HUMAN ACTIVITY RECOGNITION IN VIDEOS USING MACHINE LEARNING ALGORITHMS

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

DOCTOR OF PHILOSOPHY

In

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

By

RAHUL KUMAR

(2K19/PhD/CO/08)

Under the Supervision of

PROF. SHAILENDER KUMAR
Department of Computer Science & Engineering

Department of Computer Science & Engineering

Delhi Technological University

(Formerly Delhi College of Engineering)

November 2023

CERTIFICATE

This is to certify that the research work presented in this thesis titled "**HUMAN ACTIVITY RECOGNITION IN VIDEOS USING MACHINE LEARNING ALGORITHMS**" by Rahul Kumar (2K19/PHDCO/08) is an original contribution conducted under the guidance of Prof. Shailender Kumar from the Department of Computer Science and Engineering at Delhi Technological University.

This thesis has not been previously submitted for any other degree or diploma. Rahul Kumar has followed the prescribed Ph.D. rules and regulations throughout the writing process.

The thesis does not contain any classified information. Proper acknowledgment has been given for external sources through citations within the text and inclusion in the reference list. Direct quotations have been appropriately identified and referenced.

Date:

Prof. Shailender Kumar
Professor (Supervisor)
Department of Computer Science & Engineering
Delhi Technological University

DECLARATION

I, Rahul Kumar (2K19/PHDCO/08), hereby affirm that the research work presented in my thesis titled "**HUMAN ACTIVITY RECOGNITION IN VIDEOS USING MACHINE LEARNING ALGORITHMS**" is an original contribution, conducted under the guidance of Prof. Shailender Kumar from the Department of Computer Science and Engineering at Delhi Technological University. This thesis has not been previously submitted to any other academic institution for the purpose of obtaining a degree or diploma. Throughout the writing process, I have adhered to the prescribed Ph.D. rules and regulations set forth by the Institute.

I confirm that the thesis does not contain any classified information. Whenever I have utilized external sources, proper acknowledgment has been provided by citing them within the text and including them in the reference list. Direct quotations from external sources have been explicitly identified by quotation marks and have been appropriately referenced both in the text and the reference list.

Date:

Rahul Kumar
2K19/PhD/CO/08

ABSTRACT

Human Action Recognition research area is gaining interest due to its wide range of applications in the fields of elderly monitoring, suspicious activity monitoring, sports activity, pose estimation and health monitoring etc. The presence of a wide range of variations in normal human activities adds complexity to the recognition process. The use of automated systems is crucial in facilitating the increased utilization of cameras. These systems play a vital role in categorizing actions via the application of automated systems, namely Machine and Deep Learning. One of the primary objectives within the field of artificial intelligence is to create an automated system capable of effectively recognizing and comprehending human conduct and activities shown in video sequences. Over the last decade, several efforts have been undertaken to identify and acknowledge human activity inside visual sequences. However, this remains a formidable undertaking owing to factors such as the similarity of actions within the same class, occlusions, differences in viewpoint, and ambient circumstances.

Due to the various issues and research gaps that were discovered in vision-based action recognition during the literature review phase, different state-of-the-reference methods are reviewed with their methodologies. The researcher used multiple approaches, like handcrafted-based feature extraction and automated feature extraction approaches with deep architecture. In this thesis, the suggested methods are categorized based on the modality, feature extraction techniques, and classification approaches. This work describes the various datasets along with their specifications and limitations. The latest approaches functionality is represented with their performance parameters. This thesis includes different methods that use ML- and DL-based techniques, along with their accuracy.

Second, a Multi modal Deep Learning method is proposed for a Multiview dataset to deal with the Multiview problem. RGB, depth, and Skelton data are used in a multimodal base approach. A multi-modal based HAR approach is suggested. Depth, RGB and Skelton data are employed to evaluate the multimodal performance of the proposed approach. Depth Motion Map and Motion History Images are trained separately using the 5S-CNN model. On the other side, Skelton images are trained with the 5S-CNN and Bi-LSTM models.

In order to improve the rate of identification and correctness, the skeleton representation gets trained via the use of hybrid classification algorithms, namely the 5S-CNN and Bi-LSTM model. Next, the process of decision-level fusion is used to combine the score values obtained from three distinct movements. Ultimately, the activity of persons is determined according to how they combine value. To assess the effectiveness of the proposed 5S-CNN using the Bi-LSTM approach, an estimation is conducted.

A lightweight pre-trained model is suggested, which takes fewer parameters in comparison to other models and evaluates the model's performance on various parameters. The main objective is to assess a recent pre-tested model in HAR that takes less time and data in the training phase. These models are efficient for mobile edge devices, which require less computation power as compared to other traditional DL models. The suggested approach performs well compared to the various recent techniques.

This study describes the effectiveness of the vision transformer in action recognition. The sophisticated design of Vision Transformer models enables them to categorize activity properly. By using the UCF 50 dataset, we conducted an effective measure of these models and state-of-the-reference techniques to evaluate their respective efficacy. This research conducted a comparative study of several assessment measures, such as f1-score, precision, and recall, to assess the performance of the model. The proposed approach performs well compared to the state of the reference model.

This thesis concludes with a discussion of significant findings and prospective research directions in the domain of HAR.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENT

I am indebted to God under whose power I pursued this Ph.D. Thanks to Almighty for granting me wisdom health and strength to undertake this research task and enabling me to its completion. I pray that I can always serve him in what-ever he wants me to; and for which I need his blessings too.

I am grateful towards my supervisor, **Prof. Shailender Kumar** without whom this achievement would not have been realized. It was his valuable guidance and consistent encouragement all through my research period, which helped me to overcome the challenges that came in the way. This feat was possible only because of the unconditional support provided by his. A person with an amicableand positive temperament, he has always made himself available to clarify my doubts despite his busy schedules and I consider it as a great opportunity to do my doctoral programme under his guidance and to learn from his research experience.

My deepest thanks go to my parents for their unconditional love and support. Also, my heartiest thanks go to my family, my son Reyansh, my wife Neelam Chaudhary whose proximity, love & affection, and whose everyday prayers made it possible to complete this research work. I admire her sincere efforts in providing support at various stages.

My sincere regards to **Prof. Prateek Sharma, Vice-Chancellor, Delhi Technological University** for providing me with a platform for pursuing my PhD work. I express my gratitude to **Prof. Rajni Jindal**, DRC Chairman, Department of CSE, **Prof. Vinod Kumar**, Head, Department of CSE and **Dr R K Yadav**, Assistant Professor CSE for their kind support and providing necessary facilities to undertake this research. I take this opportunity to thankfully acknowledge **Dr Deva Nand Kamboj**, Assistant Professor ECE, DTU, all my teachers and lab mates Ankit Yadav, Aman Jolly, Rajiv Mishra, Shweta Gupta, Ananya Pandey, Anusha Chabra Abhishek and Ashish.

Rahul Kumar
2k19/PHDCO/08

Chapter 1

Introduction

Human Activity Recognition (HAR) became the focus research area in computer vision due to its wide variety of applications. Its primary applications include surveillance video, video labelling and retrieval, monitoring of patients, automation, and scene modelling, to mention a few. This chapter has explained the human activity recognition system's overview, basic terminology, conventional architecture, different problems and situations in video analysis, and numerous applications of human activity identification in everyday life. Furthermore, research problem statements, contribution, motivation, study implication and thesis overview are discussed.

1.1 Overview

In HAR, action is defined as an observable object that may be viewed from the people eye or any other sensor device. Action can be categorized into four different categories based on the person's body involved in the human activity.

- Gesture: Gestures are described as the motion of the body parts to describe tasks like sign language, head motion in negation, hand waving, etc.
- Actions: When an individual human performs, several gestures are known as actions. For example, jogging, walking and jumping.
- Interaction: Human-Human/Object is performing some action known as interaction. For example, playing guitar, washing clothes, two people conversing and cooking.
- Behaviour: Any unusual facial expression indicating a person's behaviour at that time can be normal or abnormal. A person shouting with an unpleasant facial expression shows anger.
- Group Activities: To accomplish a task by more than two people or groups of the person involved is known as a group activity. For example, playing football, military parades and protesting by groups.

Figure 1.1 represents the activity level with the complexity to recognize. Gestures and actions are fundamental to recognize. However, behaviour and interactions are intermediate. Group activities involving several people are incredibly complicated [1].

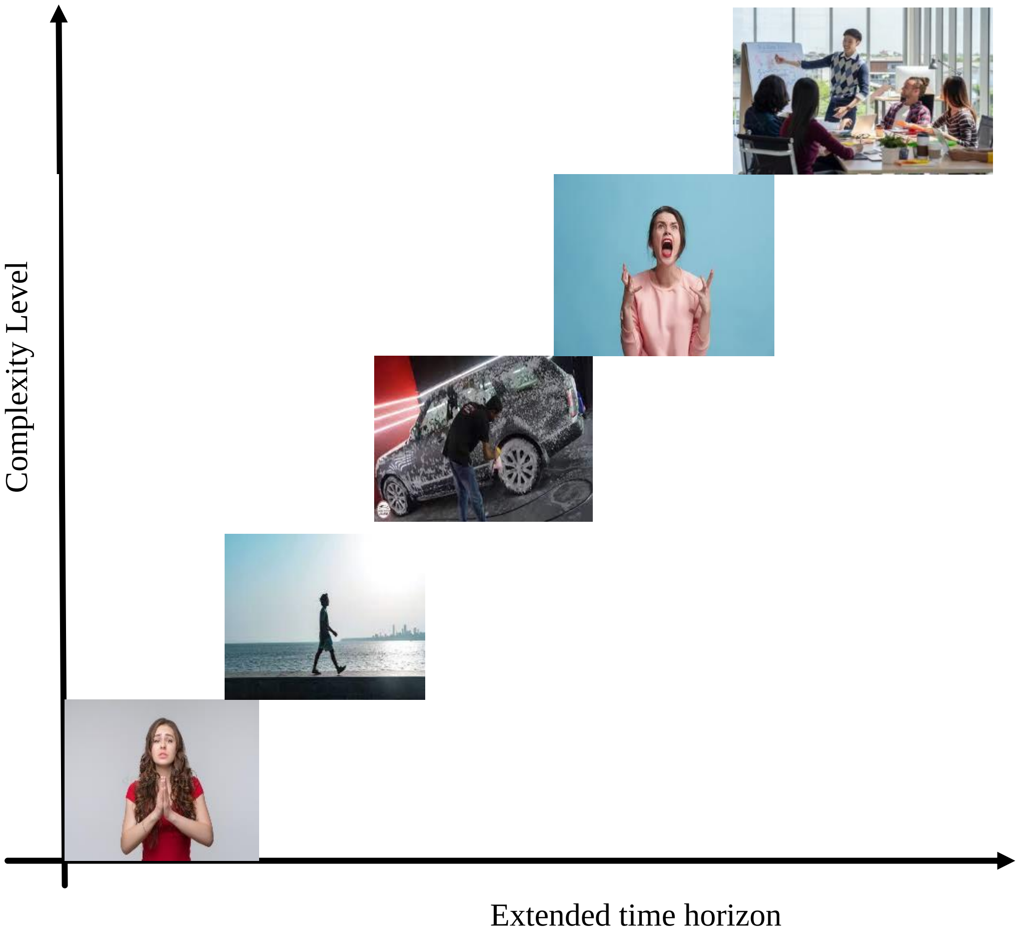


Figure 1.1 Classification of Human Activity

1.2 HAR Process

The HAR system is employed to analyze visual movements. After video data has been obtained, it is dealt with in accordance with the requirements of the which underlie application. The HAR process requires some basic steps to perform action detection like data pre-processing, feature extraction, action learning and action classification. Figure 1.2 depicts the complete step-by-step process involved in the HAR problem.

Data Pre-processing is used to improve the quality of input video sequences in order to extract robust characteristics. Background segmentation, silhouette extraction, histogram equalization, optical flow estimation, and other approaches were covered. Earlier human action recognition systems relied on processing human silhouette extraction to depict human motion. It involves background removal, frame normalization, and vector quantization in a restricted setting. The primary drawbacks of these pre-processing approaches are that they are less suitable for real-time applications and are less efficient in unconstrained contexts such as complicated and low-light circumstances.

Feature Extraction is the procedure to find robust features from the raw data to learn action efficiently. The utilization of handcrafted feature-driven approach representation is a long-used conventional feature extraction approach that has delivered impressive results in several HAR applications. The primary contribution of feature extraction is eliminating unessential information from raw data and finding spatiotemporal relationships to recognize human activity. Several handcrafted feature extraction methods are used, like interest point-based, pose-based, shape-based, trajectory-based, motion-based, etc. In interest point based, STIP[2] is used to show local image features. STIP uses extra dimension to encode images. Pre-processing, such as background segmentation or human tracking, is not required for STIP algorithms. The aforementioned attributes exhibit resistance towards scaling, rotation, and occlusion, although they do not possess viewpoint independence. In trajectory-based methods, Motion trajectories are derived by tracking joints whereby interest spots are located alongside the input videos utilizing optical flow fields[3]. Methods used in trajectory-based are Histogram of Optical Flow (HOF), Histogram of Oriented Gradient (HOG) and Motion Boundary Histogram (MBH), etc.

Handcrafted representations influence learning-based illustrations. Traditional machine learning algorithms rely on handcrafted feature representation. Deep Learning methods use automated feature extraction. Automated feature representation is dominating as compared to the handcrafted feature approach. Deep Learning methods use automated feature extraction. Automated feature representations are robust as compared to handcrafted features. Long Short-Term Memory (LSTM) or Recurrent Neural Network (RNN) works better to extract sequential information from video sequences.

In ***Action learning***, phase models are trained using the extracted features from data. Two Conventional methods for action learning are Machine Learning (ML) and Deep Learning. To circumvent the underfitting problem, these models are trained using a large quantity of data.

The ***Action Classification*** phase recognizes each action accordingly to train data. Various classifiers are used in conventional ML, like SVM, Random Forest, KNN and HMM.DL used the SoftMax layer to classify each action.

In the complete process, we can observe how DL-based methods dominate in HAR compared to ML-based methods. ML-based methods are not suitable for large-scale data.

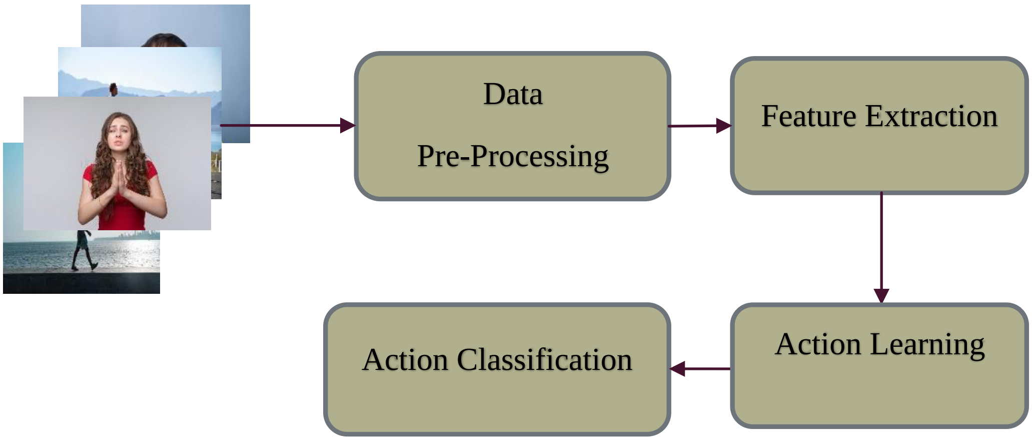


Figure 1.2 Process of Human Activity Recognition

We can suggest any novel or essential idea for any phase or collectively all steps of HAR. Because of continuous changes in the human body and surrounding conditions, every step has challenges and, for that reason, can introduce an effective method for HAR accuracy and processing speed.

1.3 Challenges in HAR

In vision-based HAR, we have several challenges due to the various aspects of the effect video stream, like background condition, view-invariant, lighting condition, camera motion, and noises in the visual stream. In vision-based HAR, we have several challenges due to the various aspects of the effect video stream, like background condition, view-invariant, lighting condition, camera motion, and noises in the visual stream. Any algorithm's effectiveness depends on the HAR dataset. If all sequences of action are captured in real conditions and manage all challenges, then this dataset will be more suitable to assess the effectiveness of the suggested approach. Accuracy is very challenging when we analyze and represent the interaction between the object and the person[1]. Furthermore, HAR approaches are still not designed to recognize numerous gestures under varying backdrop circumstances, and they are not able to deal of gesture scaling and growth [4].

1.3.1 Occlusion

Recognizing human action in the occluded visual streams is an arduous task. Other body parts or objects may also obscure the body components of the actor executing the action. When the leading actor's body parts are not clearly visible in video sequences, finding feature representations of occluded parts is very complex. In the domain of human posture assessment, occlusions may be categorized into two distinct types: self-occlusion and occlusion caused by external objects. Self-occlusion arises as a consequence of bodily parts colliding as a result of varying viewpoints, whereas other occlusions manifest when an object obstructs the visual field [5], [6].

1.3.2 Viewpoint Variation

The viewpoint and position of the camera can influence the action emergence and are very complex to recognize. The action view captured in the action dataset is the primary concern in identifying human action. In a Multiview dataset, information of view is more robust than in a single-view dataset. The unsettled camera view causes multi-view variance, which delivers a new viewpoint to all activities.

1.3.3 Intra-Class Variation and Interclass Similarity

It may be shown that various people executed the identical acts in different ways. When contemplating the 'running' motion, a person may go slowly, quickly, or leap and then run. This indicates that an action classification may include several styles performed by human motion. Similarly, the processing time of action varies according to position change. All of these elements contribute to interclass posture and appearance variances. Action classification is challenging when action classes have a nearness in video streams. Running, Jogging and walking look similar in several action datasets. Developing an automatic recognition system is very challenging because of these issues. Specific features must be captured from the action videos to overcome intraclass variation and interclass similarity issues [7].

1.3.4 Environmental Constraints

The environmental condition can change the perspective of the scene. Due to the light sources that generate shadow on the object, variation in illumination occurs. Weather and daytime circumstances have a tremendous impact on the scene and the artifacts formed; such illustration, an action captured during rain, varies radically from the identical movement taken in broad daylight or the sunset.

1.3.5 Data Imbalance

HAR methods exhibited prominent results on significant action datasets. However, it's very complex to generalize these methods on large-scale datasets. The Deep Neural Network method shows better results on large-scale action datasets. These models needed a large volume of data with labelling for learning. The importance of data is a big concern when working with deep neural models. Deep model learning is very typical due to the unavailability of labelled data.

1.3.6 Dynamic Camera

Furthermore, when the action has been captured, it's possible that there's intrinsic movement in the camera, which has a negative impact on motion features because erroneous motion patterns are added to the recordings. When the Camera is static, human action can be classified easily compared to a dynamic Camera because it adds variation in pose and illumination, which makes it more typical to detect movement.

HAR methods have various challenges. HAR methods manage all issues to give the crucial results for which they are introduced. The main aim of these methods is to provide surveillance, keep track of patients, and help the elderly together; with the growing installing prices arise some civil issues: receive by the community and privacy. Interpreting and distinguishing everyday actions in lengthy videos is a difficult task. It is because long-term recordings covering daily living activities are comprised of several complicated steps. Such tasks are challenging to describe due to their complicated structure and wide variety in how they are performed. Overlapping action is another issue because there needs to be a clear start and end time. This issue is addressed by [8]

Traditional Machine Learning (ML) algorithms classify action vaulted with a large set of actions executed in a constrained environment. When data is vast, the ML algorithm does not perform better. Controlling imbalanced data may be a challenge when employing typical

ML-based solutions. Furthermore, training using ML approaches might be affected from a sluggish learning rate, which is exacerbated by huge-scale learning (training) data, as well as a dismal recognition rate. The preponderance of the effort in ML-based HAR approaches is done via supervised learning[9]. Although this gave potential answers, one issue with this technique is that categorizing all the activities required a significant strive again to evaluate data.

Deep Learning (DL) based methods also suffer from challenges in HAR.DL approaches need a large amount of data to train the model; they perform poorly when data availability is limited. The DL method also suffers occlusion and illumination change conditions because their technique cannot capture exact features from data[9].

1.4 Application Of HAR

Researchers introduced numerous HAR models in the last decade due to various applications. HAR can be utilized in content-based video retrieval, abnormal action detection, healthcare monitoring, sports monitoring and education. Figure 1.3 shows the area of application where HAR plays a crucial role.

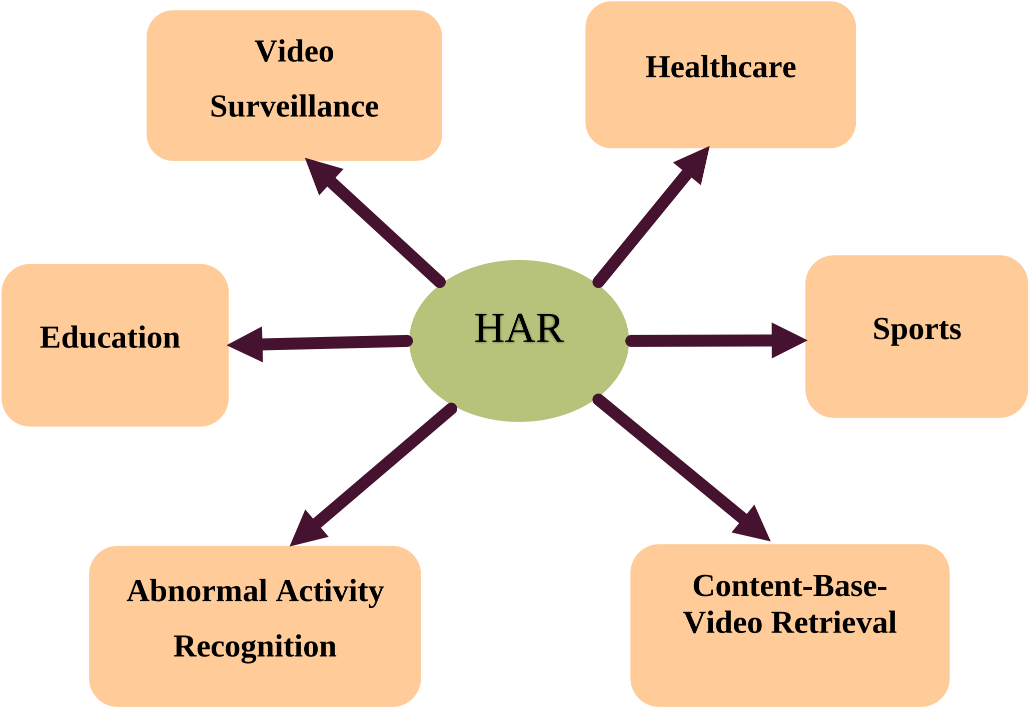


Figure 1.3 Application Area Of HAR

1.4.1 Video Surveillance

Usually, Video surveillance systems depend on a mesh of vision cameras controlled by a manual(person) observer who must be aware of activities happening in the region of camera sight. Due to the increase in the number of vision camera installation, the efficiency and effectiveness of person observers has been strained. Usually, Video surveillance systems depend on an area of vision equipped and controlled by a manual (person) observer who must be aware of activities happening in the region of camera sight. Due to the increase in the number of vision camera installation, the efficiency and effectiveness of person observers has been strained. Therefore, companies are seeking a vision-based solution to replace human base supervision. One such issue that has piqued the interest of vision experts is the automatic detection of abnormalities in a camera's field of view. A similar application includes learning activity patterns from lengthy footage to search for an activity of interest in a vast dataset. Figure 1.4 depicts the monitoring of railway station security.

Figure 1.4 Surveillance system

1.4.2 Condition of Background

The background condition matters for selecting an action, and sometimes the model does not perform correct classification due to a cluttered background. Backdrop clutter is defined as the movement patterns of background objects or people that detract from the primary activity of attention. In these instances, the foreground must be consistently split, and the foreground item must be monitored. However, such precise tracking may only be achieved on rare occasions because of low resolution and fluctuating background motion. The Weizmann dataset[10] has fewer changes in the action background compared to the KTH dataset[11]. In Visual streams, the backdrop has dynamic, static, densely occupied, weather affected and occluded variations.

1.4.3 Healthcare

Elderly care is a significant issue because they are vulnerable to sickness. Falling and abnormal behaviour are the primary concern, so continuous monitoring is necessary for older adults. Phone-based sensors can be used to monitor the behaviour of abnormal patients. Nizam et al. suggested a method for fall detection using Kinect Sensor data, which extracts a person's speed and pose-based features. The technique uses the actor's position in the subsequent frame to verify the spotting of the fall from abnormal activity. Nizam et al. suggested a technique for fall detection using Kinect Sensor data, which extracts a person's speed and pose-based features[12]. The method uses the actor's position in the subsequent frame to verify the spotting of the fall from abnormal activity.

1.4.4 Education

Action recognition from the visual plays a crucial role in education. HAR helps the institute record the behaviour of students and monitor of presence in the classroom. Various automated recognition systems are used in the classroom to monitor attendance. HAR is also used in the examination system to monitor any suspicious activity and laboratory of university and college.

1.4.5 Sports

Trainers in sports can only monitor some actions in the large videos and it is difficult for the audience to follow along with games constantly. Activity-monitoring algorithms to identify athletes' movements is referred to as human action recognition in sports. Warm-ups, fitness training, and sport-specific training, matches, or contests may all be tracked. As a result, HAR may be used to track professionals' progress (for example, jogging), assess the numerous movements done by different athletes or many performances of individual players (for example, backhand in tennis, serve in volleyball) to assist in training a technique or develop the style, and so on. Furthermore, some systematic statistical tests of a sports competition or individual athlete performances may be offered utilizing HAR[13].

1.4.6 Abnormal Activity Detection

In bus stations, City main areas, Airports and other places, security is very complex with a substantial quantity of surveillance equipment, so abnormal activity detection can be used to ensure security. A single Dynamic Oriented Graph is used to identify abnormal behaviour in three categories: person, group, and vehicle. Manual tracking for abnormal activity is difficult, so Automated HAR is needed to ensure security in crowded places. Usually, it comprises undesired behaviours such as theft, conflicts, human attacks, or any kind of violence that must be identified and controlled. Automated abnormal detected systems are trained on various datasets to identify a person's suspicious behaviour. Abnormal activity can be tracked even if the object follows the same path. For instance, a human traversing a rail corridor is uncommon, but a rail traversing a rail corridor is considered normal behaviour.

1.4.7 Content-Base Retrieval

Due to the widespread usage of multimedia devices in the modern day, video material is expanding rapidly. This information might be laborious and time-consuming to get manually. This method is referred to as Content-Based Video Retrieval (CBVR). In order to improve the user experience of video-sharing services, it has become crucial to produce the efficient indexing and archiving strategies. This consists of learning connections from unprocessed videos and summarising videos according to their content. With advancements in content-based image retrieval, content-based video summarization has attracted increased attention. The synthesis and extraction of consumer material, such as sports footage, is one of this technology's most financially feasible uses [14].

1.5 Problem Statement

Although the notion of HAR has been in use for more than two decades, cutting-edge methodologies show that the present classification techniques are inefficient[9] Based on the obstacles mentioned above, which include intra-class similarity, view variation, scales, shifting lighting conditions, clutter backdrop, and different forms of occlusions, it is necessary to design a HAR system capable of overcoming these limitations in video sequences. These obstacles include the selection of unsuitable dataset, lack of knowledge in advanced deep learning models, data collection errors, limited work on activity recognition involving more than one-person, limited work on diverse datasets, poor selection of classifiers, lack of ground truth, lack of knowledge for the selection of sensors and cameras, variation in viewpoint, execution rate, and limited work on multi-view activity recognition.

This work suggests multi-view and multimodal methods to address these issues effectively. The introduced framework used various modalities of data to understand human action. Several methods have been proposed to effectively tackle the challenge of action recognition in videos within a realistic context. We suggested a lightweight deep-learning model for HAR to address the computational problem. This model uses the pre-trained DL models for HAR.

This work also investigates the latest models for HAR, like the Vision transformer, efficient net and other models. In this work, various datasets have been explored with their limitations and the environmental conditions they captured. Compare numerous existing works with the methodologies and techniques authors use to resolve HAR issues.

1.6 Major Contribution of Thesis

The primary objective of the research study is to create and implement more sophisticated systems for recognizing human activities. These systems should be able to identify human activities at a better recognition rate. Also, the theoretical foundation for the enhancement of the performance of HAR approaches is provided by this work.

1.6.1 Theoretical Formulation

- View invariant problem in HAR detected and dealt with the multi-view model approach.
- Due to various challenges, more than the HAR single modality model is required to recognize the action accurately.
- It has been determined that the HAR model needs better identification accuracy in various complex scenarios.
- The difficulties of recognizing human movement under diverse illumination situations have been highlighted.
- Exhibited is the computation of the motion-based temporal information of a moving individual.
- Observations have been made on the performance of classifiers under different constraints of the activity performed.
- When available data is huge, the training model from scratch is complicated. This issue is addressed by transfer learning techniques to deal with vast and small datasets to avoid the overfitting problem.

- The performance of various transfer learning models was compared and evaluated on a challenging human action dataset.

1.6.2 Experimental Validation

The introduced framework's performance is evaluated using various parameters over available human action datasets. Person, several people, human-human interaction, and human interaction with objects videos are collected in complex scenarios with environmental conditions.

- A literature review was performed to identify emerging trends, challenges, limitations, handcrafted feature methods, deep/machine learning models, and publicly available datasets with their scope and future scope in HAR, which provided researchers with a brief knowledge of computer vision.
- A multimodal for HAR is developed, evaluated on a Multiview dataset, and measured in terms of effectiveness with the state of the reference models.
- Utilize a pre-trained model for action recognition, which reduces training time.
- Design a lightweight DL model using a fine-tuned pre-trained action analysis model.
- Studied various latest Transformer models for computer vision and evaluated the performance of these models in action recognition.

1.7 Motivation of Study

Over the past few years, understanding human activities in video sequences has been attached to complementary research such as human motion, segmentation based on semantics, objection identification, and domain adaptations. Currently, recognition of human actions can be learned automatically from multiple videos and applied across numerous everyday applications.

The number of uploaded videos on social media platforms such as YouTube, Facebook, and Twitter have increased rapidly in recent years. Due to the availability of low-cost, high-quality camera devices and fast internet connections in intelligent mobile phones, a large number of videos are uploaded annually to these social media platforms. Due to the vast quantity of data, there is a need for a system that can accurately analyze these videos and provide the required recommendations and solutions. Human action recognition is an essential component of these systems.

Robotics is an interdisciplinary discipline of science and engineering, as is evident. It focuses on the creation of devices that can supplant humans. Robots are a multipurpose mechanism that can be used in adverse situations such as explosives detection and de-activation or where humans cannot manage. In numerous robotics applications, human action recognition systems play a crucial role. Autonomous vehicles, which are a specific form of automaton that can govern the road situation and reduce the number of accidents, are one example. Autonomous driving necessitates accurate human pedestrian detection and body prediction, and it can avoid hazardous conditions.

Recognizing human actions in video sequences was prompted by the broader spectrum of applications, including robotics, human-computer interaction, and video surveillance.

1.8 Significance of Human Action Recognition

Intelligence-based AR is a growing field of study in computer vision and deep learning. Human action recognition's primary objective is to intelligently identify and analyze human actions from data collected by sensors, such as video sequences, depth sensors, and additional modalities. Numerous applications include monitoring and security, assisting medical care, the interaction between humans and computers, the field of robotics user-interface design, video perusing, sporting assessment, tracking of human objects, etc. Daily suspicious activities, traffic accidents, terrorist attacks, rioting, and stampedes are rising in the current social climate. Due to the vast amount of information extracted from video sequences, a HAR model effectively addresses such security concerns.

1.9 Thesis Organization

The present thesis is structured into a total of five chapters. The following section provides a concise overview of the outlines:

- **Chapter 1:** This chapter contains the introduction and background information about HAR. This chapter also discusses the general framework of HAR, the challenges associated, its applications, the process of HAR and various basic terminology. In addition, this chapter covers research problem statements, the significant contribution of the study, the motivations behind conducting the research, and the overall significance of the study.

- **Chapter 2:** This chapter briefly describes the existing methodologies and their limitations. We have reviewed various handcrafted approaches, including classification-based methods, automated feature extraction methods, and various available datasets for the HAR problem. Different modality-based practices are reviewed, like unimodal and multimodal. Furthermore, this chapter included evaluation metrics to evaluate the model's performance.
- **Chapter 3:** This chapter presents a multimodal approach for HAR. The suggested framework uses RGB, Depth and Skelton data. Multimodal HAR framework presented. After that 5S-CNN model used for training and compare with the state of the reference model.
- **Chapter 4:** This chapter describes the numerous recent deep learning models on the HAR.UCF 50 datasets were used to evaluate the performance of the suggested approach. A light-weight Deep Learning model was evaluated in this chapter to address the HAR problem. This model used recent Deep Learning models and compared them with various state-of-the-art methods.
- **Chapter 5:** In this chapter, the Vision Transformer model is used to evaluate the performance of the Human Action Recognition Approach. Vision Transformers have several variants, like Vision Transformers Large and Huge Pre-trained models require less training time than the conventional Deep Learning model because the pre-trained model has already been trained on a vast amount of data. This approach evaluates the performance of the Vision Transformer model for the HAR problem. This study helps the researcher evaluate the benefits of the latest Vision Transformer model in the domain of HAR.
- **Chapter 6**: This chapter describes the performance of the recent Transfer Learning (TL) model in the HAR problem. Various models' performance is discussed in this chapter. The UCF 11 and UCF 50 datasets are used to describe the effectiveness of these models. These models are comparable to the state of the reference model. This chapter provides a detailed view to analyze the TL model's performance and its appropriateness for the HAR problem. This chapter also discusses the use of the TL model compared to the DL model.
- **Chapter 7**: This chapter summarizes suggested approaches, limitations, and specific findings. This chapter also provides enhancement areas for the HAR problem.

Chapter 2
Literature Review

This chapter discussed the advantages and disadvantages of current state-of-the-reference procedures. The conventional hand-crafted and automated learning feature descriptors for representing human activities in video sequences have been examined. It assists us in identifying research gaps in current solutions in the relevant field. In addition, we compare publicly accessible human activity datasets. In addition, the research goals are developed based on the research gaps addressed in this thesis.

2.1 Introduction

Researchers have developed HAR from video and image data since the early 1990s. In addition, how the human visual system operates is one of the most important directions scholars have explored in AR. In a brief amount of time and at a low level, the visual system of humans is capable of receiving several observations regarding what is happening and the shape of the human body. Next, these observations are sent to the intermediate human perception system for class identification, such as taking a stroll, exercising, and racing. The human visual and perceptual system is robust and highly accurate in recognizing observed movement. In order for a computer-based recognition system to attain a comparable level of performance, researchers have exerted considerable effort over the previous several decades. However, due to the numerous challenges and issues related to human activity recognition (HAR), such as the intricate nature of the environment, variations within the same class, changes in perspective, obstructions, and the flexible structure of individuals and objects, we are still far from achieving a level of performance comparable to that of the human visual system.

When we look into the HAR process, the first step is how we can generate the best features for models to learn activity and the model could detect activity based on learning. In the last decade, numerous approaches to feature representation have been introduced. Based on these studies, we can broadly divide into two categories: Handcrafted-based action representation and automated feature learning methods. Figure 2.1 depicts the approaches for Action Recognition.

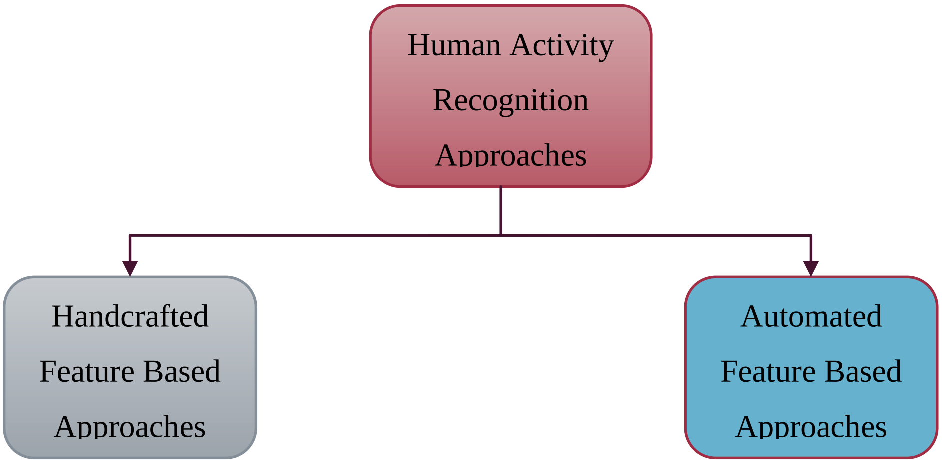

Human Activity Recognition Approaches
Handcrafted Feature Based Approaches
Automated Feature Based Approaches

Figure 2.1 HAR Approaches

2.2 Handcrafted Feature-Based Approaches

Representation of activity requires the pull out of a group of characteristics from global and local attributes. The purpose of the activity depiction challenge is to identify features that are insensitive to occlusion, changes in backdrop of images and perspective change. The essential classification of human activities Recognition techniques utilize a conventional strategy referred to as a hand-crafted feature-based solution. It has been prevalent in the HAR community over the last few decades and has produced some fascinating results with well-recognized datasets.

Approaches reliant on manually crafted attributes require human resourcefulness and preexisting knowledge to extricate unique features. These procedures require three significant steps: (1) Detection of the foreground in accordance with activity segmentation; (2) Feature selection and extraction by efficient method (3) Activity identification. Hogg [15] introduces the first step towards action detection from a video clip. To analyze human activities, a WALKER model constructed using 3D structural hierarchical modeling is presented. Rohr [16] introduced a similar method based on the linked cylindrical form to represent the limb connection for pedestrian detection. Traditional Handcrafted Feature base methods can be classified into Space-Time Base Approach, Appearance-based approach and other approaches shown in Figure 2.2.

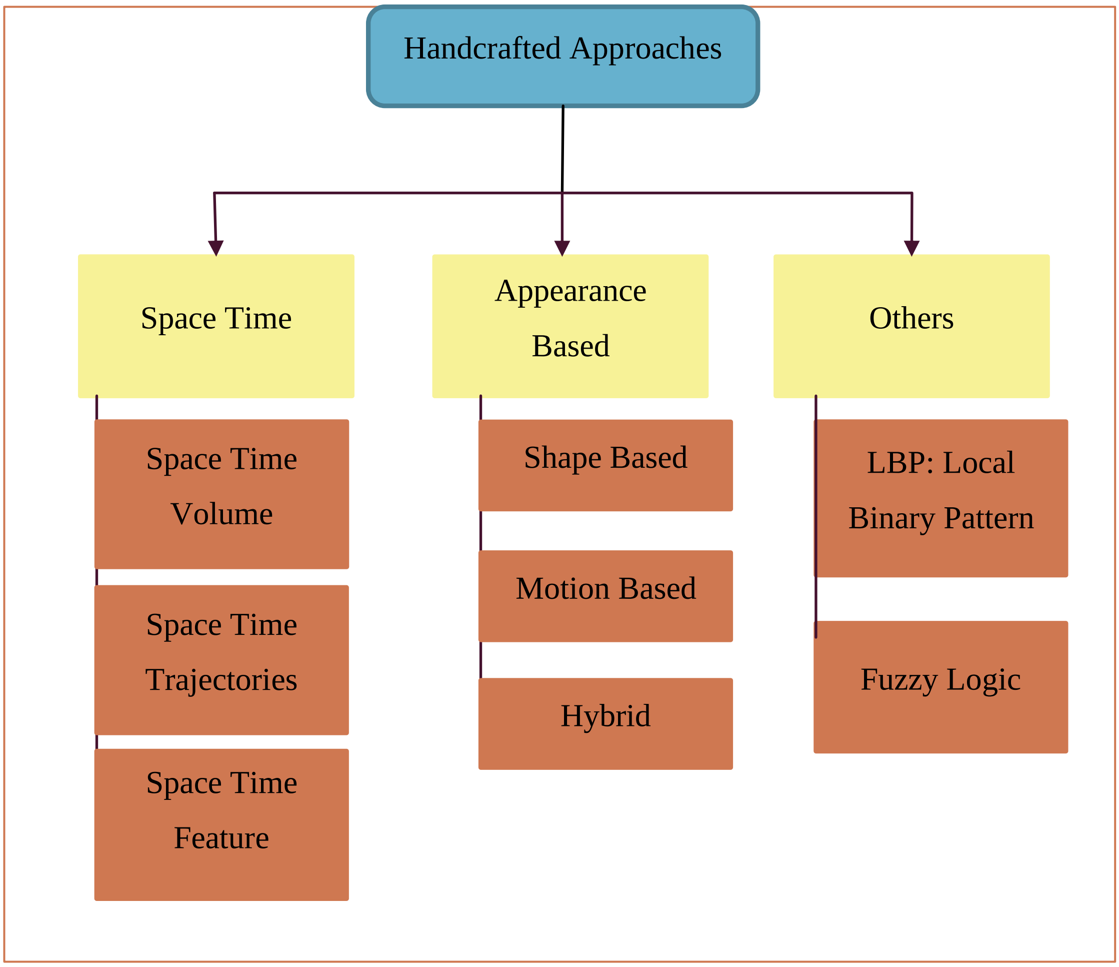


Figure 2.2 Conventional HAR Techniques based on Handcrafted

2.2.1 Space-Time Based

Using sparse and dense feature extractors in Space-Time-based methodology, local and global features are extracted from the input video. The collected characteristics are combined with methods and a vocabulary, which is then utilized to classify actions using supervised and unsupervised learning techniques. The space-time technique represents activity by considering its spatial extent, temporal evolution, and its associated characteristics. The technique is used to establish a correlation between inputs and their corresponding representative models, with the aim of discerning the activity class. As seen in Figure 2.3, it is evident that AR may be achieved by evaluating the shape similarities and appearance using these approaches.

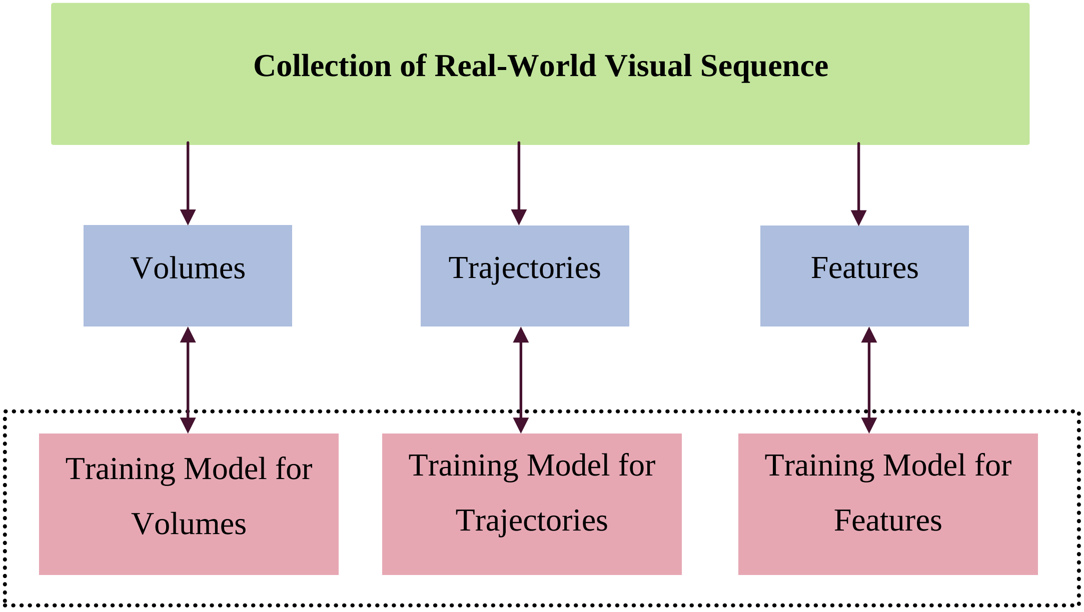

Collection of Real-World Visual Sequence
Volumes
Trajectories
Features
Training Model for Volumes
Training Model for Trajectories
Training Model for Features

Figure 2.3 Approaches for Space-Time

Space-Time Volume: The features within the space-time domain are depicted as space-time volumes (STV), which take the form of three-dimensional Spatiotemporal cuboids. The basis of STV-based approaches for action recognition is a likeness measure among two volumes. Bobick and Davis [17] did this by identifying real-time activity using two vector image components: Motion Energy Images (MEI) and Motion History Images (MHI). It develops a vector map for undetermined video frames, which is then compared to the representation corresponding to known motions. They have concentrated on the perspective of certain human motion activities, and action is taken into account across time. MEI and MHI templates both provide vital information about the object locomotive and eliminate the issue of a crowded backdrop in video streams. In [18]circumvent explicit flow computed by imposing a rank-based restrictions on spatiotemporal cuboids' the intensity information to ensure template and target consistency. This approach is validated on the Weizmann dataset and employed in sequences of videos to find behaviours of interest. It recognizes comparable actions and activities in video sequences despite visual variations caused by differences in clothes, backdrops, lighting, etc. There is no requirement for previous modelling or learning exercises. Volume-based techniques cannot function well in a congested context; they are only appropriate for basic action or gesture detection. Tian et al. [19] the gradient of the MHI template to improve activity recognition in a crowded backdrop by using the gradient. They identified the site of interest using the 2D Harris corner detector [20]at the high intensity point in the MHI template. In addition, Spatiotemporal characteristics are represented by the HOG model, and actions are identified using the GMM model. Blank et al. [21] presented the Space-Time 3D shape model of the MEI template using binary silhouettes, outperforming previous human action identification, detection, and classification methods. Their method does not need video alignment and is relevant to real-world conditions.

