Cracking the Code: Enhancing Human Activity Recognition through Deep Learning in Videos

By

Koh Jian Yong

FACULTY OF COMPUTING AND
INFORMATION TECHNOLOGY

TUNKU ABDUL RAHMAN UNIVERSITY OF
MANAGEMENT AND TECHNOLOGY
KUALA LUMPUR

ACADEMIC YEAR
2023/2024

Cracking the Code: Enhancing Human Activity Recognition through Deep Learning in Videos

By

Koh Jian Yong

Supervisor: Ts. Dr. Tan Chi Wee

A project report submitted to the
Faculty of Computing and Information Technology
in partial fulfillment of the requirement for the
Bachelor of Computer Science (Honours)

Department of Mathematical and Data Science
Faculty of Computing and Information Technology
Tunku Abdul Rahman University of Management and Technology
Kuala Lumpur

Declaration

The project submitted herewith is a result of my own efforts in totality and in every aspect of the project works. All information that has been obtained from other sources has been fully acknowledged. I understand that any plagiarism, cheating or collusion or any sorts constitutes a breach of TAR University rules and regulations and would be subjected to disciplinary actions.

_JianYong_____

Student Name

Koh Jian Yong

ID: 22WMR05347

Abstract

Human Activity Recognition (HAR) from video data has sparked considerable attention due to its applications in security monitoring, healthcare, and human-computer interaction. The goal of this research is to create a real-time and robust video-based HAR system that can recognise and categorise a wide range of human actions. The complexity and unpredictability of human movements provide problems for accurate recognition, while real-time processing necessitates efficient algorithms. The goals are to improve recognition accuracy, use high-quality video data, and optimise computing efficiency for real-time performance. The dataset used is UCF50, which contains 50 action types from genuine YouTube videos. Pose extraction employs MediaPipe for recording human postures and a LSTM network for deep learning. The phases in the project flow are as follows: pose extraction, data cleaning, model parameter tuning, and model training. The implementation of the system allows for both video upload and real-time camera use for recognition. Video data is preprocessed, activities are recognised, and the results are superimposed on the video to improve context. The relevance of this research consists in tackling issues connected to various human movements, data quality, algorithm selection, and real-time processing. It has the potential to impact security, healthcare, and human-computer interaction, offering safer surroundings, better patient care, and improved user experiences. This initiative paves the path for creative applications across several fields by merging modern computer vision methods with deep learning.

Acknowledgement

I am deeply grateful to all those who have played a role in the successful completion of my project. This endeavour has been a journey of growth, learning, and accomplishment, and I am truly appreciative of the support and guidance I have received.

First and foremost, I would like to express my sincere gratitude to my project supervisor Ts. Dr. Tan Chi Wee for his expert guidance, unwavering encouragement, and insightful mentorship. His expertise has been invaluable in shaping the direction and scope of my project, and I am thankful for his continuous support.

I am thankful to Tunku Abdul Rahman University of Management and Technology (TAR UMT) for providing me with the resources, facilities, and academic environment essential for conducting my research.

I want to acknowledge the camaraderie and collaborative efforts of my peers and classmates, whose discussions and interactions have enriched my project. Their diverse viewpoints and discussions have been intellectually stimulating.

I would like to thank my family and friends for their unwavering support, understanding, and encouragement throughout this journey. Their belief in my abilities has been a driving force behind my accomplishments.

This project has been a significant learning experience for me, and I am grateful to everyone who has been a part of it.

Table of Contents

Chapter 1

Introduction

1 Introduction

Human activity recognition (HAR) from video data has attracted substantial interest in recent years due to its potential uses in a variety of fields, including security surveillance, healthcare, and human-computer interaction. The recognition and categorization of human behaviours and movements from video footage is a task that can be difficult owing to the complexity and diversity of human motions. For HAR systems to successfully monitor and analyse human behaviours in real-world contexts, real-time processing is a critical need. The goal of this project is to create a strong and real-time video-based HAR system.

