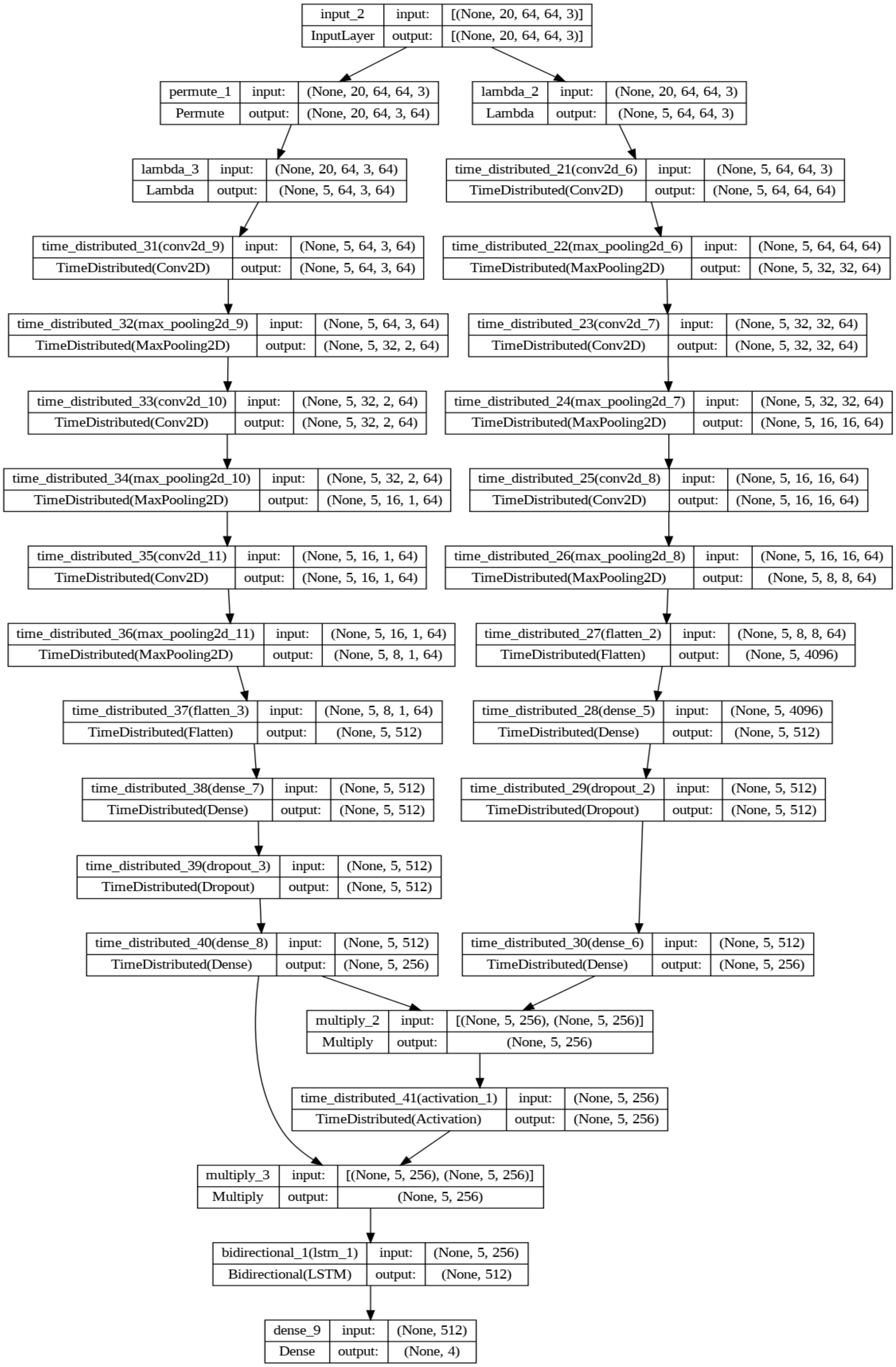

input_2 | input: | [(None, 20, 64, 64, 3)]
InputLayer | output: | [(None, 20, 64, 64, 3)]
permute_1 | input: | (None, 20, 64, 64, 3)
Permute | output: | (None, 20, 64, 3, 64)
lambda_2 | input: | (None, 20, 64, 64, 3)
Lambda | output: | (None, 5, 64, 64, 3)
lambda_3 | input: | (None, 20, 64, 3, 64)
Lambda | output: | (None, 5, 64, 3, 64)
time_distributed_21(conv2d_6) | input: | (None, 5, 64, 64, 3)
TimeDistributed(Conv2D) | output: | (None, 5, 64, 64, 64)
time_distributed_31(conv2d_9) | input: | (None, 5, 64, 3, 64)
TimeDistributed(Conv2D) | output: | (None, 5, 64, 3, 64)
time_distributed_22(max_pooling2d_6) | input: | (None, 5, 64, 64, 64)
TimeDistributed(MaxPooling2D) | output: | (None, 5, 32, 32, 64)
time_distributed_32(max_pooling2d_9) | input: | (None, 5, 64, 3, 64)
TimeDistributed(MaxPooling2D) | output: | (None, 5, 32, 2, 64)
time_distributed_23(conv2d_7) | input: | (None, 5, 32, 32, 64)
TimeDistributed(Conv2D) | output: | (None, 5, 32, 32, 64)
time_distributed_33(conv2d_10) | input: | (None, 5, 32, 2, 64)
TimeDistributed(Conv2D) | output: | (None, 5, 32, 2, 64)
time_distributed_24(max_pooling2d_7) | input: | (None, 5, 32, 32, 64)
TimeDistributed(MaxPooling2D) | output: | (None, 5, 16, 16, 64)
time_distributed_34(max_pooling2d_10) | input: | (None, 5, 32, 2, 64)
TimeDistributed(MaxPooling2D) | output: | (None, 5, 16, 1, 64)
time_distributed_25(conv2d_8) | input: | (None, 5, 16, 16, 64)
TimeDistributed(Conv2D) | output: | (None, 5, 16, 16, 64)
time_distributed_35(conv2d_11) | input: | (None, 5, 16, 1, 64)
TimeDistributed(Conv2D) | output: | (None, 5, 16, 1, 64)
time_distributed_26(max_pooling2d_8) | input: | (None, 5, 16, 16, 64)
TimeDistributed(MaxPooling2D) | output: | (None, 5, 8, 8, 64)
time_distributed_36(max_pooling2d_11) | input: | (None, 5, 16, 1, 64)
TimeDistributed(MaxPooling2D) | output: | (None, 5, 8, 1, 64)
time_distributed_27(flatten_2) | input: | (None, 5, 8, 8, 64)
TimeDistributed(Flatten) | output: | (None, 5, 4096)
time_distributed_37(flatten_3) | input: | (None, 5, 8, 1, 64)
TimeDistributed(Flatten) | output: | (None, 5, 512)
time_distributed_28(dense_5) | input: | (None, 5, 4096)
TimeDistributed(Dense) | output: | (None, 5, 512)
time_distributed_38(dense_7) | input: | (None, 5, 512)
TimeDistributed(Dense) | output: | (None, 5, 512)
time_distributed_29(dropout_2) | input: | (None, 5, 512)
TimeDistributed(Dropout) | output: | (None, 5, 512)
time_distributed_39(dropout_3) | input: | (None, 5, 512)