[22] use spatiotemporal volumes to distinguish view-variation human actions in visuals, using a similar methodology. Motion History-Volume (MHV) was created to describe view-free human activity in multi-calibrated, segmented backdrop films. In addition, the PCA (Principal Component Analysis) and LDA (Linear Discriminate Analysis) algorithms minimize the dimensions, and the Fourier transform eliminates the phase to identify the basic action classes.

[23] presented a maximum average correlation height (MACH) filter template-based approach for identifying video motion. Their system can solve the issue of intra-class variances with little computational expense. Yilmaz and Shah [24]suggested an action sketch for analyzing the 23 Spatiotemporal variations based on their differential geometrical surface features. Using the graph-theoretical technique, TV is constructed by stacking sequential contours along the time axis. Action sketch collected characteristics from the surface of STV in order to recognize activities and be invariant to perspective changes. During the years 2001 and 2008, global features-based action representation methods were the most prevalent. Nonetheless, local characteristics and deep learning-based techniques now dominate the area of action recognition research.

Space-Time Trajectories (STT): Trajectory-based techniques envision an activity as a collection of these trajectories in space-time. In these methods, a person is depicted by two-dimensional (XY) or three-dimensional (XYZ) dots that correspond to the location of their bodily joints. According to the nature of an activity, a person's joint placements alter as he or she does it. 3D XYZ or 4D XYZT STT represents these modifications. Tracking the joint location of the body allows the space-time trajectories to discriminate between various sorts of activities. This set of methodologies represents the joint locations of the body with either two-dimensional or three-dimensional points that are tracked through the frame sequences to calculate trajectories of the activity.

With these observed alterations in the posture are employed to generate the activity three-dimensional sight, that is a compilation of spatial-temporal trajectories. The above-mentioned techniques effectively recognize complex actions and resist noise, view, and/or illumination changes. Campbell and Bobick released an article [25] in the late 90s' that utilized tracing joint positions and sequences in a three-dimensional XYT plane; nine atomic movements of a ballet dancer are identified.

Rao et al. [26] presented a novel approach for learning activity that is independent of specific viewpoints, thereby avoiding the challenges associated with training models and mitigating the ambiguity inherent in action and trajectory recognition. Sheikh et al.[27] introduced a matching approach for 13 combined trajectories in a four-dimensional XYZT space. This approach has detected basic behaviours (seated, having to stand, and dancers) despite changes in perspective, anthropometry, and implementation rate. In contrast, Khan et al. [28] sought to identify complex behaviour in group activities (e.g., human marches) through a 3D polygon. A vertex of the three-dimensional polygon portrayed the representation of each group member. To obtain corner spot trajectories in a four-dimensional XYZT space, the movement of each object was continuously observed across consecutive frames and subsequently documented. The authors [29] presented a novel approach to representing human action, wherein they utilized a particle filtering tracking scheme to extract short trajectories in both spatial and temporal dimensions. A longest common subsequence algorithm was employed to conduct a comparative analysis of distinct sets of trajectories associated with various actions. Trajectory-based methods Action trajectories are calculated by monitoring joints containing points of interest with supplied image sequences using optical flow fields. Densely sampled objects are traced employing optical flow area to derive trajectories [30]. Trajectories are beneficial in situations where lengthy data is recorded.

Trajectories that are on the motion boundary are taken into consideration by the dense trajectory motion boundary-based sampling approach that was presented in [31] In addition, optical flow is employed to produce the motion boundary, and areas lacking motion boundary foregrounds are removed. Furthermore, activity detection is facilitated by a novel co-occurrence descriptor comprised of appearance and static information. Yu et al. expanded the concept of dense trajectories and integrated it with local and global motion points, resulting in a method where no background separation occurs during action recognition and dense trajectories are promptly retrieved and identified [32]. Mihir et al. [33] devised a method for dissecting visual motion into STT, which are then used by a DCS descriptor according to shear data, divergence, and curl. In addition, the VLAD coding approach is utilized for action recognition, making this technique well-suited for conducting complementary functions. Messing et al. [34] demonstrated distinguishing human motion in a high-resolution video clip using a generative mixture model based monitored critical points' motion record (history of speed). The suggested model retrieved and tracked the feature trajectories using Birchfield's implementation of the tracker KL [35], which monitored those points of interest with eigenvalues exceeding a specified range using frame-to-frame transformation and a consistency test. Human activities and interactions are detected utilizing trajectory attributes in [36]. Nevertheless, while camera motion and occlusions have existed in frame sequences, these features are not as strong; hence, sparse feature extraction is used for action identification [37]. Face recognition, scene restoration, and subspace clustering were all possible with the previous sparse coding approach.

Devanne et al. [38] advised applying a skeleton of human model to the data in order to capture the 3D coordinates of the joints and their mobility over time. Consequently, the examination of the geometric similarity among trajectories on a Riemannian manifold is simplified to the activity of action recognition. KNN has been utilized to recognize this manifold. Using Kendall's form manifold, Amor et al. [39] analyzed both static and dynamic human skeletal data recorded with a depth camera. In addition, they developed an automated method for action identification based on depth camera data.

Space-Time Features (STF): For HAR, STF based methods collect features from STV or trajectories. Frequently, these characteristics are local and include distinguishing aspects of an activity. Certain features may be classified as sparse or dense based on the attributes of STV and trajectories. Space-time features are local 3D volumetric characteristics on a scale of space-time. Mostly local features are frequently a suitable approximation for representing and perceiving the human actions, as approaches which relied on space-time features suppose that the 3-D STV is a rigid 3-D object and characterize individual action's 3-D volume by achieve with the operation of object matching. The space-time feature provides a straightforward representation of the 3D volume of human activity by eradicating object-matching issues. In [40] introduced, an event-based distance measuring method that made use of local characteristics at several temporal scales.

It requires no previous knowledge of the event model or segmentation of the background. Due to the operation of intensity gradients on several temporally scaled visual dimensions, this system's accuracy diminishes when the video's complexity increases. It is ineffective for video multiactivity recognition. Space–time point-based interest methods Space–Time Interest Point (STIP) may be used to depict a scene with local features. STIP characteristics encode images by adding a time dimension. Adding temporal domain information to spatial domain information enables the encoded picture to convey extra information about the action scene's contents and structure. By using clustering methods, STIP may be turned to saliency zones. The aforementioned characteristics exhibit invariance with respect to translation and scale yet lack invariance with respect to rotation [41]. This study [42] aims to detect and identify spatio-temporal interest points (STIPs) within multi-view images using a selective approach. This is achieved by applying local Spatial-temporal constraints to limit the search space and effectively suppress the surrounding elements. The intensity-based Space-Time Interest Point (STIP) algorithm exhibits robustness against shadows and highlights induced by disruptive photometric phenomena. Colour-based STIP outperforms intensity-based STIP. Soumitra et al. [42] introduced a facet model that makes use of STIP. The 3D Facet Model is used to identify STIP, and the 3D Haar wavelet transform is used to represent the discovered points. SVM is used as a classifier to acquire vocabulary for 200,000 descriptors for the purpose of categorization. Local space-time characteristics are retrieved using the Harris detector technique and the HOF is used as a descriptor, as described by the author in [43]. In addition, Least Square Dual SVMs are utilized instead of SVM for classification purposes, thereby accelerating the classification process by a factor of four.

Shillin et al. [44] suggested a novel approach for calculating STIP in time dimension using the Harris operator. Next, BOW and BOF are calculated and utilized for classification using the radial basis function in ada boost SVM. Using an interest point feature extraction technique, Niebles et al. [46] detected several actions in each visual sequence, resilient to scale changes and their location. Nevertheless, it does not provide perspective invariance since it represents the label of action by grouping cuboids into a series of visual code words.

Since essential action identification does not need a spatial-temporal arrangement in the local characteristics retrieved from 3D ST. In [45] author proposed a method in which the position of human joints is calculated and separated into five body segments, and additional data mining techniques are used to mines spatial–temporal pose structures for action representation. Additionally retrieved in one frame are spatial part sets corresponding to body part configurations and temporal part sets related to body part motions typical of human activities.

Including optical flow and foreground flow, Ikizler-Cinbis and Sclaroff have [46] researched the retrieval of features of different objects and individuals. To find the localization of interest spots, the retrieved features are put into instances of multiple learning frameworks (MIL). In their research, Minhas et al. [47] used the 3D dual-tree discrete wavelet transform (DTDWT) to eliminate spatiotemporal characteristics. Additionally, they utilized SIFT to remove local features. They employed a hybrid combination of the two approaches to input feature values into an extreme learning machine (ELM). While both SIFT and HOG are local feature detectors, SIFT turns an input picture into a large collection of local feature vectors and hence takes more processing resources. As a result, SIFT is incompatible with real-time applications. It was shown that the speed-up robust feature (SURF), an enhanced estimate of SIFT, runs quicker than SIFT while maintaining the quality of recognized spots [48].

Space-time features are perfectly adapted for essential datasets; for complicated datasets, a mix of features is necessary, which increases computing complexity. These restrictions may hinder real-time applications.

2.2.2 Appearance Based

The term "appearance-based approach" refers to methods of recognition that rely on an object's outward appearance; these methods may be further subdivided into "shape-based," "motion-based," and "hybrid" categories. Both 2D and 3D depth pictures are used, relying on either shape characteristics, motion features, or a mix of the two. Because to the development of depth cameras, these procedures have become much easier to implement. Because of the development of depth sensors and approaches for the estimation of real-time skeleton, skeleton-based recognition methods have also made rapid strides. Human stance, silhouettes, and appearance models are used in shape-based approaches to action identification. In 2D, people are shown as squares, whereas in 3D, they take on a more volumetric aspect. 2D (XY) and 3D (XYZ) depth pictures may be used to display the sight or viewpoint of any object, and these approaches depend on information relating to shape, motion and a combination of both.

These approaches integrate feature extraction techniques that are relevant to the feature space, such as optic flow for motion-based features and shape and contour-based features. The human body is composed of interconnecting limbs. This highlights the importance of accurately dissecting video sets to extract specific body components. Yet, the body may not be entirely visible in every video frame. So, it's a problem in the HAR procedure. Yet, several academics have presented techniques that might help fix this issue. Several action representation and identification techniques combine form and motion features. The utilization of 3D-based techniques [49] involves the creation of a human body model to represent actions. This model can be constructed using various geometric shapes such as cylinders, ellipsoids, visual hulls derived from silhouettes, or surface meshes.

Shape Base Methods: Using the human silhouette, localized features are extracted by subtracting the foreground silhouette using a segmentation algorithm[50]. The concept of positive space, which refers to the image silhouette, and negative space, which pertains to the surrounding area between the image border and the person, is commonly regarded within the realm of visual imagery [51]. The shape-based feature gives HAR information about the structure and movement of the human body, and the texture-based feature gives HAR information about motion in movies that use templates. Getting rid of the background might not be enough for outline extraction. Consequently, in [52] author utilises a texture-based segmentation approach for silhouette extraction. In the shape-based action representation approach, a person's Region of Interest (ROI) is determined using a silhouette representation. In addition to silhouette, pose invariant data are valuable for estimating human actions based on body form, and a contour-based technique leveraging multi-view key poses is employed for action detection. Shao et al. [53] introduced a method for segmenting distinct video input motions colour intensity-based and motion-based approaches. Activity was identified using dynamics and shape characteristics, with a description incorporating the local shape's spatial configuration. It is further expanded by contour point extraction from silhouette using a radial method for action representation and SVM classification. Utilizing scale-invariant silhouette characteristics, an additional method has been devised using pose-related data. Main key poses are generated by grouping these characteristics, which are then supplied into the action recognition weighted voting process[54]. Khan and Sohn [55]collected silhouette elements from videos to detect aberrant behaviours in the elderly. The R-transform is then applied to the features to generate scale- and translation-res) instant features. After the background has been processed, align silhouettes in a radial scheme, irrespective of shape and contour duration, in order to obtain the contour points of the human silhouette using binary segmentation [56]. This yields a brief overview of every feature obtained from one viewpoint. By displaying the total value associated with every radial bin, dimension reduction for each radial bin can be increased [57]. To segment and keep track objects undergoing prevalent shaped changes, the author utilized a shaped-based different levels set dynamic contoured structure in [58] to establish the object's boundary in the initial each frame. In [59] another research, the authors introduced the notion of finite-sized shape memory for the purpose of consistently retaining relevant shape information while simultaneously discarding irrelevant shape-related data. Using the technique provided by Ling Cai et al.[60], the object was tracked across low-contrast sequences by combining neighbourhood and boundary characteristics.

The main objective of the object shaped-based feature points the descriptor was to identify points of correspondence between object forms. It generates a log-polar histogram of the perimeter boundaries of a shape for each point. Edge structures closer to the reference point are sampled with greater precision than those further away, as the histogram categories expand as the radius increases. Most local feature representation techniques exhibit resilience against partial occlusions and noise.

Motion-Based Method: These methods employ features for action representation and a generic predictor for AR problems. This study focuses on the creation of a novel motion descriptor specifically designed for the representation of Multiview action depiction [61]. The histogram that represents the extent of motion is the focal point of this motion descriptor. Subsequently, a support vector machine is employed for classification purposes. Using an effective approach for identifying intelligent objects, the motion information of a moving target may be recorded. Motion tracking may be used for in-depth analysis of categorized items. Object detection and categorization compose the detection process. Exclusion of Background, optical flow, and spatiotemporal filtering may be used for object identification. Dynamic objects are recognized utilizing background subtraction via pixel-by-pixel or block-by-block differentiation of the consecutive frames and a background frame; here, motion is described by 3D Spatiotemporal data volume. The computing cost of this approach is modest, but it is subject to noise [62]. Hence, the optical flow methodology evaluates flow vectors of locomote objects to identify moving areas in pictures; nevertheless, these techniques are computationally intensive. The periodic characteristic of the visual data may be employed in motion-based processes to identify persons[62].

They use a vector picture framework, including MEI and MHI in [65], and implement a view-based technique for identifying human motions (MHI). MEI component is a binary template that emphasizes picture information places where there is motion. A region's shape can signify both the scene's activity and the viewer's perspective. MHI is utilized to demonstrate the movement of picture motion. MEI and MHI are vulnerable to problems with elimination of background. Using remanufacture and decay operators to depict MHI [63]Space–time silhouettes convey spatial information about human positions, such as location and action direction. In addition, they indicate the aspect ratio of the various bodily components at any given moment. Kliper-Gross et al. [64] created a technique for action recognition based on the essential features of movement encoding and localized variations in locomotive direction recorded using the bag-of-words methodology.

Hybrid-Based Method: Hybrid methods to action depiction Using hybrid action representation methods may improve HAR efficacy. Hybrid action representation may emerge in circumstances with comparable postures and actions, such as human activities. Hybrid approaches integrate both shape-based and motion-based information, namely by combining optic flow with silhouette-based features, to perform view-invariant action recognition. In [65], dimension reduction was also implemented using the analysis of principal components. A further approach for view-invariant action detection was introduced, which employed coarse silhouettes with radial grid-based characteristics and motion features [66]. These approaches to action representation combine shape-based and motion-based aspects. Several approaches to action recognition based on shape and motion data have been demonstrated [67]. Multiview [68] action recognition utilized coarse silhouette features, radial grid-based features, and motion features. Jiang et al.[69] employed shape-motion prototype trees to recognize human actions.

Jalal et al. [70], where input visuals are taken into account to derive multi-fused characteristics like joints in the skeleton and body shape-based features like HOG and DDS. DDS is used to capture the geometry of the whole body, and the Hidden Markov Model is used to classify actions (HMM). Nonetheless, these functionalities are implemented for simple activities. In[71] temporal segmentation to achieve action recognition. MHI is employed to describe form, whereas the Pyramid Correlogram of Oriented Gradients (PCOG) is used to describe features. Dong et al.[72] recommended a hybrid descriptor incorporating static data from the HOG descriptor and motion data from the MBH. The information obtained via dense trajectory sampling can be stored using the Vector of Locally Aggregated Descriptors (VLAD) technique, which is then used for action identification. The technique presented by Huimin et al. [73] incorporates information on morphology and motion. Identifying objects in frames by subtracting the background, extracting local characteristics such as shape and motion information, and global characteristics contour coding of a motion energy image. Utilizing a multi-class SVM classifier with a binary hierarchical structure, subsequent activities are classified.

2.2.3 Local Binary & Fuzzy Logic-Based Approaches

This course examines the two main activity recognition techniques, namely local binary pattern, in which vectors are generated using local binary patterns (LBP) and encoding geometric characteristics by their corners and edges. In contrast, fuzzy logic is based on our shared human comprehension of Boolean logic. FL mimics how humans make decisions, which encompasses all possibilities between the binary values YES and NO. In this fuzzy logic technique, rules for the activity identification of individuals from image sequences are developed.

LBP [74] is a descriptor for categorizing visual textures. Since the inception of computer vision, numerous updated versions of this descriptor [75], [76] have been used for numerous action classification-based problems. [77] illustrated a framework for HAR problem based on LBP merged with invariance of shape appearance and the patch matching. This framework was shown to be effective for AR problem when tested on numerous publicly accessible data sets. Using the LBP-TOP descriptor, another approach for action recognition was developed[78]. This framework divides the action scenario volume into numerous sub-volumes and generates the histogram of features by combining these sub-volume histograms. This method represented motion at three distinct levels: pixel (single node in the histogram), region (sub-volume histogram), and global (combination of sub-volume histograms).

Using kernel-based extreme learning machine learning (KELM), Chen et al.[79] evaluated Depth motion maps with local binary pattern action and presented a detection framework. LBP features from three distinct depth motion maps, namely front, side, and top, are fused at the feature level and then used to classify actions at the decision level using KELM. One experiment with two splits out of three is used for training, while the second experiment with one separation out of two is used for training. Several human action views have also been identified using rely on approaches. A method for Multiview human action recognition based on contour-based pose features and constant orientation LBP, followed by SVM classification, has been outlined [80]. Motion Binary Pattern (MBP) was recently introduced as a new motion descriptor for Multiview action recognition [84]. This descriptor combines VLBP with optical flow. On the Multiview INRIA Xmas Movement Acquisition Scenes (IXMAS) dataset, this method yielded identification rates of 80.55 percent.

Thanh et al.[81] established an approach centred on reconstructing texture models based on local binary patterns (LBP). LBP includes the mobility locations of the surrounding environment, and the self-similarity function is used to highlight the traced form from unconstrained films for action detection. Swarup et al.[82] put forward an approach for binary silhouette synthesis via a directed LBP comprising orientation and intensity variation information. This allows for effective Object identification. In addition, these features are combined with an Edge Orientation Histogram to produce an even more complicated feature, and SVM is then used for classification and training. Enqing et al. [83]worked with the production of a LBP by merging a SHI and an MHI to make it more proficient in holding additional motion information, coupled with using PCA as a descriptor. In article [84] author introduced a multiresolution visual analysis methodology and utilised LBP in conjunction with the Zernike moment.

For action representation and classification, conventional HAR in video systems utilize spatial or temporal information followed by a generalized classifier. Unfortunately, it is challenging to scale up these strategies to handle with the unpredictability and complication embedded in practical contexts. It is generally accepted that imprecise approaches are preferable for addressing these difficulties. The authors of this study have suggested a framework that utilizes fuzzy log-polar histograms and temporal self-similarities for the detection of human actions. The system incorporates SVM for the categorization of these actions [85].In [86] suggested a technique based on MEMS (multiple micro electromechanical systems) that retains numeric information collected by sensors of various everyday activities. Actions from everyday life are used to develop rules, which are then mapped to data using fuzzy rules in order to categorize the activities. Prudhvi et al.[87]n an inertial measuring system and smart sneakers. Four inertial measurement units are mounted bilaterally on a human's thighs and shins, and ground-notified forces are measured using an extended Kalman filter capable of identifying six fundamental activities, including getting up and down stairs, moving in both directions, sprinting, walking, and positioned. In [88] presented an intriguing method of online action detection using HMM batteries that takes into account all conceivable time constraints and action categories. Bayesian normalization is accomplished correctly by iteratively comparing the instance to the templates over sliding, overlapping temporal frames, making this method appropriate for real-time settings.

Using characteristics, in [89] author designed a high-level strategy for HAR. Attributes acquired manually based on intra-class variability are combined with data-driven attributes derived autonomously during training to create a highly informative characteristics collection. In order to assign accurate scores to the various features, a Latent SVM model is employed to learn the latent attributes.

The preponderance of HAR systems is perspective-based and can identify activity from a fixed vantage point. A real-time algorithm for identifying human actions must be able to distinguish the action from any angle. Several cutting-edge systems collect data from various cameras for this purpose. However, this approach is unrealistic due to the difficulty of calibrating multiple cameras in real-world conditions. A single surveillance footage should be the optimal solution for view-invariant action recognition. In this vein, a fuzzy logic-based technique [90] for view-invariant action identification with a single camera was developed. Before classifying viewpoints using clustering methods, the aforementioned method retrieved human outlines using the fuzzy qualitative Poisson human model for estimations of point of view. The findings suggest that the proposed approach for observing independent action recognition is highly effective.

2.3 Automatic Feature Extraction Based Method for HAR

Automated Feature Extraction is an alternative method for human activity identification instead of manually built feature-based methods. In addition, these methods may be classed as either non-deep learning or deep learning, as depicted in Figure 2.4. In recent years, DL-based algorithms have been in high demand because to their capacity for autonomous feature learning, which lowers the need for arduous human interaction. Automated learning capability offers the ideal solution and eliminates the dilemma of which attributes should be picked. The efficacy of systems for HAR systems primarily relied on the efficient and accurate depiction of data. The learning-based representation framework can automatically adequate the feasible features from the sequence of frames, introducing the concept of end-to-end learning, which entails the transformation from the input image to action detection, in contrast to a handcrafted representation-based approach, where handcrafted feature detectors and descriptors depict another activity.

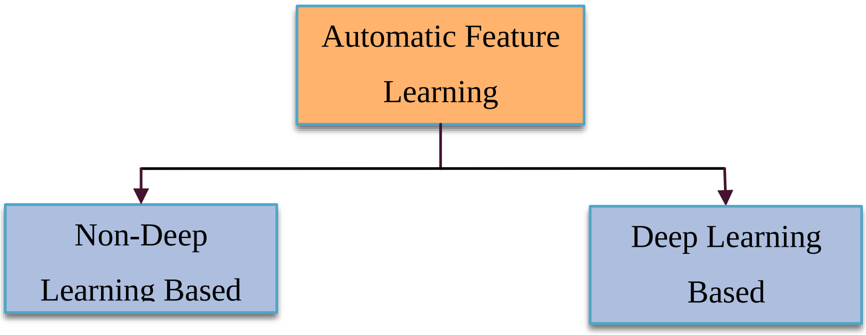


Figure 2.4 Feature Learning-based Approach

2.3.1 Non-Deep Learning Based

Generic programming and dictionary learning are non-deep learning methodologies that are employed to automate the process of feature acquisition from a given type of input data. The dictionary learning method uses a compilation of training examples to teach characteristics. Dictionary learning leads a collection of atoms so that a sparse linear combination of these atoms can effectively imitate a specific image. In contrast, deep learning approaches attempt to extract semantic feature representations via a deep network.

Dictionary learning is a kind of representation learning that involves the acquisition of sparsely represented representations. The sparse representation is suitable for classification applications involving images and videos. Dictionary learning approaches have been used in multiple applications for machine vision, including image classification and action identification [91]. The authors suggested an attribute dictionary learning technique for action recognition in information maximization. The use of the Gaussian Process (GP) model is employed in sparse representation to optimize the objective function of the dictionary. This enables the kernel locality in GP to enhance the intensity of the optimization process [92]. It was additionally capable of detecting invisible actions in videos.

Genetic programming is an effective evolutionary tactic derived from natural evolution. It can be used to tackle problems for which the solutions are unknown beforehand. Genetic programming may be used to determine the sequence of anonymous fundamental operations that would optimize the performance of a given task in recognition of human activities[93]. By combining data from smartphone sensors, genetic programming in [94] is utilized to recognize fundamental behaviors such as seated, rising, walking, and running. In addition, ML-based methods such as SVM and Naive are employed to evaluate the accuracy parameters. Feng et al.[95] suggested a genetic programming-based method for merging mobile accelerometers, gyroscopes, and magnetic sensor data. Cell phone data is also used to capture environmental raw data to identify seven fundamental action types.

2.3.2 Deep Learning Based

DL is an essential subfield of machine learning that seeks to learn multiple layers of collected data and abstraction that can actually comprehend audio, images, and text. DL-based methods can manage images/videos in their unprocessed forms and automate feature extraction, representation, and classification. These systems use trainable feature extractors and computational models with multiple processing layers to represent and recognize actions. For HAR tasks, research is being conducted on feature learning methods based on deep learning techniques due to their promising performance, robustness in feature extraction, and ability to generalize to various types of data. During the training phase, these methods intensively collect data. They require knowledge of multiple levels of abstraction and representation that will enable the extraction of features in a fully automated fashion. Methods based on deep learning may be regarded as trainable feature extractors that facilitate the identification of intricate, high-level actions. Nonetheless, the training phase's enormous computational complexity and substantial data requirements remain recurrent obstacles [1].

Generative Methods: Unsupervised learning is used to deputize for any distribution of unlabelled data. The data dimension was reduced by the new representation and conforms to its distribution. The fundamental aim of generative algorithms is to figure out the distribution of data, encompassing the attributes associated with each class, to reconstruct the authentic data distribution from the data used for the training set. When the training dataset is confined, the efficacy of the generative model is adequate [96].

Discriminative Methods: The discriminative algorithm is considered more spartan than the generative model, and when sufficient training data is available, it achieves outstanding results [96]. These supervised models use a hierarchical learning model with multiple concealed layers to classify input raw data into multiple output categories. Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN) are the most prevalent neural networks. Therefore, action labelling employs a sequential method.

2.4 Action Analysis Approaches

In addition to techniques founded on action representation, action analysis tasks are performed. While the low-level processes involved in recognizing actions may detect the motion of objects within the scene, this information is insufficient to deduce the action label. Action classification approaches are utilized for recognizing action sequences, utilizing traditional ML and DL approaches forming many of these methods. Figure 2.5 illustrates the inherent distinction between machine learning and deep learning approaches in the context of HAR methodologies.

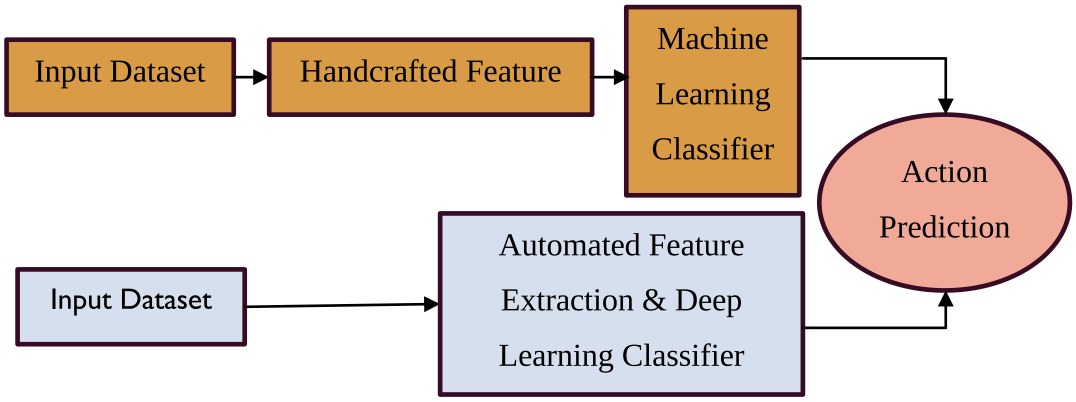

Input Dataset
Handcrafted Feature
Machine Learning Classifier
Action Prediction
Input Dataset
Automated Feature Extraction & Deep Learning Classifier

Figure 2.5 Comparison diagram showing how HAR uses both traditional machine learning methods and cutting-edge deep learning approaches

2.4.1 Machine Learning Based Approaches

ML is a subset of AI that seeks to construct an intelligent model by extracting unique characteristics that aid in spotting patterns in the input data [97]. Supervised and unsupervised machine learning (ML) methodologies exist. A mathematical model is constructed using the supervised method, which relies on the correlation between the input and output data. Without prior knowledge of the outcome, the unsupervised method seeks to discern patterns in unprocessed input data. Figure 2.6 depicts the classification method in ML.

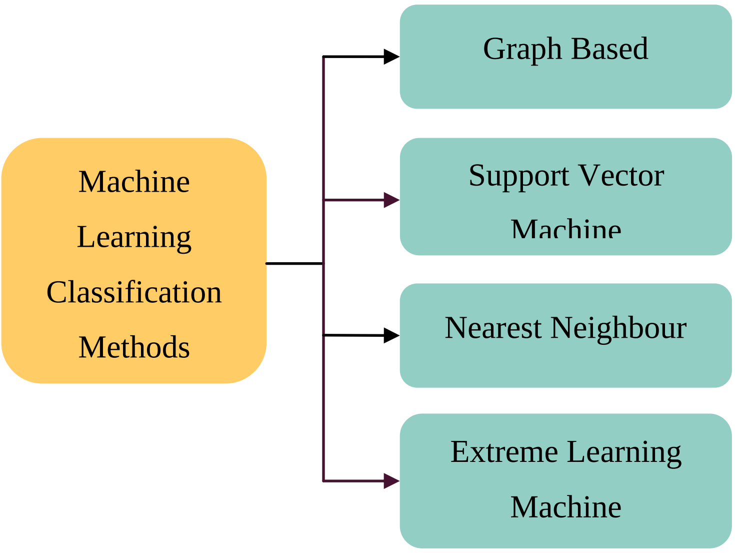


Figure 2.6 Machine Learning Classification approaches for HAR

Graph-Based Methods: HAR input properties are classified using graph-based algorithms like Random Forest (RF) and Geodesic Distance Isograph (GDI). Action identification accuracy is enhanced by using a graph of local action, with STIP features as the graph's nodes and edges denoting possible interactions.

The RF classifier is an ML technique that uses several decision trees to get a conclusion. The RF classifier for HAR can handle several inputs[98]. Ensemble learning methods are highly sought after because of their superior accuracy and noise resistance compared to individual classifiers. RF's benefits comprise less ratio of error, undertake the convergence (Overfitting removed), reduce training time (because to a concentration on fewer features) that yields higher effectiveness and less noise, and ease of use[99]. In [100] employ RF classifier and randomized DT to train depth-based features, namely APJ3D features. Joint feature APJ3D contains the joint's location and angle. In [101] the RF classifier is utilized to characterize actions in the accelerometer sensor dataset.

SVM is based on the separation of pieces of data using a hyperplane[102]. SVM algorithm is used to classify data points in a space with many dimensions. This is achieved by using a mapping technique that generates a linear decision surface for the input data, hence enabling effective classification [102]. SVM employs a kernel approach to cope with high-dimensional data and decrease computing load. Little sample sizes permit the use of SVM for HAR[73]. In [103], Author presented the BoVW model of local N-jet descriptors and vocabulary construction is achieved by integrating spatial pyramid and vocabulary compression, and two kernel-based SVM classifiers are used to classify human actions. Kim et al. [104] developed a method for optimizing and localizing human body components for a specified Region of Interest. In lieu of individual pixel classification from the input, the GDI (Geodesic Distance Iso) graph generates feature points at random, resulting in the computation of the total cost of the edges that connect two points along the shortest route. Next, the graph-cut method and SVM classifier are used to eradicate incorrectly classified feature points from the GDI graph that was previously created. In Shao et al.[71], SVM is fed the PCOG feature, descriptor of shape that generated from the MHI and MEI. A multiclass SVM with the RBF kernel is utilized to do offline training. Additionally, Sequence of the input training data are broken down into cycles throughout the span of each gesture to enhance the training technique. The Multi-class learning employing a one-versus-one SVM classifier with a polynomial kernel is utilized to classify nonlinear data [105].

Yu et al.[106] introduced the multi-class balanced Random Forest, a novel Random Forest structure. Several classes were predicted concurrently, so conserving memory and reducing computing costs. Using UT Interaction datasets, they attained a global accuracy of 91.7%. Oliver et al.[107] developed an active program that simulates the interaction between individuals and activities. The program identified and tracked human movement and to classify the motion, a feature vector is generated. After that feature vector was fed into a Hidden Markov Model, which was then utilized to activity classification. Robertson et al.[108] were the pioneers in modelling human behaviour as a sequence of stochastic acts. The recommendation in [109] was to match a human skeleton model to the data to characterise the 3D coordinates of the joints and their motion over time. Action recognition becomes a difficulty when comparing trajectory shapes on a Riemannian manifold. This manifold is identified using KNN.

Non-parametric nearest neighbour classifiers make classification decisions based on data without requiring training. Nearest Neighbour (NN) estimation is the most common non-parametric classifier. NN-image, a variation of NN, is used to classify images by comparing them to their closest image class. The results of the NN-image classifier are inferior to those of the learning-based SVM and DT classifiers [110]. In [111], the k-nearest Neighbour classifier is combined with the Relevance Vector Machine method, a kernel-based sparse model with comparable functionality to the support vector machine (SVM). RVM utilizes the Bayesian learning method and Gaussian prior for model weight estimation due to the possibility of overfitting induced by maximum-likelihood estimation of model weights. The positive values of these weights are indicative of the significance of the relevance vector in the class that pertains to the depiction of human activity. In the study by Vishwakarma et al. [112], important postures are retrieved as silhouettes and categorized using a hybrid SVM and kNN method referred to as SVM–NN. The SVM–NN methodology includes silhouette extraction and PCA for dimension reduction. The SVM's misclassified data are sent to the kNN classifier. In Xu et al.[113], skeletal features are mapped to the Li group for behaviour recognition and SVM is utilized to classify features that have been optimized by PCA after PCA preprocessing. Error value and radius value optimization improves the prediction performance of SVM.

2.4.2 Deep Learning Based Approaches

DL strategies and their automated learning capabilities have made it popular recently. CNNs and RNNs are used to automate the learning of features in deep learning approaches. Data representation, the foundational principle underlying DL, enables the generation of optimal features. It discovers undiscovered patches from unprocessed data devoid of any human intervention. DL is a subset of machine learning methods in which many layers of hierarchical information processing stages are utilized to detect patterns or pictures. It represents the amalgamation of research areas including neural networks, optimization, pattern recognition, signal processing, and image processing. Deep learning approaches deal with the issue from the start to the end, while machine learning approaches break the problem down into its component components and then integrate the findings. Due to a vast number of factors, training requires more time. During validation, deep learning techniques require significantly less time than ML-based approaches. One additional concern associated with deep learning algorithms pertains to their limited interpretability.

CNN creates maps utilizing neighbourhood-specific information. There are three processes in CNN design for extracting features: convolution, activation, and pooling. CNNs are models used for visual recognition that draw inspiration from biological processes. Variations of CNNs have been developed within this category. CNN has the capability to automatically extract intricate high-dimensional nonlinear features and comprehensively acquire knowledge of them. Ji et al.[114] looked into the utilization of CNNs in visuals for HAR. Author managed video frames as static pictures and used CNN to identify activities inside each frame. For the KTH dataset, the 3D CNN application gave an average precision of 90.2%. In contrast, 2D-CNN [115] is exclusively concerned with spatial domain convolution. According to Baccouche et al.[116] , the CNN convolution work is executed in the space and time domains. Due to this rationale, the 3D-CNNis capable of accepting input in the form of space-time volumes. Subsequently, the Long Short-Term Memory (LSTM) model training is conducted utilizing the extracted features obtained from the 3D-CNN. A 3D-CNN enables spatiotemporal information extraction from the input video. The posture-based features are extracted from a 3D-CNN in the study of Huang et al[117]. This network combines 3D pose, 2D appearance, and motion stream information. Incorporating colour joint features into the 3D CNN model introduces a significant level of complexity. As a solution, a 15-channel heatmap is generated, and convolution operations are performed on each map. Convolution networks for space and time levels are merged at the SoftMax level, whereas[118] studies produced superior results by fusing the final convolutional layer and class prediction layer can improve the performance of model. Two pre-trained ImageNet models take advantage of the space for 2-Dimensional features and temporal fusion. Madhuri et al.[119] suggested a DL-based framework using accelerometer sensors worn on the wrists of four distinct participants. After preprocessing, the CNN classifier extracts characteristics and classifies fundamental forearm motions. CNN-Bi-LSTM is trained in [120] to get the temporal information from visual data with the support of RGB frames. Late fusion is implemented at the network's decision level to offer end-to-end learning. Slow fusion hinges on linking two frames in both spatial and temporal dimensions. On the other hand, delayed fusion extracts motion-related information at the global level by combining two CNNs that are connected at a fully connected layer and apply them to two distinct frames.

Recurrent neural networks (RNN) have been extensively utilised to handle sequences because they retain the previous state and send it to the subsequent unit. To make them simpler to use, RNNs such as the Long Short-Term Memory Network (LSTM) and the Gated Recurrent Unit (GRU) are utilized. RNN differs from other models in that its artificial neurons are time-linked. When the outputs of one layer filter back into the inputs of an earlier layer, a recurrence is established. Du et al. demonstrated how a hierarchical RNN model can be utilised to develop a skeleton-based action recognition system. In addition, their recommended method was compared to five others deep RNN architectures that were derived from it. Their experiments utilized the MSR Action 3-D, the Berkeley MHAD dataset, and the HDM-05 dataset. The RNN tree framework was presented by Li et al. [121] as a model for adaptive learning for skeleton-based HAR. Their system classified activities according to category and employed RNNs arranged in a tree-like structure. Their method exhibited an accuracy score of 89.2% when applied to a novel dataset called the 3D-SAR-140 dataset.

Skeletal patterns are considered low-level features by numerous LSTM and RNN-based algorithms, which accept unprocessed skeletal data as inputs. Therefore, these models are not capable enough to derive high-level characteristics effectively[122]. However, CNN-based techniques are effective for image-based recognition applications[123]. They can immediately transform raw skeletal data into imagery and maintain spatiotemporal information. Due to differences in perspective and appearance, the effectiveness of such systems may not be realistic. CNN may be used with LSTM to account for spatial and temporal behaviour. To take advantage of LSTM + CNN is more vital to LSTM and suitable for 3D datasets[124]. Feature extraction, network training, and score fusion is used for an AR problem. CNN and LSTM are used to input skeleton-based properties of the spatial and temporal domains into the network. Temporal features relate to Time and trajectory, whereas spatial features refer to the relative location and distance between joints. Employing late fusion, many of these traits are fused together [124].

Correlational Convolutional LSTM was devised by Majd and Safabakhsh [125] by extending a pre-existing LSTM module to account for spatial and motion information in addition to building temporal connections (C^{2}LSTM) and represented. Their study was evaluated using two widely recognized comparison datasets, UCF-101 and HMDB-51 with achieved performance of network 92.3% and 61.0% (accuracy) respectively. Qi et al. [126] provided a novel method for creating a semantic RNN called stagNet, presenting the recognition of group and individual actions. The authors expanded upon their existing semantic network model by including the time dimension between video frames via the use of a structural RNN. The Volleyball dataset was evaluated with 90.5% group activity and 83.5% individual action using this method.

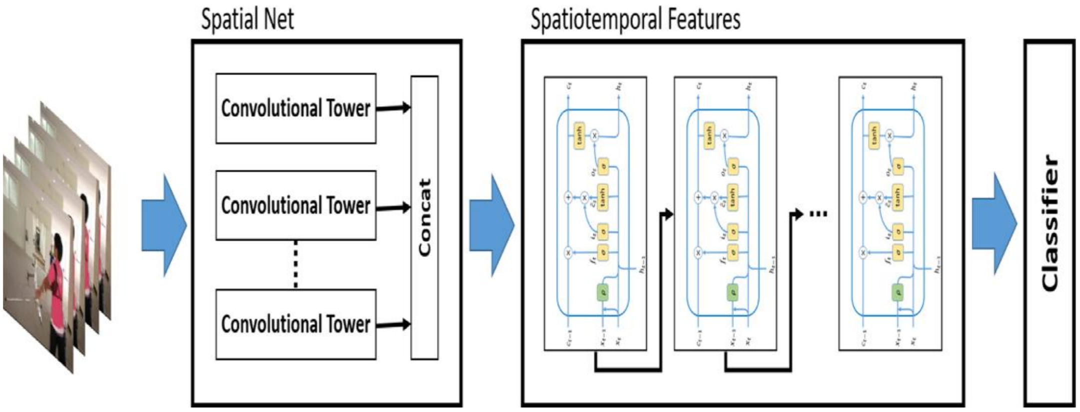


Figure 2.7 C^2 LSTM framework for HAR problem [125]

Every frame of image is input into a convolutional tower consisting of convolutional and pooling layers. Then, all tower outputs are concatenated and sent to the C^2 LSTM layer. A classifier, often SoftMax, classifies the retrieved characteristics into a label at the conclusion [125].

In [182] authors suggested utilizing the Quaternion Spatiotemporal (QST) CNN and LSTM, which appears to be referred to as QST-CNN-LSTM, on RGB data by incorporating its spatio-temporal information. LSTM is employed to distinguish between consecutive visual frames. The method extracts the regions of motions and outperforms on the UCF-11 and UCF- sports datasets. The authors of [185] demonstrate that already-trained weights can influence the learning of a model and that this issue can be resolved using a Bi-LSTM model. They propose employing an attention mechanism to rank human behaviours within an animated sequence. Using the RGB D mode, the following method generates a series of atomic 3D streams. On the basis of these 3D streams, the RNN is employed to predict categorization actions. This model is compatible with all input modalities, including RGB, RGBD, and depth, and it represents spatial-temporal links and long-term motion dependencies [127]. RNN networks like LSTM or GRU work exceptionally well in sequential data modelling applications when the dimensionality of the data is low. However, when high dimensional data, such as video content and text, is examined, these models do not perform well because of the enormous memory requirements throughout the input-to-hidden layer transition and the computational power complexity.

Veeriah et al. [128] provide an approach called Differential RNN (DRNNT using Spatio-temporal representation to depict actions. In contrast, the Back-Propagation-Through-Time

(BPTT) method is employed to train the network. Cross-validation accuracy is reported by Veeriah et al.[128] using random 16 participants for training and the remaining subjects for testing. The utilization of deep LSTM networks can facilitate the end-to-end action identification process by acquiring knowledge of feature co-occurrences derived from skeletal joints.

Auto-encoders receive data representation by unsupervised learning, with a focus on low-dimensionality data. Auto-encoders learn via hidden layers through the method known as encoding–decoding. An SAE (Stacked Autoencoder) is produced when these autoencoders are layered. Every layer is a fundamental auto-encoder model. The authors of [129] performed ongoing detection algorithm using auto-encoders with CNN, with CNN learning frame-level description. The autoencoder, therefore, does sequencing learning and feature dimension reduction. Then, contrary to CNN, another set of researchers employed similar approaches for HAR and autoencoders for anomalous action recognition, which avoids missing label concerns by learning spatiotemporal properties from data[130].

Gu et al.[131] developed a Stacked Denoising Auto-Encoders-based deep learning algorithm for Locomotion Action Recognition (LAR)[132]. SAE is a commonly used strategy for the unsupervised acquisition of new, valuable traits. Stacked Auto-Encoders strengthen the learned characteristics. In[133] combined a pair of auto-encoders and appended an additional SoftMax layer. The initial auto-encoder (AE) acquired knowledge using 561 features, allowing the subsequent AE to learn a compacted code of 5 features using 80 attributes. Backpropagation was performed on the network consisting of two AEs and the SoftMax classifier.