1.1 Project Background

Video-based HAR is a computer method that uses a sequence of images (video frames) to automatically recognise what human activity is being or was done. Developments in computer vision and machine learning have resulted in tremendous advances in human activity recognition in recent years. Video-based HAR systems are commonly used for surveillance, healthcare monitoring, and increasing human-computer interaction (Siddiqi et al., 2014). These systems focus on collecting important information from video data and using effective algorithms to classify and recognise activities.

However, HAR systems have a number of problems that limit their usefulness. Because of the complexity and diversity of human actions, effectively identifying and classifying behaviours in movies is challenging (Sharma et al., 2022). Lighting, camera angles, and other environmental elements all have an impact on recognition performance. Furthermore, real-time processing is a key difficulty that necessitates an efficient and effective way to process video input and making timely predictions.

1.2 Problem Statement

Human movements are diverse and complicated, making it difficult to effectively recognise and classify behaviours in video data. People do tasks in a variety of ways, which can be impacted by factors such as age, physical ability, cultural background, and personal habits (Sharma et al., 2022). Individuals' motions may also vary over time as a result of physical problems, injuries, or adopting new behaviours. This variety is a substantial challenge for HAR systems, which must be adaptive and capable of capturing subtleties and variances in human movements in order to effectively interpret and classify activities.

The system's recognition performance is directly affected by the quality of the visual data utilised for HAR (Preksha, 2021). Low-resolution or noisy video footage might impede accurate feature extraction and make distinguishing between distinct activities difficult. Similarly, the video frame rate is critical in representing the temporal dynamics of human motions. Inadequate frame rate can result in information loss and impair the system's capacity to detect small motions. As a result, employing high-quality video data with acceptable resolution and frame rate is critical for improving HAR system accuracy and dependability.

Many HAR applications, like security monitoring and healthcare, require real-time processing. Timely detection of suspicious activity or possible threats is critical in security monitoring for averting incidents and protecting public safety. Similarly, real-time monitoring of patient movements and behaviours in healthcare settings can help in the early discovery of anomalies or changes that may necessitate rapid medical intervention. Real-time recognition, on the other hand, creates substantial hurdles. The computational complexity of real-time video data processing, along with the necessity for rapid replies, demands the development of efficient and responsive HAR systems. To maintain system efficacy and responsiveness, efficient algorithms and optimisation approaches must be used to balance computing economy with precise recognition in real-time circumstances.

1.3 Objective

Objective 1: To develop a robust Human Activity Recognition (HAR) system capable of accurately recognizing and classifying a diverse range of human behaviours, despite variations in human mobility with at least 70% accuracy

Objective 2: To enhance the performance of a Human Activity Recognition (HAR) system by incorporating at least 720p video data

Objective 3: To improve the computational efficiency and responsiveness of a Human Activity Recognition (HAR) system, enabling real-time recognition and classification of human behaviours in various applications

1.4 Significance of the study

The importance of deploying HAR in video surveillance cannot be overstated. For starters, it immediately addresses the essential demand for increased public safety and security. The technology can assist avoid events and possible threats by properly recognising and classifying suspicious behaviours in real-time, assuring the well-being of persons and the protection of public areas. This is especially important in today's society, when security concerns are becoming more widespread.

Second, real-time monitoring of patient movements and behaviours may considerably aid medical practitioners in the healthcare business. The technology can give vital insights into the patient's status by continually following and analysing their actions, recognising any strange or worrying behaviours immediately. This allows for early intervention and proper medical reactions, resulting in better patient care and perhaps saving lives.

The project extends the frontiers of HAR capabilities by tackling the issues connected with the diversity and complexity of human movements, video data quality, algorithm selection, and real-time processing. This not only improves the accuracy and efficacy of existing HAR systems, but also prepares the way for future advancements and applications in a variety of fields.

Furthermore, the suggested approach allows for intuitive and natural human-computer interface, allowing people to communicate with technology via gestures and actions. This has the potential to transform human-computer interactions, making them more fluid and user-friendly. The system can recognise and respond to human movements

by merging machine learning and deep learning techniques, bringing up new opportunities for enhanced user experiences and creative applications in sectors such as gaming, virtual reality, and robotics.