TimeDistributed(Dropout) | output: | (None, 5, 512)
time_distributed_40(dense_8) | input: | (None, 5, 512)
TimeDistributed(Dense) | output: | (None, 5, 256)
time_distributed_30(dense_6) | input: | (None, 5, 512)
TimeDistributed(Dense) | output: | (None, 5, 256)
multiply_2 | input: | [(None, 5, 256), (None, 5, 256)]
Multiply | output: | (None, 5, 256)
time_distributed_41(activation_1) | input: | (None, 5, 256)
TimeDistributed(Activation) | output: | (None, 5, 256)
multiply_3 | input: | [(None, 5, 256), (None, 5, 256)]
Multiply | output: | (None, 5, 256)
bidirectional_1(lstm_1) | input: | (None, 5, 256)
Bidirectional(LSTM) | output: | (None, 512)
dense_9 | input: | (None, 512)
Dense | output: | (None, 4)

Figure 3.2: Dimension at each Layer of the Proposed Architecture

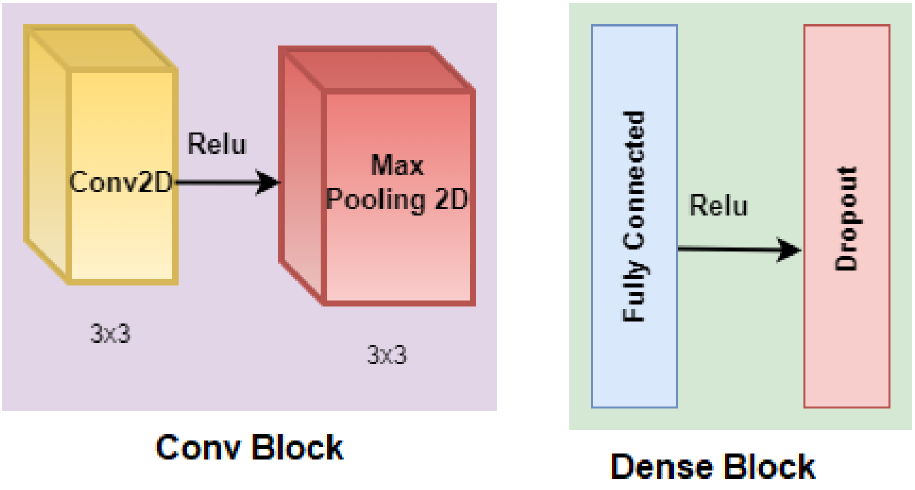


Figure 3.3: Conv and Dense Blocks of the Proposed Architecture

3.2 Dense Block

The Dense Block layers are applied to both Spatial and Temporal streams before flattening. The first fully connected layer uses 512 units and a ReLU activation function to transform the feature maps into feature vectors. The second layer is a dropout layer that uses a rate of 0.5 to drop out some of the units to prevent overfitting randomly. The third layer is another fully connected layer that uses 256 units and a ReLU activation function to reduce the feature vectors' dimensions further. The code output is a sequence of 256-dimensional feature vectors, one for each RGB frame and optical flow image for spatial and temporal streams, respectively. It can be seen in Figure 3.3.

Table 3.1: List of formulae for Operations in Conv and Dense Blocks

Operation	**Formula**
Conv2D	$\mathrm{z}^l = h^{(l-1)} \cdot W^l$
Max Pooling 2D	$\mathrm{h}^l_{xy} = max_{i=0.5,j=0.5} h^{(l-1)}(x+i)(y+j)$
Fully Connected Layer	$\mathrm{z}_l = W_l \cdot h_{(l-1)}$
ReLU	$\mathrm{ReLU}(\mathrm{z}_i) = max(0, z_i)$
Softmax	$\mathrm{softmax}(\mathrm{z}_i) = \frac{e^{zi}}{\sum_j e^{zj}}$

The attention mechanism applies a softmax function to the attention map over the temporal dimension, resulting in a set of weights that emphasise each feature's importance in the temporal stream. The resulting features are multiplied element-wise with the attention weights and passed through a Bidirectional LSTM layer with 256 units. Finally, the output of the LSTM layer is passed through a fully connected layer with the number of units equal to the number of classes, followed by a softmax activation function to produce the final classification.