Generative Adversarial Network: Generative approaches employ unsupervised learning to discover data representations from any sort of unlabelled data. They were used for producing synthetic data by learning the characteristics of each type from the actual data. In the current era, there is a profusion of unlabeled data that lacks value; however, generative methodologies have enabled the processing of such data. GAN has been adapted by a group of scholars to the preliminary estimation of human movement to reduce motion distortion issues and foretell future motion [134].

2.5 Hybrid Model for Activity Recognition

Using both handcrafted features and Deep network models, hybrid models combine the advantages of both approaches. Simonyan et al.[135] devised a dual-stream CNN-based framework by decomposing visual information into spatial and temporal domains, then learning a CNN on optic flow. Alternative approaches, such as optic flow stacking, trajectory stacking, and bidirectional optic flow, have been proposed by the researchers. To compare the accuracy of classification, activity datasets are utilized for two-stream training.

In [136] used RNNs to identify activity by employing joint angles that varied over time as well as were expressed by a multimodal features matrix. In their model, RNN-based activity detection and recognition showed improvement than HMM- and DBN-based models on the Microsoft Research Cambridge Research-12 (MSRC-12) dataset. In [137] a hybrid approach was evaluated employing a deep convolutional network and a genetic algorithm. Initial population undergoes fitness assessment and crossover mutation, which results in the production of CNN classifiers. Combining scores from many action banks using a genetic algorithm is utilized for label prediction.

In[138], a unique technique to human activity recognition is described. This method focuses on visual analysis and extraction of features. GMM and KF methods are used to track human motions in the Mobility features (Figure 2.8). Using a model RNN with a Gated Recurrent Unit, the remaining attributes are derived from every visual attribute of each frame in the video sequence. This innovative method's primary advantages are analyzing and extracting every video frame and time-based characteristics. This hybrid model plays a crucial role in improving the performance of HAR problem. The suggested method is evaluated on the complex UCF-Sports, UCF-101, and KTH datasets. A mean of 96.3% is observed while evaluating the performance on the KTH dataset. Researchers have also employed combinations of diverse deep learning models to reach more remarkable performance in HAR.

A CNN and RNN hybrid model are described in [139] employing CNN for extracting neighbourhood relationships and RNN for discovering temporal associations among signal modalities. Initially the Fourier transform is used for the extraction of features, afterwards follows a series of layers of CNN for learning local dependencies. RNN layers, and lastly, the issue type determines the output layer. The model is considered unified since it has been suggested and assessed for both regression and classification issues. In [140], CNN and RNN are combined to create the hybrid deep model Deep Conv-LSTM, which exploits the capabilities of CNN to extract differentiating features automatically and RNN to learn the complicated temporal dynamics of time series. A variety of DNN architectures were examined by Kanjo et al.[141], which encompassed hybrid models that utilized sensor data from multiple channels. To evaluate models, the Env Body Sens dataset is leveraged.The precision value of the MLP model was 72.9%, the CNN model was 78.6%, and the CNN + LSTM model was 94.7%.

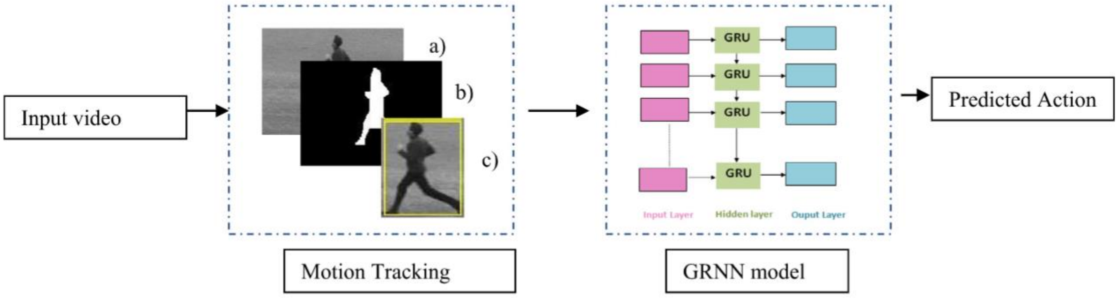


Figure 2.8 Shows the process flow of the action recognition prediction framework [138]

In motion capture, authors employ GMM and KF to: a) provide the frame of a runner in the KTH dataset; and b) monitor the mobility of an object. b) the GMM-based background subtraction approach for detecting human motion. c) The human motion bounding boxes derived from the Kalman Filter monitoring method. The GRNN model has three layers: an input layer, a hidden layer, and an output layer.[138] .

Yang et al. [144] proposed an effective approach for activity identification using depth maps. The depth maps also provide information on mobility and body contours. Using this approach, they created a Depth Motion Map (DMM) by projecting data onto the three orthogonal planes of a complete video clip. The histogram of oriented gradients (HMM) was then derived from each DMM corresponding to the video stream. For the MSRAction3D dataset, it has been determined that the given technique outperforms the previous method. In addition, they assessed the minimal number of frames necessary for the proposed technique to identify human activity. Then they determined that thirty to thirty-five frame sequences are enough for obtaining comparable findings.

2.6 Modality Based Human Action Recognition

Methodologies of human actions Depending on the type of the detector, divides [142]recognition into two primary categories: Unimodal techniques that treat an action as a collection of visual properties and enable identification from single-modal data. Many modalities are used as a result of the employment of multi-modal approaches, which incorporate data from several sources.

2.6.1 Unimodal

Approaches for unimodal in nature human activity identification identify human activities using information from only a single modality.

RGB data comprises three streams and is an essential component. Compared to the skeletal and depths modality, RGB data has several key features, including form, colour, texture, and flow. Owing to these qualities, several texture-based or shape-based feature extraction approaches, such as 2DCNN for single image-based spatial feature extraction or 3DCNN for volume-based spatial-temporal feature extraction, might also be directly applied to it. Activity recognition datasets are easily accessible to the public, which is another incentive to pick RGB data. Khan et al.[143] also proposed an activity recognition fusion scheme based on DNN features and multi-view. The VGG-16 pre-trained network is used to retrieve the DNN features, while the horizontal and vertical gradients and vertical direction features are used to extract the multi-view features. After obtaining all the features, the optimal set of features has been retrieved with the support of three parameters: mutual information, relative entropy and strong correlation coefficient (SCC). The final optimal feature set is then provided to the Naïve Bayes classifiers for accurate activity recognition. The presented approach has been implemented on the five popular activity recognition datasets namely: KTH, IXMAS, YouTube, UCF sports and HMDB51 datasets and achieved 97.0%, 95.2%, 99.4%, 98% and 93.7% recognition accuracy, respectively. Li et al. [144] postulated that a spatio-temporal attention network (STA) accumulates the discriminant features at the segmenand stream levels since a 3DCNN is limited to using equal frames for collecting spatial and temporal data from a specific video sequence. The suggested STA network can readily discern spatial and temporal aspects concurrently, enhancing the 3D convolutions' ability for learning. The described method has been applied to the HMDB51, UCF101, and

THUMOS 2014 activity recognition datasets and obtained 81.4% and 98.4% precision on HMDB51 and UCF101, respectively.

Wang et al. [145] introduced a lightweight activity detection framework built around CNN LSTM nets and a temporal wise-attention model for use with RGB movies. CNN was initially employed in this configuration for detecting objects from the rest of the scene. Subsequently, two LSTM structures were implemented on two separate CNN layers—the completely connected layer and the pooling layer—in order to get mobility patterns.

Next, a temporally aware attention model was used to learn the most significant characteristics from the video frames, followed by an optimization module to comprehend the internal relationships between these LSTM models. The provided method was assessed on three prominent datasets, namely UCF Sports, UCF101, and UCF11, and obtained recognition rates of 91.89 percent, 84.10 percent, and 98.76 percent, respectively.

Depth data is a different kind of modality resistant to varying lighting conditions, partial occlusion, texture and colour, and has no better resolution than RGB data. The pros of depth data are that it is trustworthy and offers correct estimations of silhouette and 3D structural details in terms of the locations of the skeleton's body joints.

As color-texture characteristics are absent from the depth data, the CNN, which relies heavily on texture characteristics, becomes less discriminative. Especially in comparison to RGB data, specific depth data is relatively limited and insufficient for learning discriminative representations directly from pixels. These models may be susceptible to overfitting while developed and tuned with inadequate data.

Using depth data, Wang et al.[146] applied ConvNet to distinguish human activities. They used depth motion maps (DMMs) derived from depth data using three distinct methodologies. First, the virtual camera is spun to capture the various perspectives. Second, Pseudo RGB pictures are generated from DMMs using optimal encoding to identify the space-time characteristics contained in the edges and textures, and third, three ConvNets are applied independently to the colour DMMs generated in the previous phase. The provided technique is assessed on three public datasets, including MSRAction3D, UTKinectAction3D, and MSRAction3D, and outperforms previous methods in terms of accuracy. Le et al.[147] introduced an activity recognition approach based on depth data. Using several 2D planes, they translated 3D data to 2D data. Subsequently, they collected dense trajectories with discriminative characteristics from each 2D plane. Several trained classifiers are used to integrate the projected scores in order to get a conclusion. A greedy-based algorithm is applied to discover the best collection of trained classifiers. The suggested approach is evaluated on the MSRAction3D dataset and produces superior results compared to baseline methods that typically do not use multi-projection-based features.

Liu et al.[148] developed a multi-view and hierarchical categorization framework for 3D human activities. In this framework, a 3D picture of human activity was projected onto three coordinating planes so that a 2D image could be obtained for each plane and inserted into the appropriate three subnets. When the number of layers in the subnet rises, the structure of the subnets is fused hierarchically, and the final structures of the depth video are fed into a neuron with a single layer. Using two publicly available datasets, the performance of the offered method is examined, and the findings illustrate the computational efficiency of the suggested model.

To make use of local Spatio-temporal characteristics and collaborative depiction of classifiers using regularised least squares, Liang et al. [149] introduced a technique for activity detection using a multi-layered depth motion map and multi-scale histograms of oriented gradient (HOG) descriptors. The described process correctly describes an activity sequence's local temporal motion and structural alterations. In addition, they proposed an analytical approach for collaborative representation utilized to reduce recognition computing costs. For MSRAction3D and MS Gesture datasets, the efficacy of the recommended method has been assessed.

The skeleton data is distinctive from RGB and depth maps and provides the location of the human body's joints. The locations of the skeletal joints are regarded as 3D data and offer a rich collection of features for activity detection in video sequences. Skeletal data is insensitive to variations in lighting, camera viewpoint, and movement pace. The ultimate focus of the skeleton-based technique is to turn the skeleton pattern into a two-dimensional representation with colour and texture, from which spatiotemporal patterns for activity detection may be retrieved. Du et al. suggested a Deep convolutional neural network for skeleton-based activity identification in [150]. They generated a vector by putting the joint coordinates of each skeleton sequences into a matrix and reordering them in chronological order. The produced matrix is then transformed into an image and normalized for further processing. CNN is then used to extract characteristics and recognize the final image.

Hou et al. [151] have provided an efficient approach for obtaining spatial-temporal information through skeletal optical spectrum color texture images (SOS). Subsequently, a Convnet is used to the picture acquired in the previous stage to identify the discriminative characteristics for robust activity identification.
The experimental outcomes on three publicly accessible datasets indicate the effectiveness of the provided method. The Skeleton Net architecture was developed by Ke et al.[152] using deep learning and 3D skeleton-based activity identification. Initially, the body-part characteristics of each frame were retrieved. These characteristics are independent of scaling, translation, and rotation variables when compared to the original coordinates. To extract temporal information, retrieved features are translated into pictures and inserted into a deep learning network developed for this purpose. The proposed deep learning net has two components; the first portion extracts generic characteristics, while the second section extracts discriminative and dense representations for robust activity identification. The conveyed technique has been evaluated on three activity recognition datasets, including CMU, SBU Kinect interaction, and NTU RGB+D datasets, and achieves the highest recognition accuracy.

Recurrent Neural Network (RNN) is ideally adapted for acquiring temporal features through 3D skeletal joint locations. This technique extracts appearance and motion characteristics and feeds into an RNN for temporal evolution. [153] presented an end-to-end hierarchical RNN method capable of learning extended temporal sequences for 3D skeleton-based activity recognition. Based on the human's physical structure, the skeleton is separated into five pieces and then integrated into five subnets. As the number of network layers increases, the subnet removes the representation from the input and combines them for the uppermost layer. The expression of the skeleton sequence is then input into a single-layer neuron, whose output represents the final choice. Lastly, five distinct RNN designs based on our model are compared to illustrate the efficacy of our proposed strategy. In addition, they reach the suggested technique to three more available datasets. The experimental outcomes of the recommended process outperform the cutting-edge techniques.

Zhu et al.[154] demonstrated an end-to-end fully associated deep LSTM network for activity identification, employing 3D skeleton joints to learn long-term temporal sequences automatically. As joint co-occurrences intrinsically describe human actions, a unique regularisation approach has been developed to understand the co-occurrence characteristics from the 3D skeletal data. In addition, to quickly learn the deep LSTM network, the author presented a unique dropout method that may operate in parallel with cells, gates, and LSTM neurons to perform more sophisticated activity identification tasks. The effectiveness of the suggested framework was assessed on the three HAR action datasets. Shahroudy et al. [155] devised a part-aware(P) LSTM model for HAR as an alternative to maintaining long-term memory in the cell. In the P-LSTM paradigm, memory is distributed across part-based cells. A series of experiments determined that the context of part of the body and the outcome of the P-LSTM unit are more effective for complicated activity recognition tasks.

Zhang et al. [156] introduced a semantics-guided neural network, a simple but effective 3D skeleton-based human activity identification (SGN) technique. To enhance the feature representation, a combined semantics at a high level has been established. The authors offered two modules, including a module at the frame level and a model at the joint level, to develop the connections between the joints in a single frame or a series of frames. The given method performed better on the NTU60, SYSU, and NTU120 datasets. Huang et al. [127] also proposed the view transform graph attention recurrent network (VT+GARN) technique, which employs skeleton characteristics for a more accurate spatial-temporal representation. The provided method eliminates the influence of perspective change on the joint locations of the spatial-temporal skeleton and quickly learns the activity representation for classification. Extensive trials conducted on three activity recognition datasets, including NTU RGB+D, North-western-UCLA, and UWA3DII, indicate the effectiveness of the proposed approach. Ahad et al. [157] obtained kinematics posture features (KPF) from 3D skeleton joint locations using a KPF extractor. This technique established the Linear Joint Position Feature (LJPF), Angular Joint Position Feature (AJPF), and the slope among bones portions. LJPF and AJPF were merged for every visual frame to understand the movement pattern in the temporal domain.

A linear SVM, Conv RNN, and CNNRNN framework are employed for classifying. The suggested method is used to evaluate five standard datasets, including UTKinectAction3D, Florence 3D, MSR3DActionPairs, Kinect Activity Recognition, and Office datasets.

2.6.2 Multimodal

Activity recognition employing multi-modalities integrates the characteristics of many modalities, such as RGB, Depth, and Skeleton information, to execute activity detection tasks reliably and effectively. Duan et al.[158] introduced an RGBD-based activity recognition. They exhibited an interoperable consensus voting network with two streams for RGB data and a 3D depth-saliency ConvNet network for depth data. The 2SCVN network integrates the long-term and short-term structure of RGB sequences, and depth maps are additionally coupled with RGB sequences to lessen the influence of the backdrop. The given technique excels on the Chalearn Iso GD dataset and obtains the highest recognition accuracy of 96.74 percent. Zho et al. [159] presented a three-dimensional convolutional neural network (3DSTCNN) architecture for activity identification utilizing depth and skeletal data. Each independent stream is supplied the original depth map, depth motion maps derived from depth data, and 3D skeleton sequence for global space-time feature learning. Utilising depth and skeleton data, all streams have been designed to learn space and temporal aspects effectively. The suggested method is assessed using three activity recognition datasets, including UTD-MHAD, MSRAction3D, and UT-Kinect Action.

Singh et al. [160] unveiled a multi-modal human activity recognition technique that employs RGB, Depth, and 3D joint coordinate information. Deep bottleneck multimodal feature fusion (D-BMFF) is a technique that uses all available data concurrently. The 3D joint coordinates are converted to an RGB skeleton MHI (RGB-SklMHI) before being combined with RGB and depth frames. The recovered features are then used to train a deep neural network. M-DCA is used to connect the characteristics obtained from the bottleneck layer right before to the top layer. Then, a multiclass SVM is employed to classify the features into several activity classes. The provided method is assessed on four activity recognition datasets, including UTKinectAction3D, SBU Interaction, CAD-60, and Florence 3D.

Wu et al. [161] suggested a DL-based numerous stream framework that can retrieve several visual features leveraging CNN for multimodal extraction of features. The discovered feature data is supplied into an LSTM model, which subsequently integrates this knowledge for HAR after utilizing this information to identify long-term temporal fluctuations in the data.

Mukherjee et al. [162] proposed utilizing dynamic images created by extracting motion data from RGB and depth images independently and then joining them. Two segments of the Resnet-101 network are utilised to complete the task, resulting in a sparser matrix of video data. Zhang et al. [163] have also developed a semantics-based multi-stream DNN for action attribute training, motion detection, and zero-shot AR. Furthermore, the approach integrates joint learning and semantics in the context of graph regularisation with adaptive moment estimation optimization. For activity monitoring, the author in [164] employed temporal and posture-based data and proposed a Dual-stream framework based on skeletal and RGB data. Skeleton details offer insights into the positioning of the human body, whereas RGB data offers valuable time information for assessing human actions, enhancing the action recognition and process depicted in Figure 2.9.A comparison between various approaches for HAR is represented in Table 2.1.

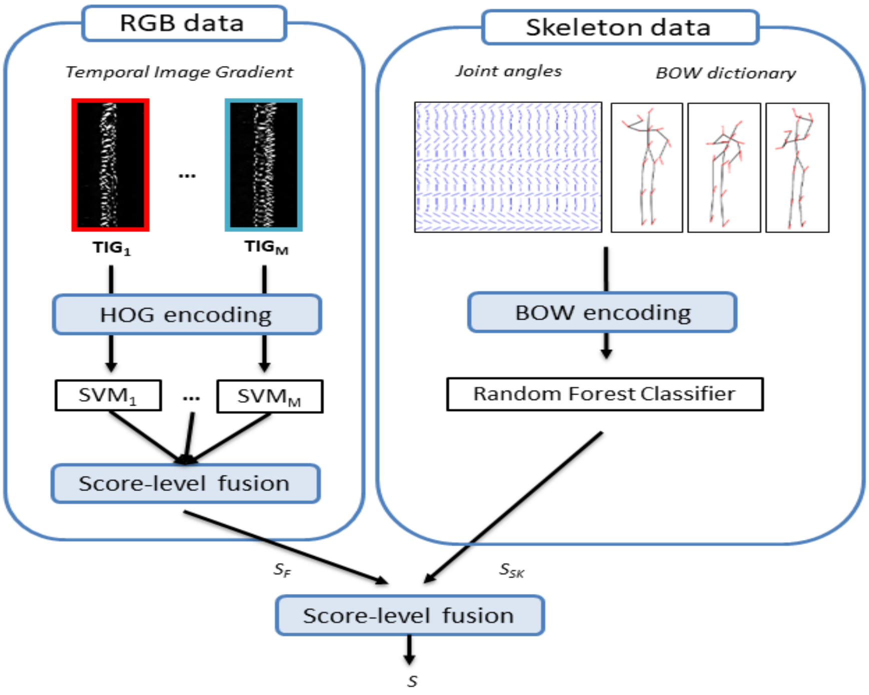


Figure 2.9 Multimodal Human Action Recognition Process [164]

Table 2.1 Methods for Recognizing Human Activity: Evaluation and Comparison.

Reference	Dataset	Result (in %)	Strengths	Weakness	Brief Study
[89]	UIUC Olympic Sport	89.7 74.38	Problems caused by data are optimised by picking the most discriminative attributes.	It is challenging to carry out complex activities as characteristics.	K-Nearest Neighbour serves as a classifier to learn high-level semantic features automatically from training data and explicitly input data.
[165]	KTH	97.2	The use of LS-TSVM speeds up the identification procedure by 4 x.	Instead of enhanced efficiency using fold, an evolutionary method with an adjustable penalty should be utilised.	The HOF and the Harris Detector System are employed for feature extraction, using least square Twin SVM for classifying actions.
[72]	KTH Weizmann YouTube	95.4 97.8 86.2	Enhanced overall accuracy due to the combination of HOG and MBH, which results in a hybrid descriptor.	Using PCA renders the results unstable and uncontrollable; hence, they are not combined.	STIP extraction is performed on videotape with a hybrid feature descriptor, including static and mobility info; a VLAD is utilized as the video encoder.

[166]	Weizmann	94.26	It is capable of differentiating multiple actions with excellent recognition results.	Not suited for changes in perspective, point scaling, or point spinning.	Histogram employs Directed LBP and MHI with Spatiotemporal information and SVM functions as the classifier.
[79]	MSRAction3D: MSRGesture3D:	87.9 94.6	It's not required more effort to execute in real time since its processing rates is 30 fps.	Close activities such as hand grab and hand chuck cannot be properly categorised.	At the feature and decision levels, LBP and DMM from three angles, namely front, side, and top, are combined. The use of kernel extreme learning machine classification.
[167]	KTH UCF YouTube:	93.3 72.07	Capable of recognising events in an unrestricted environment with a complicated context	Dynamic background remains a classified challenge.	Trajectories and LBP are used to characterise the spatial information of moving parts of a tracked body, while SVM is used as a classifier.
[33]	Hollywood2 HMDB51	62.5 52.1	Extraction of trajectories is facilitated by a straightforward and simple technique.	No extrinsic visual cues are used, resulting in no enhancing approach efficacy.	Vision mobility is split into space-time trajectories that are then employed for descriptive purposes by the DCS descriptor. Moreover, the VLAD coding approach is employed for AR.

[67]	KTH Weizmann i3Dpost Ballet IXMAS	95.5 100 92.92 93.25 85.5	This fusion strategy aims to provide a large number of unique feature vectors, which results in action modelling that is robust and devoid of disturbance.	Not suitable for dynamic background and higher complexity.	HAR based on the merging of SDEG of human postures and orientation of key poses of human silhouettes, which is conducted sequentially but individually, is described.
[168]	MSRAction3D-Test1 MSRAction3D-Test2 MSRAction3D-Cross-subject test	95.8 97.8 83.3	It is then suggested to use Accumulated Motion Energy (AME) to conduct informative frame selection, which may eliminate noisy frames and minimize computing costs.	These methods are helpful at close range and in specialised contexts because their high cost and precise requirements are not suitable in low light conditions.	Eigen Joints-based approach for action recognition using NBNN classifier. Eigen Joints' compact and discriminative frame format effectively captures the features of static posture, motion between successive frames, and overall dynamics relative to the neutral position.
[80]	Own dataset Video Web i3DPost WVU	99 99 98.33 99.33	The technique takes less computational resources to perform.	Neither trying to bend nor resting motions are fully acknowledged	From silhouettes, mutually scale invariant contour-based posture structures and rotationally stable LBP are retrieved, and then SVM is used as a classifier.

[112]	KTH Weizmann	96.4 100	The issue of a lower recognition ratio in varying environmental circumstances has been resolved.	This method is less efficient when an item is obscured.	The Spatio-temporal form fluctuations of human silhouettes are conveyed by dividing the principal poses of the silhouettes into a specified number of grids and cells, resulting in a representation devoid of noise. The measurement of grid and cell parameters facilitates the modeling of feature vectors. This calculation of grid and cell parameters is further organized to preserve the temporal order of the silhouettes.
	Berkeley MHAD	99.24	A membership function that facilitates categorization can be achieved with a minimal distance of three units from the ground.	Evaluated on one dataset	The membership function is constructed by utilizing a Convolutional Neural Network (CNN) with fuzzy inputs derived from motion capture data and the distances between the ground and

					the left hand, right hand, and pelvis.
[169]	Weizmann KTH	89.2 93.97	The approach addresses the issue of ongoing activity recognition in visual frames.	Instead of HMM, the time sequential model should be used.	A hybrid technique using CNN to autonomously obtain features from visuals & the Viterbi algorithm for training. HMM is employed to perform classification.
[170]	UCF101 HMDB51	93 70.2	Apply for relatively small datasets to minimize the learning parameters to a minimum while maintaining performance.	Not suitable for complex action.	The proposed approach involves utilizing the Temporal Pooling Pyramid technique in the FCNN framework. Additionally, the integration of spatiotemporal and fuse strategies is employed to construct dual-stream fuse networks, which incorporate manually generated features.
[171]	JHMDB MPII Cooking	79.5 71.4	The RGB and flow modalities are merged, which results in an assessment of human	There is no CNN posture detection for each body	CNN description is based on poses and contains information on mobility and appearances for

			position that is more error resistant.	component, including arms, legs, etc.	every component, including both automatic and manual information
[172]	UCF101 HMDB51	95.5 72.5	Despite their easy implementation compared to RGB and optical flow images, dynamic images exhibit remarkable effectiveness.	Incapable of handling rapid changes in a particularly complex video stream, dynamic graphics operate with a defined window size	Four stream networks receive dynamic visuals comprised of all the video content in one frame, RGB, and optical flow, as well as rank pooling representing temporal data.
[173]	MCAD North-western-UCLA:	86.9%, 88.9%	They calculated the inter-view correlation by taking the pairwise dot product of the output of the LSTM network for each view, which is advantageous in multi-view scenario.	This model requires lots of computation.	A method for multi-view action detection in which frame-level features are extracted and fed to conflux LSTM. The correlation coefficient is then calculated using view-dependent pattern recognition and action categorization.
[174]	Activity-Net	39.37	The number of activities and sub-activities in each video are also found right away.	This approach is not suitable for prolonged sequences of videos.	A four-step method for identifying activity comprising frames conversion, human body detection, action recognition, and

					occurrence time of action using two stream data (i.e., RGB image and optic flow) via a CNN-based network.
[175]	Activity Net THOMUS	53.5 65.6	Transformer networks help figure out where in time an action is happening in a video.	The method is to use video features that have already been extracted, which has been done in many other ways.	It includes information about multiscale feature representations and local self-attention and sends it to a decoder so that AR can be done.
[176]	Own -Dataset	90.89	Single-person activity recognition with better accuracy.	Recognizing multi-person action may not work correctly.	With MSR Kinect sensor V2, 25 dissimilar joints can be found. Activity Recognition Using a Hybrid CNN+LSTM Conceptual Framework
[177]	NTU RGB + D	83.6	Not perform adequately when distinguishing identical activities	Performs better than other approaches that are not based on deep neural networks.	3D FCNN framework with depth images as inputs.

2.7 Dataset for HAR

The primary assets for measuring the precision of the proposed technique are datasets. Many reference datasets, including a vast array of activities, have been presented so far. The decision of a dataset affects the selection of an appropriate method for human activity recognition. These datasets incorporate RGB, depth, skeleton data, or any combination thereof. In overall, these datasets were recorded with various sensors, including RGB cameras, Kinect depth sensors, and gyroscope sensors. Several regularly used recognition of human activity datasets have been evaluated in this research. This section describes numerous experimental and assessment datasets used in the study.

2.7.1 KTH Dataset

This dataset has several video clips. It consists of six classes of human actions: jogging, walking, sprinting, boxing, hand waving, and hand applauding [11]. Each action category contains 100 sequences featuring 25 actors performing acts in four distinct settings: outdoors, outdoors with varying attire, indoors with varying illumination, and outdoors with differing dimensions. The frame rate is 25 per second, and the resolution is 120 by 160 pixels. There is an excellent difference in video quality, time, and angles of view.

2.7.2 Weizmann Dataset

The Weizmann dataset [178]contains 90 clips of actions performed by humans, which include ten activities such as: 'walking', 'bending', 'jumping', etc. One of ten distinct individuals carries out each of these 10 acts. The collection contains silhouettes as well as static backdrops. The lack of perspective flexibility prevents the simulation of many real-world occurrences. The videos have a resolution of 144 by 180 pixels and 25 fps.

2.7.3 IXMAS Dataset

IXMAS dataset[179] includes activities collected from five perspectives. 11 individuals carry out 14 classes of activity. The actions may be performed in any orientation with respect to the camera setup. The viewpoint, backdrop, and lighting is all static. Due to this, the models developed expressly for this dataset perform poorly in several real-world contexts. This collection also contains silhouettes and voxel representations of the objects it has.

2.7.4 UCF 101

The UCF 101 dataset[180] contains more than 13000 RGB video clips with 101 action types belonging to 25 distinct groups, with 5–7 videos per group. All activities may be categorised into five broad categories: human–human interaction, human–object interaction, physical motion, sports, and musical performance. The UCF 101 offers actual action films instead of produced action videos, enhancing the entire recognition task.

2.7.5 UCF Sports

The UCF sports [181]action dataset includes 140 + visual sequences and ten categories of sports acts, such as 'diving,' 'kicking,' 'weight-lifting,' 'horseback riding,' 'golf swinging,' 'running, "skating," 'wielding a baseball bat,' and 'walking,' were retrieved from activities featured on broadcast television. Thus, it is somewhat more challenging to deal with since it consists of a natural pool of behaviours occurring in various locations and perspectives. The video resolution is 480 by 720 pixels.

2.7.6 MSR Dataset

The MSR action dataset was created to assess the performance of recognition methods in scenarios characterized by clutter and dynamic backgrounds. This dataset consists of more than 15 distinct classes and includes over 60 video recordings, which multiple actors conducted. Various movements encompass high arm waves, hand clapping, running, and other similar actions. The dataset was divided into two subsets: the MSR action dataset and the MSR action II dataset. Including depth films or 3D films in educational curricula presents a significant challenge in terms of anticipation despite the meticulous annotation of frames for each action.

2.7.7 HMDB 51

he HMDB51 dataset, as described by Kuehne et al. [161] comprises 51 distinct categories, each containing a minimum of 100 videos. The dataset consists of 6849 instances of various actions sourced from multiple origins. Five distinct types of action can be identified. The database in question is widely acknowledged as one of the most exceptional resources for recognizing human activities.

2.7.8 Activity Net

It was presented by Heilbron et al. (2015) at the varsity of Del Norte, Colombia (http://activity-net.org/)[182] This colossal dataset aims to give a comprehensive taxonomy of videos depicting the actions of humans and to include an extensive range of subjects for each activity. These recordings were gathered from social video-sharing sites such as YouTube in recent years. The Activity Net dataset is separated into three divisions based on application area, including classification of untrimmed clips, classification of trimmed videos, and detection of activity in all untrimmed videos. The untrimmed video collection includes 2,7801 labeled films with 203 activity categories. The pruned activity dataset comprises 203 activity groups with an average of 193 samples per category to predict the label of solitary activity videos.

2.7.9 MSR action pairs (3D action pairs)

This dataset [183]emphasizes the two most essential elements, skeletal trajectories for comparable activities and the association between sets of motions. Within this dataset, combinations of behaviors; such as Pick up and place down, are chosen based on the premise that form signals may be similar, but their correlations vary. Ten individuals carry out six pairings of activities. Pick up/set down a carton, Raise or position a crate, push or pull a chair, don or remove a hat, don or remove a rucksack and then attach or remove a poster. Each movement is executed three times, and five actors are utilized throughout training for testing and recuperating.

2.7.10 PKU-MMD

It emphasizes activity recognition in long and complex continuous sequence data and multimodal action analysis. This dataset [184]was obtained using the Kinect 2.0 sensor. This collection is comprised of both short and long video streams, respectively. Phase 1 consists of 1076 3–4-minute-long video clips captured at 30 frames per second. In Phase 1, 66 actors perform 51 action courses, divided into 41 commonplace acts and 10 human contact actions. Phase 2 includes 2,000 1–2-minute films captured at 30 frames per second. Phase 2 consists of sixty actors performing forty-nine action lessons. This collection comprises RGB images, depth maps, skeletal joints, infrared sequences, and RGB visuals.

2.8 Evaluation Metrices

Human Activity Recognition incorporates several performance criteria taken from a variety of categorization domains. Many evaluation criteria from other categorization domains have been developed and used to the identification of human activities. In this subsection, we provide widely used measures such as accuracy, precision, recall, etc., based on [185]. Before summarising these measures, we define the following terms: The most prevalent performance indicators are.

Accuracy: Overall, instances that have been accurately classified constitute accuracy. It is estimated by dividing the ratio of accurate classifications by the whole number of categories.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N}$$

Precision: It is also known as Positive Prediction Value (PPV) and refers to the probability that a detected instance of activity will really occur. Similarly, (1 -precision) influences the likelihood that the recognizer would misidentify an observed action. This is mathematically stated as:

$$Precision = \frac{T_P}{T_P + F_P}$$

F-Measure: It establishes the harmonic mean between accuracy and recall. It provides information on the test's reliability. Hence, F measure simultaneously influences the classifier's precision and robustness. Its highest value is 1 and its lowest is 0. This is technically stated as:

$$F - Measure = 2 * \frac{Precision}{Recall}$$

The Area Under Curve (AUC) is commonly utilised to tackle binary classification problems. The estimation pertains to the probability that the classifier will assign a higher rating to a randomly chosen positive sample compared to a randomly chosen negative sample. The AUC represents the integral of the curve obtained by plotting the False Positive Rate (also known as Specificity) against the True Positive Rate (also known as Sensitivity) over a range of thresholds from 0 to 1.

Confusion Matrix: It provides a summary of the prediction results and a detailed description of the model's performance and is also known as the error matrix. The confusion matrix depicts classification errors and their corresponding categories. Each row of the matrix represents the anticipated class, while each column represents the actual class, or vice versa.

The ROC curve, or precision-recall rate, compares the genuine positive rate to the true negative rate (FPR). Since the ROC curve is dependent on the number of True Negative classes, it is only applicable to detection models and cannot be utilised with imbalanced datasets, which are typical in deep learning-based human behaviour.

Where T_P : True Positive,F_P : False Positive,T_N : True Negative F_N: False Negative.

Chapter 3

Multimodal Human Action Recognition System

Recognizing human activity poses a significant challenge, particularly when confronted with the complexities of multiple actions and scenarios. This paper presents a proposed human action recognition (HAR) approach using a multi-view, multi-modal framework. The motion illustration of each Depth motion map, motion history images, and skeleton images is generated from data obtained from RGB-D sensor depth, RGB, and skeleton data. Following the motion representation, each individual motion is trained separately using a 5-stack convolutional neural network (5S-CNN). The recognition rate and precision are enhanced through the training of the skeleton representation using a hybrid classifier consisting of a 5S-CNN and Bi-LSTM. The score values of three motions are combined using decision-level fusion in a sequential manner. Human activity is identified based on the fusion value. In order to assess the effectiveness of the proposed 5S-CNN with the Bi-LSTM method, we utilise the UTD-MHAD dataset in our experimental analysis. The findings indicate that the proposed human activity recognition (HAR) method exhibited superior performance compared to other established methods in the field.

3.1 Introduction

HAR is a promising area for computer vision research. Common applications include interactions between humans and computers, healthcare settings, visual investigation, etc. The device records video for the present HAR procedure. Primarily, spatial-temporal characteristics are employed for identification[186]. As a result of the rapid development of imaging technology, depth cameras such as Microsoft Kinect cameras are currently utilised. This camera can record RGB, depth, and skeletal information. Deep and skeletal knowledge have the advantage of being less sensitive to changes in illumination conditions than conventional RGB information. Prior methods of recognizing actions based on depth information made use of explicitly constructed descriptors. The advantages of CNNs prepared exclusively for DMMs are outlined in [146], [187]. A DMM is an unambiguously shaped movement depiction image constructed from rudimentary depth outlines. This is similar to the development of MHI and optical stream illustrations derived from RGB outlines. [188], [189] describes the autonomous use of skeletal data for action recognition based on meticulously crafted element descriptors. In [190], an attempt is made to generate surface images from skeleton joint clusters.

DCNN can naturally distinguish features from data in the multiple phases processes of the visual cortex[191], [192] and are used for tasks related to image characterization, recognition, division, discovery, and recovery (ConvNets) employ techniques to effectively expand the capacity of organizations to encompass a significant number of boundaries. Additionally, ConvNets retrieve boundaries from an extensive repository of named data. The incorporation of RGB, depth, and skeletal data at the individual level has led to notable advancements in Human Activity Recognition (HAR) within the past few years. Nevertheless, experts still face difficulties in effectively incorporating these captivating viewpoints into various approaches to augment recognition further. The recognition of human action was achieved in [193] through the utilization of RGB and depth data. At both the component level and the chosen level, skeletal information and idleness information are intertwined [194] . In order to zero in on HAR, depth movement maps with skeleton information are combined within [195]. To enhance the efficacy of the HAR system, a Multiview multimodal HAR system is proposed in this methodology.

3.2 Related Work

A considerable proportion of scholars focus on the identification of human actions by employing multiple modalities. The following article presents a discussion of various research studies.

In [195]sing a deep learning algorithm incorporating multiple modalities, including RGB and Skeleton data. The first step involves the extraction of MHI and MEI from the input RGB video. Subsequently, three distinct perspectives of the skeletal depiction were obtained through skeleton intensity. Following the motion extraction procedure, the extracted characteristics undergo training via LSTM. Ultimately, the decision is executed based on the value of the SoftMax score. The efficacy of this methodology was evaluated through experimentation on three prominent datasets, specifically UTD-MHAD, CAD-60, and NTU-RGB + D120.

In addition, Wang et al.[187] aimed to advance the field of Human Activity Recognition (HAR). This task's achievement was facilitated by utilizing a weighted hierarchical Dirichlet multinomial model (WHDMM) in conjunction with a three-channel deep convolutional neural network (3 ConvNets). Using a 3D point, an initially distinct perspective of depth can be captured. This information is used to educate ConvNets. The AHDMM is then constructed from temporal dimensions. In fact, they utilised three types of datasets for simulation purposes. Chen et al.[196] sought to design HAR by integrating skeleton, RGB, and depth information. Similarly, Escobedo et al. [197] created HAR utilizing skeleton and depth data.

Gaglio et al. aimed to construct HAR using ML-based methods [198]. In this instance, skeleton joint data was utilised. The CNN stream based HAR system developed by Khaire et al.[199] Videos are used to generate MHI, DMM, and skeleton images, which are three types of motion representation images. After data extraction, each piece of data was individually trained using CNN classifier. From the outcomes of the classification, a final score was calculated. On account of the score value, recognition has taken place. Using three distinct datasets, the proposed method's efficacy was evaluated.

In addition, Guo et al.[200] devised human activity recognition by combining Wi-Fi and video data. They influence how Wi-Fi signals transmit human activity data that was previously optically robust. To validate this creative concept, they devise a practical system for HAR and compile a dataset containing both video clips and Wi-Fi Channel State Info of human activities. Video highlights were extracted using a 3D convolutional neural network, while radio highlights were extracted using measurable calculations. After combining video and radio elements, an old-fashioned direct assist support vector machine was utilised as the classifier.

Similar to this, Tran et al.[201] sought to develop a method to identify gesture-based recognition. Here, multi-modal streams were used to accomplish the recognition procedure. Three varieties of stream depth, RGB, and optical flow are utilised for the recognition procedure. These streams are incorporated into the process of feature extraction. Then, the classifier is supplied with these characteristics to classify a distinct activity. For simulation, various gesture datasets were utilized. Nie et al. intended to establish emotion recognition with a multi-layer LSTM classifier in [202].

This classifier makes it easy to identify the activity. Due to the interval between video frames, this method cannot precisely detect the presence of motion. Khowaja et al. [203] intended to construct HAR using cross-modal learning at a fundamental level. In this case, RGB and optical flow were employed. UCF101 and HMDB51 datasets are employed for experimental analysis.

3.3 Problem statement

The researcher utilised the benefits of the skeleton joint and RGB video in [22]. In the proposed method, the authors learned CNN using RGB data and processed skeleton data using CNN and LSTM networks. After integrating the features, the authors presented four convolutional layers and three fully connected layers, resulting in a significant computation complexity despite the improved recognition accuracy. In order to minimize computational complexity and maximize the learning process, this methodology makes the following contributions.

- At the outset, we generate Depth Map Models (DMMs), Multimodal Human Interaction (MHI) representations, and skeletal images using the depth, RGB, and skeletal data obtained from the RGB-D sensor.
- Individual CNNs with 5 convolutional layers are trained using the extracted images. 5S-CNN is trained on skeleton and DMM images for multi-views such as top, front, and side.
- While The CNN network is implemented to acquire spatial information, Bi-LSTM is used to train skeleton images for temporal dependence. After extracting features from each Skelton image view using 5S-CNN, these features are combined and provided as input to the Bi-LSTM.
- The WPM is utilized to combine the output score generated by each model. According to the combined results, the actions of persons is apparent.
- Accuracy, F-score, precision, and recall are used to assess the effectiveness of the proposed scheme.

3.4 Proposed Model

Identifying human activity from video footage poses challenges due to factors such as partial opacity, background congestion, visual impairments, variations in size, alterations in appearance, and variations in illumination. Recognition systems play a crucial role in various domains, such as facilitating human-computer communication, enhancing video surveillance systems, and enabling automata to analyze and interpret human behaviour. In the past, the utilization of Human Activity Recognition (HAR) predominantly relied on the Motion History image and depth map exclusively. The provided system does not offer the Recognition system with optimal precision. In order to address the aforementioned concern, this study presents a novel approach to automatic human action recognition that incorporates multiple views and modalities. In this context, recognition systems employ three modalities, namely MHI, DMMs, and skeleton images. Figure 3.1 depicts the overarching concept of the proposed approach. The suggested approach extracts derived DMMs, MHIs, and skeleton images through an RGB-D sensor's depth, RGB, and skeletal data.

DMMs and MHI are independently trained by a 5S-CNN. Similarly, 5S-CNN gets used for training the Multiview skeleton images, and the trained output is inputted into the input of Bi-LSTM.

Furthermore, the output scores of each model are combined using WPM. Fusion's output enables the identification of human activity.

3.4.1 Constructing Motion Representation Images

The primary objective of this section is to extract motion representation images (features) from every image input for HAR. In this paper, MHI, skeleton images, and DMMs are derived from RGB data, skeletal data, and the depth of an RGB-D sensor.

a) Constructing MHI

MHI is a simple and reliable method for representing video motion. It provides transient information regarding the movement of an image within a video. Local kinetic density correlates to the MHI pixel intensity. MHI distributes the time scale of human gestures because it can be encrypted multiple times over a range. When the intensity of an image's pixels is low, the motion occurred in the past. If the luminosity of an image's pixels is high, the movement occurred very recently.

Eqn estimates the MHI. 3.1:

$$M\tau_{(a,b,k)} = \begin{cases} \tau & if\ \theta_{(a,b,k)} = 1 \\ max(0, M\tau_{(a,b,k-1)} - \delta & otherwise \end{cases} \qquad (3.1)$$

Where $\Theta_{(a,\, b,\, k)}$ →denotes the Occurrence of motion or object in the current frame in a video; k →time; (a, b) → position of pixel; δ→ decay parameter.