Real-time video-based Human Activity Recognition (HAR) revolutionises the way we perceive and understand human actions. By enabling automated recognition of human activities, it liberates us from the limitations of human operators and reduces the potential for human error (Kamthe & Patil ,2018). This advanced technology empowers us to effectively monitor and interpret vast amounts of video data, unlocking new possibilities for efficient surveillance and analysis. Not only does real-time video-based HAR streamline operations, but it also offers significant cost savings by reducing the reliance on additional manpower. Embracing this cutting-edge approach allows organisations to allocate resources more efficiently and focus on other critical aspects of their operations.

In conclusion, the importance of video-based HAR resides in its capacity to increase safety and security, healthcare monitoring, and intuitive human-computer interface. The project will have a beneficial influence on numerous sectors by tackling the highlighted difficulties and accomplishing the defined objectives, leading to a safer and more efficient society.

1.5 Overview of the thesis structure

This thesis is organised into five parts, namely Introduction, Literature Review, Methodology, Results and Discussion, and Conclusion.

The Introduction section serves as an introduction to our project, where we outline the problem we aim to solve and establish the objectives of our research. Additionally, we delve into the societal impact of our project, highlighting its potential contributions to society.

Within the Literature Review, existing scholarly works that are relevant to our project are thoroughly examined. This comprehensive review enables us to gain a comprehensive understanding of the current knowledge on the topic, identify applicable ideas, methodologies, and pinpoint areas that require further research.

The Methodology section delves into the approach we have chosen to undertake our project. We provide a detailed description of the method we have employed and present a flowchart illustrating the program's workflow and implementation process.

In the Results and Discussion section, we present the outcomes of our project and engage in an in-depth analysis to comprehend the reasons behind the obtained results. This section allows us to evaluate the significance and implications of our findings, facilitating a comprehensive understanding of the project's outcomes.

Finally, in the Conclusion, we draw overall conclusions based on our project's outcomes and discuss the findings in relation to the initial objectives. This concluding section provides a summary of the entire thesis and highlights the key contributions and implications of our research.

Chapter 2

Literature Review

2 Literature Review

2.1 Image Processing

Image processing is a way of performing operations on an image in order to improve it or extract important information from it. It is a sort of signal processing in which the input is an image and the output might be an image or image characteristics/features (Leemets & Vajakas, n.d.). It applies algorithms and mathematical models to handle and analyse digital pictures. It aims to improve picture quality, extract relevant information from images, and automate image-based operations (Mostafa & Hegazy, 2021).

Image processing technologies are classified into two types: analogue image processing and digital image processing. Hard copies, such as prints and pictures, can benefit from analogue image processing. When applying these visual approaches, image analysts employ a variety of interpretive foundations. Digital image processing techniques aid in the alteration of digital pictures through the use of computers. Pre-processing, augmentation, and presentation, as well as information extraction, are the three main processes that all sorts of data must go through when employing digital techniques (Leemets & Vajakas, n.d.). There are few image processing techniques which are median filtering, histogram equalisation, gaussian smoothing and many more.