A predetermined number of feature maps are generated during the convolution process and passed into the max pooling layer, which generates pooled feature maps using the feature maps acquired from the convolution layer preceding it. This process continues until we reach the third max pooling layer before the pooled feature map is transferred into the following convolution layer. The last max pooling layer's pooled feature map is flattened and put into the fully connected layers. Once the model has been trained through numerous iterations of forward and backward propagation, a prediction can be produced. The formulae for all the steps are mentioned in Table 3.1.

The proposed model is evaluated on standard benchmark datasets for action recognition, like UCF11, UCF50, UCF101 and NTU RGB, and compared its performance with state-of-the-art methods. The attention-based model proposed for action recognition exhibits promising outcomes, suggesting its potential to enhance the accuracy and efficiency of existing methods.

CHAPTER 4

Experiments & Results

This chapter presents the details of our experiments to validate the proposed frameworks. UCF11 [10], UCF50 [18], UCF101 [23], and NTU RGB [20] are four different widely used datasets on which the architecture is trained and tested to determine the efficiency of the suggested approach. The experimental findings demonstrated that our approach could surpass the majority of earlier techniques regarding recognition rate.

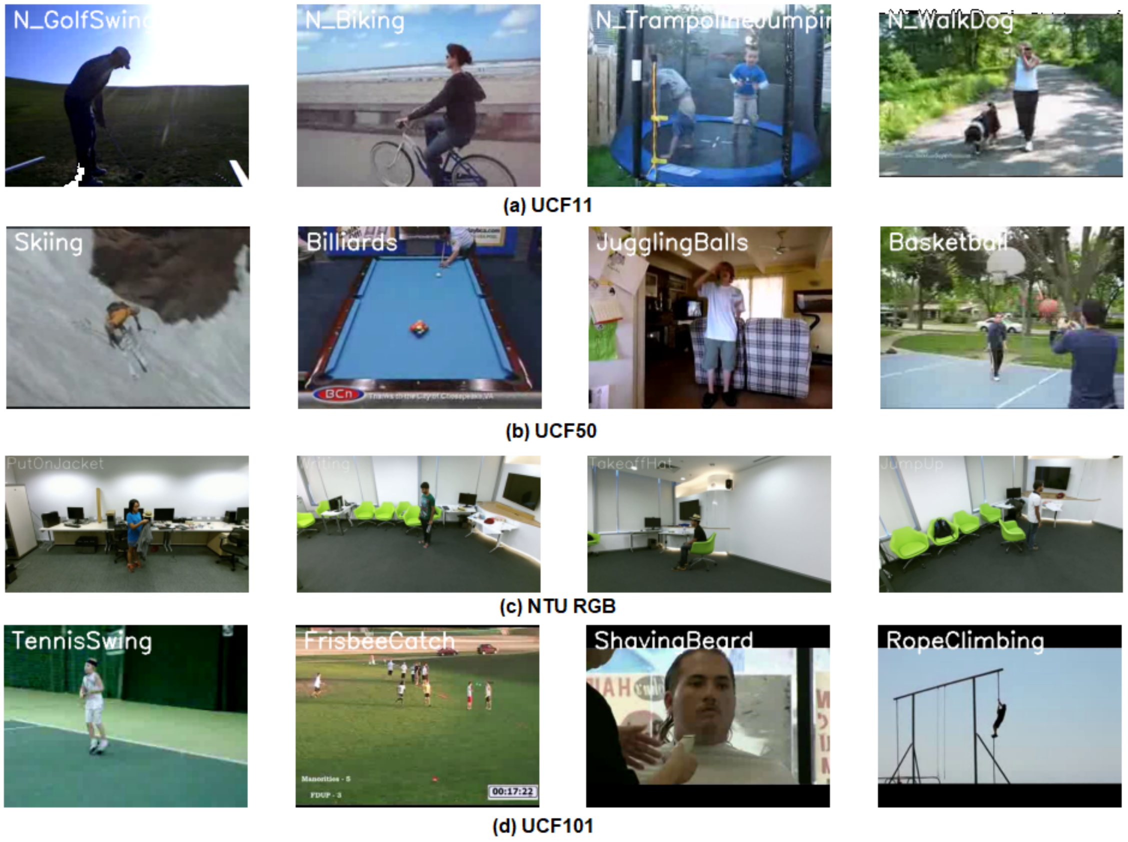
4.1 Datasets

The experiments are performed on widely used video action datasets: UCF11, UCF50, UCF-101, and NTU RGB.

- **UCF11 [10]** is a difficult dataset for recognising actions in videos because of changes in lighting, a crowded background, and camera movements. The 1600 movies in the UCF11 dataset are divided into eleven action categories, including shooting, leaping, riding, swimming, etc. All of the videos were shot at 30 frames per second (fps) rates.
- **UCF50 [18]** is an action recognition data set with 50 action categories comprising 6676 real videos. It consists of 199 average number of frames per Video.
- **UCF101 [23]** is an action recognition data set of realistic action videos collected from YouTube, having 101 action categories and a total of 13,320 videos.
- **NTU RGB [20]** contains 60 action classes and 56,880 video samples. As this is a large dataset of 136 GB, we refined it to 60 action classes and a total of

12,240 videos. The resolutions of RGB videos are 1920x1080. The dataset is captured by three Kinect V2 cameras concurrently.