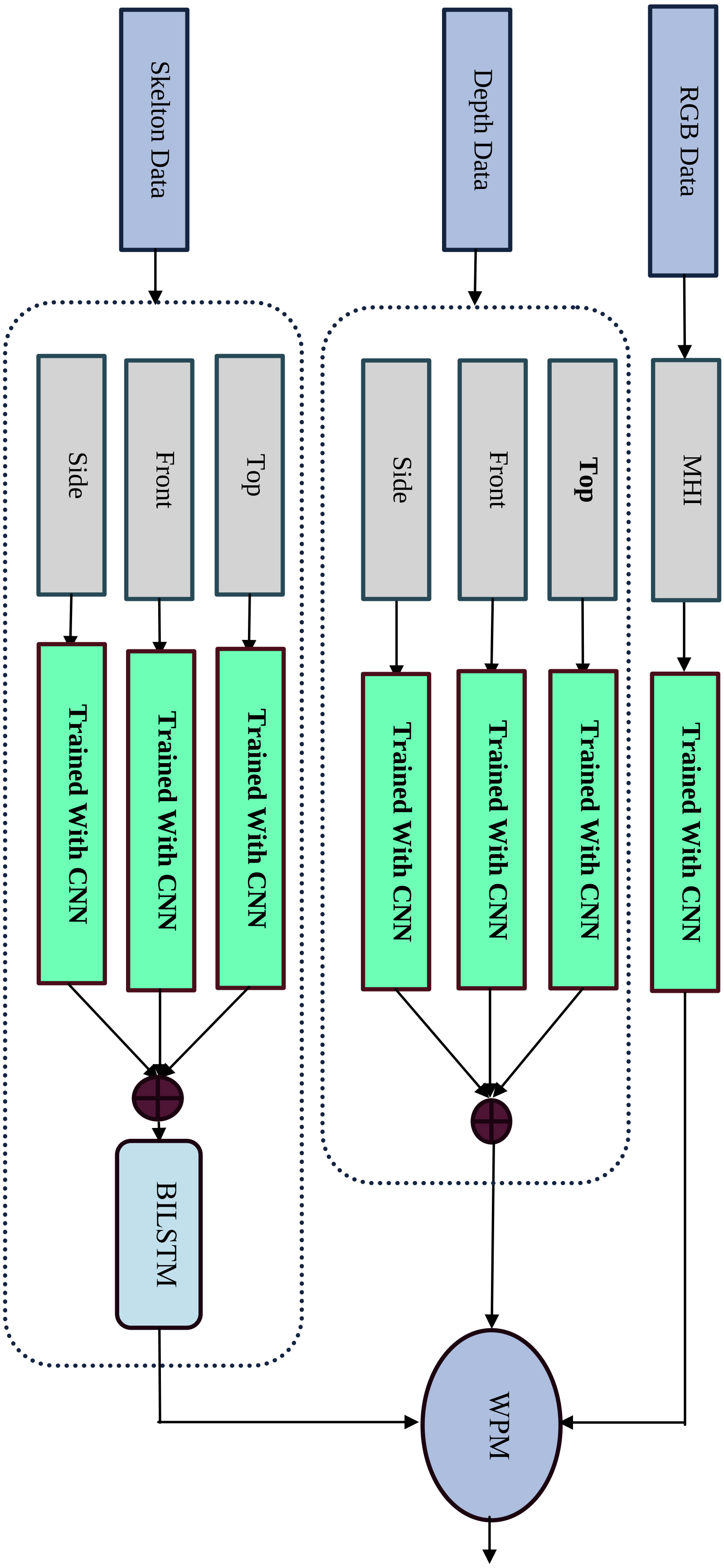


Figure 3.1 Architecture of Proposed Methodology

The transient movement is controlled by the parameter τ. MHI is typically derived from a binary image extracted from the frame using a threshold ξ.

$$\theta_{(q,b,k)} = \begin{cases} 1\ if\ D_{(a,b,k)} \geq \xi \\ 0\ \ otherwise \end{cases} \tag{3.2}$$

$$D_{(a,b,k)} = |I(k) - I(a, b, k\ \pm \Delta)| \tag{3.3}$$

Where I (a, b, k) represent the intensity value.
This methodology sets the threshold value Δ as 1 for all the experiments. The final output of MHI is obtained from $M\tau_{(a,b,t)}$.

b) Depth Motion Map Construction

Through the use of an RGB-D sensor, an image's depth map is acquired. This 3D structure captures the structure's form. The depth frame is projected onto three orthogonal Cartesian planes to extract additional body shapes. The ROI of each extended map is then defined as the standardized bounding box of a closer view (e.g., a non-zero region) with a fixed **dimension. This standardization can reduce intra-class variations, such as the Material** heights and operating lengths of various materials performing the same action. Each frame generates three perspectives, including front, side, and top views, denoted by Map^F, Map^S, and $Map^{T,}$ respectively. The kinetic energy of each frame is then calculated by subtracting two consecutive maps and applying a threshold. Kinetic energy refers to the elements or locations that are in motion at any given time interval. This is a clear indication of the sort of action being carried out.

The DMM of the complete video can be estimated as follows:

$$DMM_{iu} = \sum_{i=1}^{N-1}(\left|Map_u^{j+1} - Map_u^j\right| > \xi) \tag{3.4}$$

Where v ϵ {f, s, t} →projection view, $\mathrm{Map}_\mathrm{u}^\mathrm{i}$→Projected map of the i_{th} frame, N→ Number of frames, $\left|Map_u^{j+1} - Map_u^j\right| > \xi$→ Binary Map of the motion image.

c) Skelton images

The skeletal shots are primarily used in the process of activity recognition. The count of N-number frames and K-number joints is available in the bone data. The quantity of action frames varies from function to function. The amount of frames varies between the various subjects. The number of joints throughout the activity is frequently unaltered. Each frame contains a combination of three-dimensional coordinate values denoted by x, y, and z. Consider each joint in the skeleton to be a three-dimensional vector suggested by J_i, where i represents the active joints and J_i is a three-dimensional vector. Since the person can be seen anywhere within the sensor's field of view, it is necessary to organize the integration space in order to accommodate any joint.

The hip centre point is the first normalization frame in the aforementioned study. The following equations are used to normalize the joint Eqn: (3.5)– (3.9).

$$D_i(S) = L_i(S) - L_0(1) \tag{3.5}$$

$$D_{max} = max(D_i(S)) \quad \forall \in joints \quad s \in frames \tag{3.6}$$

$$D_{min} = min(D_i(S)) \quad \forall \in joints \quad s \in frames \tag{3.7}$$

$$d_i(S) = (D_i(S) - D_{min})/(D_{max} - D_{min}) \tag{3.8}$$

Where $L_i(S)$→Coordinate the i_{th} joint in the S_{th} frame; $D_i(S)$→vector distance $L_i(S)$ and $J_0(1)$; $d_i(p)$→ The length of the gap between the vectors $L_i(S)$ and $J_0(1)$ normalized; L (1) → These are the coordinates of the centre of the hip in the first image.

In order to make skeletal photos from skeletal photographs, the skeleton images are first separated into five segments, which include the trunk, the left and right arms, and the left and right legs. Each segment has dimensions of 160 x160 by 5, corresponding to the top, side, and bottom views. The extent that corresponds to this picture is 160 pixels by 160 pixels. Let the image of the size x and y corresponding to the vith view of the body part S seen in the frame be denoted by the letters A x, y, vi S_{th}.

$$P_x(top)_{st} = h_1 + k_1 + d_i(t)_x \tag{3.9}$$

$$P_x(side)_{st} = h_3 + k_3 + d_i(t)_y \tag{3.10}$$

$$P_x(front)_{pq} = h_1 + k_5 + d_i(q)_x \tag{3.11}$$

$$P_y(top)_{st} = h_2 + k_2 + d_i(t)_z \tag{3.12}$$

$$P_y(side)_{st} = h_2 + k_4 + d_i(t)_z \tag{3.13}$$

$$P_y(front)_{st} = h_4 + k_6 + d_i(t)_y \tag{3.14}$$

Constants are the values of h1, h2, h3, h4, h5, and h6, as well as k1, k2, k3, k4, k5, and k6. The values of h1, h2, h3, h4, h5, and h6 are utilized to determine the center of the skeleton picture.

The values k1, k2, k3, k4, k5, and k6 are derived based on the distance between the joints and the stretch of the joint feature that has been provided. The values used for normalization might change depending on which joints are being examined.

3.4.2 Training with 5S-CNN Classifier on RGB and Depth Data

After the RGB, depth, and skeleton data have been generated, they are trained using a 5S-CNN classifier. In this experiment, the MHI and DMM are both trained on a 5S-CNN classifier. On the other hand, a hybrid of 5S-CNN and Bi-LSTM classifiers is used to train the skeleton data. In this case, each of the visual frames undergoes its own individual training using a 5S-CNN classifier. The training procedure is shown quite well in Figure 3.2. An input layer, five convolutional layers, two pooling layers, and a fully connected layer make up the proposed CNN's four layers. The final phase is a wholly interconnected one.

The arrangement of these levels is determined by the functions that they perform. The following provides a detailed explanation of the functioning of each layer of the proposed CNN architecture:

Convolution Layer: Convolution occurs in the first layer of a CNN network. The input image's feature map is created by using this layer. In this study, a 5×5 filter is used to create a feature map. We provide a precise mathematical description of the convolution layer in Eqn 3.15.

$$B_i^b = \sum_{j \in F_i} B_j^{b-1} \otimes \xi_{ij}^b + L_i^b \qquad (3.15)$$

Where "$\otimes$" stands for the Convolution operator; The variable "ξ_{ij}^b" denotes the weight value,"i" represents the filter of the "b$_{th}$" convolutional layer; L_i^b denotes the bias of i$_{th}$ filter of b$_{th}$ convolution layer and the representation of the activation map is depicted as B_i^b.

Pooling Layer: In CNNs, the pooling layer is a typical component used for image recognition. Its goal is to compress the width and height of the input volume while maintaining the essential details. The input volume is partitioned into non-contiguous areas, and a summary statistic, such as the maximum or average value, is calculated for each section.

The resulting output volume has reduced spatial dimensions, which can help to reduce the computational complexity of subsequent layers in the network. Following the implementation of the convolutional layer is the pooling layer. This layer's purpose is to reduce the dimensions of the feature map, thereby reducing the network's computational cost.

On each feature map, a pooling method is implemented. Within the context of this academic article, maximal pooling is utilized. The following equation represents the combined map.

$$P_i = \max_{j \in R_j} F_i \qquad (3.16)$$

Fully Connected Layer: The layer that is ultimately linked to the input of the fully linked layer is provided with the reduced feature map. This layer serves as the classifier for the other layers. This neural network's completely connected layer has been created using a feed-forward neural network.

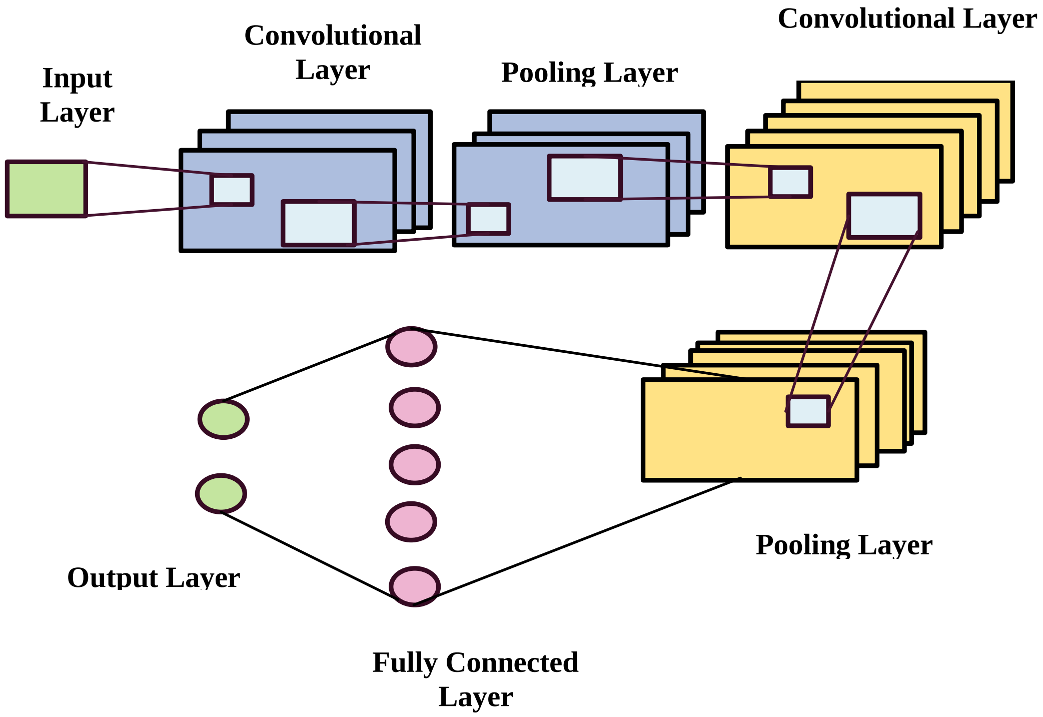


Figure 3.2 CNN Architecture

3.4.3 Training Skeleton Data with 5SCNN and Bi-LSTM

After skeleton data extraction, the three-view data are provided to the 5S-CNN and then the Bi-LSTM classifier. The 5S-CNN concept has already been explained in this section. The output of 5S-CNN is fed into the input of the Bi-LSTM classifier. Bi-LSTM is a sequential processing model that comprises of two LSTMs. The first is utilized in the forward direction and the second in the reverse order. Bi-LSTMs maximize the quantity of network-accessible data in an efficient manner. Figure 3.3 depicts the structure of a Bi-LSTM.

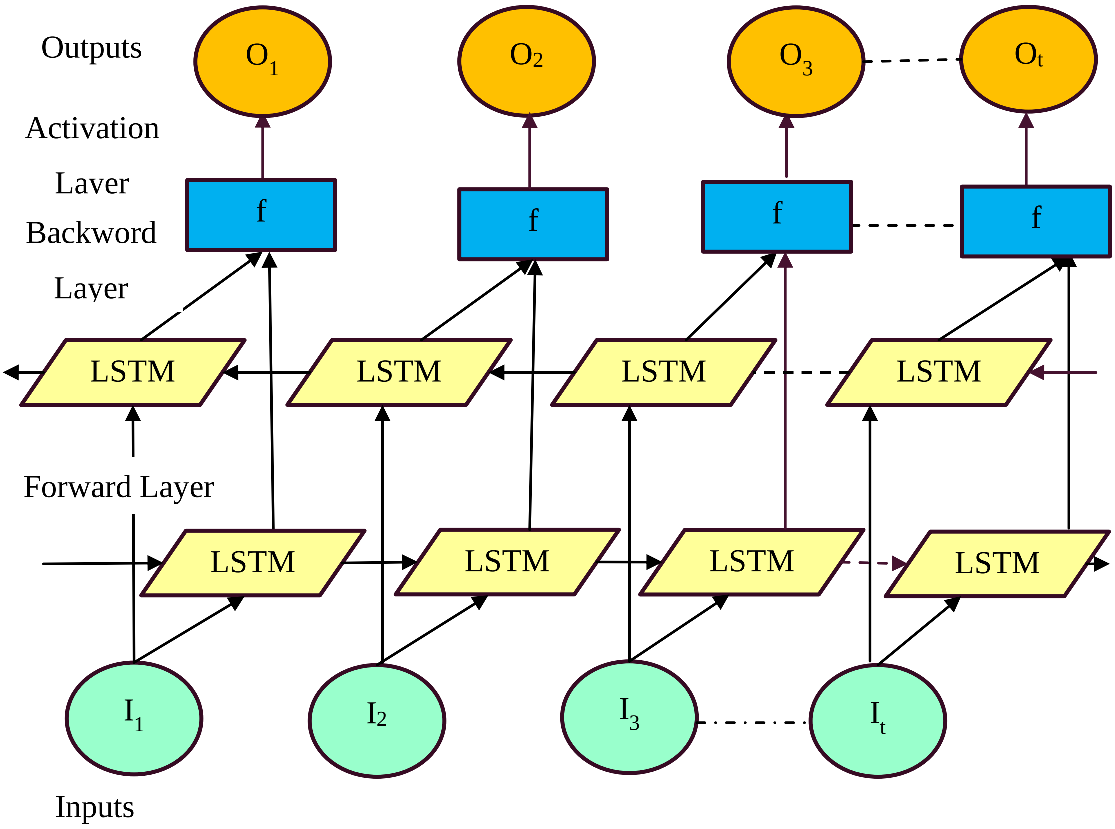


Figure 3.3 Bi-LSTM Architecture

$$I_t = ([m_{t-1}, v_t]W_I + D_I)\sigma \tag{3.17}$$

$$F_t = ([m_{t-1}, v_t]W_F + D_F)\sigma \tag{3.18}$$

$$O_t = ([m_{t-1}, v_t]W_O + D_O)\sigma \tag{3.19}$$

$$\widetilde{U_t} = tanh([m_{t-1}, v_t]W_U + D_U)\sigma \tag{3.20}$$

$$U_t = \widetilde{U_t} \times I_t + U_{t-1} \times F_t \tag{3.21}$$

$$m_t = tanh(U_t) \times O_t \tag{3.22}$$

Where tanh→ denotes the hyperbolic tangent function; $\boldsymbol{\sigma}$→sigmoid activation function; v_t→ Vector input; I_t→ output of the input; F_t→ output of forget gate; O_t→ output of output gates at time t; d→ bias value; W→ weight of the control gates;$\widetilde{U_t}$ →Ccurrent state of input; U_t and h_t→ update state and output at time t.

In the standard LSTM, the picture is encrypted in only one direction. Two LSTMs, also known as two-way LSTMs (B-LSTMs), can be utilized as bidirectional encoders. By incorporating the CAT mechanism to improve the state transmission system, this Bi-LSTM network can tackle the long-distance dependence problem of conventional RNs.

The Bi-LSTM generates a series of hidden states in response to an input picture by using the image as input. The hidden state of the output that corresponds to the input value. The following is how it is determined:

$$m_t = LSTM(m_{t-1}, I_t) \tag{3.23}$$

$$\acute{m}_t = LSTM(m_{t-1}, I_t) \tag{3.24}$$

$$M_t = [m_t, \acute{m}_t] \tag{3.25}$$

where the vector representing the concealed layer moving towards a positive orientation at time t is denoted by m_t, and the vector of the hidden layer moving in the direction of negative at time t is represented by $\acute{m}_t$. The vector representing the result at time t is denoted by the letter $M_{t.}$ In the end, we were successful in obtaining the score value. This value of the score is utilised for the next stage of processing.

3.4.4 Decision Level Fusion

Afterward the learning procedure, the overall result was assigned to each input. Combining the calculated score values using WPM. The suggested WPM is utilized for making decisions primarily. Let MHIs, DMM_F, DMM_T, DMM_S, SK_T, SK_F, SK_S represent the MHI, DMM, and skeleton joint score values. These results function as the WPM decision-making criteria. Here, the number of classes is variable. On the basis of the results' parameters (scores), the optimal solution (class) is selected and classified. The score value is determined by Eqn 3.26.

$$WPM^S = Max[MHI_S^1 \times DMM_F^2 \times DMM_T^3 \times DMM_S^4 \times SK_T^5 \times SK_F^6 \times SK_F^7 (3.26)$$

Based on the score's worth, we can figure out the proper human action.

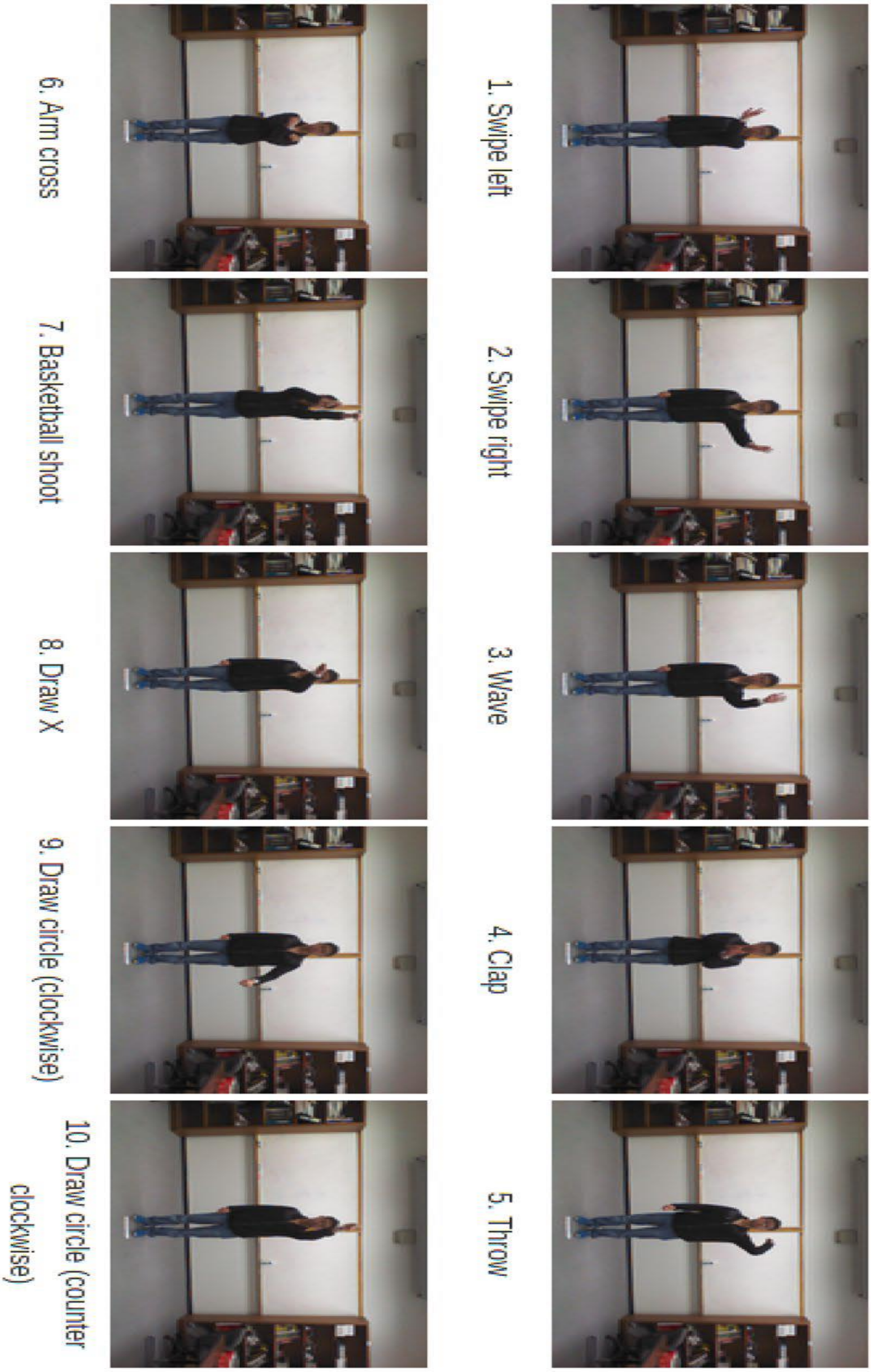
3.5 Result Analysis

The following section presents an analysis of the results obtained from the HAR system that was proposed.

3.5.1 Description of Dataset

UTD-MHAD is the most recent functional database. There are twenty-seven human activities in this database. A subsurface camera and a passive sensor worn on the body are used to collect these database recordings. This database contains RGB and deep videos, as well as skeleton position data and transition signals. This database is utilised by the training and examination processes. For the training process, unusual models are utilised, while samples are used for testing.

This dataset was gathered utilizing Kinect. There are numerous variations within each division. This dataset was collected as part of HAR research by combining depth and inertial sensor data. It includes 27 distinct activities performed by eight individuals (four females and four males). Each issue repeated each action four times. Figure 3.4 illustrates a set of frames extracted from the UTD-MHAD dataset, showcasing ten distinct action classes.



1. Swipe left

2. Swipe right

3. Wave

4. Clap

5. Throw

6. Arm cross

7. Basketball shoot

8. Draw X

9. Draw circle (clockwise)

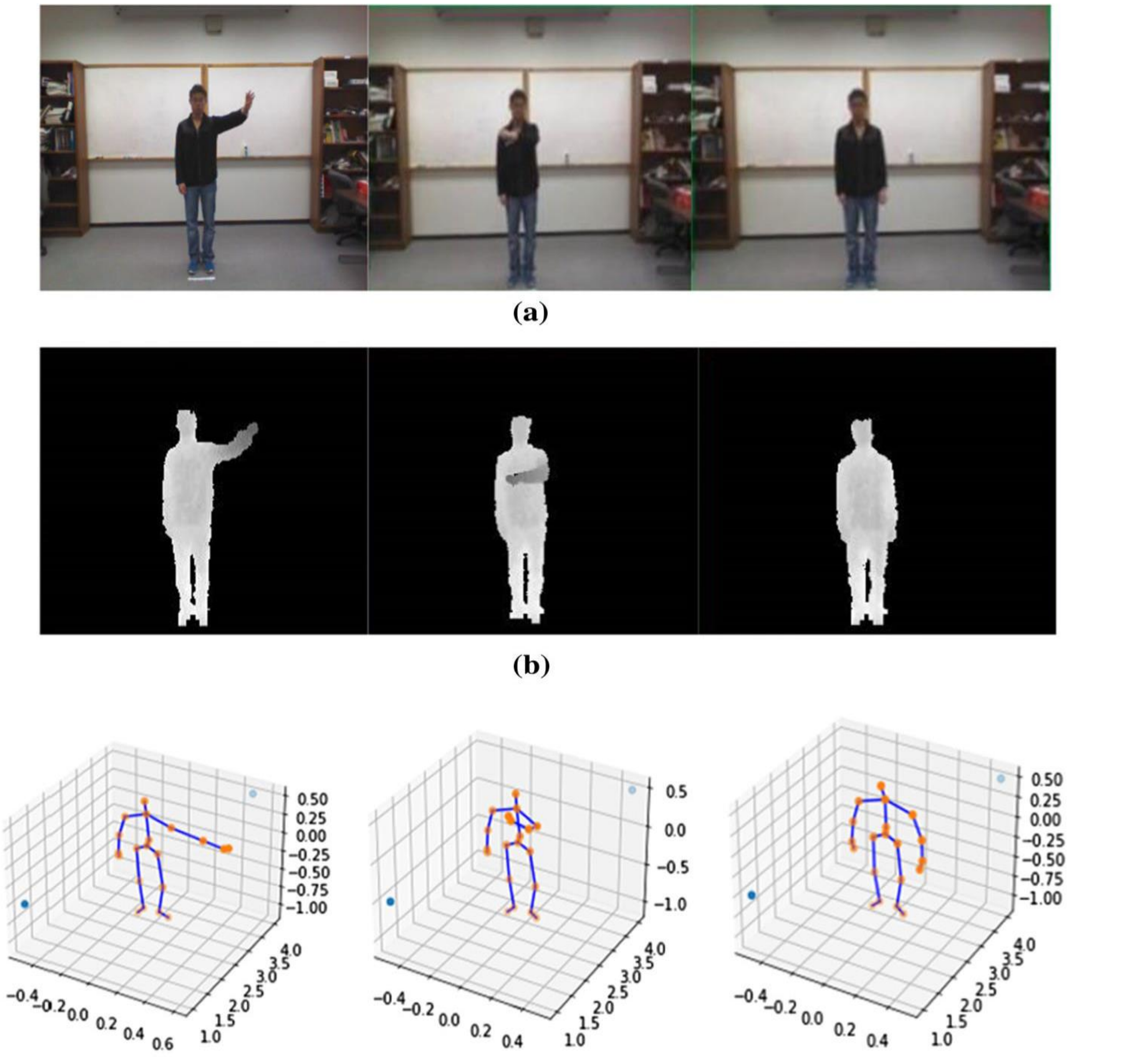
10. Draw circle (counter clockwise)

Figure 3.4 Sample Frames of Dataset

3.5.2 Experimental Setup

The primary objective of the proposed methodology is to identify and classify human activities depicted in video or image data. The present study employs three modalities, namely MHI, DMMs, and skeletons, for the purpose of activity recognition.

The sequence of the dataset is shown in Figure 3.5 and the confusion matrix is shown in Figure 3.6. The AUC for three distinct actions is shown in Figure 3.7. The trajectory is drawn between the rate of false positives and the rate of genuine positives. Here, the area under the ROC curve for boxing is 0.984; for archery, it is 0.9231, and for the tennis movement, it is 0.9573.



(a)

(b)

(c)

Figure 3.5 Dataset modality sequence, (a) RGB frames, (b) depth frames and (c) skeleton sequences

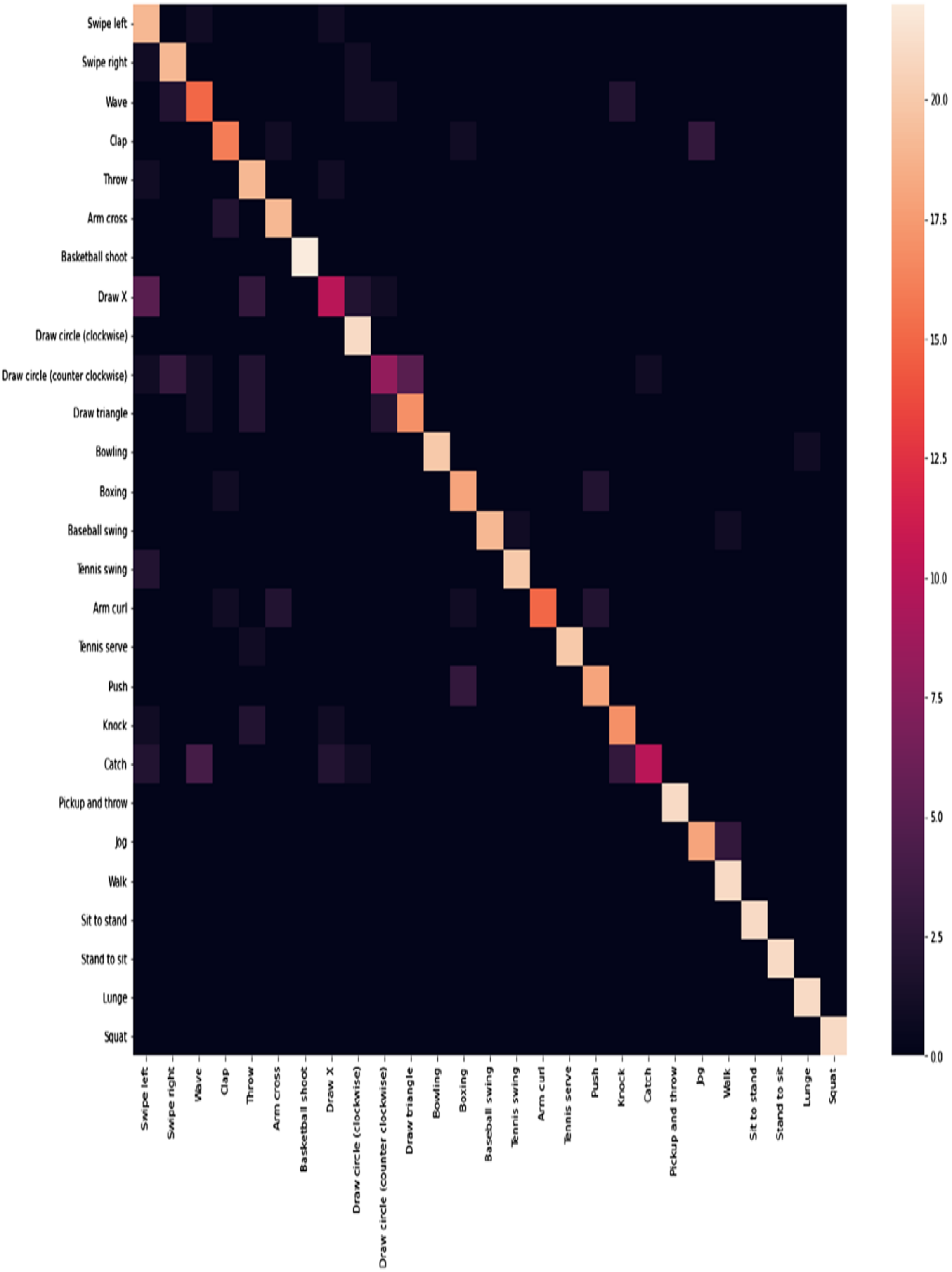


Figure 3.6 Presented Methodology Confusion Matrix

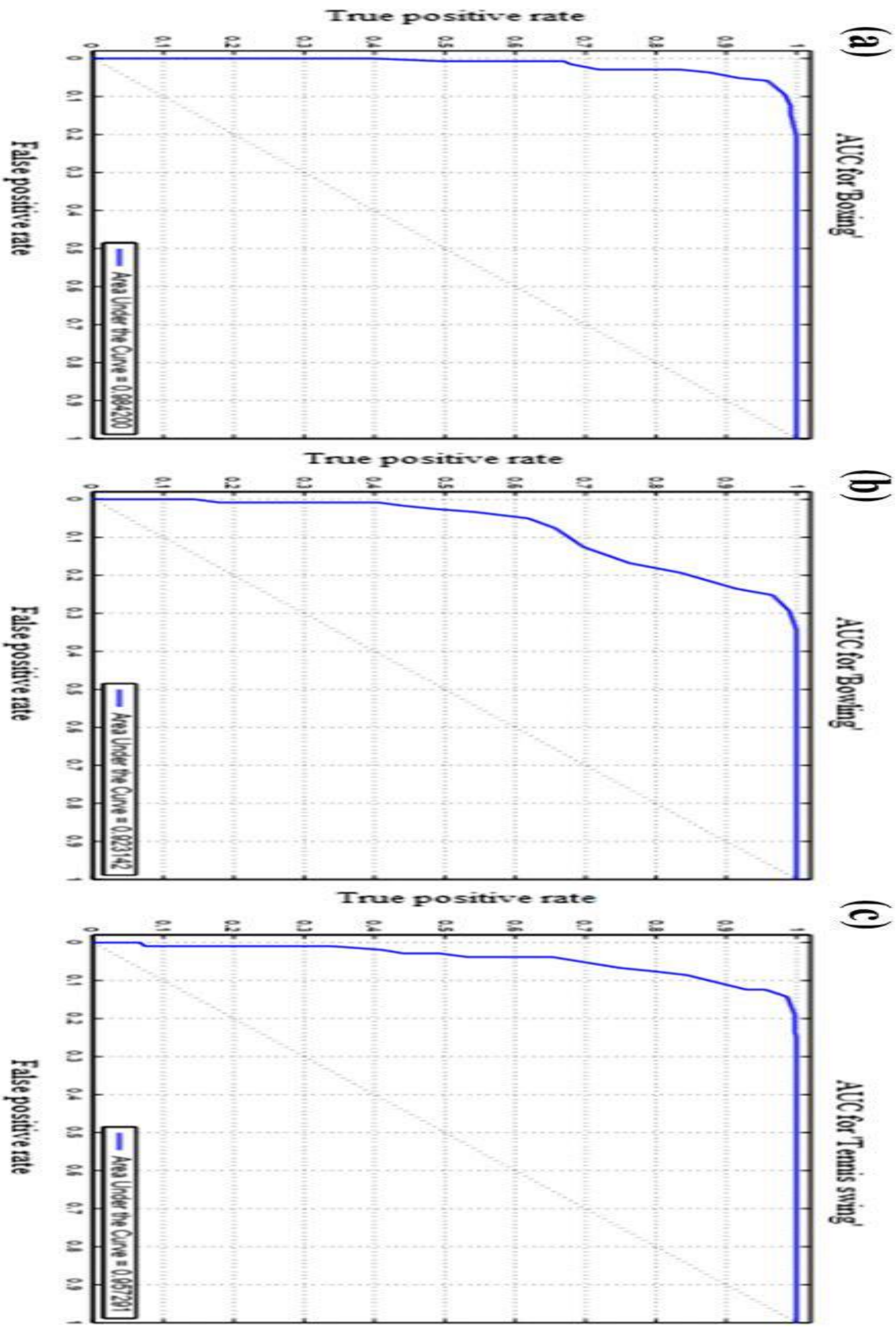


Figure 3.7 AUC for three actions (a) boxing, (b) bowling, and (c) tennis swing

Figure 3.8 and Figure 3.9 depict the accuracy and loss of the training dataset and validation data for 80 epochs. The graph demonstrates that the validation process achieved an accuracy of 96.2% and a loss of 0.45%.

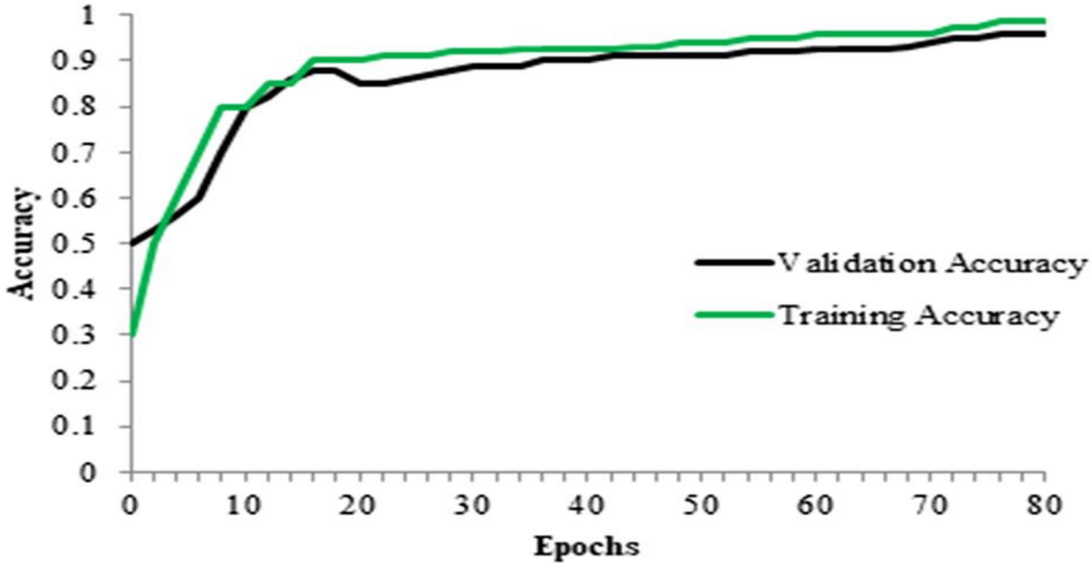


Figure 3.8 Accuracy versus Epoch

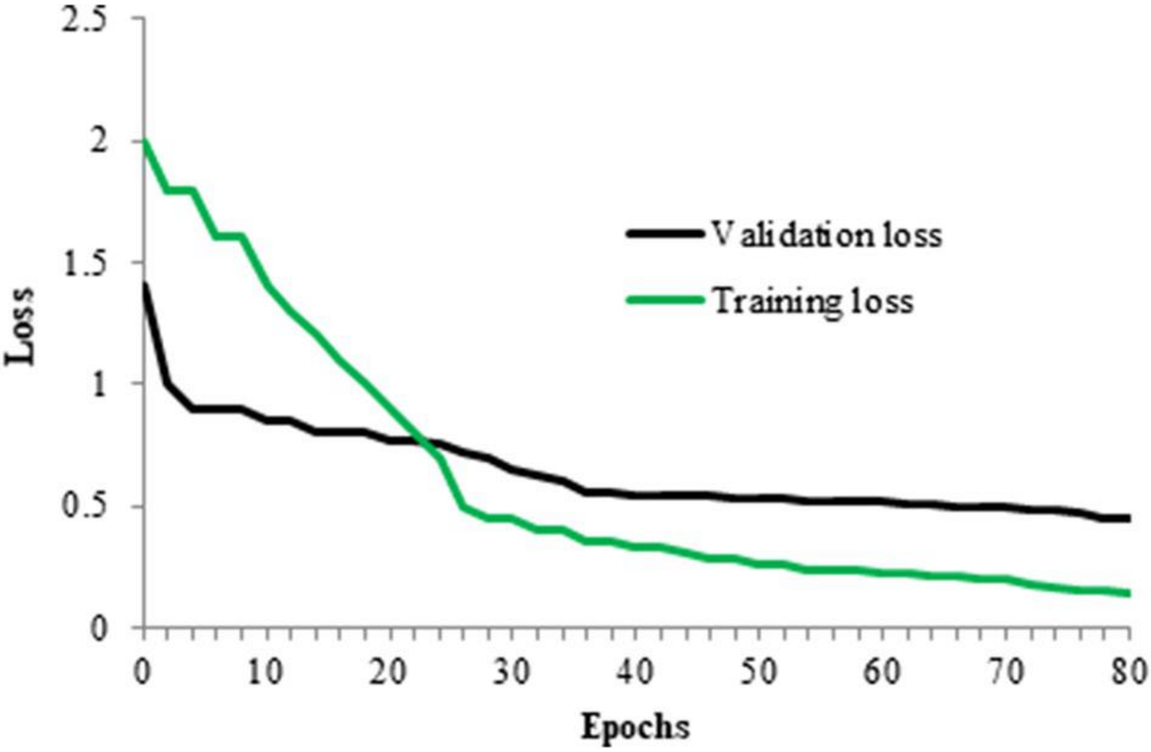


Figure 3.9 Validation and Training Loss per Epoch

3.5.3 Comparative Results

To illustrate the effectiveness of the recommended approach, the performance of our task has been compared to that of three different classifiers: SVM-based HAR, KNN-HAR, and CNN-based HAR. Predictions were made using a novel 5S-CNN + Bi-LSTM classifier in this study. Precision, accuracy, recall, and recognition rate were assessed to determine the performance.

The accuracy of the proposed methodology is analyzed in relation to the variation of training and testing data size, as presented in Figure 3.10. The recognition process employs three distinct methods, MHI, DMMs, and skeletons, as per the recommended approach. Features are extracted from each technique. The aforementioned characteristics are trained through specific classifiers.

The 5S-CNN classifier is used for MHI and DMM training. The skeleton image is trained similarly using the 5S-CNN and Bi-LSTM classifiers. The application of this hybrid procedure increases the level of accuracy. In order to demonstrate the efficacy of the proposed method, we undertake a comparative analysis of our algorithm against the K-Nearest Neighbour Classifier (KNN), Support Vector Machine (SVM), and CNN Classifier. The methods in this program are trained utilizing a uniform classifier. As depicted in Figure 3.10, the proposed approach attained a peak precision of 96.2%. Specifically, the KNN-based recognition yielded an accuracy of 79%, while the CNN-based recognition techniques resulted in accuracies of 83% and 86%, respectively. Moreover, the analysis of the suggested approach's performance with respect to accuracy is presented in Figure 3.11.

As depicted in Figure 3.11, the proposed approach yielded a peak precision of 96.5%. This outcome surpasses the recognition performance of K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Convolutional Neural Network (CNN) based methods.

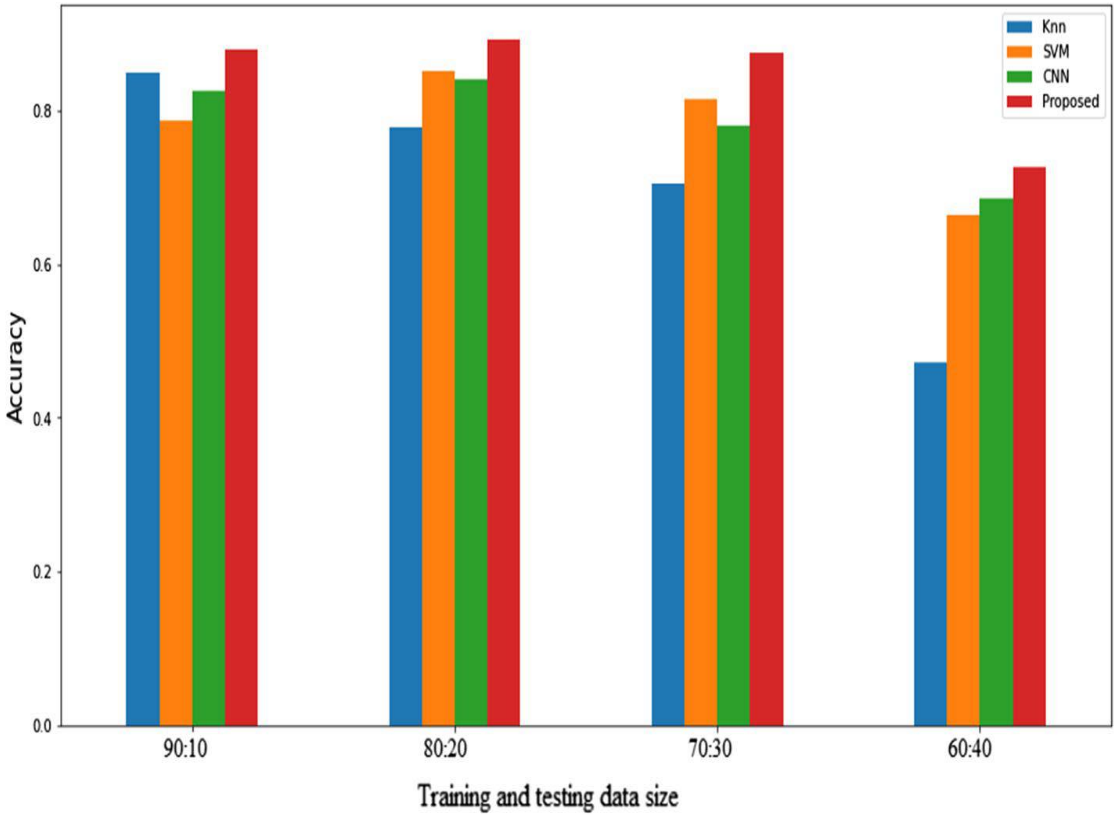


Figure 3.10 Analysis of performance based on Accuracy

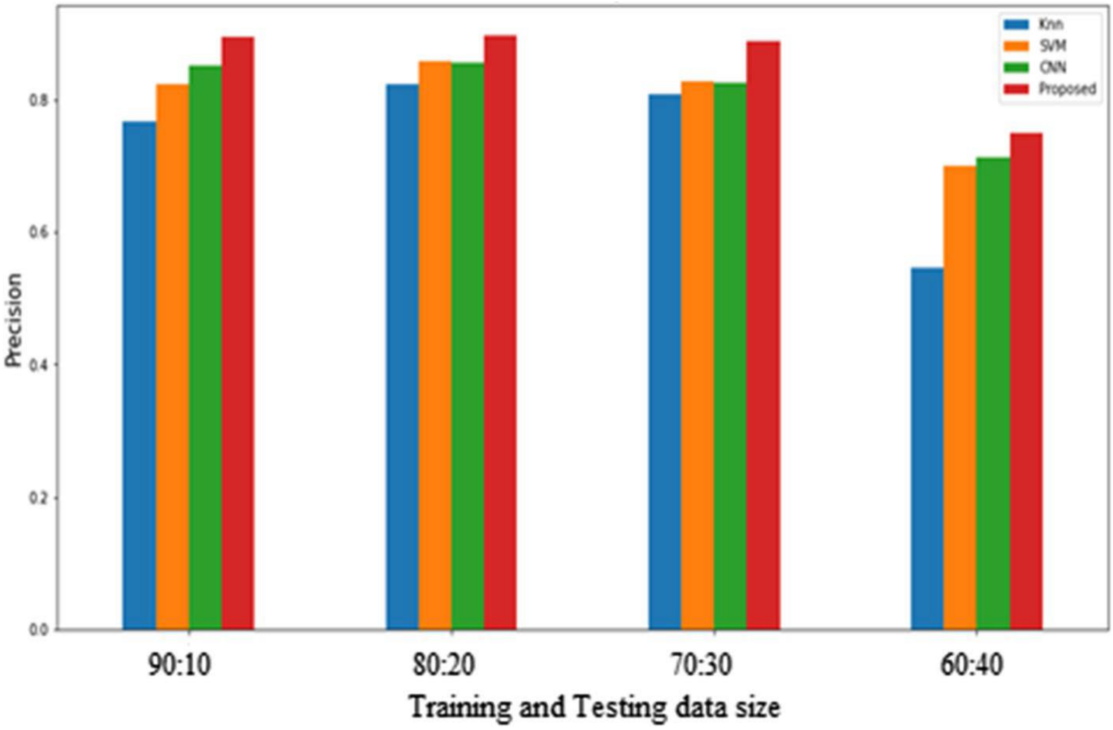


Figure 3.11 Analysis of performance based on Precision

The observed result can be ascribed to the application of a hybrid 5S-CNN and dual-LSTM-based skeletal training methodology, along with a three-step fusion procedure. The performance of the suggested approach is analyzed with respect to the size of recall in Figure 3.12. An effective organization possesses the most extraordinary capacity for recall.