The visual representation of sample actions from all datasets is given in Figure 4.1.



(a) UCF11

(b) UCF50

(c) NTU RGB

(d) UCF101

Figure 4.1: Sample Action Videos from different Datasets: (a) UCF11. (b) UCF50. (c) NTU RGB. (d) UCF101.

4.2 Efficiency and Accuracy

To validate the effectiveness of the proposed method, a series of extensive experiments are conducted. These experiments aimed to assess the performance of the method in recognizing and classifying activities in videos. The results of these experiments are presented in the form of accuracy measurements.

In Figure 4.2, the accuracy of individual activities from the UCF11 dataset is depicted. The UCF11 dataset is a widely used benchmark dataset for action recognition, containing videos of eleven different human actions. The figure showcases the accuracy achieved by the proposed method for each specific activity in the dataset.

Similarly, Figure 4.3 showcases the accuracy results for the UCF50 dataset. The UCF50 dataset is a more extensive dataset consisting of fifty different action categories. The figure illustrates the accuracy attained by the proposed method for each activity in this dataset.

Additionally, Figure 4.4 presents the accuracy outcomes for the NTU RGB dataset. The NTU RGB dataset is a dataset specifically designed for human action recognition, containing videos captured by multiple RGB cameras. The figure demonstrates the accuracy achieved by the proposed method for individual activities within this dataset.

To further evaluate the performance of the proposed method, Table 4.1 provides a comparison of the accuracy results with other state-of-the-art methods. The table highlights the accuracy achieved by the proposed method alongside the accuracy achieved by different existing methods that are considered the current state-of-the-art. This comparison allows researchers and practitioners to assess the effectiveness of the proposed method in relation to other leading approaches in the field.

By conducting these experiments and presenting the results in figures and tables, the aim is to demonstrate the effectiveness and competitive performance of the proposed method in action recognition. These results provide evidence of the method's capability to accurately classify activities in videos, as well as its potential to outperform or be on par with other state-of-the-art methods in the field.

Table 4.1: Accuracy Comparison with different state-of-the-art methods

Model	**UCF11**	**UCF50**	**UCF101**	**NTU RGB**
Xu et al.[27]	91.5%	94.6%	91.5%	83.8%
Muhammad et al.[14]	94.6%	96.5%	86.6%	80.2%
Dai et al.[4]	96.9%	92.1%	77.9%	87.7%
Proposed Method	**98.3%**	**97.1%**	**92.1%**	**89.5%**

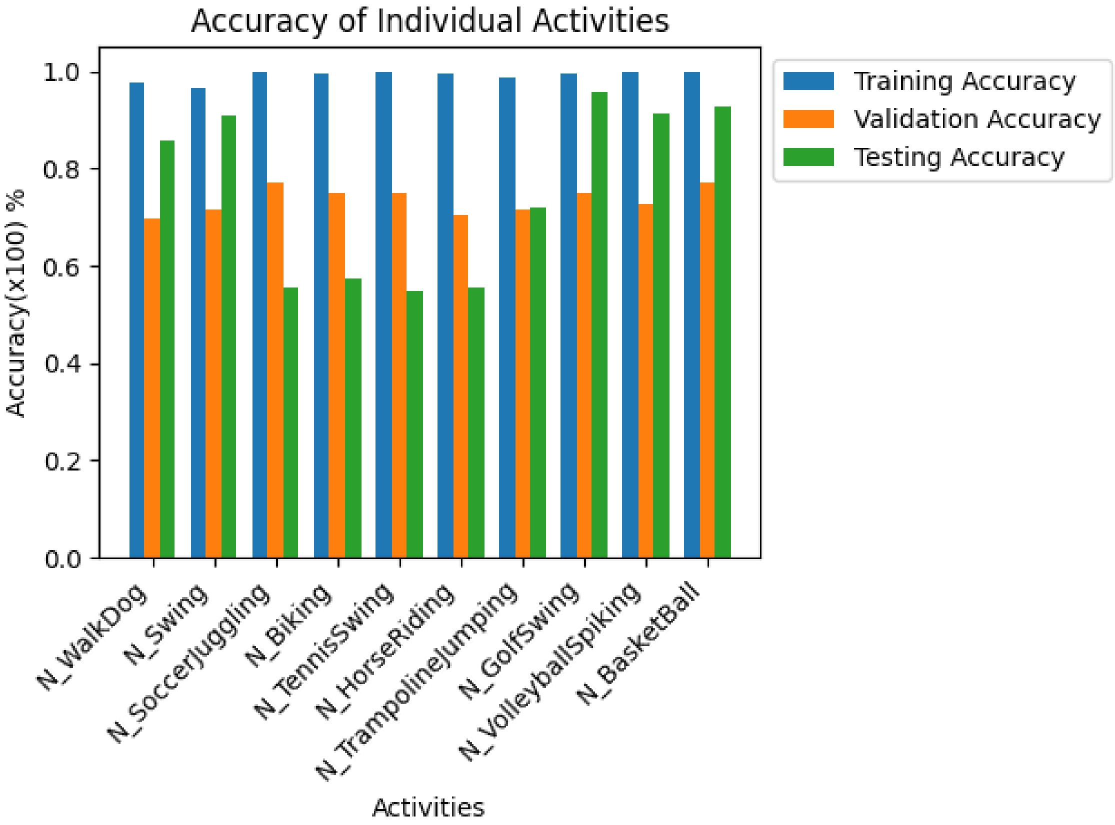


Figure 4.2: Accuracy for Individual Activities for UCF11

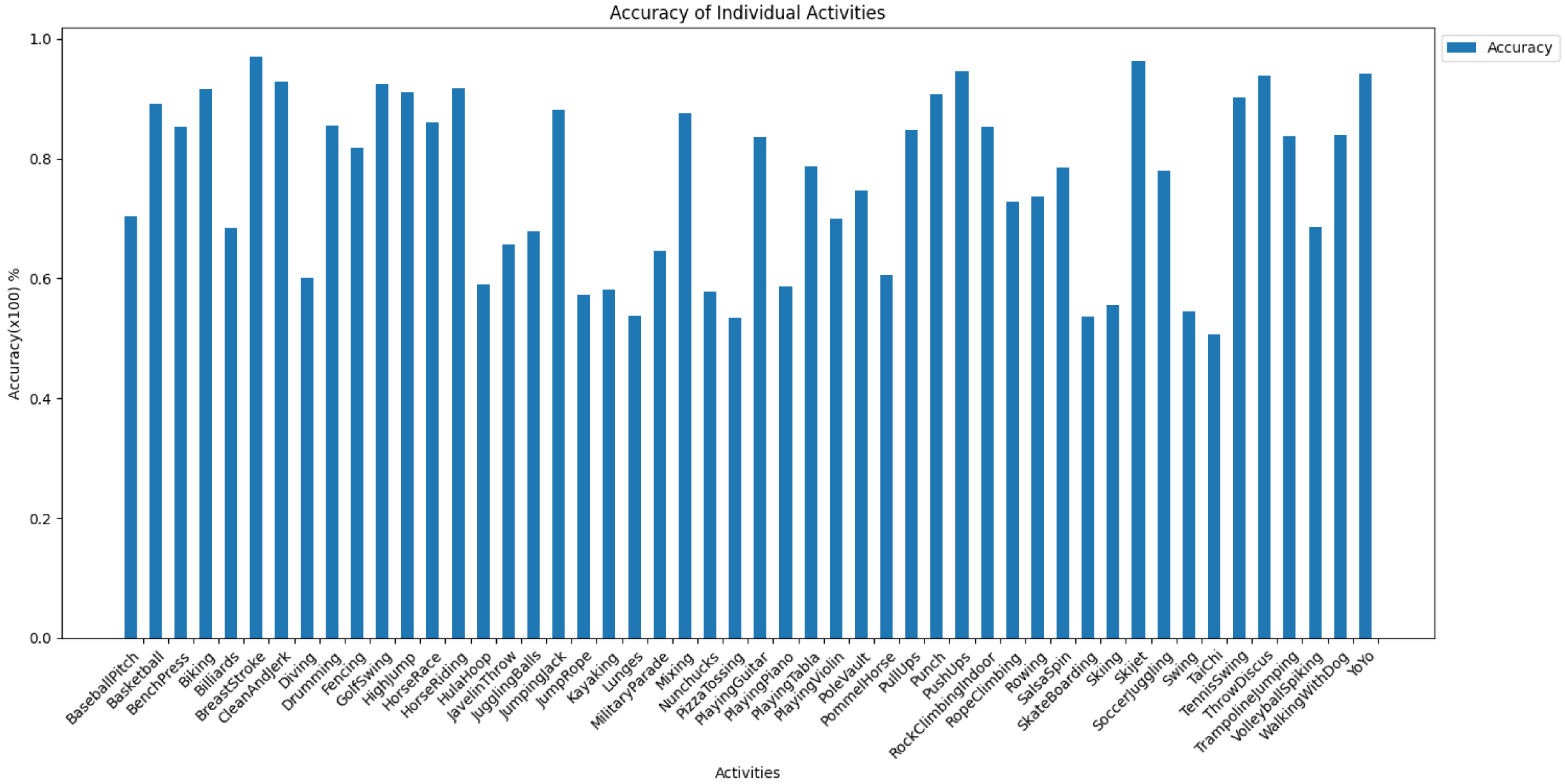


Figure 4.3: Accuracy for Individual Activities for UCF50

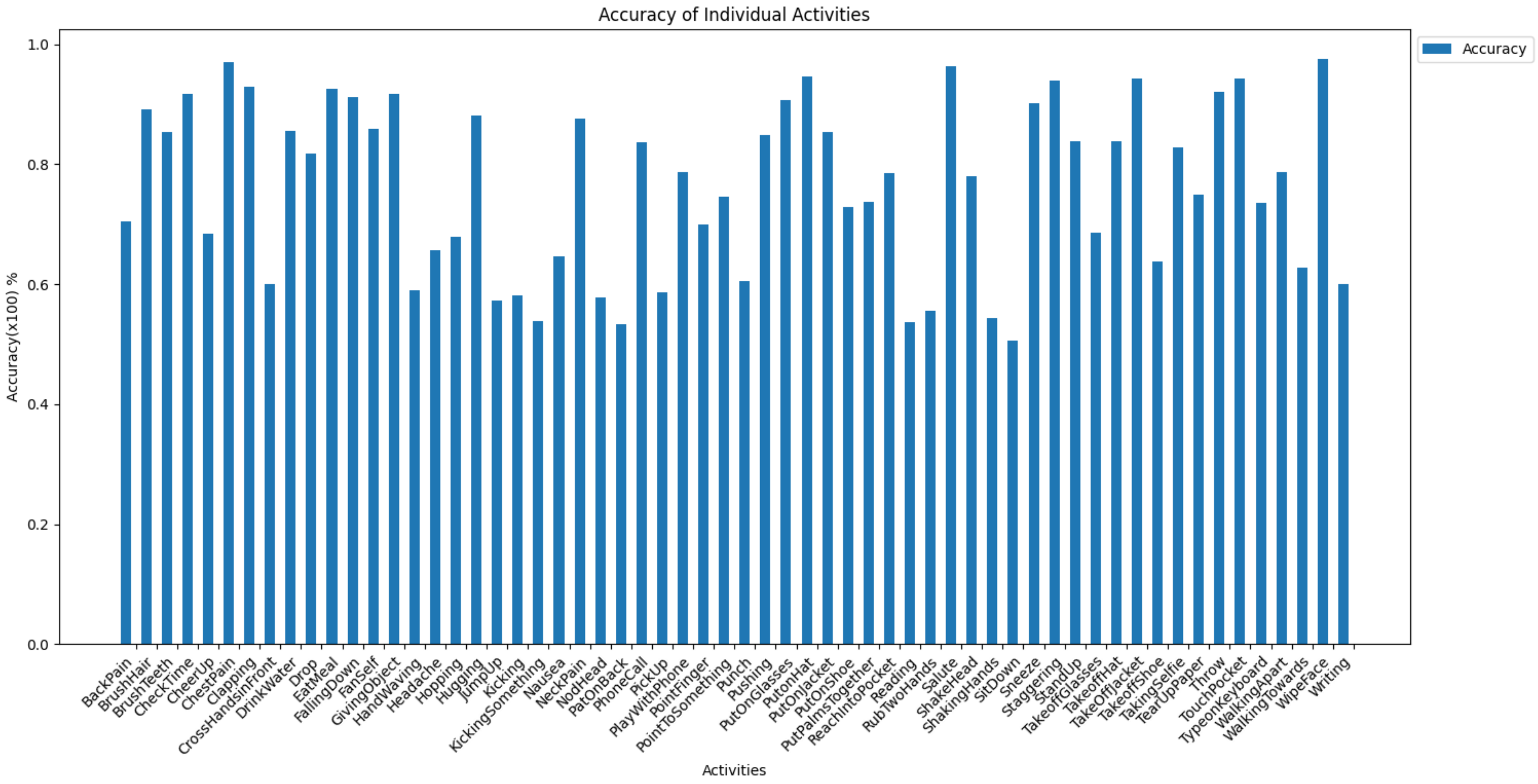


Figure 4.4: Accuracy for Individual Activities for NTU RGB

Figure 4.5 shows the graphical representation of the accuracy comparison of the proposed methodology with different state-of-art existing approaches [27, 14, 4] for Human activity recognition from videos.

As it can be clearly observed from Figure 4.5, our method outperforms the existing methods on four