As indicated by Figure 3.12, the prescribed approach entails extracting a maximum of 95%, which exceeds the values obtained through alternative methodologies. The human performance recognition performance, based on authorization rate, utilizing a multi-view multi-model approach, is presented in Figure 3.13. The measurement of the approval ratio is based on the value of the fusion score. The fusion process involves the integration of MHIs, DMMs, and skeletons. As indicated by Figure 3.13, our proposed methodology yields a peak recognition rate of 0.6. Specifically, the SVM-based recognition achieves a recognition rate 0.53, while the CNN-based recognition attains a recognition rate of 0.52. Additionally, the CNN-based recognition yields a recognition rate of 0.46.

The employment of 5S-CNN and Bi-LSTM-based training and integration techniques has resulted in higher accreditation rates in comparison to alternative methodologies. Furthermore, the analysis of time complexity is presented in Figure 3.14.The quantification of the amount of time an algorithm takes to execute in relation to the input's length is referred to as the algorithm's time complexity. Upon analysis of Figure 3.14, it can be observed that the proposed method exhibits a longer execution time in comparison to other techniques. This phenomenon can be attributed to the utilization of a hybrid approach. The duration of processing has no impact on the overall accuracy of recognition.

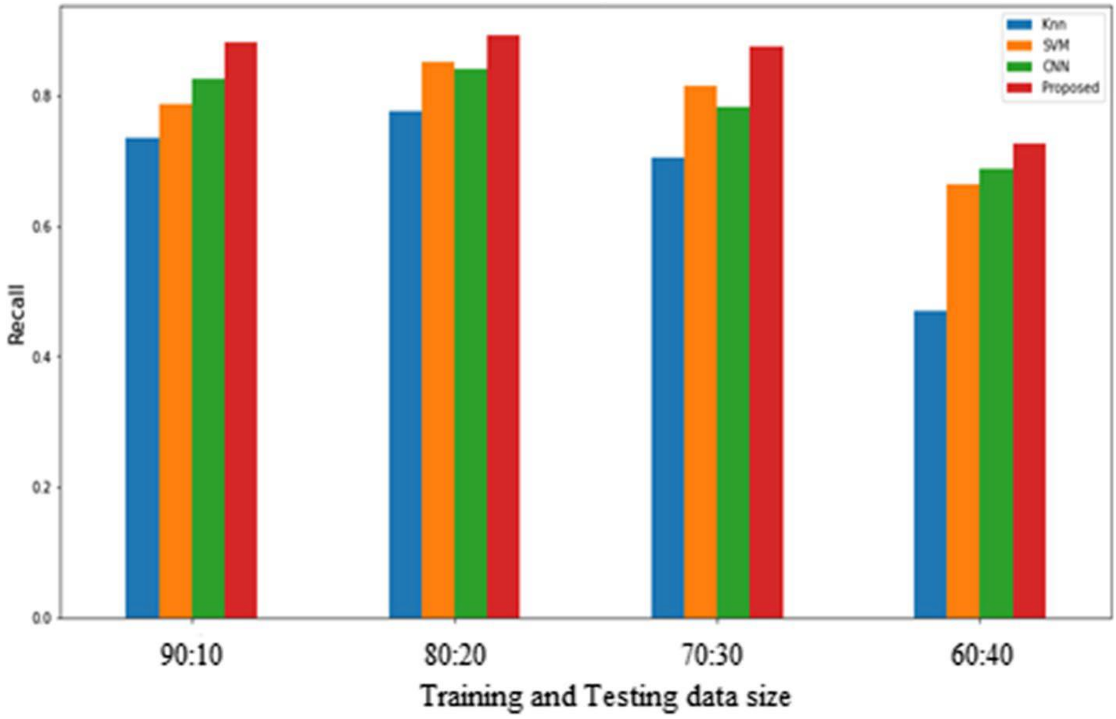


Figure 3.12 Analysis of performance based on Recall

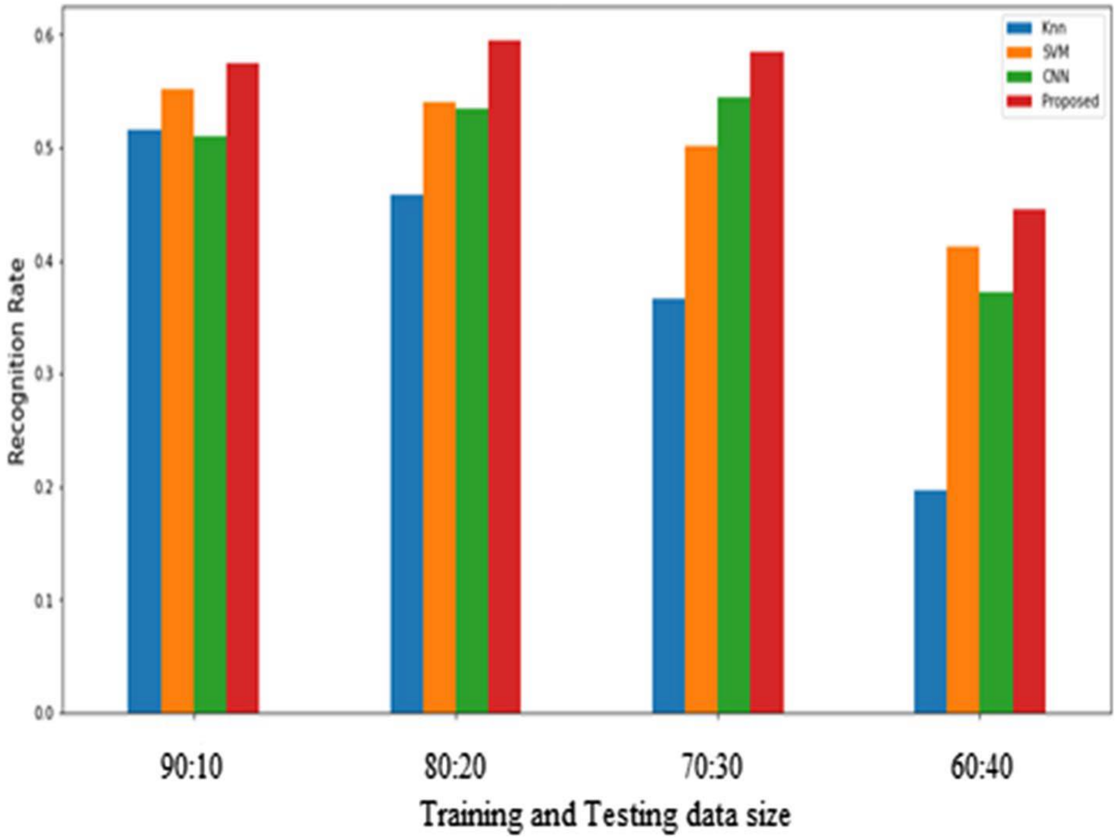


Figure 3.13 Analysis of performance based on Recognition Rate

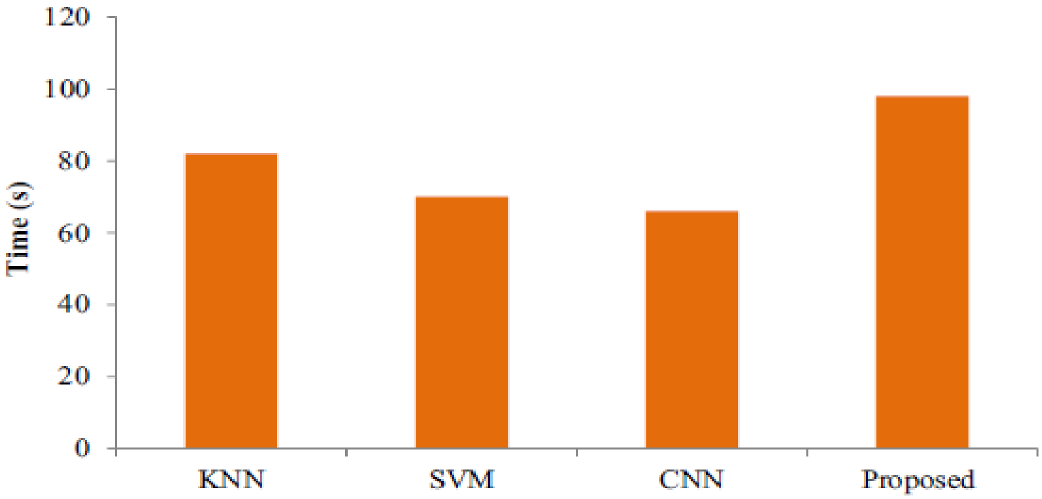


Figure 3.14 Complexity-based Analysis of proposed Algorithm

3.5.4 Comparison with State of the Reference Methods

In order to demonstrate the efficacy of the proposed approach, a comparative analysis was conducted between the suggested algorithm and alternative methods, as presented in Table 3.1. Within this section, we undertake a comparative study of our research with previously published works, including those of Chen et al. [204], Bulbul et al. [205], Annadani et al. [206], and Escobedo and Camara [197]. According to the data presented in the table, the precision of the suggested methodology is 96.2%. The recognition of human action is conducted in [204] through the utilization of depth and inertial sensors.

This study employed the collaborative representation classifier (CRC) for recognition purposes. The UTD-MHAD dataset is used in this study. The aforementioned approach proves to be efficacious in the recognition of human actions across multiple modalities. The authors of reference [205] employ a fusion strategy that integrates three distinct types of features to address the AR challenge. The present study involves the extraction of three-dimensional mathematical models (DMMs) through the utilisation of front, side, and top projection views obtained from input images or videos. Two decision-level fusions were developed in this context. The authors of reference [206]introduced a method for activity recognition based on skeleton data. The recognition process in [197]employs intensity, depth, and skeleton joints.

Approaches suggested in. [204], [205], [206], and [197] are among the finest multimodal human activity recognition programs currently available. In addition, they employed MHI, DMMs, and skeletal representation. We have therefore chosen to evaluate the performance of our proposed procedure to these. Table 2 compares the suggested framework with the previously mentioned method. When analyzing Table 3.1, our proposed method achieved a maximal precision of 96.2%, compared to 79.1% for [204], 88.4% for [205], 86.12% for [206], and 84.2% for [197]. The rationale behind our proposed strategy lies in its integration of three distinct training methodologies and a hybrid approach. The existing process was limited to utilizing only one or two methods. The results demonstrate that the proposed method yielded superior outcomes to alternative methods.

Table 3.1 Comparison with Prior Approaches.

Approaches	Accuracy (%)
Chen et al., [204] (Depth + Inertia)	79.1
Bulbul et al.[205] (Depth)	88.4
Annadani et al.[206] (Skeleton)	86.12
Escobedo and Camara [197] (RGB + Depth + Skeleton)	84.4
Proposed (RGB + Depth + Skeleton)	**96.2**

3.6 Discussion

The primary objective of this study is to propose a framework for human activity recognition that utilizes multiple views and models. The problem of human actions is addressed by integrating the multiple view notes obtained from the RGB-D sensor for recognition. The first step involves the extraction of features from the input image. Subsequently, each individual feature is trained using separate classifiers. The mathematical expression for each phase is depicted. The experimental findings provide evidence that the suggested approach is suitable for diverse data methodologies and achieves superior outcomes compared to other existing studies. In the future, the focus of the Recognition procedure will be on the optimization of processes and the establishment of distinct classifications.

Chapter 4

Novel Light-Weight Deep Learning Model for Human Action Recognition in Videos

Human Action Recognition (HAR) from a video feed has recently attracted considerable attention from computer vision researchers. Due to its extensive applications, including health surveillance, home automation, and tele-immersion, among others. However, it is still susceptible to human differences, occlusion, illumination variations, and complex backgrounds. The evaluation criteria rely on the accurate execution of the features collection method and learning data. The accomplishment of Deep Learning (DL) has produced numerous impressive outcomes, including neural networks.

Nonetheless, an efficient classifier must have a robust features vector in order to provide the class label. Features are the fundamental element of any data set. Indeed, feature extraction may affect the algorithm's efficacy and computational expense. For this research framework, we extracted features from an image sequence using pre-trained deep learning models VGG19, Dense Net, and Efficient Net and classified each action using the SoftMax layer. UCF50 action dataset consists of 50 sections and measures performance using precision, recall, f1-score, and AUC score. Testing model accuracy yielded VGG19-90.11, DenseNet-92.57, and EfiicientNet-94.25.

4.1 Introduction

In HAR, an action is an observable object that the human eye or a sensor device can detect. An activity such as strolling necessitates constant focus on a person in the field of vision. Actions can be divided into four categories based on the body parts required to perform them. [1].

- Gesture is determined by facial expression. Not requiring any form of action or verbal communication.
- Action consisted of walking, playing, and striking.
- Interaction: human-object interaction and human interaction, such as a salutation. An example of interaction is a hug.
- Group activity: a group activity occurs when more than two actions occur, such as the combination of gesture and interaction. There are two or more actors involved in acting.

Over the past two decades, HAR has become integral to computer vision research. Based on a compilation of observations, HAR is intended to detect and identify the activities of one or more individuals. This can be performed for any number of individuals. This subsequently necessitated the advancement of human –computer interaction. Numerous researchers from around the world are drawn to this field of study due to its broad spectrum of applications. Among its most prominent applications are surveillance video, labelling and retrieving imagery, health monitoring, automation, and environmental modelling [9]. Human activities have an inherent hierarchical structure that denotes their numerous levels, which can be classified into three categories. First, there is a fundamental atomic element, and these action primitives represent progressively complex human actions. The actions/activities level is the second level after the action primitive level. The utmost level of categorization for human activities is complex interactions. Each category is adequately broad to warrant its field of study. This is primarily due to the unpredictability and ambiguity of actual human behaviour. HAR confronts numerous obstacles and impediments.

HAR has emerged as a crucial component of computer vision research during the last decade. Based on an accumulation of observations, HAR is intended to detect and identify the activities of one or more individuals. This can be carried out for any number of people. Consequently, advancements in human–computer interaction were required. Due to its extensive range of applications, this field of study attracts many scholars from all over the globe [9] Its most prominent applications include surveillance video, image labelling and retrieval, health monitoring, automation, and environmental modelling. The inherent hierarchical structure of human activities denotes their numerous levels, which can be divided into three categories. These action primitives represent increasingly complex human actions. The actions/activities level follows the action primitive level as the second level. Complex interactions are the highest classification level for human activities. Each category is sufficiently comprehensive to necessitate its own academic discipline. This is attributable primarily to the unpredictability and ambiguity of actual human behaviour. HAR faces numerous obstacles and hindrances.

In the following step, actions are taught and recognized using extracted features. Essential components of action learning and identification incorporate learning new models introduced by the obtained features, determining which features are relevant to which action classes, and evaluating those features with the aid of classifiers.

Machine Learning (ML) and Deep Learning (DL) techniques are two of the most prominent approaches to addressing the HAR issue. Conventional Artificial Intelligence takes a more deterministic approach, requiring the user to layout, prescribe, and refine extracted features and define action. We anticipate the deep neural network (DNN) by employing the latter strategy. Using the latter approach, on the other hand, we predict that the DNN will independently solve all the attributes by simulating the human brain [9][207].

ML-based methods such as random forest (RF), Bayesian networks (BN), Markov models (MM), and support vector machine (SVM) have been used for decades to solve HAR-related problems such as debris background, noise problems, and class similarity problems. In contexts with limited data inputs and stringent restrictions, conventional ML algorithms have performed admirably. Due to pre-processing stages with handcrafted features, machine learning algorithms are time-consuming and require special consideration; they must be improved. If the data size is substantial. DL has made significant progress in recent years. This is because research on deep learning has yielded outstanding results in various disciplines, including object detection and action recognition, frame classification, and natural language processing.

DL substantially reduces the effort required to select the appropriate features compared to conventional ML algorithms, and its structure is suitable for unsupervised learning and reinforcement learning. Consequently, the number of proposed deep learning based HAR frameworks has increased. Figure 4.1 represents the ML-based and DL-based process to deal with the HAR problem.

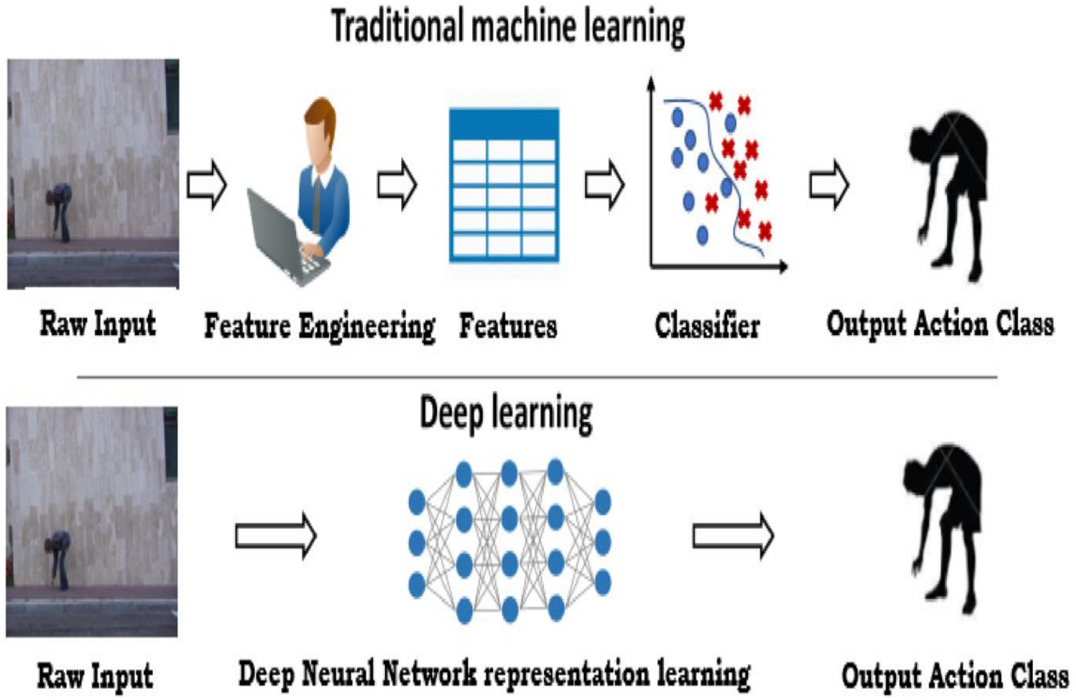


Figure 4.1: A graphical representation of the conventional ML methods and the cutting-edge DL methods employed for HAR [207]

4.2 Literature Work

In HAR, significant progress has been made. Due to its vast array of applications, numerous opportunities exist to enhance the prediction of human behaviour. In the past decade, numerous manually produced and automatically learned feature-based techniques for human activity detection in visuals have been developed. Earlier designs for human activity identification relied on handcrafted characteristics that were primarily concerned with small atomic activities that appear relevant to actual applications [208].

Despite producing highly accurate models, the primary drawback of these techniques is that they require extensive data pre-processing and are difficult to generalise in practise.

Following the success of convolutional neural networks (CNNs) in text and visual classification, numerous spatiotemporal approaches have been developed for video activity analysis; these algorithms can automatically train and classify from raw RGB video [135]. Shuiwang Ji et al.[209] developed a 3D convolution technique for extracting spatial and temporal video data for action recognition. Consequently, the proposed architecture generates multiple data channels from the video sequence and applies convolution and subsampling independently to each channel. Gu et al. proposed a DL-based technique for identifying locomotive motions pertinent to indoor localization and navigation. Their solution utilized stacked denoising auto-encoders that automatically learned data characteristics, reducing the need to construct the pertinent features [213] explicitly[210]. The proposed research framework is said to have greater precision than another classifier. Aubry et al. [211] devised a new method for determining what action is by analyzing RGB (Colour model) video. First, the human skeleton must be extracted by removing the motion from the video. This extraction was performed using Open Pose [212] a Deep Neural Network (DNN) employing an identification method that obtains a 2-D skeleton with 18 known joints from each body object. In another scenario, motion patterns are converted into RGB images using an image classifier. Channels R, G, and B are used to store motion information. It produces an RGB image for use in an action sequence. Future neural networks that classify frames may be able to identify actions.

Dai et al. [216] proposed a dual-stream model that employs an attention-based LSTM structure for localizing action within visual frames. They claimed to have resolved the issue of disregarding visual attention. With the UCF11 dataset, the accuracy of the architecture was 96.9%. The accuracy of the UCF Sports dataset was 98.6%, while the accuracy of the j-HMDB dataset was 76.3%. Using a hierarchical RNN model, Du et al. [213]developed a skeleton-based framework for recognizing actions based on a skeleton. In addition, five distinct deep RNN designs based on their suggested methodologies were evaluated (compared). They utilised the MSR Action-3D dataset, the Berkeley MHAD dataset, and the HDM05 dataset throughout their evaluation.

The Correlational Convolutional LSTM was created by Majd and Safabakhsh [214] by incorporating spatial and motion information into a pre-existing LSTM module and establishing temporal links.

On the widely used benchmark datasets UCF101 and HMDB51, on which their work was evaluated, they attained correctness rates of 92.3% and 61.0%, respectively. To recognize group and individual activities, Qi et al. [215] developed stag-Net, an alternative method for constructing a semantic RNN.

Using a structural RNN, they added a fourth dimension to their semantic network model: time. 90.5% of the Volleyball dataset was completed as a team endeavor and 8.5% as an individual effort using this method.

Huang et al. [216] extract posture-based characteristics from a 3D convolutional neural network (ConvNet) by fusing 3-D Pose, 2-D appearance, and motion information. It is anticipated that the computation-intensive nature of 3-D CNNs derived from color-joint frame features will be a factor. To reduce noise, we apply convolution to each of the 15 channels of the heatmap. Wang et al. used the (BN-inception) network architecture in Inception and Batch Normalisation [217]. As with two-stream networks, the method described above employs RGB variation frames (to simulate appearance change) and optical flow fields in conjunction with RGB and optical flow frames (to inhibit background motion).

In [218], the author employed the GCN with a channel attention strategy for joints and graph pooling networks. The SGP design ultimately incorporated the human skeletal network and enhanced the convolution. Use of kernel-receptive regions to acquire specific human body data. The proposed SGP method has the potential to significantly enhance GCNs' ability to collect data based on motion characteristics while reducing computation costs.
The study [223] employed context stream and fovea stream designs. The context channel receives frames with half their original resolution, whereas the fovea channel receives the central region with full resolution. Using each video as a compilation of short, fixed-length segments, the research teaches a model to identify three pattern classes: Early, Late, and Slow Fusion. CNN may produce single-frame animations by combining time and space in several methods.

Singh et al. [219] proposed a highly connected ConvNet with RGB frames as the top layer for identifying human activities-term LSTM. Individual DMI are utilized to train the ConvNet model's lowest layer.

The ConvNet-Bi-LSTM model is trained from inception for RGB frames to enhance the features of the pre-trained CNN. In contrast, the uppermost layers of the pre-trained ConvNet are modified to extract temporal information from video feeds. Using a late fusion technique that follows the SoftMax layer, the decision layer combines features to produce a value with a higher degree of precision by combining them after the SoftMax layer. Four RGB-D (depth) datasets consisting of single-person and multi-person activities are utilized to evaluate the efficacy of the proposed methodology.

4.3 Proposed Approach

The DL model for HAR indicates the effect of each activity's classification. We discussed several deep-learning models, their classification accuracy, and how they operate. To train a DL model from scratch requires lots of computation power. Compared to transfer learning models, learning models are trained. They are learned using the massive quantity of ImageNet data [220]. ImageNet contains over 1 million images that can be used to train transfer learning models. This research paper classified each action using various transfer learning models and compared them to cutting-edge techniques. This study compared numerous transfer learning models to human action recognition. Human Action Recognition model with the pre-trained deep learning model is depicted in Figure 4.2.

Transfer learning (TL)-based methodologies are evaluated using Dense Net [221]. Dense Net neural networks were selected due to their innovative techniques for contending with vanishing or growing gradients and their unique architecture, which enables one layer to learn from the feature maps of antecedent layers, thereby enabling feature reuse. VGG[190] is also trained using a transfer-learning-based HAR method due to its extremely deep architecture, which is achieved by employing minuscule (3 3) filters. Due to their complexity, VGG models frequently undergo gradient eruptions. We utilised VGG models with batch normalization layers to solve this issue to maintain gradient control. Efficient Net [222] method is also used to assess framework performance.

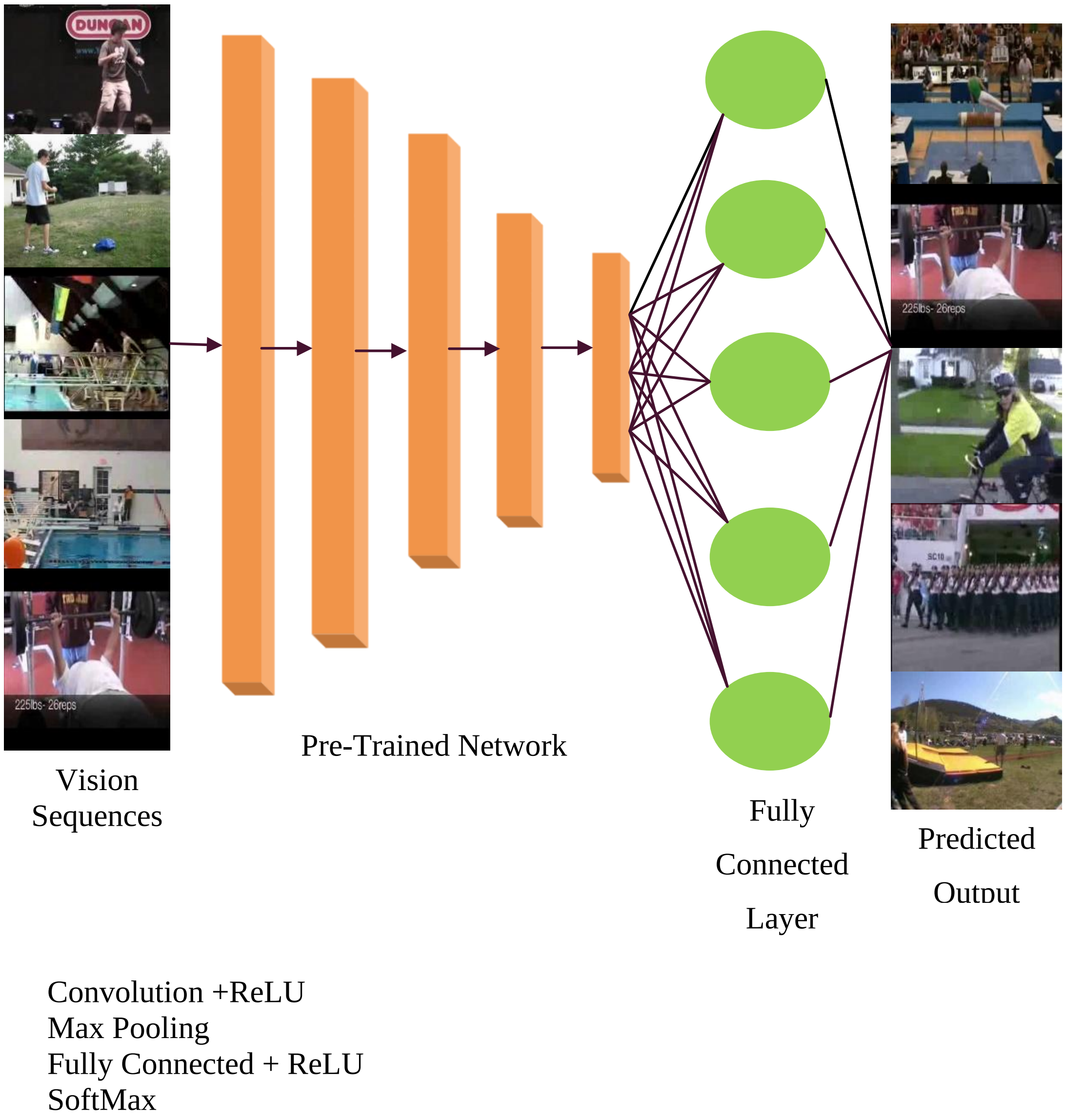


Figure 4.2 Proposed Methodology Architecture

4.3.1 Dense Net

A dense Convolution neural network (Dense Net) [221] interconnects each successive network layer via feed-forward; this extensive interconnection has garnered it the moniker Dense Net. The data is initially routed through a Conv2D layer with a large filter size, followed by a dense block that creates dense connections with all subsequent layers. Each layer of a Dense Net receives new inputs from all preceding layers and broadcasts its feature maps to all succeeding layers.

4.3.2 VGG

VGG [190] is a CNN architecture that we incorporated into the TL-based strategy to recognize human actions. The images provided to VGG for training are 512 by 512 pixels (224, 224, 3) with a strict aspect ratio of 224, 224, 3. These images have been processed with convolutional layers containing 3-by-3-pixel filters. After particular conv2D layers, five max-pooling layers perform spatial pooling. Following a set of convolutional layers are dense layers with full connection and a SoftMax prediction layer. Figure 4.3 displays the VGG19 architectural design, in which conv represents the convolution layer, the pool is the pooling layer, and FC is the completely connected layer.

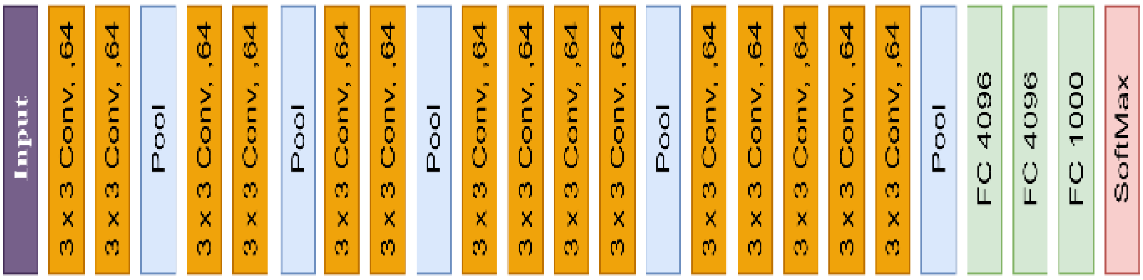


Figure 4.3 Architecture of VGG 19 [190]

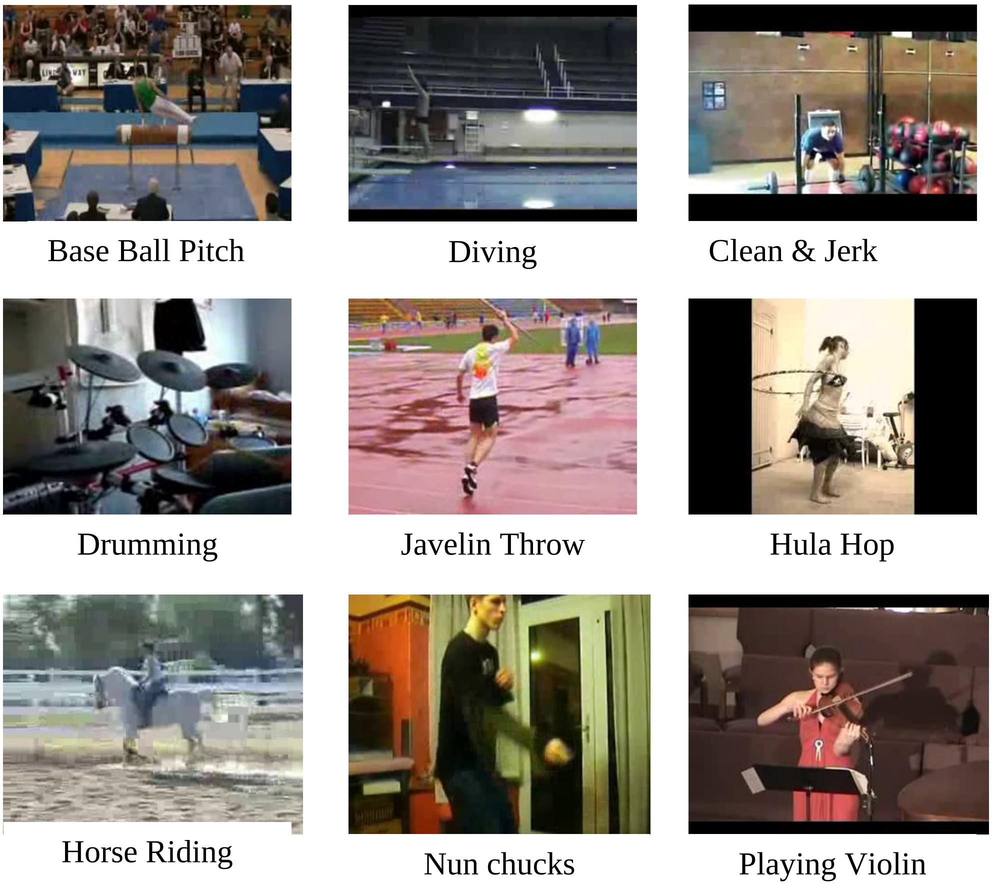
4.3.3 Efficient Net

Efficient Net [222] is an architectural and scaling strategy for convolutional neural networks that employs a compound coefficient uniformly scale all depth/width/resolution parameters. In contrast to current practice, which scales these elements arbitrarily, the Efficient Net scaling technique modifies network dimension (Breadth, Depth and Resolution) uniformly using a set of predefined scaling factors.

Efficient Internet [222] It is an exceptional CNN network with great parameter estimation efficiency and speed. Efficient Net [222] proposed a simple and complex scaling strategy to scale up CNN models more systematically by uniformly scaling network features such as depth, breadth, and resolution. Efficient Net [222] was also used as a spatial feature extraction network in classification applications. The Efficient Net family included seven CNN models labelled EfcientNet-B0 through EfcientNet-B7. EfcientNet-B0 surpassed Resnet-50[22] with fewer parameters and FLOPs (floating-point operations per second) accuracy, demonstrating that EfcientNet-B0 can extract features efficiently.

4.3.4 Description of Dataset

The UCF50 [223] dataset was utilised to assess model performance. Reddy et al. offered this dataset in 2012. For video gathering, online sites such as YouTube are employed. All videos are shot in a realistic setting and are not from a controlled environment. This dataset has been updated from the UCF11 dataset. It includes 50 activity lessons such as basketball, shooting, tabla, riding, violin, and so on. There are 6618 videos in all, covering a wide range of activities from general sports to daily life activities. Each activity class is divided into 25 homogenous groups, with at least four films assigned to each activity. The same person, background, or viewpoint may appear in many films in the same genre. Figure 4.4 depicts action snippets from the UCF 50 dataset.



Base Ball Pitch Diving Clean & Jerk

Drumming Javelin Throw Hula Hop

Horse Riding Nun chucks Playing Violin

Figure 4.4 Dataset Frames [223]

4.4 Experimental Results

To classify each activity, we used three pre-trained deep-learning models—Dense Net, VGG19, and Efficient Net. We utilised pre-trained deep learning to extract value from the data collected from enormous datasets such as ImageNet. The transfer learning method trains a neural network on a new domain by transferring information from a previously trained model. The UCF50 action dataset, which comprises numerous image categories, is evaluated. In this approach, we assessed the effectiveness of multiple DL models on the dataset and compared their accuracy to that of state-of-the-art methods. First, frames from each action video group were extracted and input into a pre-trained deep learning model. The confusion matrix for recognising 50 activities from the UCF 50 dataset using the

VGG19 model, Dense Net 161, and Efficient Net b7, is depicted in Figure 4.5, Figure 4.6, and Figure 4.7.

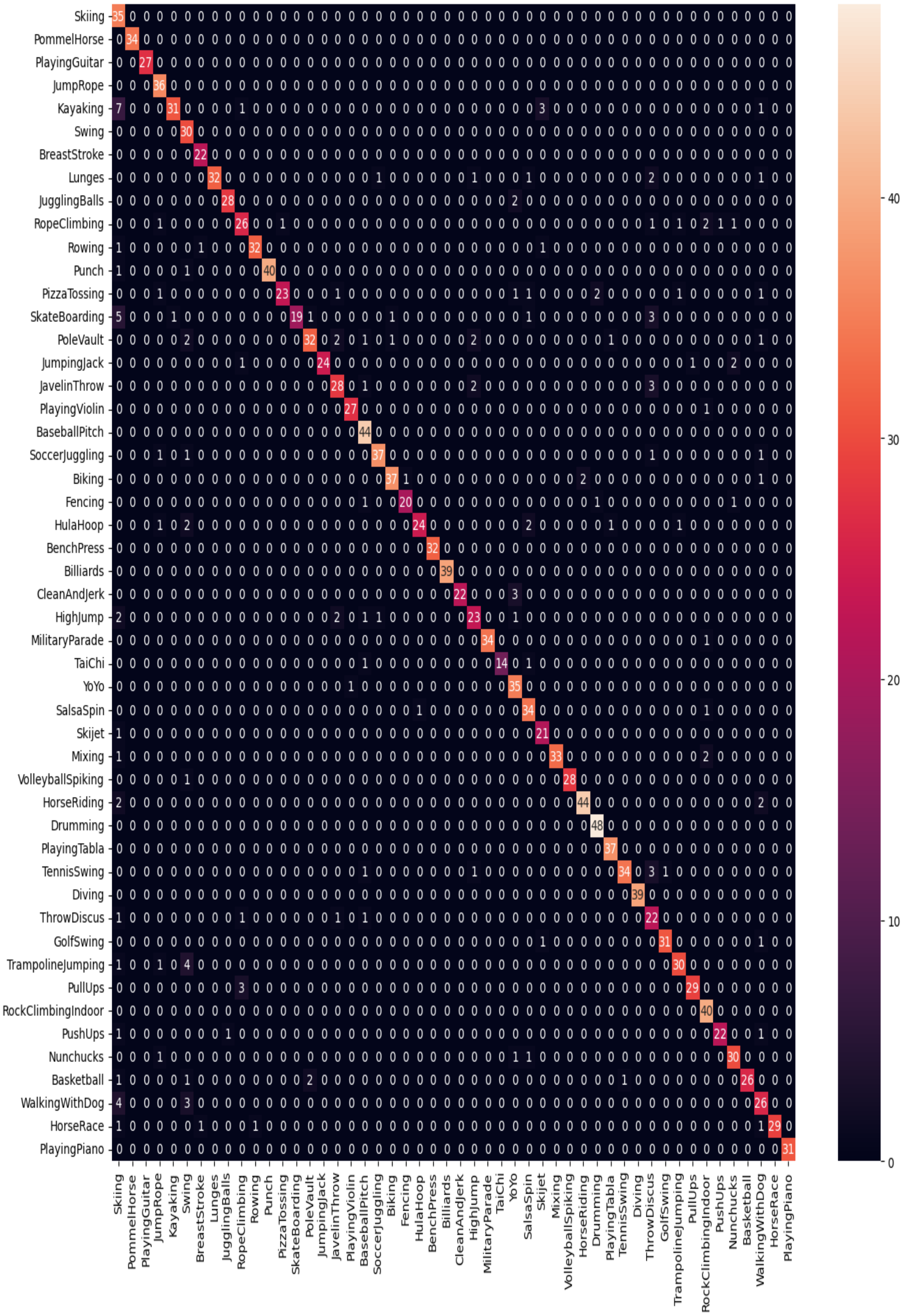


Figure 4.5 Confusion Matrox for VGG 19 Model based Technique

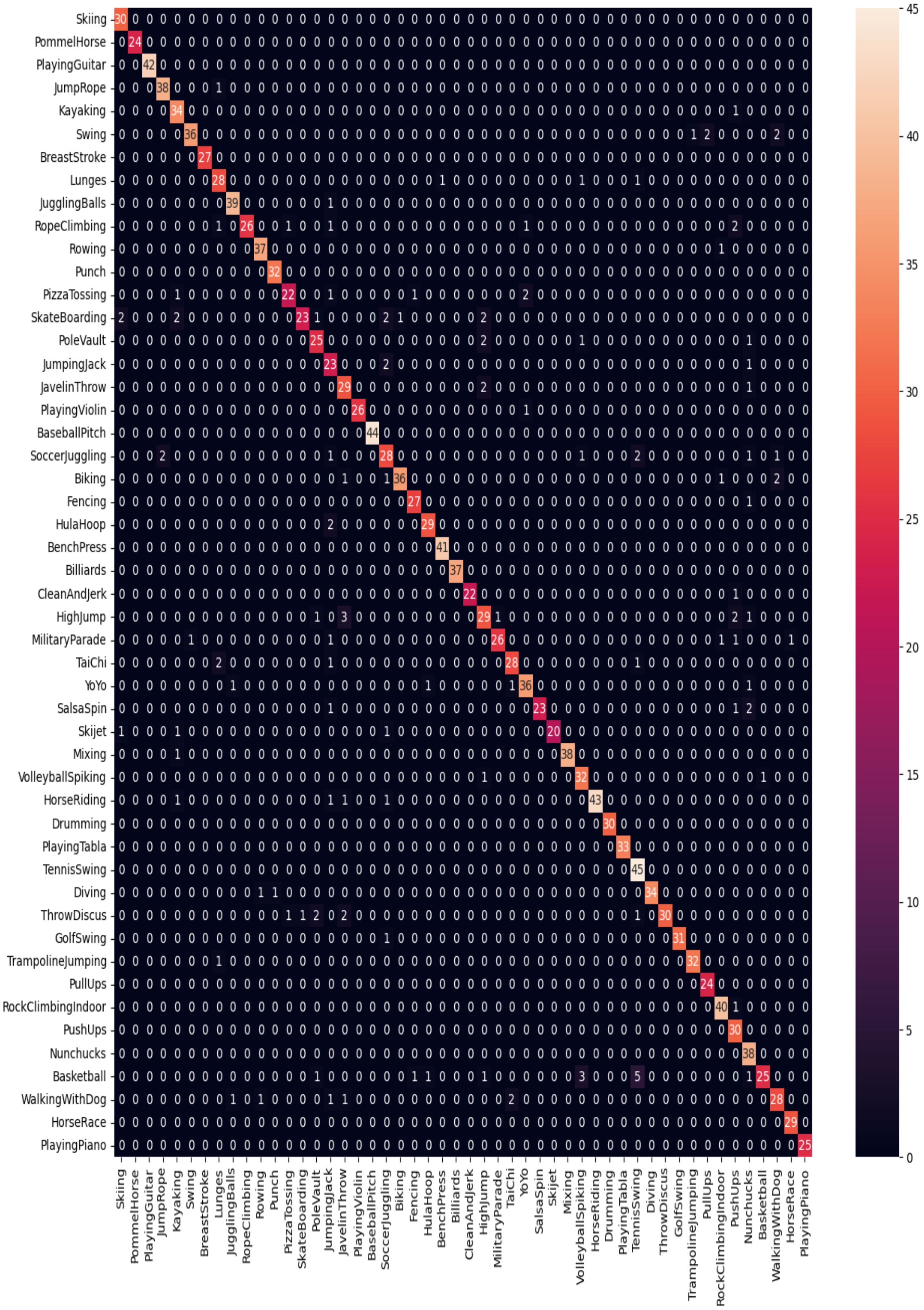
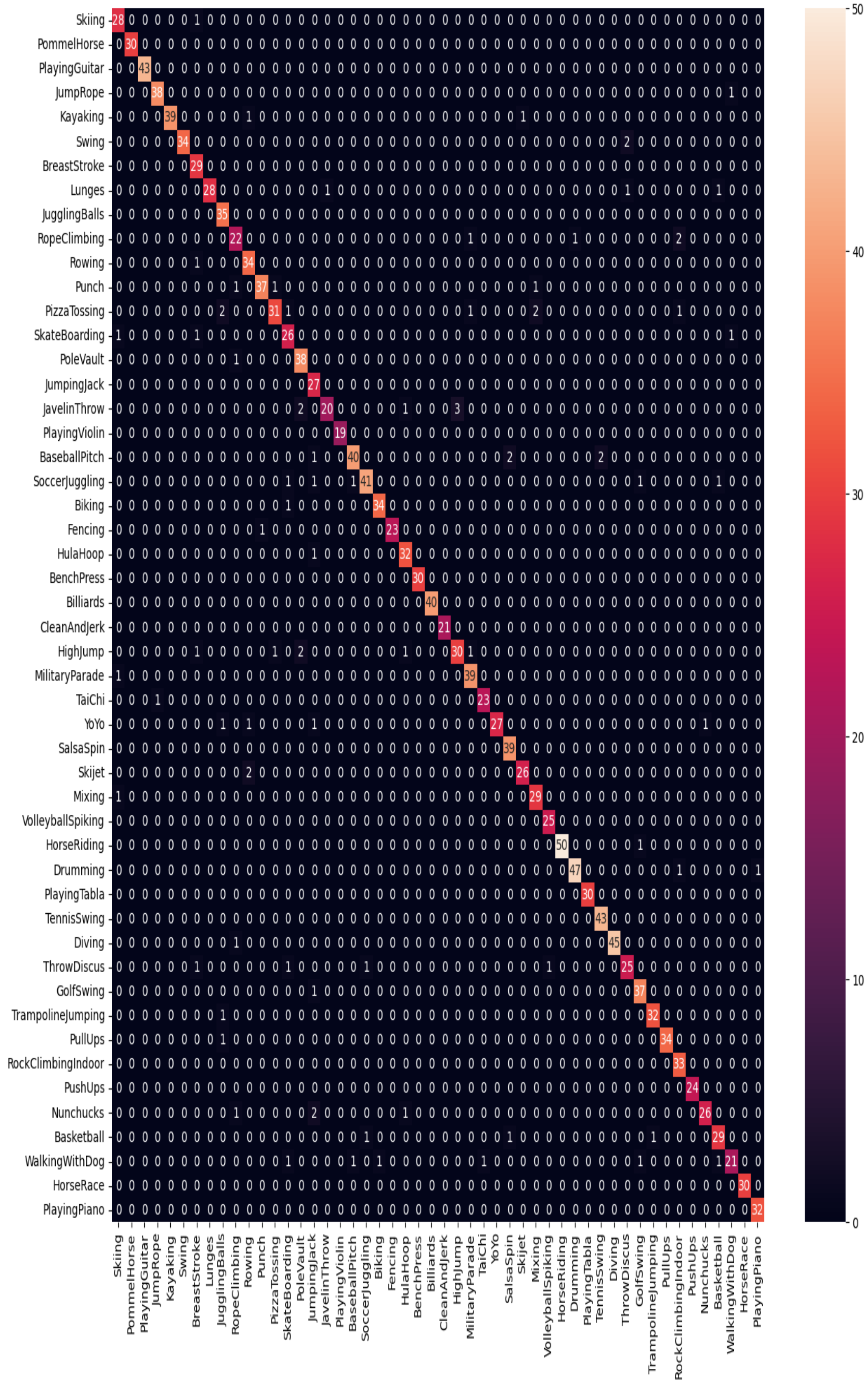