benchmark datasets: UCF11, UCF50, UCF101, and NTU RGB, achieving state-of-the-art accuracies of 98.3%, 97.1%, 92.1%, and 89.5%, respectively.

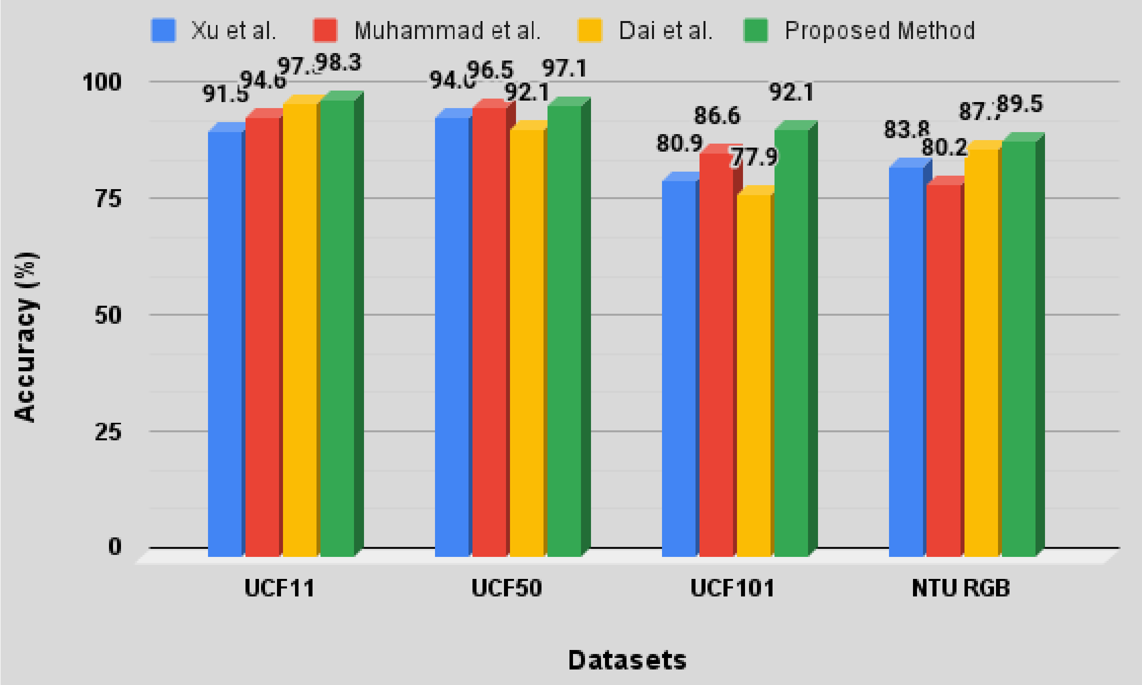


Figure 4.5: Accuracy Comparison with different state-of-the-art methods

4.3 Confusion Matrix

The confusion matrix is created to facilitate comparisons between distinct classes in different datasets. The confusion matrix for the UCF11 dataset is shown in Figure 4.6, and it can be seen that the accuracy distribution of our proposal is reasonably uniform, and the recognition accuracy rates of all classes are greater.

Figure 4.7 demonstrates the confusion matrix for the UCF50 dataset, which provides a clear understanding of the proposal's capacity to discriminate between various action classes. The figure shows that movements like pushups, golf swings, trampoline jumps, and punches have higher accuracy values, while others like soccer juggling, playing the violin, and pommel horse have lower accuracy values. Comparatively, our idea can improve the ability to distinguish between several related motions.

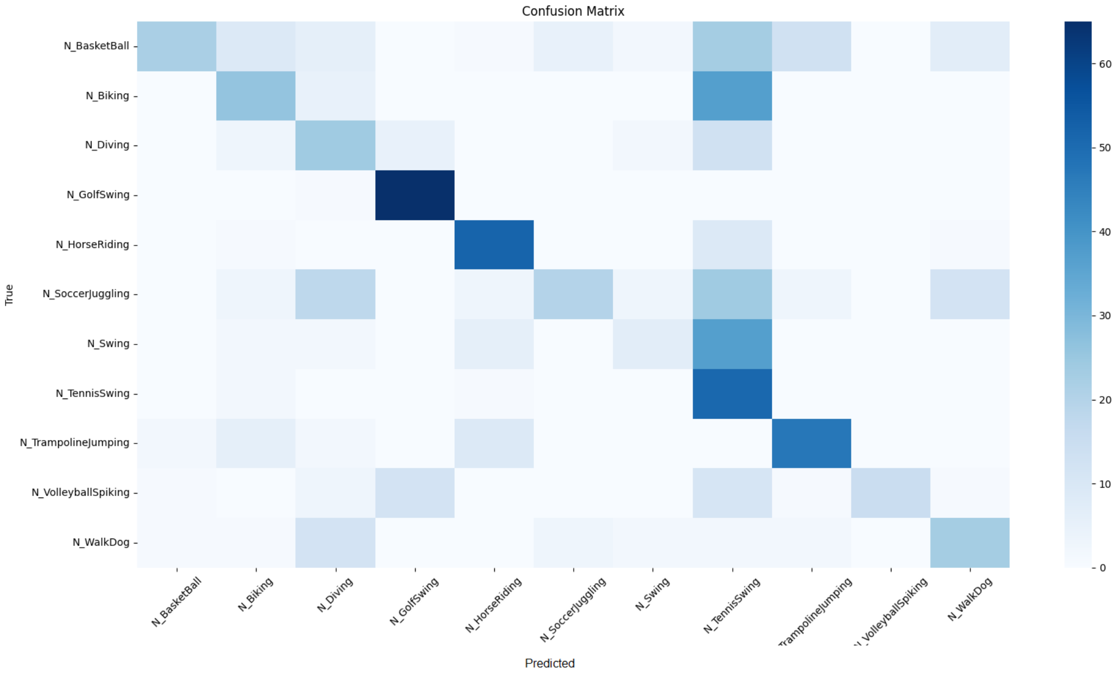


Figure 4.6: Confusion Matrix of UCF11 dataset

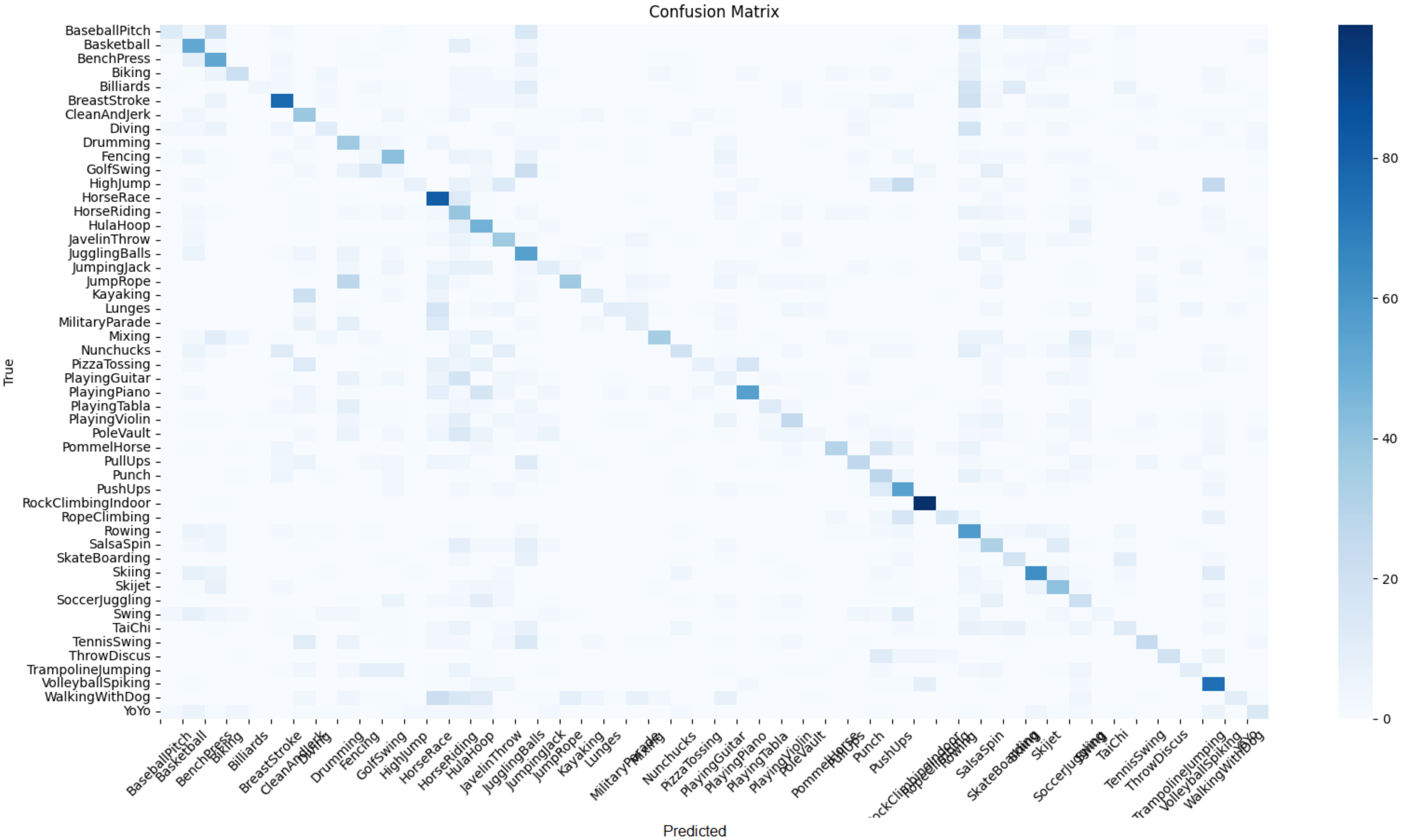


Figure 4.7: Confusion Matrix of UCF50 dataset

CHAPTER 5

Conclusion & Future Scope

We have presented a novel method for Human Activity Recognition using Two-stream Attention Based Bi-LSTM Networks (TAB-BiLSTM) in RGB videos. Our method leverages both the spatial and temporal information of human actions by using a CNN and an optical flow network as feature extractors for the spatial and temporal streams, respectively. Our method also employs an attention mechanism that can dynamically adjust the weights of the features based on both streams, enhancing the performance of the Bi-LSTM network. Our method outperforms well as compared to the existing state-of-the-art methods on four benchmark datasets: UCF11, UCF50, UCF101, and NTU RGB, achieving state-of-the-art accuracies of 98.3%, 97.1%, 92.1%, and 89.5%, respectively. Our method can be applied to applications requiring human action recognition, such as video surveillance, human-computer interaction, sports analysis, and health care.

In the future, the work can be extended so that multiple different activities in a single frame can be detected (recognized). Also, this work can be extended and used for Anomaly Detection in Video Surveillance.

References

[1] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy. Activity recognition with evolving data streams: A review. *ACM Computing Surveys (CSUR)*, 51(4):1–36, 2018.

[2] J. K. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[3] M. H. Arshad, M. Bilal, and A. Gani. Human activity recognition: Review, taxonomy and open challenges. *Sensors*, 22(17):6463, 2022.

[4] C. Dai, X. Liu, and J. Lai. Human action recognition using two-stream attention based lstm networks. *Applied soft computing*, 86:105820, 2020.

[5] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108:107561, 2020.

[6] S. Herath, M. Harandi, and F. Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

[7] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim. Robust human activity recognition from depth video using spatiotemporal multi-fused features. *Pattern recognition*, 61:295–308, 2017.