Figure 4.6 Confusion Matrox for Dense Net Model based Technique



	Skiing	PommelHorse	PlayingGuitar	JumpRope	Kayaking	Swing	BreastStroke	Lunges	JugglingBalls	RopeClimbing	Rowing	Punch	PizzaTossing	SkateBoarding	PoleVault	JumpingJack	JavelinThrow	PlayingViolin	BaseballPitch	SoccerJuggling	Biking	Fencing	HulaHoop	BenchPress	Billiards	CleanAndJerk	HighJump	MilitaryParade	TaiChi	YoYo	SalsaSpin	Skijet	Mixing	VolleyballSpiking	HorseRiding	Drumming	PlayingTabla	TennisSwing	Diving	ThrowDiscus	GolfSwing	TrampolineJumping	PullUps	RockClimbingIndoor	PushUps	Nunchucks	Basketball	WalkingWithDog	HorseRace	PlayingPiano
Skiing	28	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PommelHorse	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PlayingGuitar	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JumpRope	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Kayaking	0	0	0	0	39	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Swing	0	0	0	0	0	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
BreastStroke	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Lunges	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
JugglingBalls	0	0	0	0	0	0	0	0	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RopeClimbing	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0
Rowing	0	0	0	0	0	0	1	0	0	0	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Punch	0	0	0	0	0	0	0	0	0	1	0	37	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PizzaTossing	0	0	0	0	0	0	0	0	2	0	0	0	31	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
SkateBoarding	1	0	0	0	0	0	1	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
PoleVault	0	0	0	0	0	0	0	0	0	1	0	0	0	0	38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JumpingJack	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
JavelinThrow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	20	0	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PlayingViolin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BaseballPitch	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	40	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0
SoccerJuggling	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
Biking	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Fencing	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HulaHoop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BenchPress	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Billiards	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CleanAndJerk	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HighJump	0	0	0	0	0	0	1	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	1	0	0	0	30	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MilitaryParade	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TaiChi	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
YoYo	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
SalsaSpin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Skijet	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mixing	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
VolleyballSpiking	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HorseRiding	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Drumming	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0	0	1	0	0	0	0	0	1
PlayingTabla	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0	0	0	0	0	0	0	0
TennisSwing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	43	0	0	0	0	0	0	0	0	0	0	0	0
Diving	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	0	0	0	0	0	0	0	0	0	0	0
ThrowDiscus	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0
GolfSwing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0
TrampolineJumping	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	0	0	0	0	0	0	0	0
PullUps	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0	0
RockClimbingIndoor	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	0	0	0	0	0	0
PushUps	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	0	0	0	0	0
Nunchucks	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	0	0	0
Basketball	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	29	0	0	0
WalkingWithDog	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	21	0	0
HorseRace	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0
PlayingPiano	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32

Figure 4.7 Confusion Matrox for Efficient Net Model based Technique

The classification result for the UCF 50 activity dataset is displayed as a Confusion matrix. The majority of activities are classified accurately and with high precision. On the UCF50 action dataset, Table 4.1 contrasts frameworks evaluating matrices utilizing TL techniques. During the implementation phase, the captured frames were partitioned into training, validation, and testing phases. Figure 4.8 & Figure 4.9 represent the comparison of performance of models in HAR problem. Comparison with a variety of contemporary techniques, as shown in Table 4.2.

Table 4.1 Comparison of various Light-Weight DL Models.

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
VGG19	90.11	91.92	90.34	90.53
Dense Net 161	92.57	93.06	92.45	92.43
Efficient Net b7	**94.25**	**94.92**	**94.79**	**94.71**

Table 4.2 Comparison with State of the reference methods.

Researcher	**Dataset**	**Accuracy (%)**
L. Zhang et al[224]	UCF50	88.0
H. Wang et al[225]	UCF50	89.1
Q. Meng et. al[226]	UCF50	89.3
Ahmad Jalal et. al[227]	UCF50	90.48
VGG19_bn	UCF50	**90.11**
Dense Net 161	UCF50	**92.57**
Efficient Net_b7	UCF50	**94.25**

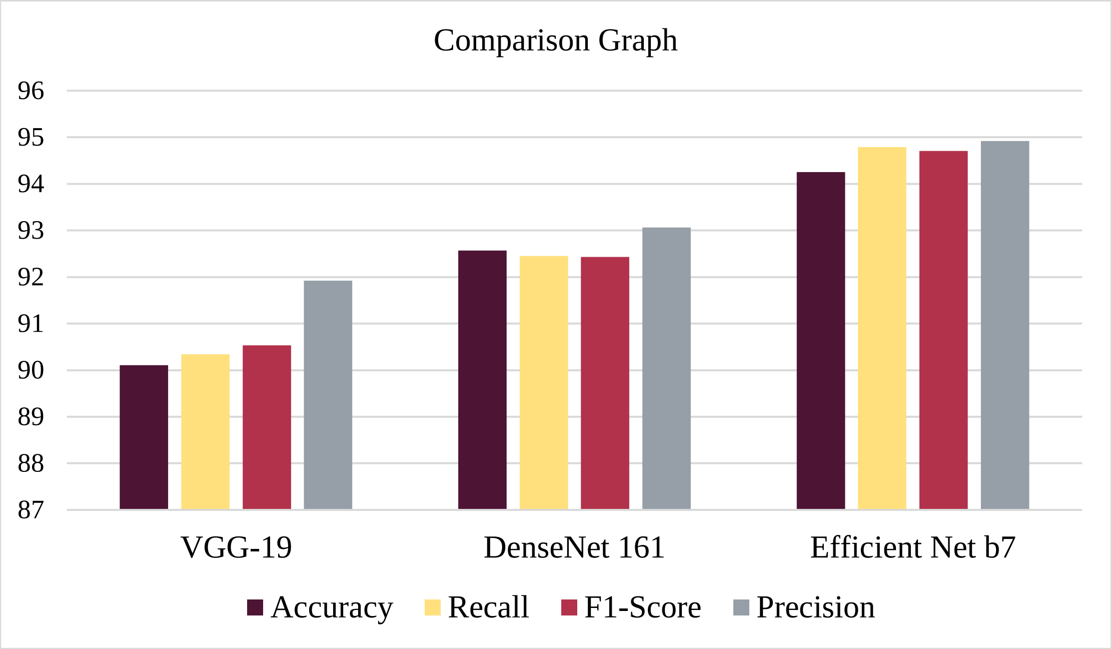


Figure 4.8 Comparison of Various DL-based Models

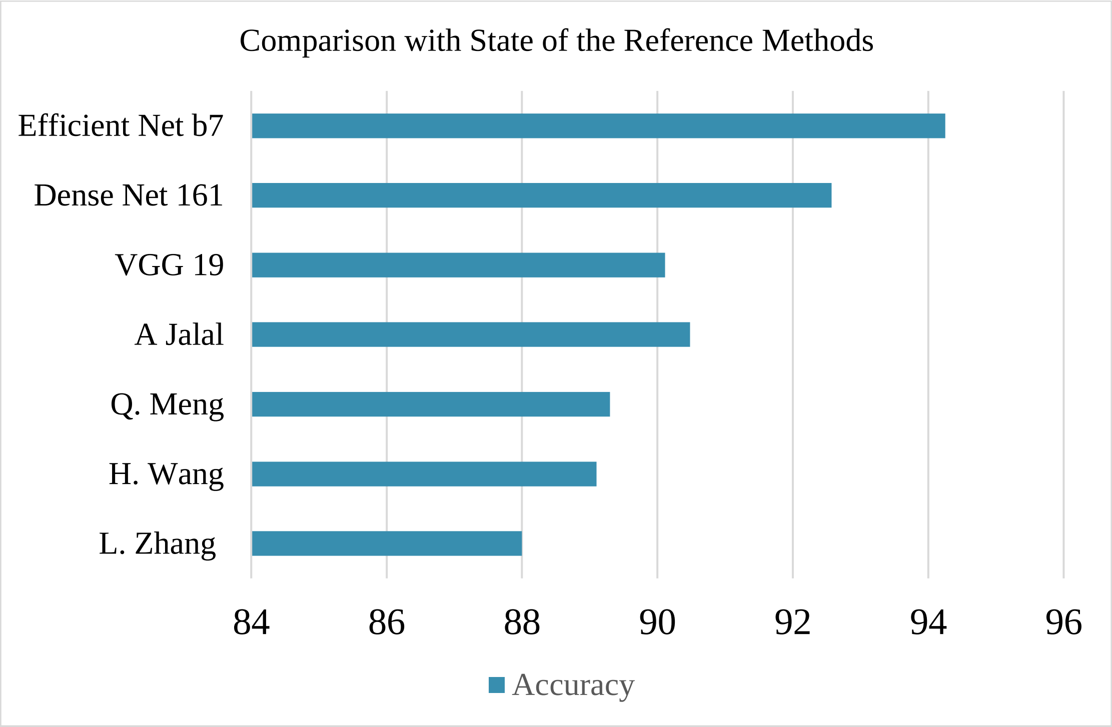


Figure 4.9 Performance comparison with State of the Reference Models

4.5 Discussion

Deep learning models that have been pre-trained are used to classify human actions from the UCF 50 action dataset. UCF50 action dataset comprises 50 distinct action categories organized into 25 groups, with at least four recordings per group. Multiple evaluation metrics, namely precision, recall, F1 score, and AUC score, were employed to evaluate the models' precision, effectiveness, and overall performance. Models VGG19, Dense Net 161, and Efficient Net classify each action in the dataset. This study also contrasted cutting-edge techniques applied to the UCF50 dataset. The performance of these pre-trained deep learning models is superior to current best practices. With 94% accuracy, Efficient Net outperforms other pre-trained deep learning models.

Chapter 5

Vision Transformer-based Human Action Recognition in Videos

Human Action Recognition (HAR) has attracted the interest of computer vision researchers due to its wide range of applications, including surveillance, behaviour detection, sports action monitoring, and elderly monitoring. As a result of the vast quantity of data, the Deep Learning-based method is more prevalent in HAR than the Machine Learning-based method. This investigation investigated the numerous Deep Learning and pre-trained Deep Learning models in HAR. In the pre-trained model, it is unnecessary to train it from the beginning because it has already been trained on massive amounts of data. This study examined the recent pre-trained Deep Learning model for accurately classifying actions. This study assists the researcher in assessing the value of the most recent Vision Transformer model in the field of HAR. This study employs the UCF 50 action dataset to evaluate the effectiveness of the Vision Transformer model in HAR. We obtained 94.70% accuracy on the UCF 50 action dataset using a Vision Transformer model variant.

5.1 Introduction

Due to its wide range of applications, including video retrieval, security, behaviour monitoring, patient activity, elderly surveillance, and defense, HAR is a highly active research field. Data pre-processing, feature extraction (pose-based, shape-based), a learning algorithm, and action classification are the steps that comprise HAR. Real-time Action Recognition (AR) is a popular research topic in HAR for security considerations. In real-time, we must train the model on fewer available data; after that, the amount of data grows swiftly and requires extensive computation. Previously, the researcher used a classification method based on Machine Learning (ML) to identify the action. Due to the vast quantity of data, ML-based methods are not superior. Due to the ineffectiveness of the ML-based approach on massive data sets, the researcher employed Deep Learning (DL) techniques to identify action in video sequences. DL-based methods necessitate more computational capacity to train a model on massive data sets.

CONET serves a crucial function in computer vision for identifying images and patterns. Other deep learning models utilized by the research community include Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GAN). Traditional ML-based methodologies are outperformed by the DL-based strategy [228]. The four action levels of HAR are Gesture, Interaction, Group Activity, and Action.

Due to view-invariant, low-resolution, and variable-length sequences in datasets, HAR continues to struggle with action classification. Several machine learning (ML)-based algorithms have been developed, but each has its limitations. Some methods work well with a specific data set, brief videos, a single view, and impractical conditions. The optimal HAR model can overcome obstacles and perform better in every circumstance. Compared to ML methods, DL methods are better equipped to manage large amounts of data due to their need to perform better when data sets are enormous. Researchers are continually refining techniques for reliably recognizing actions. In addition, some multimodal approaches to learning specific activities are introduced. In the multimodal approach, we fed the model numerous modalities, such as RGB, Depth, and Skelton data. In a multimodal approach, fusion techniques (early fusion, delayed fusion, and late fusion) combine features obtained from various modalities. Data Pre-Processing, Object Segmentation, Feature Extraction, Data Training, and Classification are typical HAR process steps [228]. The DL-based method employs an automated feature extraction technique, whereas the ML-based method employs a pose- and shape-based handcrafted feature approach. Figure 5.1 depicts the processing step involved in HAR.

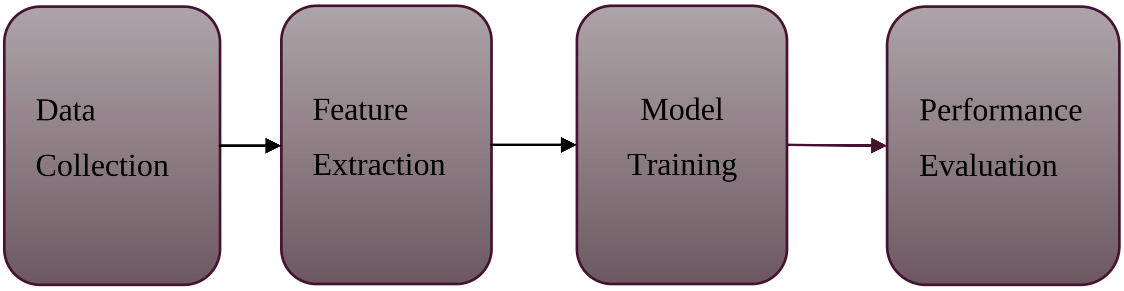


Figure 5.1 HAR Complete Process

The DL method was successful in image classification, object detection, and action recognition. Deep Learning provides a variety of architectures, such as CONET, which uses the most compelling feature extraction method compared to the handcrafted feature extraction method [229]. Numerous frameworks rely on CONET to address the HAR issue. Recurrent Neural Network (RNN) is an additional DL framework, whereas Links between nodes create a directed graph along a time sequence, allowing for dynamic activity[230] . The primary contribution of this study is to classify action using the most recent Vision Transformer model and compare it to contemporary methods.

5.2 Related Work

Deep architecture CONET for feature extraction and sequential pattern learning using the LSTM architecture in [231]. Automatic Learning of the characteristics proposed by Ji et al. [114]. 3-D CONET was utilized for HAR. The inflexible architecture frames and optical flow input base dual stream model proposed by Simoyan et al. [135] and input fed to the pre-trained DL model render this model inadequate for variable-length segments. CONET is utilized for FE and action classification of KTH 1 or KTH 2 datasets in the two-step approach[116]. Ijjina and Mohan created a deep hybrid model by combining homogeneous CONETs, which achieved 99.68% precision on the UCF 50 dataset [232]. Liang et al.[233] investigated a highly unsupervised learning model for human segmentation tasks. Human mask inference based on video context and segmentation network learning based on CONET were iterated until mutual improvement was no longer feasible. PASCAL VOC 2012 passed rigorous testing with 81.8% accuracy.

Safaei M. et al. [234] proposed a sophisticated CONET-based method for predicting future actions and identifying the form and position of the image's most prominent components. They trained a single framework for every move using a one-versus-all strategy and achieved an accuracy of 76.1 percent. Pose-based characteristics from the 3D CONET network can unify 3D posture, 2D appearance, and motion flow during [216] capture. Complexity will result from the extraction of joint colour features for the 3D CONET; consequently, a 15-channel heatmap is generated, and convolution is performed in each map. Feichtenhofer et al. [118] present a two-stream convolution system for combining dual temporal and spatial streams, wherein RGB information (spatial) and optical flow (motion) are simultaneously modelled and estimates are aggregated in the final layers. Due to optical flow, this network cannot capture long-term motion; another disadvantage of the spatial CONET stream is that its efficacy depends on a randomly chosen image from the input video. Due to background confusion and perspective adjustments, complications arise.

Shi et al.[235] recommend an additional two-stream network to enhance the efficacy of skeletal joint-based HAR. The Adaptive Graph Convolutional Network (AGCN) network processes a two-stream input that includes joint and bone information. The network is comprised of a hierarchical arrangement of these fundamental components. The soft-max layer has been incorporated into the ultimate product. Ullah et al. [129] perform HAR on the system using real-time visuals captured by a camera in motion. CONET, a technique based on deep learning, extracts autonomously frame-level characteristics. Jian et al. describe a method for ROI extraction using a Fully Convolutional Network (FCN) in their paper [236]. CONET is utilized to ascertain the posture probability of each frame. Key-frame extraction is performed using the probability difference between adjacent frames. The variation-aware key-frame extraction method takes into account the frame with the highest likelihood of a key pose, as determined by CONET. The Central frame is chosen if multiple frames yield the same probability value for the critical pose.

Using a CONET and an LSTM recurrent network, a DL-based solution for 3D temporal poses identification difficulties. They begin by training the CONET before modifying the combination (CONET+ LSTM)[237]. CONET extracts the pertinent data, which the LSTM then utilises to classify the target action classes. This innovative training strategy outperforms conventional single-stage training. On limited data sets, the outcomes exceed a number of contemporary methods. [238] created a unique HAR model with three primary stages: pre-processing, background removal, and classification. Their method significantly improved the results for five action classes extracted from the INRIA and KTH datasets, namely sprinting, strolling, bounding, standing, and reclining. [191] presented a multi-resolution CONET architecture for feature connectivity in the time domain to capture local Spatiotemporal data. This method is being evaluated on the current "YouTube 1 million videos dataset" with 487 action sequences of classes. The authors noted that the focalized design of CONET accelerated the complexity of training. Their action categorization rate for large datasets increased to 63.9%, but their recognition rate for UCF101 remains at 63.3%, which is insufficient for such a crucial task as action recognition.

Bilen et al. [239] investigate the feature maps of a pre-trained model for the representation of dynamic image video. At the tuning stage, a rank pooling operator and an approximation rank pooling layer were introduced to aggregate the mappings of all frames into a singular dynamic image representing the entire film. Due to their extensive feature representation infrastructure, deep learning-based algorithms can reliably unearth concealed patterns in visual data. On the other hand, training requires a massive amount of data and a high level of computer processing capacity. Wu et al. [240] introduced an MPCA-Net technique for classifying human actions based on deep learning. This strategy utilizes tensor interaction to increase the recognition rate. It has three layers: projection dictionaries, projection encoder, and pooling. The effectiveness of the proposed method is evaluated using UCF11, various medical imaging datasets, and UCF sports action datasets. Majd et al.[214] proposed a correlational convolutional LSTM (C2 LSTM) for dealing with surveillance video data's spatial and movement knowledge. Conv-LSTM is a connectivity combination model proposed in[241] for violence detection.

This architecture fed the CONET network into the LSTM network for feature analysis and action classification. Jaoudi et al. proposed a novel method for motion tracking based on human observation and the extraction of spatial features from video sequences. Gaussian mixture model (GMM) and Kalman filter (KF) algorithms were used to detect and extract moving individuals, while Gated Recurrent Neural Networks were used to collect data in each frame and predict human activity. They evaluated their methodology with datasets from UCF Sports, UCF101, and KTH, achieving 89.01%, 89.30%, and 96.30%, respectively [242]

5.3 Proposed Methodology

With their models, CONET, LSTM, GAN, and Autoencoder, Deep Learning-based methods play a crucial role in action recognition and attain superior results than ML-based methods. DL-based methods are able to manage vast quantities of data. However, training from inception for each model is extremely complex, so numerous Transfer Learning-based methods propose models that have already been trained on a massive quantity of data. The proposed framework evaluated the effectiveness of Vision Transformer for action recognition problems in video sequences. Figure 5.2 depicts how VT detects a class of corresponding frames from sequence of frames.

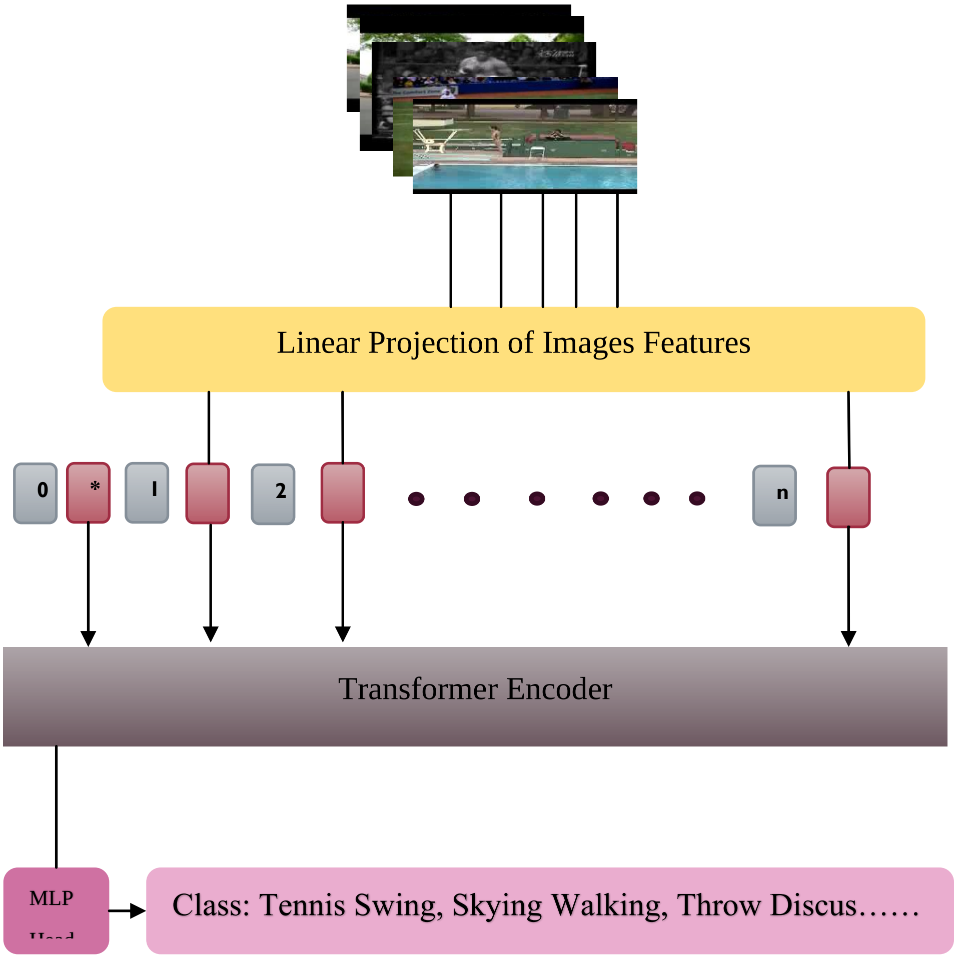

Linear Projection of Images Features
0
*
1
2
n
Transformer Encoder
MLP Head
Class: Tennis Swing, Skying Walking, Throw Discus……

Figure 5.2 Action Transformer architecture for action class recognition [243]

5.3.1 Vision Transformer

The standard Transformer is given a 1D sequence of token embeddings. To manage 2-D images, the Transformer reshapes required images $x\,\epsilon\mathbb{R}^{H\times W\times C}$ into an order of flattened two-dimensional spots $x_p\epsilon\mathbb{R}^{N\times P^2\times C}$, where image resolution is represented by (H, W) and channel count is represented by C. Each image's resolution displays from (P, P) and the total number of patches represented by $N = HW/P^2$, which also serves as the Transformer's essential input order span.

Because the Transformer employs a constant latent vector size D across all its phases, they use a trainable linear projection to align the areas and convert them to D dimensions. The first layer of the VT sequentially reflects the refined regions into a domain that is shorter [243]. The characteristics resemble appropriate basis functions for a low-dimensional representation of each patch's delicate structure. The patch representations are then enhanced with a learned position embedding following the projection.

In location embedding similarity, the model is instructed to express distance within the visual, i.e., closer regions have more comparable position embeddings. The row-column structure is also evident, with identical fundamental characteristics within each row/column. Due to self-attention, VT can acquire data throughout the entire visual, including at the lowest levels.

Positional Embedding: Positional embedding adds spatial information to the sequence. Due to the fact that the model is unaware of the spatial relationship between tokens, it may be beneficial to provide additional information to indicate this. This is typically a learned embedding, or tokens are assigned weights from two sine waveforms with high frequencies, which is sufficient for the model to learn that these tokens have a positional relationship.

Transformer Encoder: A transformer encoder consists of a series of encoding layers stacked on top of one another. Each encoder layer is comprised of two sublayers: a Multi-Headed Self-Attention (MHSA) head and a Multi-Layer Perceptron (MLP) head. Each sub-layer includes a layer normalization (LN) and a residual link to the next sub-layer.

Classification: Vision transformers frequently add an additional learnable [class] token to the sequence of embedded patches for classification, which reflects the class parameter of an entire image and its state after transformer encoding. The [class] token contains latent information and accumulates additional knowledge about the sequence for categorization purposes via self-attention. ViT [243] also investigated the alternative of aggregating output tokens but found no discernible performance difference. However, they found that the learning rates must be altered between the two variants: [class] token and average pooling.

5.3.2 Dataset

Using the UCF 50 action dataset, the efficacy of the VT model was evaluated. Reddy et al.[223] presented this dataset in 2012. Realistic action videos are compiled from online sources like YouTube and have a realistic setting. This dataset contains 50 distinct action classes, such as playing tabla, hurling a baseball, using a yo-yo, strolling a dog, and launching a discus. The dataset contains approximately 100 brief videos for each class, encompassing a wide range of camera motion, object appearance and posture, object scale, perspective, congested background, illumination conditions, and other variables. Figure 5.3 represents the dataset's visual frames of action.

Diving

Biking

Baseball Pitch

Drumming

Pullup

Punching

Figure 5.3 Sample Frame of Datase**t** [223]

5.4 Implementation Details & Result

Transfer learning is a weight initialization technique in which the neural network is not trained with stochastic weights and biases but with ImageNet-learned weights and biases. The network is then permitted to learn and improve its parameters using HAR vision datasets as training data. In this manner, it is fine-tuned for the categorization objective. We classified each action in the UCF50 dataset using the VT model and compared these classifications to current best practices. The VT variants are pre-trained deep learning models trained on the massive ImageNet dataset [220].

These models can reliably classify each action without needing to be trained from inception. This study compares the precision of various variants of VT on the UCF 50 dataset. This study extracted image sequences for each action category and used them to train a neural network. Compared to the numerous Transfer Learning and DL models, VT performs better. Network. Compared to the numerous Transfer Learning and DL models, VT performs better. We generate the confusion matrix based on the result to evaluate the accuracy. In Figure 5.4

represent the classification result through vit_b_16 and Figure 5.5 illustrates the result through vit_l_16. Figure 5.6 depicts the various evaluation matrices to assess the performance of the vision transformer in the HAR problem.

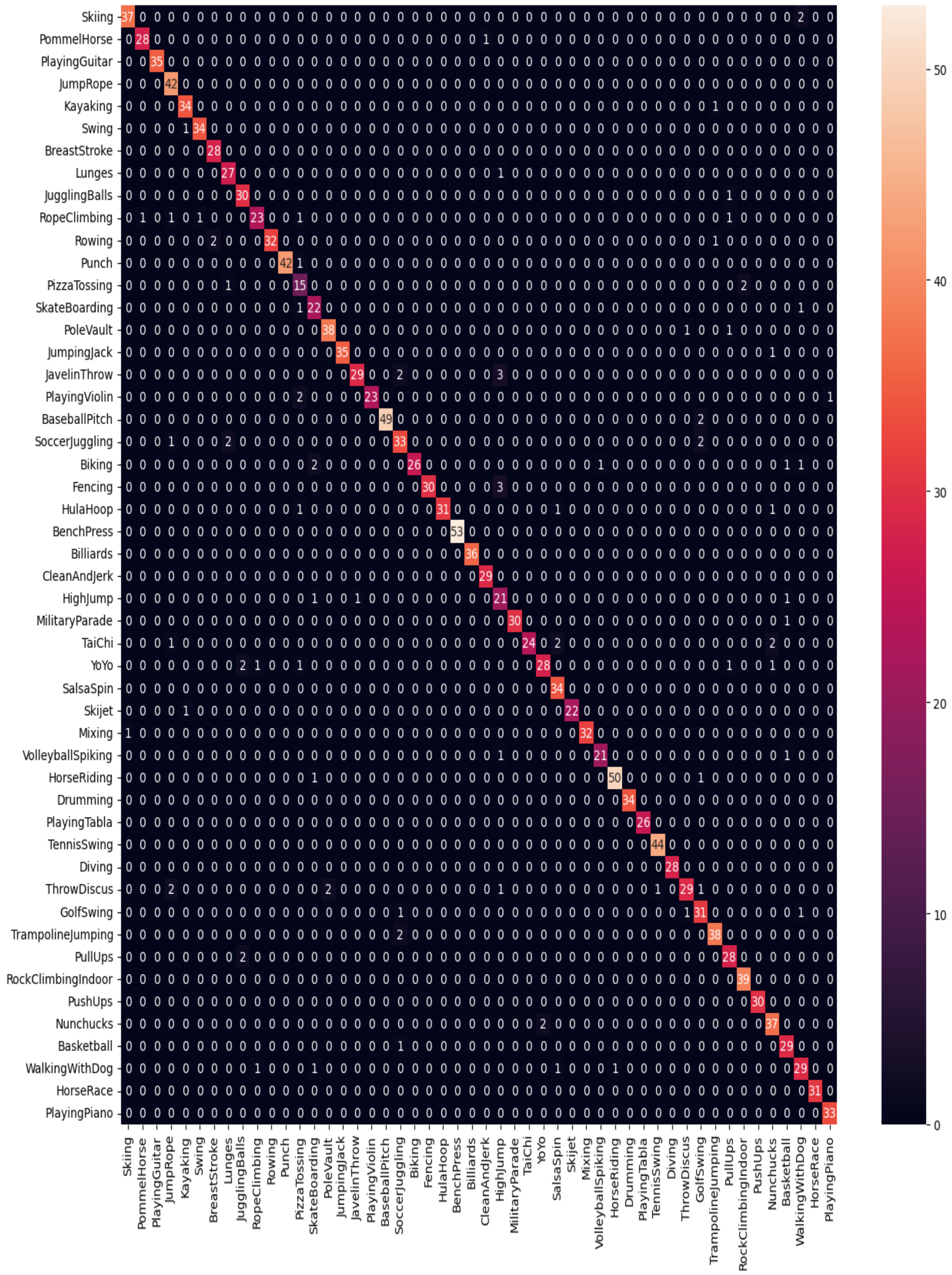


Figure 5.4 A confusion matrix for the classifying of the UCF 50 activity datasets employing the VT_B_16 model

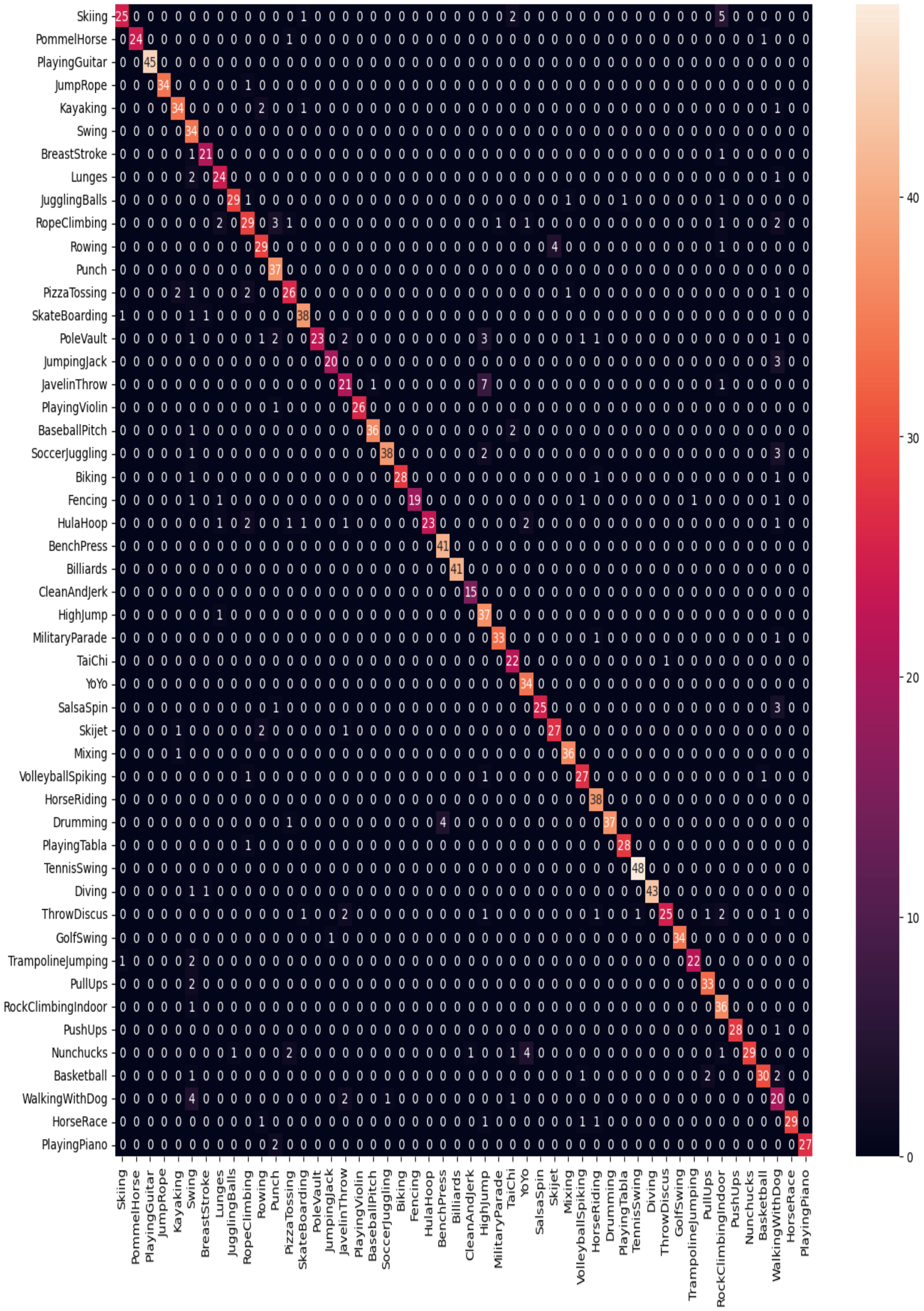


Figure 5.5 A confusion matrix for the classifying of the UCF 50 activity datasets employing the VT_l_16 model

Table 5.1 entails the various performance evaluation matrices of Vision Transformer models & compares model's performance based on multiple evaluation matrices.

Table 5.1 Comparison of performance between Vision Transformer models.

Model Name	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Vision Transformer (vit_b-16)	94.70	94.36	94.17	94.11
Vision Transformer (vit_l_16)	89	91.07	89.47	89.69

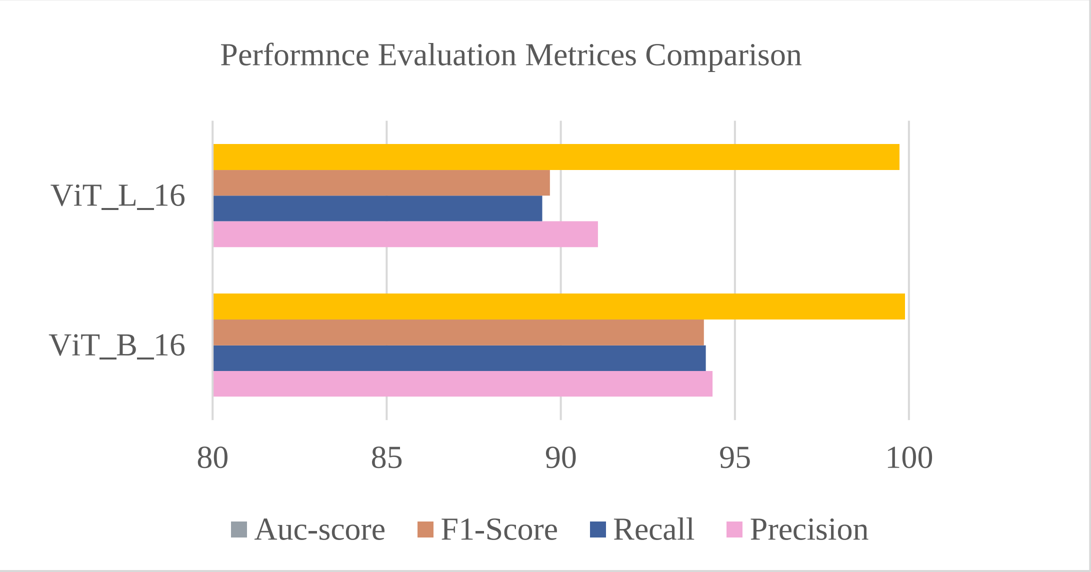


Figure 5.6 Compares several evaluation parameters

In this study, the UCF 50 action dataset has been utilized to evaluate the accuracy, precision, and recall of the VT model. VT models are pre-trained models that can effectively classify each action. We implemented VT models on GPU to evaluate these models' efficacy in HAR. This investigation employed VT variants VT_b_16 and VT_l_16, corresponding to the base or large model. There are sixteen by sixteen input sizes. The small model has 12 layers with

32 million parameters, while the large model has 24 layers with 307 million parameters. Table 5.2 depicts the comparison with other methodologies.

Table 5.2 Comparison with State of the Reference Models.

Reference Model	**Accuracy (in %)**
Action Bank [244].	76.4
Multi-Channel Descriptor [245]	83.3
Global-Spatio Temporal Feature[246]	70.1
Wang et al[247]	91.7
Vision Transformer (VT_b_16) [243]	**94.70**
Vision Transformer (VT_l_16) [243]	**89.00**

5.5 Discussion

This methodology examines the VT model in the HAR domain. Due to their intricate architecture, VT models are able to classify actions effectively. We compared the efficacy of these models to that of state-of-the-art methodologies using the UCF 50 data set. To determine its efficacy, this study contrasted various model evaluation metrics (f1-score, precision, recall, etc.). With an accuracy of 94.70%, the VT model outperforms other suggested methods on the UCF 50 action dataset.

Complex and Multiview datasets can be evaluated in the future. These models can be utilised for online action classification and the identification of complex actions. Numerous Transformer variants with fewer parameters were proposed, demanding less time and computing power. These models are also utilized for rapid response in surveillance data.

Chapter 6
Transfer Learning Models Effectiveness for Human Action Recognition

This chapter explores the recent Transfer Learning (TL) models in HAR. These models' performance is evaluated based on precision, recall and AUC score. The UCF 11 and UCF 50 datasets are used to assess the performance of TL models.

6.1 Introduction

Pre-trained models have gained the attention of researchers in recent times in HAR. The conventional computer vision and ML approaches utilize handcrafted feature methods. Handcrafted feature methods are not robust and efficient in challenging conditions, like lighting, occlusion, and environmental conditions. Transfer learning models perform better than models trained from scratch on the HAR problem. TL models can be trained quickly compared to other models. These models have been used to develop HAR models for new domains or actions with less effort.

6.2 Transfer Learning Models

TL models refer to a category of DL models that undergo pre-training on a substantial dataset of visual and image sequences for a specific problem, followed by fine-tuning on comparatively less data of frames and image sequences for a unsimilar problem. There are two options available: either the final layers of the model can be removed, or they can be frozen. The process of freezing involves inhibiting the modification of the weights within those layers during the training phase, hence preserving the acquired knowledge from the original task. TL models are trained on large datasets to perform specific tasks. To perform action recognition, you may modify the model (Figure 6.1).