[8] I. Jegham, A. B. Khalifa, I. Alouani, and M. A. Mahjoub. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901, 2020.

[9] X. Ji, Q. Zhao, J. Cheng, and C. Ma. Exploiting spatio-temporal representation for 3d human action recognition from depth map sequences. *Knowledge-Based Systems*, 227:107040, 2021.

[10] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1996–2003. IEEE, 2009.

[11] X. Liu, Y. Li, and R. Xia. Adaptive multi-view graph convolutional networks for skeleton-based action recognition. *Neurocomputing*, 444:288–300, 2021.

[12] D. C. Luvizon, D. Picard, and H. Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5137–5146, 2018.

[13] D. C. Luvizon, D. Picard, and H. Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2752–2764, 2020.

[14] K. Muhammad, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, V. H. C. de Albuquerque, et al. Human action recognition using attention based lstm network with dilated cnn features. *Future Generation Computer Systems*, 125:820–830, 2021.

[15] T. N. Nguyen, S. Lee, H. Nguyen-Xuan, and J. Lee. A novel analysis-prediction approach for geometrically nonlinear problems using group method of data handling. *Computer Methods in Applied Mechanics and Engineering*, 354:506–526, 2019.

[16] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin. Video-based human action recognition using deep learning: a review. *arXiv preprint arXiv:2208.03775*, 2022.

[17] M. Ramanathan, W.-Y. Yau, and E. K. Teoh. Human action recognition with video data: research and evaluation challenges. *IEEE Transactions on Human-Machine Systems*, 44(5):650–663, 2014.

[18] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine vision and applications*, 24(5):971–981, 2013.

[19] A. Sarabu and A. K. Santra. Human action recognition in videos using convolution long short-term memory network with spatio-temporal networks. *Emerging Science Journal*, 5(1):25–33, 2021.

[20] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[21] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[22] L. Song, G. Yu, J. Yuan, and Z. Liu. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation*, 76:103055, 2021.

[23] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[24] A. Stergiou and R. Poppe. Multi-temporal convolutions for human action recognition in videos. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2021.

[25] S. Sun, Z. Kuang, L. Sheng, W. Ouyang, and W. Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

[26] P. Wu, S. Chen, and D. N. Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11385–11395, 2020.

[27] J. Xu, R. Song, H. Wei, J. Guo, Y. Zhou, and X. Huang. A fast human action recognition network based on spatio-temporal features. *Neurocomputing*, 441:350–358, 2021.

[28] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1112–1121, 2020.