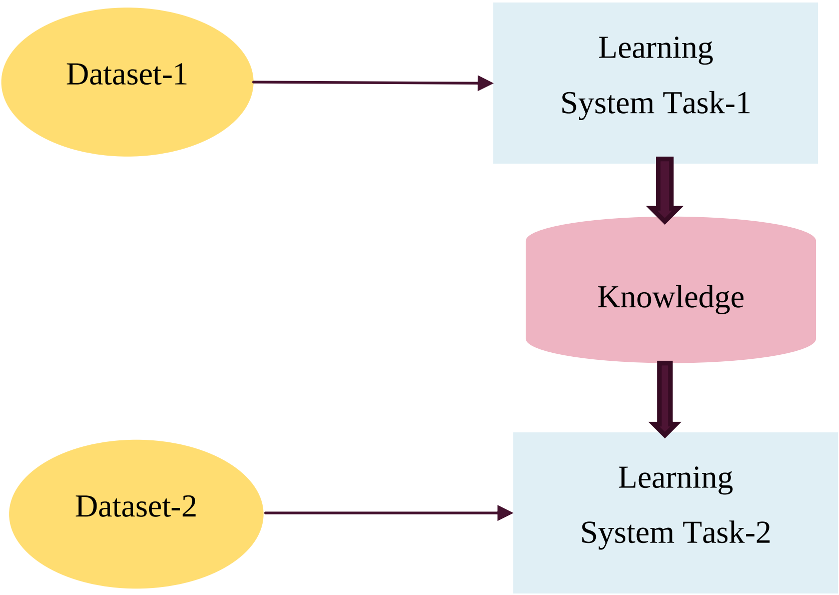


Figure 6.1 Transfer Learning visualization

6.2.1 Mobile Net

MobileNetV3 [248] refers to a collection of convolutional neural network designs specifically developed to facilitate fast and lightweight deep learning tasks. These architectures are particularly well-suited for deployment on mobile and embedded devices. The design is a progression from the previous MobileNetV1 and MobileNetV2 models, exhibiting enhanced performance and efficiency. The network design philosophy of MobileNetV3 encompasses the use of two distinct design methods, namely the "small" and "large" architectures. The smaller models have been designed and optimized to cater to low-latency applications, prioritizing reduced response times. Conversely, the larger models have been specifically developed with the primary objective of maximizing accuracy.

6.2.2 ResNet 50

ResNet 50[249] is a DNN with 50 convolutional layers. This neural network is classified as a residual neural network, characterized by its utilization of skip connections to enhance the network's ability to acquire more effective representations of the input data. Skip relationships enable the neural network to acquire knowledge straight from the preceding layer, mitigating disappearing gradients. The network's depth allows it to effectively catch intricate details present in visuals.

6.2.3 VGG 19

VGG-19[250] consists of 19 layers, 16 of which are convolutional and three wholly connected. At the time, VGG-19's use of deep layers was a notable feature that contributed to its efficacy in recognizing complex patterns and features in images. Small 3 × 3 filters with a stride of 1 and a padding of 1 are utilized in VGG-19's convolutional layers. Using modest filters throughout the network enables it to learn a hierarchical structure of features.

Figure 6.2 depicts the process of HAR using pre-trained models. A pre-trained model learns features from image sequences and categorizes the action class.

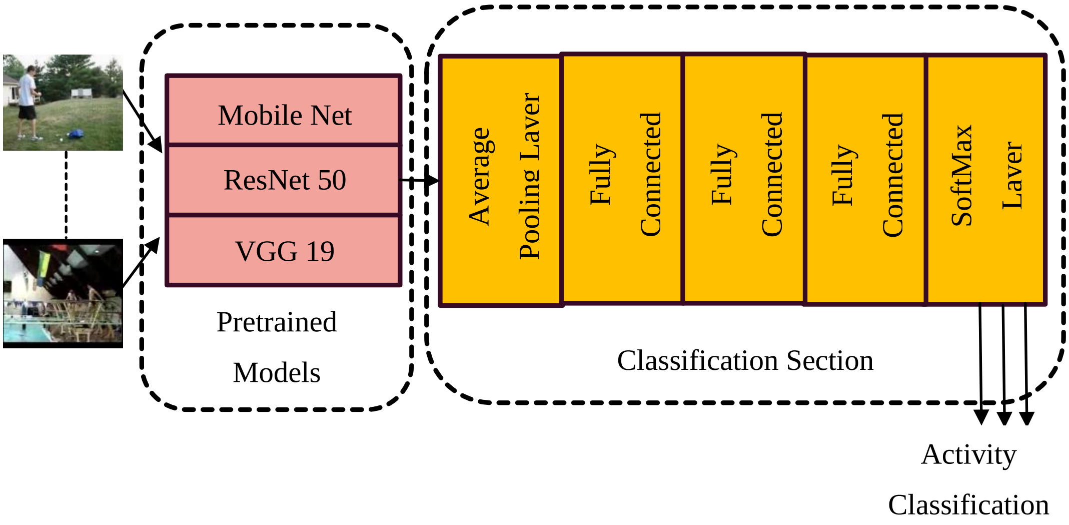
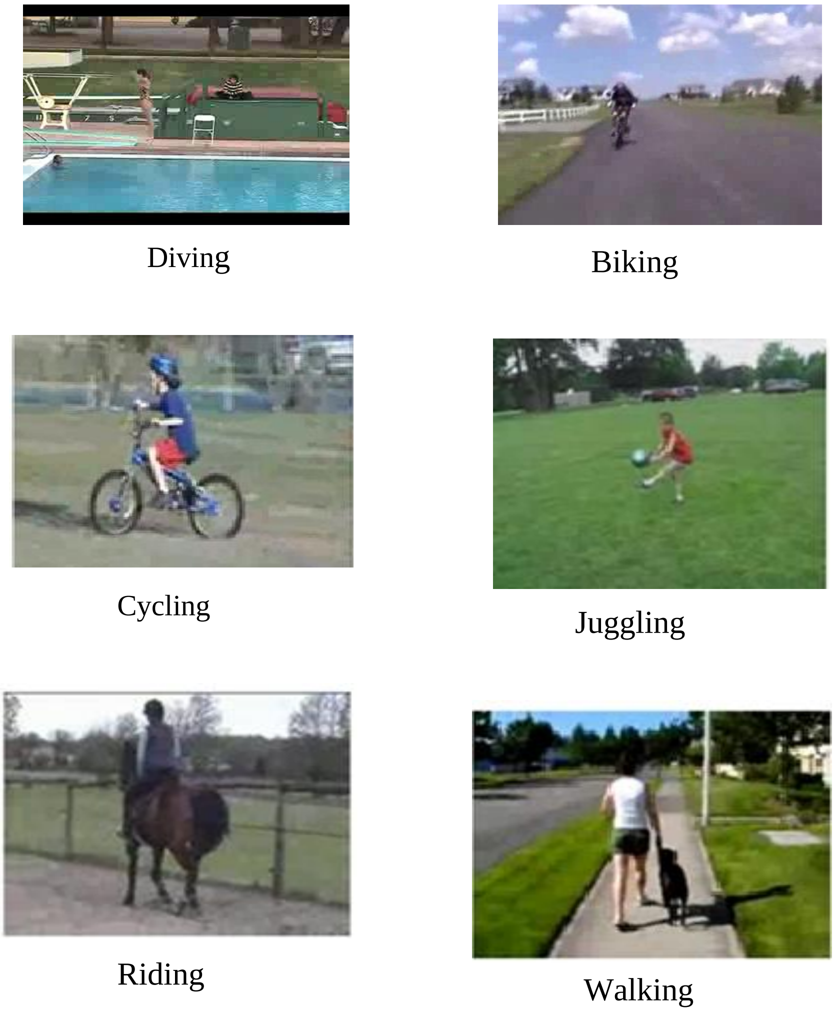


Figure 6.2 Transfer Learning model based HAR Problem

6.3 Datasets

UCF 11[251] and UCF 50[252] datasets are employed to assess the effectiveness of these models. Videos are gathered for these datasets from online resources and have realistic environments. UCF 11 contains eleven distinct daily activities. On the contrary, UCF 50 comprises 50 different action groups. The UCF 11 dataset has 1600 videos encompassing various activity categories. UCF 50 dataset has 50 distinct action classes, such as playing tabla, baseball pitch, yo-yo, walking with the dog, and throwing discus. The dataset contains about 100 short videos on every class, with a broad range of camera motion, object look and posture, object scale, perspective, cluttered backdrop, illumination conditions, etc. The UCF 11 action frame is represented in Figure 6.3.



Diving

Biking

Cycling

Juggling

Riding

Walking

Figure 6.3 UCF 11 action image sequences [251]

6.4 Performance Analysis

TL models’ performance is evaluated on various parameters like classification ratio matrix, accuracy, precision, recall and F1-Score. Figure 6.4, Figure 6.5 and Figure 6.6 represent the confusion matrix for classification of UCF 11 dataset actions**.** Figures represent the confusion matrix on the UCF-11 dataset from TL models. The confusion matrix can determine how models are capable of classifying activity accurately.

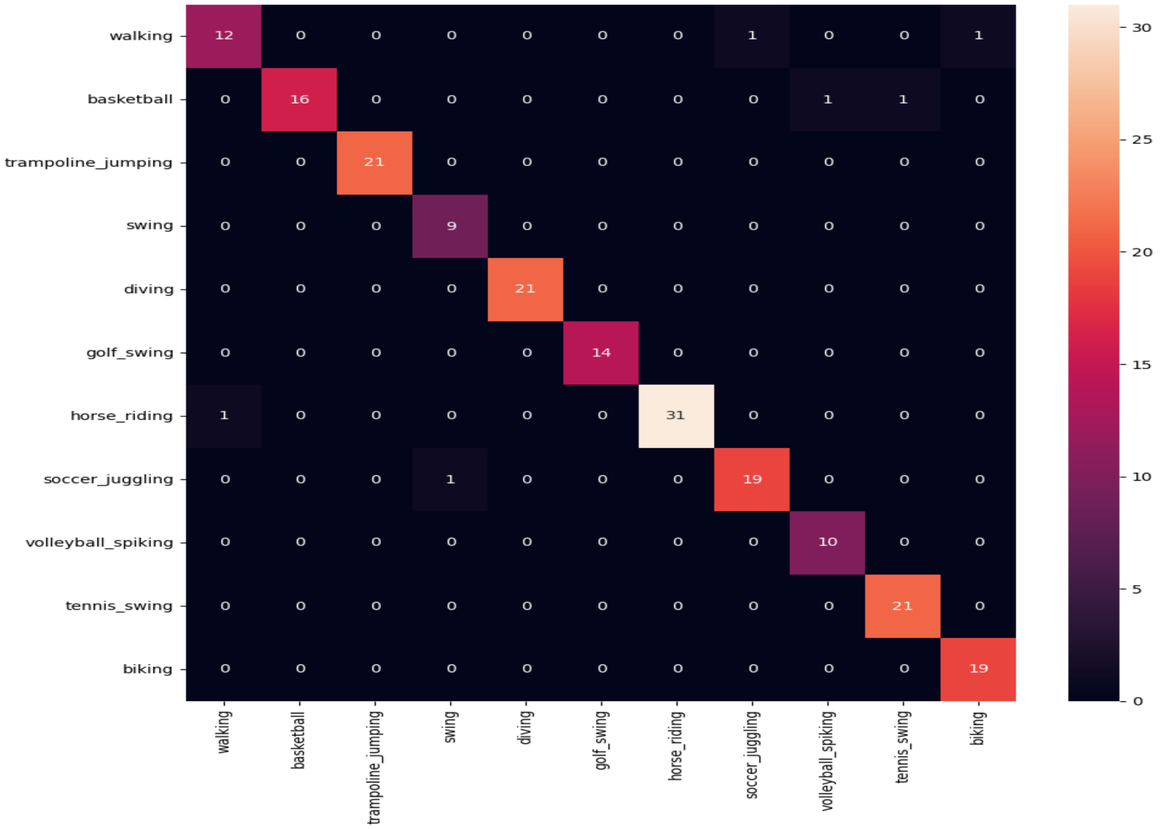


Figure 6.4 Confusion matrix for the MobileNet V3 model

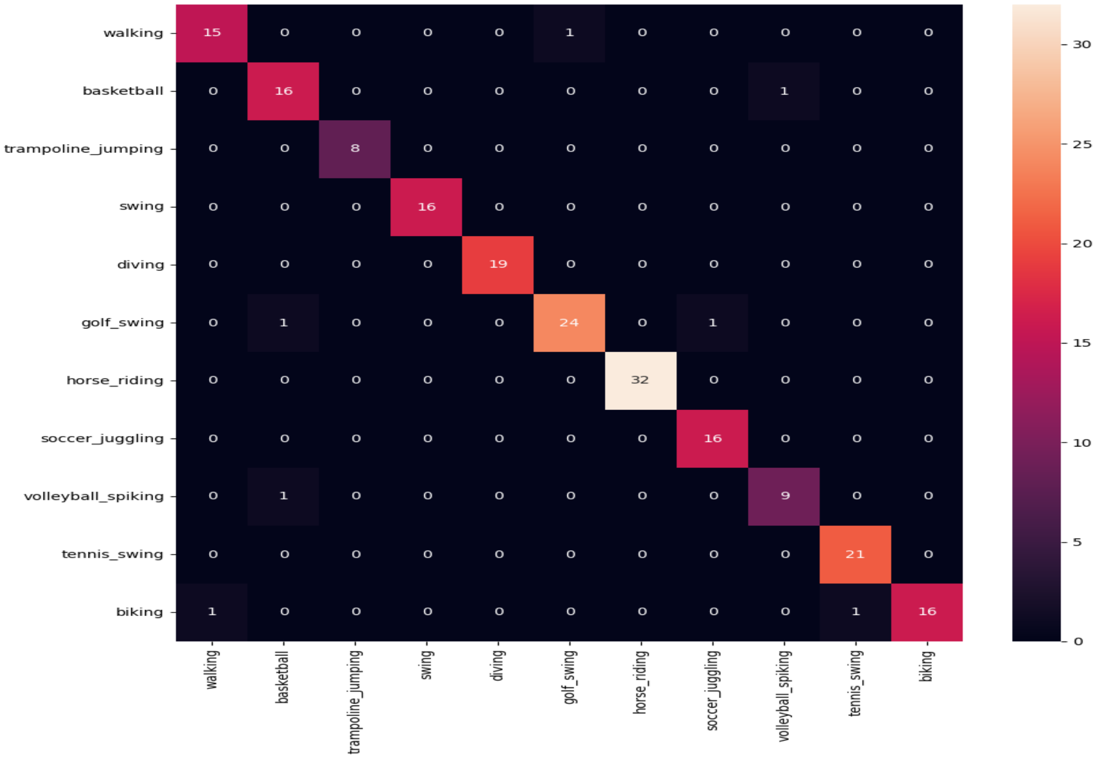


Figure 6.5 Confusion matrix for the ResNet 50 model

	walking	basketball	trampoline_jumping	swing	diving	golf_swing	horse_riding	soccer_juggling	volleyball_spiking	tennis_swing	biking
walking	10	0	0	1	0	0	0	2	0	0	1
basketball	0	15	0	0	0	0	0	1	1	1	0
trampoline_jumping	0	0	20	0	0	0	1	0	0	0	0
swing	0	0	0	9	0	0	0	0	0	0	0
diving	0	0	0	0	20	0	0	0	0	0	1
golf_swing	0	0	0	0	0	14	0	0	0	0	0
horse_riding	0	1	0	0	0	0	30	1	0	0	0
soccer_juggling	0	0	0	0	0	2	0	18	0	0	0
volleyball_spiking	0	0	0	0	0	0	0	0	10	0	0
tennis_swing	0	0	0	0	0	0	0	1	0	20	0
biking	0	0	0	0	0	0	0	0	0	0	19

Figure 6.6 Confusion matrix for the VGG 19 model

Table 6.1 is used to determine the effectiveness of TL models in the HAR problem and these are compared on different performance evaluation matrices. After that, Table 6.2 is used to compare TL models with the state of the reference model to evaluate the effectiveness of the HAR problem.

Table 6.1 Performance comparison of various TL models on the UCF 11 dataset.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
VGG 19	93.69	92.99	93.11	92.64
ResNet 50	95.98	96.24	96.27	96.18
Mobile Net V3	96.66	96.24	96.95	96.50

Table 6.2 Comparison of various reference models on the UCF 11 dataset.

Reference Model	**Accuracy (in %)**
Dense Trajectories[253]	84.2
Visual Attention Model [254]	84.9
Two Stream LSTM [255]	92.2
VGG 19	**93.69**
ResNet 50	**95.98**
Mobile Net V3	**96.66**

Figure 6.7**,** Figure 6.8 and Figure 6.9 represents the confusion matrix for the UCF 50 action dataset. The confusion matrix is used to describe the classification model's effectiveness.

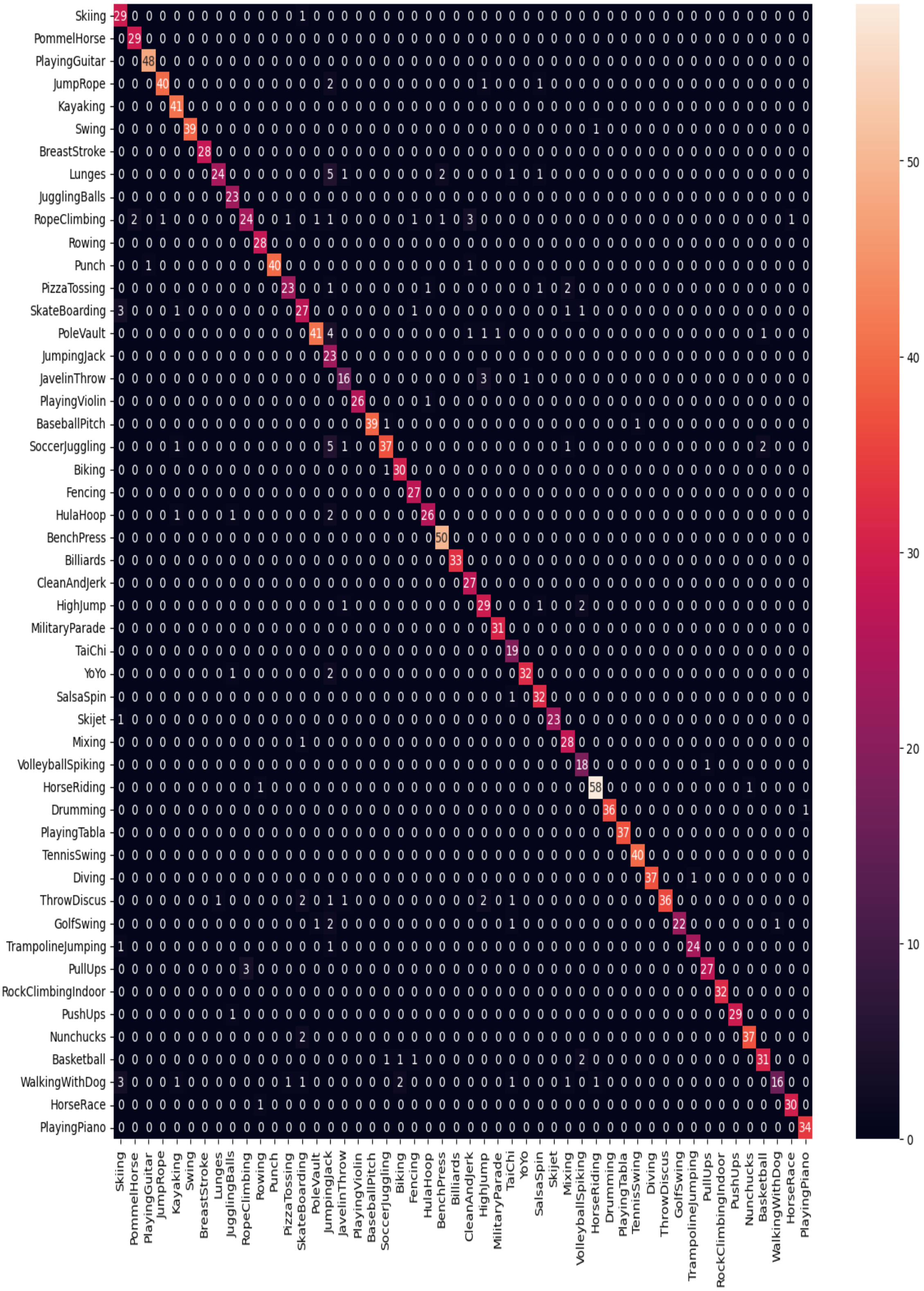


Figure 6.7 Confusion matrix for the MobileNet V3 model for the UCF 50 dataset

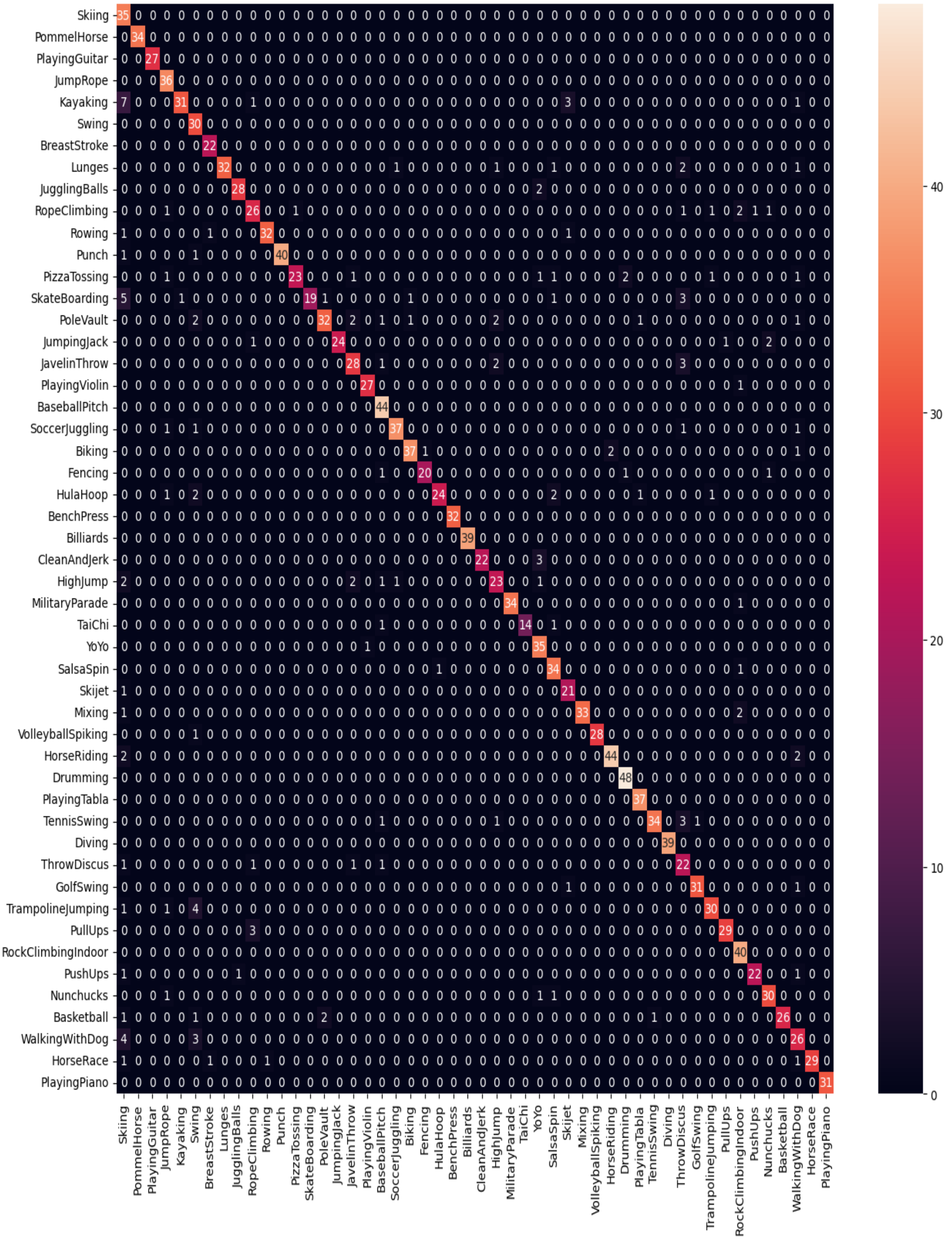


Figure 6.8 Confusion matrix for the VGG 19 model for the UCF 50 dataset

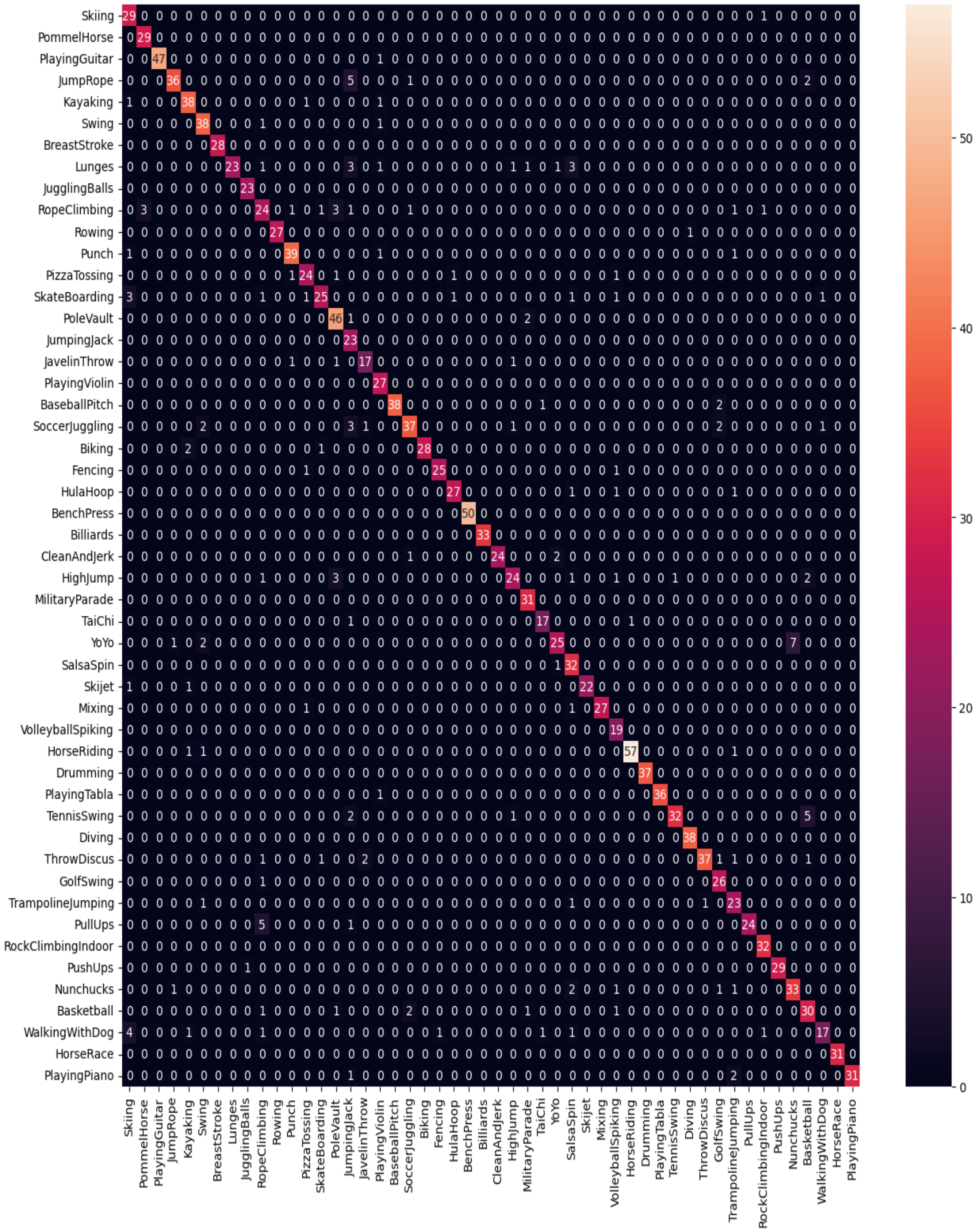


Figure 6.9 Confusion matrix for the ResNet 50 model for the UCF 50 dataset

Table 6.3 represents the comparison of TL models on numerous performance evaluation matrices to assess the effectiveness on the UCF 50 dataset. These models are also compared with the state of the reference model in Table 6.4 to evaluate the effectiveness for HAR problem.

Table 6.3 Performance comparison of various TL models on the UCF 50 dataset.

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
ResNet 50	90.0	90.56	90.2	89.88
Mobile Net	92.27	92.65	92.60	92.1
VGG 19	83.92	83.90	82.81	81.32

Table 6.4 Comparative analysis with reference models on the UCF 50 dataset.

Reference Model	**Accuracy**
Global-Spatio Temporal Feature[246]	70.1
Convolutional Long Short Term (CLSTDN) [256]	80.8
Motion Modalities [226]	89.3
VGG19	**90.11**
ResNet 50	**90.00**
Mobile Net V3	**92.27**
Dense Net 161	**92.57**
Efficient Net_b7	**94.25**

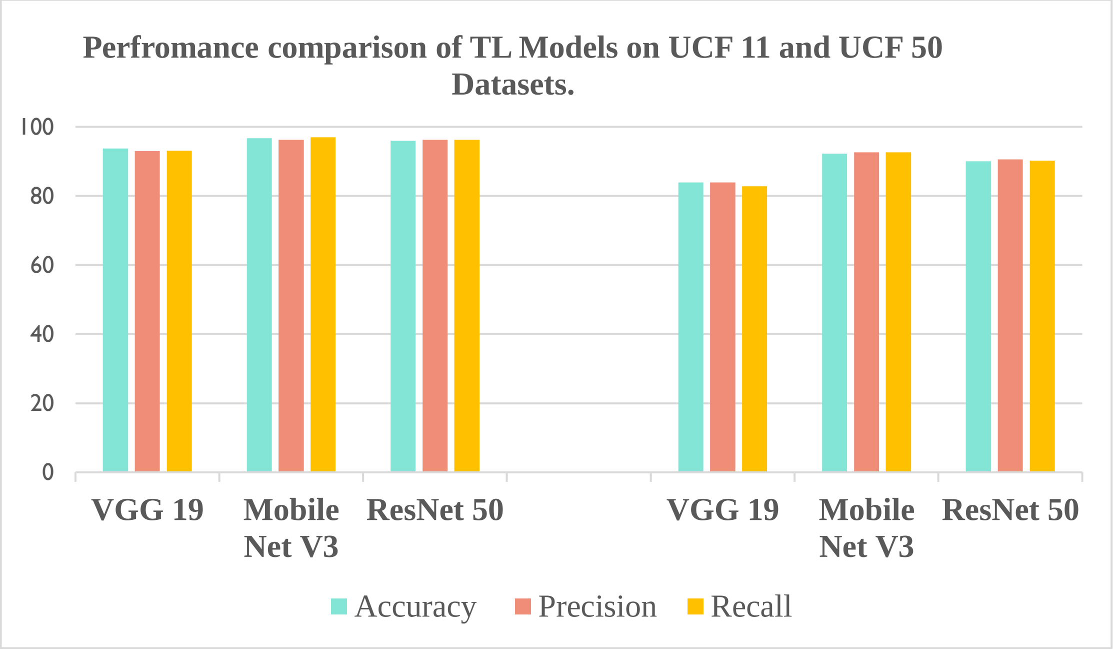


Figure 6.10 Performance comparison of TL models on UCF 11 & 50 datasets

6.5 Discussion

This chapter uses several recent TL models to perform the human action recognition task. To evaluate their performance, these models are trained on the UCF 11 and UCF 50 action datasets. The purpose is to compare these models' performance on the HAR problem to predict action with less computing power. These models do not require more training time and have fewer parameters compared to other DL models. TL models are faster and smaller than CNN models. These models are suitable for mobile devices that require low computational resources. These models perform better as compared to several reference models with low computational power requirements.

Chapter 7

Conclusion and Future Scope

The following section presents a comprehensive overview of the suggested works, significant results, contributions, and limitations. Additionally, we outline potential avenues for future research in the area of human activity recognition in videos, encompassing both short-term and long-term perspectives.

7.1 Conclusion

This study covers essential modules, including handcrafted-based feature extraction, feature representation, summarized action dataset and classification of activity with the support of various methods. This study summarizes the past research on vision-based human activity detection systems. This study has described a general framework for AR. Numerous handcrafted feature extraction methods are described with comparative analysis. The fusion of features is also summarized. We have also represented various classification methods such as ML, DL, hybrid, and some pre-trained DL approaches. A considerable amount of data is needed to train, so hybrid DL methods and pre-trained DL methods are represented with improvement. This study presented a comparative analysis of the latest approaches for AR. We have summarized various available datasets. Evaluation matrices to test model performance are also discussed.

- It has been found that recognizing action in a single still image poses more significant challenges than analyzing action in video sequences. Recognizing action in a still image may be perceived as more challenging compared to video analysis. This is because the method does not consider the temporal variations, illumination variation, and alignment of the images.
- Various automatic feature extraction methods elaborated with their pros and cons. This study covers the traditional framework for HAR with feature extraction methods and classification techniques like ML-based or DL-based.
- This study enlists various multimodal, hybrid models and various methods for HAR non-deep learning-based or Deep learning-based.

- According to the data variation, HAR has been categorized into two categories: Unimodal and Multimodal. These methods discuss in detail, with their approach, how they process human action. Multimodal techniques gain popularity due to their robustness; they utilize multiple cues, which is why they require a lot of computational power.
- This study explains the type of activity according to the complexity level with benchmark action datasets and various states of the reference model analysis.
- Several methods have been explored in this study to recognize human actions of the individual person, crowd behaviour, and gesture of human and observe that the space-time approach provides better results for the gesture.
- Multiview modality is presented to solve the view-invariant issue with the use of a multimodal approach. This approach uses 5-S CNN, which is trained with various modalities of action sequence. For better representation of action, MHI and DMM are constructed. Bi-LSTM is employed to learn sequential information from video sequences.
- Self-occlusions, noise, and error evident in the video have a significant impact on single-modality-based approaches, particularly in skeleton coordinates. Therefore, the RGB skeleton MHIs are used to eliminate errors in the estimation of three-dimensional coordinates.
- A lightweight deep learning model has been proposed that provides better accuracy in less time. Various pre-trained models are employed to recognize human action and compare it with the state of the reference model. To train a model from scratch requires a lot of time compared to a pre-trained model.
- Efficient recognition of human action in less time is significant and fully fills the need for action recognition from video surveillance data. In this study, various latest pre-trained methods were used to develop an efficient HAR framework.
- This study also discusses the Vision Transformer models, which are trained on huge amounts of data. The architecture of Vision Transformer is profound. Using the UCF 50 action dataset, the efficacy of the transformer framework in HAR was evaluated. Comparing the precision of the proposed method to the current state of the reference model.

7.2 Future Directions

The primary goal of HAR (Human Activity Recognition) system techniques is to identify and classify activities seen in films autonomously. The majority of solutions, including handcrafted features, are built and acknowledged as a means of activity recognition via the construction of action templates using a specified collection of movies. Numerous concerns remain unaddressed, yielding unsatisfactory outcomes. The structure of the action template is susceptible to human errors in manually created feature descriptions. Consequently, it is essential to direct our attention towards novel expansions for the most recent iteration of action templates. It is crucial for us to exhibit heightened awareness, particularly regarding the real-time implementations of computer vision systems that systematically use action templates on streaming video in an iterative manner.

It has been noticed that the training of deep ConvNet architectures needs a substantial amount of labelled data in order to mitigate the issue of overfitting in the model. Moreover, the accuracy of action recognition systems is compromised as a result of imprecise labelling. However, it is feasible to use a combination of annotated and unannotated data in order to train the models. Hence, it is essential to develop Convolutional Neural Network (ConvNet) architectures that possess the capability to extract features from both annotated and unannotated data for the purpose of action recognition.

Future research might potentially prioritize the development of resilient human identification methodologies that provide enhanced performance in scenarios characterized by low light, crowded backdrops, changing light conditions, and noisy data. However, the majority of models discussed in the existing literature have primarily been evaluated in indoor situations. In order to enhance the reliability and effectiveness of these models, it is essential to subject them to real-time environmental circumstances and include relevant characteristics. Furthermore, in order to improve performance, it will be imperative to investigate novel deep learning methods such as the vision transformer-trained DL technique and the lightweight DL method, for the effective detection of Human Activity Recognition from inadequately labeled data in the foreseeable future. The use of 3D CNN has the potential to enhance the performance of CNN due to its ability to exploit spatiotemporal features, hence conferring an advantageous edge. Labelled data gathering is often characterized by its significant financial and temporal demands. While deep generative models such as Autoencoders (AEs) and Generative Adversarial Networks (GANs) have the capability to use unstructured data, their direct applicability to Human Activity Recognition (HAR) is limited. In prospective scenarios, there exists the potential for the use of multi-person recognition techniques. The potential for future study lies in the ability to make predictions about forthcoming events by analyzing current circumstances. Ensemble-based methodologies might be advantageous in scenarios when an excessive amount of data is being handled.

Further investigation is required in order to adequately and efficiently understand the significance of various classifier systems. Considering the presence of overlapping activities is an additional issue to be taken into consideration. Individuals have the potential to engage in simultaneous locomotion and gesturing, as an illustrative instance.

This research included many disciplines of augmented reality (AR) in order to contribute to varied methodologies, datasets, constraints, issues, and prospects within this field.

List of Publications:

1. Rahul Kumar, Shailender Kumar, "Survey on artificial intelligence-based human action recognition in video sequences," *Opt. Eng.* 62(2), 023102 (2023), Doi: 10.1117/1.OE.62.2.023102.
2. Kumar, R., Kumar, S. Multi-view Multi-modal Approach Based on 5S-CNN and BiLSTM Using Skeleton, Depth and RGB Data for Human Activity Recognition. *Wireless Pers Commun* (2023). https://doi.org/10.1007/s11277-023-10324-4.
3. Kumar, R., Kumar, S. A survey on intelligent human action recognition techniques. *Multimedia Tools Appl* (2023). https://doi.org/10.1007/s11042-023-17529-6.
4. Rahul Kumar, Shailender Kumar "ETBMF: Efficient Transformer-Based Multimodal Framework for Human Action Recognition in Videos" (Under Review in Engineering Applications of Artificial Intelligence).
5. R. Kumar and S. Kumar, "Light-Weight Deep Learning Model for Human Action Recognition in Videos," *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2023, pp. 1-6, Doi: 10.1109/ISCON57294.2023.10111975.
6. R. Kumar and S. Kumar, "Effectiveness of Vision Transformers in Human Activity Recognition from Videos," *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, Gharuan, India, 2023, pp. 593-597, Doi: 10.1109/InCACCT57535.2023.10141761.

References:

[1] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, “Vision-based human activity recognition: a survey,” *Multimed Tools Appl*, vol. 79, no. 41–42, pp. 30509–30555, Nov. 2020, doi: 10.1007/s11042-020-09004-3.

[2] I. Laptev, “On Space-Time Interest Points,” 2005.

[3] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., 2013, pp. 3551–3558. doi: 10.1109/ICCV.2013.441.

[4] S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *Artif Intell Rev*, vol. 43, no. 1, pp. 1–54, Jan. 2015, doi: 10.1007/s10462-012-9356-9.

[5] W. Liu and T. Mei, “Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective,” *ACM Comput Surv*, Apr. 2022, doi: 10.1145/3524497.

[6] Z. Liu, L. Zhou, H. Leung, and H. P. H. Shum, “Kinect posture reconstruction based on a local mixture of Gaussian process models,” *IEEE Trans Vis Comput Graph*, vol. 22, no. 11, pp. 2437–2450, Nov. 2016, doi: 10.1109/TVCG.2015.2510000.

[7] M. Berchtold, M. Budde, H. Schmidtke, and M. Beigl, “An Extensible Modular Recognition Concept that Makes Activity Recognition Practical UNIVERSITÄTSBIBLIOTHEK BRAUNSCHWEIG An Extensible Modular Recognition Concept that Makes Activity Recognition Practical,” 2010. [Online]. Available: http://www.digibib.tu-bs.de/?docid=00033761

[8] F. Negin, M. Koperski, C. F. Crispim, F. Bremond, S. Cosar, and K. Avgerinakis, “A hybrid framework for online recognition of activities of daily living in real-world settings,” in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016*, Institute of Electrical and Electronics Engineers Inc., Nov. 2016, pp. 37–43. doi: 10.1109/AVSS.2016.7738021.

[9] P. Pareek and A. Thakkar, “A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications,” *Artif Intell Rev*, vol. 54, no. 3, pp. 2259–2322, Mar. 2021, doi: 10.1007/s10462-020-09904-8.

[10] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as Space-Time Shapes.”

[11] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing Human Actions: A Local SVM Approach *.”

[12] Y. Nizam, M. N. H. Mohd, and M. M. A. Jamil, “Human Fall Detection from Depth Images using Position and Velocity of Subject,” in *Procedia Computer Science,* Elsevier B.V., 2017, pp. 131–137. doi: 10.1016/j.procs.2017.01.191.

[13] K. Host and M. Ivašić-Kos, “An overview of Human Action Recognition in sports based on Computer Vision,” *Heliyon*, vol. 8, no. 6. Elsevier Ltd, Jun. 01, 2022. doi: 10.1016/j.heliyon.2022.e09633.

[14] Yong Rui, “Image Retrieval: Current Techniques, Promising Directions, and Open Issues”.

[15] D. Hogg, “Model-based vision: a program to see a walking person.”

[16] “rohr1994”.

[17] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *IEEE Trans Pattern Anal Mach Intell*, vol. 23, no. 3, pp. 257–267, Mar. 2001, doi: 10.1109/34.910878.

[18] E. Shechtman and M. Irani, “Space-time behavior based correlation,” in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, IEEE Computer Society, 2005, pp. 405–412. doi: 10.1109/CVPR.2005.328.

[19] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, “Hierarchical filtered motion for action recognition in crowded videos,” *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 3, pp. 313–323, May 2012, doi: 10.1109/TSMCC.2011.2149519.

[20] C. Harris and M. Stephens, “A Combined Corner and Edge Detector,” British Machine Vision Association and Society for Pattern Recognition, Apr. 2013, pp. 23.1-23.6. doi: 10.5244/c.2.23.

[21] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 1395–1402. doi: 10.1109/ICCV.2005.28.

[22] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” *Computer Vision and Image Understanding*, vol. 104, no. 2-3 SPEC. ISS., pp. 249–257, Nov. 2006, doi: 10.1016/j.cviu.2006.07.013.

[23] Mikel D. Rodriguez, Ahmed Javed, and Shah Mubarak, *2008 IEEE Conference on Computer Vision and Pattern Recognition.*

[24] A. Yilmaz and M. Shah, “Actions sketch: A novel action representation,” in *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, IEEE Computer Society, 2005, pp. 984–989. doi: 10.1109/CVPR.2005.58.

[25] L. W. Campbell and A. F. Bobick, “Recognition of human body motion using phase space constraints,” in *IEEE International Conference on Computer Vision*, IEEE, 1995, pp. 624–630. doi: 10.1109/iccv.1995.466880.

[26] C. Rao and M. Shah, “View-Invariance in Action Recognition.”

[27] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 144–149. doi: 10.1109/ICCV.2005.90.

[28] S. M. Khan and M. Shah, "Detecting Group Activities using Rigidity of Formation *," 2005.

[29] A. Oikonomopoulos, I. Patras, M. Pantic, and N. Paragios, "Trajectory-Based Representation of Human Actions."

[30] H. Wang, A. Kläser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, 2011, pp. 3169–3176. doi: 10.1109/CVPR.2011.5995407.

[31] X. Peng, Y. Qiao, Q. Peng, and X. Qi, "Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition," in *BMVC 2013 - Electronic Proceedings of the British Machine Vision Conference 2013*, British Machine Vision Association, BMVA, 2013. doi: 10.5244/C.27.59.

[32] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "LNCS 7576 - Trajectory-Based Modeling of Human Actions with Motion Reference Points," 2012.

[33] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2555–2562. doi: 10.1109/CVPR.2013.330.

[34] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2009, pp. 104–111. doi: 10.1109/ICCV.2009.5459154.

[35] "lucas_bruce_d_1981_1".

[36] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "LNCS 7576 - Trajectory-Based Modeling of Human Actions with Motion Reference Points," 2012.

[37] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition." [Online]. Available: http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html

[38] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold," *IEEE Trans Cybern*, vol. 45, no. 7, pp. 1340–1352, Jul. 2015, doi: 10.1109/TCYB.2014.2350774.

[39] B. Ben Amor, J. Su, and A. Srivastava, "Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories," *IEEE Trans Pattern Anal Mach Intell*, vol. 38, no. 1, pp. 1–13, Jan. 2016, doi: 10.1109/TPAMI.2015.2439257.

[40] L. Zelnik-Manor and M. Irani, "Event-Based Analysis of Video *."

[41] I. Laptev, “On Space-Time Interest Points,” 2005.

[42] S. Samanta and B. Chanda, “Space-time facet model for human activity classification,” *IEEE Trans Multimedia*, vol. 16, no. 6, pp. 1525–1535, Oct. 2014, doi: 10.1109/TMM.2014.2326734.

[43] K. Mozafari, J. A. Nasiri, N. M. Charkari, and S. Jalili, “Action recognition by local space-time features and least square twin SVM (LS-TSVM),” in *Proceedings - 1st International Conference on Informatics and Computational Intelligence, ICI 2011*, 2011, pp. 287–292. doi: 10.1109/ICI.2011.55.

[44] M. Shillin Bella, R. Bhanumathi, and G. R. Suresh, “HUMAN ACTION RECOGNITION USING LOCAL SPACE TIME FEATURES AND ADABOOST SVM.” [Online]. Available: http://www.ijret.org

[45] C. Wang, Y. Wang, and A. L. Yuille, “An approach to pose-based action recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 915–922. doi: 10.1109/CVPR.2013.123.

[46] N. Ikizler-Cinbis and S. Sclaroff, “LNCS 6311 - Object, Scene and Actions: Combining Multiple Features for Human Action Recognition,” 2010.

[47] R. Minhas, A. Baradarani, S. Seifzadeh, and Q. M. Jonathan Wu, “Human action recognition using extreme learning machine based on visual vocabularies,” *Neurocomputing*, vol. 73, no. 10–12, pp. 1906–1917, Jun. 2010, doi: 10.1016/j.neucom.2010.01.020.

[48] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa, “Activity representation by SURF-based templates,” *Comput Methods Biomech Biomed Eng Imaging Vis*, vol. 6, no. 5, pp. 573–583, Sep. 2018, doi: 10.1080/21681163.2017.1298472.

[49] Z. Jiang, Z. Lin, and L. Davis, “Recognizing human actions by learning and matching shape-motion prototype trees,” *IEEE Trans Pattern Anal Mach Intell*, vol. 34, no. 3, pp. 533–547, 2012, doi: 10.1109/TPAMI.2011.147.

[50] L. Wang and D. Suter, “Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model.”

[51] S. A. Rahman, S. Y. Cho, and M. K. H. Leung, “Recognising human actions by analysing negative spaces,” *IET Computer Vision*, vol. 6, no. 3, pp. 197–213, May 2012, doi: 10.1049/iet-cvi.2011.0185.

[52] D. K. Vishwakarma and R. Kapoor, “Hybrid classifier based human activity recognition using the silhouette and cells,” *Expert Syst Appl*, vol. 42, no. 20, pp. 6957–6965, Jun. 2015, doi: 10.1016/j.eswa.2015.04.039.

[53] L. Shao, L. Ji, Y. Liu, and J. Zhang, “Human action segmentation and recognition via motion and shape analysis,” *Pattern Recognit Lett*, vol. 33, no. 4, pp. 438–445, Mar. 2012, doi: 10.1016/j.patrec.2011.05.015.

[54] Ieee and Ieee, *2011 IEEE International Conference on Computer Vision Workshops.*

[55] Z. A. Khan and W. Sohn, “Abnormal Human Activity Recognition System Based on R-Transform and Kernel Discriminant Technique for Elderly Home Care,” 2011.

[56] A. A. Chaaraoui and F. Flórez-Revuelta, “Optimizing human action recognition based on a cooperative coevolutionary algorithm,” *Eng Appl Artif Intell*, vol. 31, pp. 116–125, 2014, doi: 10.1016/j.engappai.2013.10.003.

[57] A. A. Chaaraoui and F. Flórez-Revuelta, “A Low-Dimensional Radial Silhouette-Based Feature for Fast Human Action Recognition Fusing Multiple Views,” *Int Sch Res Notices*, vol. 2014, pp. 1–11, Oct. 2014, doi: 10.1155/2014/547069.

[58] J. Chiverton and M. Mirmehdi, “CHIVERTON, ET AL. : ON-LINE LEARNING OF SHAPE INFORMATION On-line Learning of Shape Information for Object Segmentation and Tracking.”

[59] J. Chiverton, X. Xie, and M. Mirmehdi, “Automatic bootstrapping and tracking of object contours,” *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1231–1245, Mar. 2012, doi: 10.1109/TIP.2011.2167343.

[60] L. Cai, L. He, T. Yamashita, Y. Xu, Y. Zhao, and X. Yang, “Robust contour tracking by combining region and boundary information,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 12, pp. 1784–1794, Dec. 2011, doi: 10.1109/TCSVT.2011.2133550.

[61] S. Y. Chun and C. S. Lee, “Human action recognition using histogram of motion intensity and direction from multiple views,” *IET Computer Vision*, vol. 10, no. 4, pp. 250–256, Jun. 2016, doi: 10.1049/iet-cvi.2015.0233.

[62] M. Paul, M. E. Haque, and S. Chakraborty, “Human detection in surveillance videos and its applications-a review,” 2013. [Online]. Available: http://asp.eurasipjournals.com/content/2013/1/176

[63] A. F. Bobick and J. W. Davis, “The Recognition of Human Movement Using Temporal Templates.”

[64] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, “Motion Interchange Patterns for Action Recognition in Unconstrained Videos.” [Online]. Available: http://www.openu.ac.il/home/hassner/projects/MIP/

[65] M. Ahmad and S. W. Lee, “HMM-based human action recognition using multiview image sequences,” in *Proceedings - International Conference on Pattern Recognition*, 2006, pp. 263–266. doi: 10.1109/ICPR.2006.630.

[66] S. Pehlivan and D. A. Forsyth, “Recognizing activities in multiple views with fusion of frame judgments,” *Image Vis Comput*, vol. 32, no. 4, pp. 237–249, 2014, doi: 10.1016/j.imavis.2014.01.006.

[67] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, “A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics,” *Rob Auton Syst*, vol. 77, pp. 25–38, Mar. 2016, doi: 10.1016/j.robot.2015.11.013.

[68] S. Pehlivan and D. A. Forsyth, “Recognizing activities in multiple views with fusion of frame judgments,” *Image Vis Comput*, vol. 32, no. 4, pp. 237–249, 2014, doi: 10.1016/j.imavis.2014.01.006.

[69] Z. Jiang, Z. Lin, and L. Davis, “Recognizing human actions by learning and matching shape-motion prototype trees,” *IEEE Trans Pattern Anal Mach Intell*, vol. 34, no. 3, pp. 533–547, 2012, doi: 10.1109/TPAMI.2011.147.

[70] 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 Recognit*, vol. 61, pp. 295–308, Jan. 2017, doi: 10.1016/j.patcog.2016.08.003.

[71] L. Shao, L. Ji, Y. Liu, and J. Zhang, “Human action segmentation and recognition via motion and shape analysis,” *Pattern Recognit Lett*, vol. 33, no. 4, pp. 438–445, Mar. 2012, doi: 10.1016/j.patrec.2011.05.015.

[72] D. Xing, X. Wang, and H. Lu, “Action Recognition Using Hybrid Feature Descriptor and VLAD Video Encoding.”

[73] H. Qian, Y. Mao, W. Xiang, and Z. Wang, “Recognition of human activities using SVM multi-class classifier,” *Pattern Recognit Lett*, vol. 31, no. 2, pp. 100–111, Jan. 2010, doi: 10.1016/j.patrec.2009.09.019.

[74] T. Ojala, M. Pietikäinen, and D. Harwood, “Performance evaluation of texture measures with classification based on Kullback discrimination of distributions,” in *Proceedings - International Conference on Pattern Recognition*, Institute of Electrical and Electronics Engineers Inc., 1994, pp. 582–585. doi: 10.1109/ICPR.1994.576366.

[75] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans Pattern Anal Mach Intell*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.

[76] G. Zhao and M. Pietikäinen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” *IEEE Trans Pattern Anal Mach Intell*, vol. 29, no. 6, pp. 915–928, Jun. 2007, doi: 10.1109/TPAMI.2007.1110.

[77] J. Luo and H. Qi, “Motion Local Ternary Pattern for Distributed Multi-View Human Action Recognition,” 2012.

[78] V. Kellokumpu, G. Zhao, and M. Pietikäinen, “Human activity recognition using a dynamic texture based method,” in *BMVC 2008 - Proceedings of the British Machine Vision Conference 2008*, British Machine Vision Association, BMVA, 2008. doi: 10.5244/C.22.88.

[79] C. Chen, R. Jafari, and N. Kehtarnavaz, “Action recognition from depth sequences using depth motion maps-based local binary patterns,” in *Proceedings - 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2015*, Institute of Electrical and Electronics Engineers Inc., Feb. 2015, pp. 1092–1099. doi: 10.1109/WACV.2015.150.

[80] A. K. S. Kushwaha, S. Srivastava, and R. Srivastava, “Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns,” *Multimed Syst*, vol. 23, no. 4, pp. 451–467, Jul. 2017, doi: 10.1007/s00530-016-0505-x.

[81] T. P. Nguyen, A. Manzanera, N.-S. Vu, and M. Garrigues, “LNCS 8259 - Revisiting LBP-Based Texture Models for Human Action Recognition.”

[82] S. K. Dhar, M. M. Hasan, and S. A. Chowdhury, “Human activity recognition based on Gaussian mixture model & directive local binary pattern,” in *ICECTE 2016 - 2nd International Conference on Electrical, Computer and Telecommunication Engineering*, Institute of Electrical and Electronics Engineers Inc., Mar. 2017. doi: 10.1109/ICECTE.2016.7879568.

[83] E. Chen, S. Zhang, and C. Liang, “Action Recognition Using Motion History Image and Static History Image-based Local Binary Patterns,” *International Journal of Multimedia and Ubiquitous Engineering*, vol. 12, no. 1, pp. 203–214, Jan. 2017, doi: 10.14257/ijmue.2017.12.1.17.

[84] M. Khare and M. Jeon, “Multi-resolution approach to human activity recognition in video sequence based on combination of complex wavelet transform, Local Binary Pattern and Zernike moment,” *Multimed Tools Appl*, vol. 81, no. 24, pp. 34863–34892, Oct. 2022, doi: 10.1007/s11042-021-11828-6.

[85] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, “An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity,” *EURASIP J Adv Signal Process*, vol. 2011, 2011, doi: 10.1155/2011/540375.

[86] M. Nii, K. Nakai, T. Fujita, and Y. Takahashi, “Action estimation from human activity monitoring data using soft computing approach,” in *Proceedings - 3rd International Conference on Emerging Trends in Engineering and Technology, ICETET 2010*, 2010, pp. 434–439. doi: 10.1109/ICETET.2010.149.

[87] P. T. Chinimilli, S. Redkar, and W. Zhang, “Human activity recognition using inertial measurement units and smart shoes,” in *Proceedings of the American Control Conference*, Institute of Electrical and Electronics Engineers Inc., Jun. 2017, pp. 1462–1467. doi: 10.23919/ACC.2017.7963159.

[88] R. Vezzani, M. Piccardi, and R. Cucchiara, “An efficient Bayesian framework for on-line action recognition,” in *Proceedings - International Conference on Image Processing, ICIP*, IEEE Computer Society, 2009, pp. 3553–3556. doi: 10.1109/ICIP.2009.5414340.

[89] J. Liu, B. Kuipers, and S. Savarese, “Recognizing human actions by attributes,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* IEEE Computer Society, 2011, pp. 3337–3344. doi: 10.1109/CVPR.2011.5995353.

[90] C. H. Lim and C. S. Chan, “Fuzzy qualitative human model for viewpoint identification,” *Neural Comput Appl*, vol. 27, no. 4, pp. 845–856, May 2016, doi: 10.1007/s00521-015-1900-5.

[91] F. Zhu, L. Shao, J. Xie, and Y. Fang, “From handcrafted to learned representations for human action recognition: A survey,” *Image Vis Comput*, vol. 55, pp. 42–52, Nov. 2016, doi: 10.1016/j.imavis.2016.06.007.

[92] Q. Qiu, Z. Jiang, and R. Chellappa, “Sparse dictionary-based representation and recognition of action attributes,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 707–714. doi: 10.1109/ICCV.2011.6126307.

[93] L. Liu, L. Shao, X. Li, and K. Lu, “Learning spatio-temporal representations for action recognition: A genetic programming approach,” *IEEE Trans Cybern*, vol. 46, no. 1, pp. 158–170, Jan. 2016, doi: 10.1109/TCYB.2015.2399172.

[94] F. Xie, A. Song, and V. Ciesielski, “LNCS 7835 - Human Action Recognition from Multi-Sensor Stream Data by Genetic Programming.” [Online]. Available: http://www.rmit.edu.au/compsci

[95] F. Xie, A. Song, and V. Ciesielski, “Genetic programming based activity recognition on a smartphone sensory data benchmark,” in *Proceedings of the 2014 IEEE Congress on Evolutionary Computation, CEC 2014*, Institute of Electrical and Electronics Engineers Inc., Sep. 2014, pp. 2917–2924. doi: 10.1109/CEC.2014.6900635.

[96] J. Dou, Q. Qin, and Z. Tu, “Robust visual tracking based on generative and discriminative model collaboration,” *Multimed Tools Appl*, vol. 76, no. 14, pp. 15839–15866, Jul. 2017, doi: 10.1007/s11042-016-3872-6.

[97] M. Maniruzzaman *et al.*, “Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers,” *J Med Syst*, vol. 42, no. 5, May 2018, doi: 10.1007/s10916-018-0940-7.

[98] I. Ar and Y. S. Akgul, “Action recognition using random forest prediction with combined pose-based and motion-based features,” in *ELECO 2013 - 8th International Conference on Electrical and Electronics Engineering*, IEEE Computer Society, 2013, pp. 315–319. doi: 10.1109/eleco.2013.6713852.

[99] U. M. Nunes, D. R. Faria, and P. Peixoto, “A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier,” *Pattern Recognit Lett*, vol. 99, pp. 21–31, Nov. 2017, doi: 10.1016/j.patrec.2017.05.004.

[100] L. Gan and F. Chen, “Human action recognition using APJ3D and random forests,” *Journal of Software*, vol. 8, no. 9, pp. 2238–2245, 2013, doi: 10.4304/jsw.8.9.2238-2245.

[101] Y. Liu, Institute of Electrical and Electronics Engineers, and IEEE Circuits and Systems Society, *ICNC-FSKD 2017 : 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery : Guilin, Guangxi, China, 29-31 July, 2017*.

[102] C. Cortes, V. Vapnik, and L. Saitta, “Support-Vector Networks Editor,” Kluwer Academic Publishers, 1995.

[103] B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzàlez, and F. Xavier Roca, “A selective spatio-temporal interest point detector for human action recognition in complex scenes,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 1776–1783. doi: 10.1109/ICCV.2011.6126443.

[104] H. Kim *et al.*, “Weighted joint-based human behavior recognition algorithm using only depth information for low-cost intelligent video-surveillance system,” *Expert Syst Appl*, vol. 45, pp. 131–141, Mar. 2016, doi: 10.1016/j.eswa.2015.09.035.

[105] S. Nazir, M. H. Yousaf, and S. A. Velastin, “Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition,” *Computers and Electrical Engineering*, vol. 72, pp. 660–669, Nov. 2018, doi: 10.1016/j.compeleceng.2018.01.037.

[106] Noboru. Babaguchi *et al.*, *MM’12 : proceedings of the 20th ACM International Conference on Multimedia : October 29 - November 2, 2012, Nara, Japan*. ACM, 2012.

[107] N. M. Oliver, B. Rosario, and A. P. Pentland, “A bayesian computer vision system for modeling human interactions,” *IEEE Trans Pattern Anal Mach Intell*, vol. 22, no. 8, pp. 831–843, 2000, doi: 10.1109/34.868684.

[108] N. Robertson and I. Reid, “A General Method for Human Activity Recognition in Video.”

[109] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, “3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold,” *IEEE Trans Cybern*, vol. 45, no. 7, pp. 1340–1352, Jul. 2015, doi: 10.1109/TCYB.2014.2350774.

[110] IEEE Staff and IEEE Staff, *2008 IEEE Conference on Computer Vision and Pattern Recognition.*

[111] A. Oikonomopoulos, I. Patras, and M. Pantic, “Spatiotemporal salient points for visual recognition of human actions,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 3, pp. 710–719, Jun. 2006, doi: 10.1109/TSMCB.2005.861864.

[112] D. K. Vishwakarma and R. Kapoor, “Hybrid classifier based human activity recognition using the silhouette and cells,” *Expert Syst Appl*, vol. 42, no. 20, pp. 6957–6965, Jun. 2015, doi: 10.1016/j.eswa.2015.04.039.

[113] D. Xu, X. Xiao, X. Wang, and J. Wang, "Human action recognition based on Kinect and PSO-SVM by representing 3D skeletons as points in lie group," in *ICALIP 2016 - 2016 International Conference on Audio, Language and Image Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Feb. 2017, pp. 568–573. doi: 10.1109/ICALIP.2016.7846646.

[114] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional neural networks for human action recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 1, pp. 221–231, 2013, doi: 10.1109/TPAMI.2012.59.

[115] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.

[116] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential Deep Learning for Human Action Recognition."

[117] S. C. Huang, "An advanced motion detection algorithm with video quality analysis for video surveillance systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 1–14, Jan. 2011, doi: 10.1109/TCSVT.2010.2087812.

[118] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," Apr. 2016, [Online]. Available: http://arxiv.org/abs/1604.06573

[119] M. Panwar *et al.*, *CNN Based Approach for Activity Recognition Using a Wrist-Worn Accelerometer*. 2017. doi: 10.0/Linux-x86_64.

[120] T. Singh and D. K. Vishwakarma, "A deeply coupled ConvNet for human activity recognition using dynamic and RGB images," *Neural Comput Appl*, vol. 33, no. 1, pp. 469–485, Jan. 2021, doi: 10.1007/s00521-020-05018-y.

[121] W. Li, L. Wen, M. C. Chang, S. N. Lim, and S. Lyu, "Adaptive RNN Tree for Large-Scale Human Action Recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, Institute of Electrical and Electronics Engineers Inc., Dec. 2017, pp. 1453–1461. doi: 10.1109/ICCV.2017.161.

[122] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.09130

[123] T. Akilan, Q. M. J. Wu, A. Safaei, and W. Jiang, *A Late Fusion Approach for Harnessing Multi-CNN Model High-level Features*. 2017. doi: 10.0/Linux-x86_64.

[124] C. Li, P. Wang, S. Wang, Y. Hou, and W. Li, "Skeleton-based Action Recognition Using LSTM and CNN," Jul. 2017, [Online]. Available: http://arxiv.org/abs/1707.02356

[125] M. Majd and R. Safabakhsh, “Correlational Convolutional LSTM for human action recognition,” *Neurocomputing*, vol. 396, pp. 224–229, Jul. 2020, doi: 10.1016/j.neucom.2018.10.095.

[126] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. Van Gool, “StagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 549–565, Feb. 2020, doi: 10.1109/TCSVT.2019.2894161.

[127] Z. Luo, B. Peng, D.-A. Huang, A. Alahi, and L. Fei-Fei, “Unsupervised Learning of Long-Term Motion Dynamics for Videos,” Jan. 2017, [Online]. Available: http://arxiv.org/abs/1701.01821

[128] V. Veeriah, N. Zhuang, and G.-J. Qi, “Differential Recurrent Neural Networks for Action Recognition,” Apr. 2015, [Online]. Available: http://arxiv.org/abs/1504.06678

[129] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, “Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments,” *Future Generation Computer Systems*, vol. 96, pp. 386–397, Jul. 2019, doi: 10.1016/j.future.2019.01.029.

[130] Y. S. Chong and Y. H. Tay, “Abnormal Event Detection in Videos using Spatiotemporal Autoencoder,” Jan. 2017, [Online]. Available: http://arxiv.org/abs/1701.01546

[131] F. Gu, K. Khoshelham, S. Valaee, J. Shang, and R. Zhang, “Locomotion Activity Recognition Using Stacked Denoising Autoencoders,” *IEEE Internet Things J*, vol. 5, no. 3, pp. 2085–2093, Jun. 2018, doi: 10.1109/JIOT.2018.2823084.

[132] P. V. Ca, L. T. Edu, I. Lajoie, Y. B. Ca, and P.-A. M. Ca, “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion Pascal Vincent Hugo Larochelle Yoshua Bengio Pierre-Antoine Manzagol,” 2010.

[133] B. Almaslukh, B. Almaslukh, J. Almuhtadi, and A. Artoli, “An Effective Deep Autoencoder Approach for Online Smartphone-Based Human Activity Recognition Online Smartphone-Based Human Activity Recognition View project An Effective Deep Autoencoder Approach for Online Smartphone-Based Human Activity Recognition,” 2017. [Online]. Available: https://www.researchgate.net/publication/323019783

[134] R. Cui, G. Hua, and J. Wu, “AP-GAN: Predicting skeletal activity to improve early activity recognition,” *J Vis Commun Image Represent*, vol. 73, Nov. 2020, doi: 10.1016/j.jvcir.2020.102923.

[135] K. Simonyan and A. Zisserman, “Two-Stream Convolutional Networks for Action Recognition in Videos.”

[136] S. U. Park, J. H. Park, M. A. Al-Masni, M. A. Al-Antari, M. Z. Uddin, and T. S. Kim, “A Depth Camera-based Human Activity Recognition via Deep Learning Recurrent Neural

Network for Health and Social Care Services," in *Procedia Computer Science*, Elsevier B.V., 2016, pp. 78–84. doi: 10.1016/j.procs.2016.09.126.

[137] E. P. Ijjina and K. M. Chalavadi, "Human action recognition using genetic algorithms and convolutional neural networks," *Pattern Recognit,* vol. 59, pp. 199–212, Nov. 2016, doi: 10.1016/j.patcog.2016.01.012.

[138] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," *Journal of King Saud University - Computer and Information Sciences,* vol. 32, no. 4, pp. 447–453, May 2020, doi: 10.1016/j.jksuci.2019.09.004.

[139] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "DeepSense: A Unified Deep Learning Framework for Time-Series Mobile Sensing Data Processing," Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.01942

[140] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors (Switzerland),* vol. 16, no. 1, Jan. 2016, doi: 10.3390/s16010115.

[141] E. Kanjo, E. M. G. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion,* vol. 49, pp. 46–56, Sep. 2019, doi: 10.1016/j.inffus.2018.09.001.

[142] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, Mar. 2019, doi: 10.1016/j.inffus.2018.06.002.

[143] M. A. Khan *et al.*, "Human action recognition using fusion of multiview and deep features: an application to video surveillance," *Multimed Tools Appl*, 2020, doi: 10.1007/s11042-020-08806-9.

[144] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-Temporal Attention Networks for Action Recognition and Detection," *IEEE Trans Multimedia,* vol. 22, no. 11, pp. 2990–3001, Nov. 2020, doi: 10.1109/TMM.2020.2965434.

[145] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human Action Recognition by Learning Spatio-Temporal Features with Deep Neural Networks," *IEEE Access,* vol. 6, pp. 17913–17922, Mar. 2018, doi: 10.1109/ACCESS.2018.2817253.

[146] P. Wang, W. Li, Z. Gao, C. Tang, J. Zhang, and P. Ogunbona, "Convnets-based action recognition from depth maps through virtual cameras and pseudocoloring," in *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*, Association for Computing Machinery, Inc, Oct. 2015, pp. 1119–1122. doi: 10.1145/2733373.2806296.

[147] C. Q. Le, T. D. Ngo, D. D. Le, S. Satoh, and D. A. Duong, "Human Action recognition from depth videos using multi-projection based representation," in *2015 IEEE 17th International*

Workshop on Multimedia Signal Processing, MMSP 2015, Institute of Electrical and Electronics Engineers Inc., Nov. 2015. doi: 10.1109/MMSP.2015.7340879.

[148] X. Liu, Y. Li, and Q. Wang, “Multi-View Hierarchical Bidirectional Recurrent Neural Network for Depth Video Sequence Based Action Recognition,” *Intern J Pattern Recognit Artif Intell*, vol. 32, no. 10, Oct. 2018, doi: 10.1142/S0218001418500337.

[149] C. Liang, E. Chen, L. Qi, and L. Guan, “Improving action recognition using collaborative representation of local depth map feature,” *IEEE Signal Process Lett*, vol. 23, no. 9, pp. 1241–1245, Sep. 2016, doi: 10.1109/LSP.2016.2592419.

[150] Y. Du, Y. Fu, and L. Wang, “Skeleton based action recognition with convolutional neural network,” in *Proceedings - 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015*, Institute of Electrical and Electronics Engineers Inc., Jun. 2016, pp. 579–583. doi: 10.1109/ACPR.2015.7486569.

[151] Y. Hou, Z. Li, P. Wang, and W. Li, “Skeleton Optical Spectra-Based Action Recognition Using Convolutional Neural Networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, Mar. 2018, doi: 10.1109/TCSVT.2016.2628339.

[152] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, “SkeletonNet: Mining Deep Part Features for 3-D Action Recognition,” *IEEE Signal Process Lett*, vol. 24, no. 6, pp. 731–735, Jun. 2017, doi: 10.1109/LSP.2017.2690339.

[153] Y. Du, Y. Fu, and L. Wang, “Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition,” *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3010–3022, Jul. 2016, doi: 10.1109/TIP.2016.2552404.

[154] W. Zhu *et al.*, “Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks,” Mar. 2016, [Online]. Available: http://arxiv.org/abs/1603.07772

[155] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis,” Apr. 2016, [Online]. Available: http://arxiv.org/abs/1604.02808

[156] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, “Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition,” Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.01189

[157] M. A. R. Ahad, M. Ahmed, A. Das Antar, Y. Makihara, and Y. Yagi, “Action recognition using kinematics posture feature on 3D skeleton joint locations,” *Pattern Recognit Lett*, vol. 145, pp. 216–224, May 2021, doi: 10.1016/j.patrec.2021.02.013.

[158] J. Duan, S. Zhou, J. Wan, X. Guo, and S. Z. Li, “Multi-Modality Fusion based on Consensus-Voting and 3D Convolution for Isolated Gesture Recognition,” Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.06689

[159] C. Zhao, M. Chen, J. Zhao, Q. Wang, and Y. Shen, “3D behavior recognition based on multimodal deep space-time learning,” *Applied Sciences (Switzerland),* vol. 9, no. 4, Feb. 2019, doi: 10.3390/app9040716.

[160] T. Singh and D. K. Vishwakarma, “A deep multimodal network based on bottleneck layer features fusion for action recognition,” *Multimed Tools Appl,* vol. 80, no. 24, pp. 33505–33525, Oct. 2021, doi: 10.1007/s11042-021-11415-9.

[161] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, X. Xue, and J. Wang, “Fusing Multi-Stream Deep Networks for Video Classification,” Sep. 2015, [Online]. Available: http://arxiv.org/abs/1509.06086

[162] S. Mukherjee, L. Anvitha, and T. M. Lahari, “Human Activity Recognition in RGB-D Videos by Dynamic Images,” Jul. 2018, [Online]. Available: http://arxiv.org/abs/1807.02947

[163] C. Zhang, Y. Tian, X. Guo, and J. Liu, “DAAL: Deep activation-based attribute learning for action recognition in depth videos,” *Computer Vision and Image Understanding,* vol. 167, pp. 37–49, Feb. 2018, doi: 10.1016/j.cviu.2017.11.008.

[164] A. Franco, A. Magnani, and D. Maio, “A multimodal approach for human activity recognition based on skeleton and RGB data,” *Pattern Recognit Lett,* vol. 131, pp. 293–299, Mar. 2020, doi: 10.1016/j.patrec.2020.01.010.

[165] K. Mozafari, J. A. Nasiri, N. M. Charkari, and S. Jalili, “Action recognition by local space-time features and least square twin SVM (LS-TSVM),” in *Proceedings - 1st International Conference on Informatics and Computational Intelligence, ICI 2011,* 2011, pp. 287–292. doi: 10.1109/ICI.2011.55.

[166] S. M. M. Ahsan, J. K. Tan, H. Kim, and S. Ishikawa, “Histogram of spatio temporal local binary patterns for human action recognition,” in *2014 Joint 7th International Conference on Soft Computing and Intelligent Systems, SCIS 2014 and 15th International Symposium on Advanced Intelligent Systems, ISIS 2014,* Institute of Electrical and Electronics Engineers Inc., Feb. 2014, pp. 1007–1011. doi: 10.1109/SCIS-ISIS.2014.7044905.

[167] T. P. Nguyen, A. Manzanera, N. S. Vu, and M. Garrigues, “Revisiting LBP-based texture models for human action recognition,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2013, pp. 286–293. doi: 10.1007/978-3-642-41827-3_36.

[168] X. Yang and Y. Tian, “Effective 3D action recognition using EigenJoints,” *J Vis Commun Image Represent,* vol. 25, no. 1, pp. 2–11, Jan. 2014, doi: 10.1016/j.jvcir.2013.03.001.

[169] J. Lei, G. Li, J. Zhang, Q. Guo, and D. Tu, “Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model,” *IET Computer Vision,* vol. 10, no. 6, pp. 537–544, Sep. 2016, doi: 10.1049/iet-cvi.2015.0408.

[170] S. Yu, Y. Cheng, L. Xie, and S. Z. Li, "Fully convolutional networks for action recognition," *IET Computer Vision*, vol. 11, no. 8, pp. 744–749, Dec. 2017, doi: 10.1049/iet-cvi.2017.0005.

[171] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.03607

[172] H. Bilen, B. Fernando, E. Gavves, and A. Vedaldi, "Action Recognition with Dynamic Image Networks," *IEEE Trans Pattern Anal Mach Intell*, vol. 40, no. 12, pp. 2799–2813, Dec. 2018, doi: 10.1109/TPAMI.2017.2769085.

[173] A. Ullah, K. Muhammad, T. Hussain, and S. W. Baik, "Conflux LSTMs Network: A Novel Approach for Multi-View Action Recognition," *Neurocomputing*, vol. 435, pp. 321–329, May 2021, doi: 10.1016/j.neucom.2019.12.151.

[174] D. Deotale *et al.*, "HARTIV: Human activity recognition using temporal information in videos," *Computers, Materials and Continua*, vol. 70, no. 2, pp. 3919–3938, 2022, doi: 10.32604/cmc.2022.020655.

[175] C. Zhang, J. Wu, and Y. Li, "ActionFormer: Localizing Moments of Actions with Transformers," Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.07925

[176] I. U. Khan, S. Afzal, and J. W. Lee, "Human activity recognition via hybrid deep learning based model," *Sensors*, vol. 22, no. 1, Jan. 2022, doi: 10.3390/s22010323.

[177] A. Sánchez-Caballero *et al.*, "3DFCNN: real-time action recognition using 3D deep neural networks with raw depth information," *Multimed Tools Appl*, vol. 81, no. 17, pp. 24119–24143, Jul. 2022, doi: 10.1007/s11042-022-12091-z.

[178] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 1395–1402. doi: 10.1109/ICCV.2005.28.

[179] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3 SPEC. ISS., pp. 249–257, Nov. 2006, doi: 10.1016/j.cviu.2006.07.013.

[180] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," Dec. 2012, [Online]. Available: http://arxiv.org/abs/1212.0402

[181] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH: A spatio-temporal maximum average correlation height filter for action recognition," in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2008. doi: 10.1109/CVPR.2008.4587727.

[182] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding."

[183] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716–723. doi: 10.1109/CVPR.2013.98.

[184] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, “PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding,” Mar. 2017, [Online]. Available: http://arxiv.org/abs/1703.07475

[185] D. Minnen, T. Starner, P. Lukowicz, T. Westeyn, and J. A. Ward, “Performance metrics and evaluation issues for continuous activity recognition," in Performance Metrics for Intelligent Systems Self-sustainable sensing and communication architecture View project Neurolive View project Performance Metrics and Evaluation Issues for Continuous Activity Recognition.” [Online]. Available: https://www.researchgate.net/publication/228748046

[186] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: A review,” *ACM Computing Surveys*, vol. 43, no. 3. Apr. 2011. doi: 10.1145/1922649.1922653.

[187] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. O. Ogunbona, “Action Recognition from Depth Maps Using Deep Convolutional Neural Networks,” *IEEE Trans Hum Mach Syst*, vol. 46, no. 4, pp. 498–509, Aug. 2016, doi: 10.1109/THMS.2015.2504550.

[188] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3D skeletons as points in a lie group,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Sep. 2014, pp. 588–595. doi: 10.1109/CVPR.2014.82.

[189] A. Bouzerdoum and Institute of Electrical and Electronics Engineers, *2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA) : Wollongong, New South Wales, Australia : 25-27 November 2014*.

[190] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[191] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. F. Li, “Large-scale video classification with convolutional neural networks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Sep. 2014, pp. 1725–1732. doi: 10.1109/CVPR.2014.223.

[192] J. Luo, W. Wang, and H. Qi, “Spatio-temporal feature extraction and representation for RGB-D human action recognition,” *Pattern Recognit Lett*, vol. 50, pp. 139–148, Dec. 2014, doi: 10.1016/j.patrec.2014.03.024.

[193] C. Chen, R. Jafari, and N. Kehtarnavaz, “Improving human action recognition using fusion of depth camera and inertial sensors,” *IEEE Trans Hum Mach Syst*, vol. 45, no. 1, pp. 51–61, Feb. 2015, doi: 10.1109/THMS.2014.2362520.

[194] N. El Din El Madany, Y. He, and L. Guan, “Human action recognition via multiview discriminative analysis of canonical correlations,” in *Proceedings - International Conference on Image Processing, ICIP*, IEEE Computer Society, Aug. 2016, pp. 4170–4174. doi: 10.1109/ICIP.2016.7533145.

[195] P. Verma, A. Sah, and R. Srivastava, “Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition,” *Multimed Syst*, vol. 26, no. 6, pp. 671–685, Dec. 2020, doi: 10.1007/s00530-020-00677-2.

[196] C. Chen, R. Jafari, and N. Kehtarnavaz, “Fusion of depth, skeleton, and inertial data for human action recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., May 2016, pp. 2712–2716. doi: 10.1109/ICASSP.2016.7472170.

[197] E. Escobedo and G. Camara, “A New Approach for Dynamic Gesture Recognition Using Skeleton Trajectory Representation and Histograms of Cumulative Magnitudes; A New Approach for Dynamic Gesture Recognition Using Skeleton Trajectory Representation and Histograms of Cumulative Magnitudes,” *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2016, doi: 10.1109/SIBGRAPI.2016.35.

[198] S. Gaglio, G. Lo Re, and M. Morana, “Human Activity Recognition Process Using 3-D Posture Data,” *IEEE Trans Hum Mach Syst*, vol. 45, no. 5, pp. 586–597, Oct. 2015, doi: 10.1109/THMS.2014.2377111.

[199] P. Khaire, P. Kumar, and J. Imran, “Combining CNN streams of RGB-D and skeletal data for human activity recognition,” *Pattern Recognit Lett*, vol. 115, pp. 107–116, Nov. 2018, doi: 10.1016/j.patrec.2018.04.035.

[200] J. Guo, H. Bai, Z. Tang, P. Xu, D. Gan, and B. Liu, “Multi modal human action recognition for video content matching,” *Multimed Tools Appl*, vol. 79, no. 45–46, pp. 34665–34683, Dec. 2020, doi: 10.1007/s11042-020-08998-0.

[201] T. H. Tran, H. N. Tran, and H. G. Doan, “Dynamic Hand Gesture Recognition from Multi-modal Streams Using Deep Neural Network,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2019, pp. 156–167. doi: 10.1007/978-3-030-33709-4_14.

[202] W. Nie, Y. Yan, D. Song, and K. Wang, “Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition,” *Multimed Tools Appl*, vol. 80, no. 11, pp. 16205–16214, May 2021, doi: 10.1007/s11042-020-08796-8.

[203] S. A. Khowaja and S. L. Lee, “Hybrid and hierarchical fusion networks: a deep cross-modal learning architecture for action recognition,” *Neural Comput Appl*, vol. 32, no. 14, pp. 10423–10434, Jul. 2020, doi: 10.1007/s00521-019-04578-y.

[204] C. Chen, R. Jafari, and N. Kehtarnavaz, “UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor,” in *Proceedings - International Conference on Image Processing, ICIP*, IEEE Computer Society, Dec. 2015, pp. 168–172. doi: 10.1109/ICIP.2015.7350781.

[205] M. F. Bulbul, Y. Jiang, and J. Ma, "DMMs-Based Multiple Features Fusion for Human Action Recognition," *International Journal of Multimedia Data Engineering and Management*, vol. 6, no. 4, pp. 23–39, Oct. 2015, doi: 10.4018/ijmdem.2015100102.

[206] Y. Annadani, D. L. Rakshith, and S. Biswas, "Sliding Dictionary Based Sparse Representation For Action Recognition," Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.00218

[207] P. K. Singh, S. Kundu, T. Adhikary, R. Sarkar, and D. Bhattacharjee, "Progress of Human Action Recognition Research in the Last Ten Years: A Comprehensive Survey," *Archives of Computational Methods in Engineering*, vol. 29, no. 4, pp. 2309–2349, Jun. 2022, doi: 10.1007/s11831-021-09681-9.

[208] A. Ladjailia, I. Bouchrika, H. F. Merouani, N. Harrati, and Z. Mahfouf, "Human activity recognition via optical flow: decomposing activities into basic actions," *Neural Comput Appl*, vol. 32, no. 21, pp. 16387–16400, Nov. 2020, doi: 10.1007/s00521-018-3951-x.

[209] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition." [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[210] Fuqiang Gu, Kourosh Khoshelham, and Shahrokh Valaee, *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC) : 8-13 Oct. 2017.*

[211] S. Aubry, S. Laraba, J. Tilmanne, and T. Dutoit, "Action recognition based on 2D skeletons extracted from RGB videos," *MATEC Web of Conferences*, vol. 277, p. 02034, 2019, doi: 10.1051/matecconf/201927702034.

[212] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," Dec. 2018, [Online]. Available: http://arxiv.org/abs/1812.08008

[213] Y. Du, W. Wang, and L. Wang, "Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition."

[214] M. Majd and R. Safabakhsh, "Correlational Convolutional LSTM for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, Jul. 2020, doi: 10.1016/j.neucom.2018.10.095.

[215] M. Qi, Y. Wang, J. Qin, A. Li, J. Luo, and L. Van Gool, "StagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 2, pp. 549–565, Feb. 2020, doi: 10.1109/TCSVT.2019.2894161.

[216] Y. Huang, S.-H. Lai, and S.-H. Tai, "Human Action Recognition Based on Temporal Pose CNN and Multi-Dimensional Fusion."

[217] Wang Limin *et al.*, *Computer Vision – ECCV 2016*, vol. 9912. in Lecture Notes in Computer Science, vol. 9912. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-46484-8.

[218] Y. Chen *et al.*, “Graph convolutional network with structure pooling and joint-wise channel attention for action recognition,” *Pattern Recognit*, vol. 103, Jul. 2020, doi: 10.1016/j.patcog.2020.107321.

[219] T. Singh and D. K. Vishwakarma, “A deeply coupled ConvNet for human activity recognition using dynamic and RGB images,” *Neural Comput Appl*, vol. 33, no. 1, pp. 469–485, Jan. 2021, doi: 10.1007/s00521-020-05018-y.

[220] O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.0575

[221] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Institute of Electrical and Electronics Engineers Inc., Nov. 2017, pp. 2261–2269. doi: 10.1109/CVPR.2017.243.

[222] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” May 2019, [Online]. Available: http://arxiv.org/abs/1905.11946

[223] K. K. Reddy and M. Shah, “Recognizing 50 human action categories of web videos,” *Mach Vis Appl*, vol. 24, no. 5, pp. 971–981, 2013, doi: 10.1007/s00138-012-0450-4.

[224] L. Zhang and X. Xiang, “Video event classification based on two-stage neural network,” *Multimed Tools Appl*, vol. 79, no. 29–30, pp. 21471–21486, Aug. 2020, doi: 10.1007/s11042-019-08457-5.

[225] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, “A Robust and Efficient Video Representation for Action Recognition,” *Int J Comput Vis*, vol. 119, no. 3, pp. 219–238, Sep. 2016, doi: 10.1007/s11263-015-0846-5.

[226] Q. Meng, H. Zhu, W. Zhang, X. Piao, and A. Zhang, “Action recognition using form and motion modalities,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 16, no. 1s, Apr. 2020, doi: 10.1145/3350840.

[227] A. Jalal, I. Akhtar, and K. Kim, “Human posture estimation and sustainable events classification via Pseudo-2D stick model and K-ary tree hashing,” *Sustainability (Switzerland)*, vol. 12, no. 23, pp. 1–24, Dec. 2020, doi: 10.3390/su12239814.

[228] S. Vishwakarma and A. Agrawal, “A survey on activity recognition and behavior understanding in video surveillance,” in *Visual Computer*, Oct. 2013, pp. 983–1009. doi: 10.1007/s00371-012-0752-6.

[229] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks.” [Online]. Available: http://code.google.com/p/cuda-convnet/

[230] A. Graves, A. Mohamed, and G. Hinton, “Speech Recognition with Deep Recurrent Neural Networks,” Mar. 2013, [Online]. Available: http://arxiv.org/abs/1303.5778

[231] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, “Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features,” *IEEE Access*, vol. 6, pp. 1155–1166, Nov. 2017, doi: 10.1109/ACCESS.2017.2778011.

[232] E. P. Ijjina and C. Krishna Mohan, “Hybrid deep neural network model for human action recognition,” *Applied Soft Computing Journal*, vol. 46, pp. 936–952, Sep. 2016, doi: 10.1016/j.asoc.2015.08.025.

[233] X. Liang *et al.*, “Learning to Segment Human by Watching YouTube,” Oct. 2017, [Online]. Available: http://arxiv.org/abs/1710.01457

[234] M. Safaei and H. Foroosh, “Single Image Action Recognition by Predicting Space-Time Saliency,” May 2017, [Online]. Available: http://arxiv.org/abs/1705.04641

[235] L. Shi, Y. Zhang, J. Cheng, and H. Lu, “Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition.” [Online]. Available: https://github.com/lshiwjx/2s-AGCN

[236] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object Detection via Region-based Fully Convolutional Networks,” May 2016, [Online]. Available: http://arxiv.org/abs/1605.06409

[237] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, “Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition,” *Pattern Recognit*, vol. 76, pp. 80–94, Apr. 2018, doi: 10.1016/j.patcog.2017.10.033.

[238] G. Sulong and A. Mohammedali, “RECOGNITION OF HUMAN ACTIVITIES FROM STILL IMAGE USING NOVEL CLASSIFIER 1,” *J Theor Appl Inf Technol*, vol. 10, no. 1, 2015, [Online]. Available: www.jatit.org

[239] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, “Dynamic Image Networks for Action Recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Computer Society, Dec. 2016, pp. 3034–3042. doi: 10.1109/CVPR.2016.331.

[240] J. Wu, S. Qiu, R. Zeng, Y. Kong, L. Senhadji, and H. Shu, “Multilinear Principal Component Analysis Network for Tensor Object Classification,” *IEEE Access*, vol. 5, pp. 3322–3331, 2017, doi: 10.1109/ACCESS.2017.2675478.

[241] Swathikiran Sudhakaran and Oswald Lanz, *Learning to Detect Violent Videos using Convolutional Long Short-TermMemory*.

[242] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 4, pp. 447–453, May 2020, doi: 10.1016/j.jksuci.2019.09.004.

[243] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[244] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1234–1241. doi: 10.1109/CVPR.2012.6247806.

[245] F. Shi, E. Petriu, and R. Laganiere, "Sampling strategies for real-time action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2595–2602. doi: 10.1109/CVPR.2013.335.

[246] G. Somasundaram, A. Cherian, V. Morellas, and N. Papanikolopoulos, "Action Recognition Using Global Spatio-Temporal Features Derived from Sparse Representations."

[247] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A Robust and Efficient Video Representation for Action Recognition," *Int J Comput Vis*, vol. 119, no. 3, pp. 219–238, Sep. 2016, doi: 10.1007/s11263-015-0846-5.

[248] A. Howard *et al.*, "Searching for MobileNetV3," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 1314–1324, May 2019, doi: 10.1109/ICCV.2019.00140.

[249] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", Accessed: Sep. 27, 2023. [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[250] "VGG-19 Explained | Papers With Code." Accessed: Aug. 16, 2023. [Online]. Available: https://paperswithcode.com/method/vgg-19

[251] Liu Jingen, Luo Jiebo, and Shah Mubarak, *Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE Conference on : dates: 20-25 June 2009.* IEEE, 2009.

[252] K. K. Reddy and M. Shah, "Recognizing 50 Human Action Categories of Web Videos," *Machine Vision and Applications Journal (MVAP), September*.

[253] H. Wang, A. Kläser, C. Schmid, L. Cheng-Lin, and C.-L. Liu, "Action Recognition by Dense Trajectories," pp. 3169–3176, 2011, doi: 10.1109/CVPR.2011.5995407ï.

[254] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action Recognition using Visual Attention," Nov. 2015, Accessed: Sep. 27, 2023. [Online]. Available: https://arxiv.org/abs/1511.04119v3

[255] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream LSTM: A deep fusion framework for human action recognition," in *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, Institute of Electrical and Electronics Engineers Inc., May 2017, pp. 177–186. doi: 10.1109/WACV.2017.27.

[256] S. Saif, E. D. Wollega, and S. A. Kalevela, “Spatio-Temporal Features based Human Action Recognition using Convolutional Long Short-Term Deep Neural Network,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, p. 2023, Accessed: Sep. 27, 2023. [Online]. Available: www.ijacsa.thesai.org