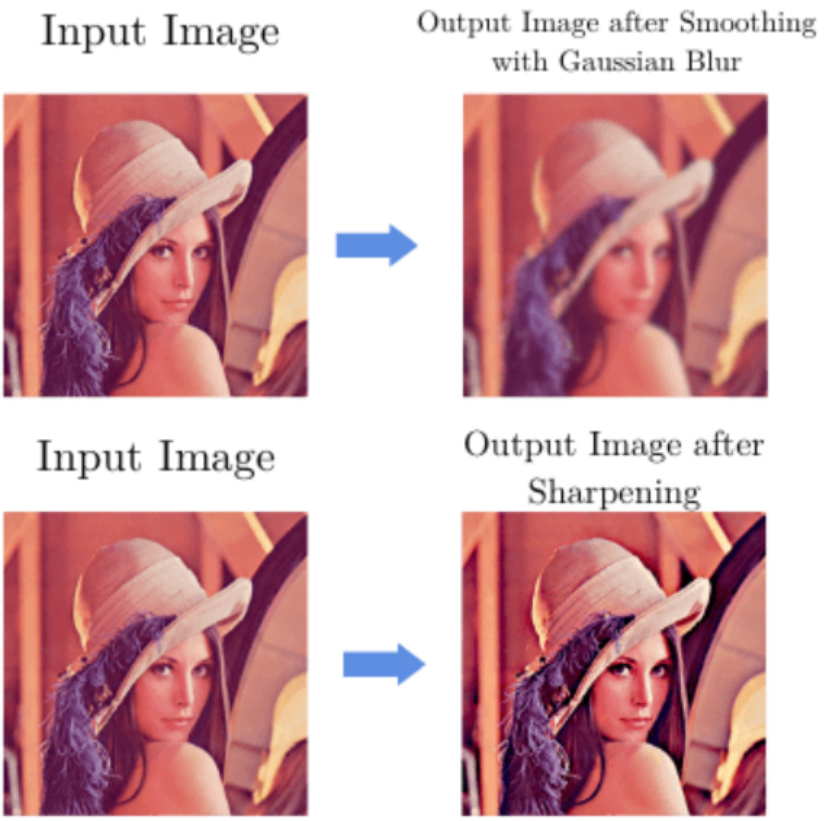


Figure 2.2: Example of Image Processing

2.1.1 Image Processing Techniques

Edge Detection

Edge detection is a critical component of image processing that has a direct impact on image segmentation quality. Edge detection approaches include Sobel, Prewitt, Roberts, the Canny method, and the line Hough transform (Zhou et al., 2021). These approaches can be used to compare the benefits and scenes of various procedures. In 2015, Vikram Mutneja introduced the use of Fuzzy Logic in edge detection. Fuzzy rules provide other benefits, such as the ability to adjust the edge thickness by adding new rules or changing parameters, making it a flexible framework that may be used at any moment to create edge detection processes. The fuzzy edge detection approach has a flexible framework and is less complicated (Mutneja, 2015).

Image Enhancement

Image enhancement is an essential aspect of image processing because it helps to improve picture quality by highlighting important information and eliminating unnecessary information in the image (Qi, 2022). The purpose of image enhancement techniques is to increase the quality and features of an image so that the image's significant information can be retrieved easily (Verma, 2020).

Colour Image Processing

Colour image processing is a large and dynamic discipline concerned with improving the quality and features of digital images (Qi, 2022). A colour image consists of three different but linked channels: red, green, and blue (RGB) (Huang, 2023). Huang proposed quaternion-based colour image processing approaches as a viable alternative to directly describing colour images as vectors or matrices. In colour image processing, quaternion-based algorithms offer various potential advantages. Creating new colour spaces based on quaternion representations might boost the accuracy and efficiency of colour image processing methods (Huang, 2023). Using quaternion representations to undertake data augmentation techniques may increase the resilience and generalisation of colour image dataset-trained models. Furthermore, Quaternion Convolutional Neural Networks (QCNNs) may learn features from colour pictures more effectively than typical CNNs by using quaternion representations.

Image Restoration

Image restoration is a critical component of computer vision that seeks to forecast and replace the pixels of missing pictures in order to generate acceptable visual effects (Liu, 2022). It has a wide range of applications, including special effects creation in cinema and television, picture editing, digital cultural heritage preservation, and virtual reality. Zhai et al.'s research on deep learning-based real-world image restoration provides a comprehensive overview of critical benchmark datasets, image quality assessment methods, and four major categories of deep learning-based image restoration methods, including CNN, GAN, Transformer, and MLP. The paper emphasises the most recent breakthroughs and advances in each network architecture category and provides a comparative examination of representative state-of-the-art picture restoration methods.

2.2 Computer Vision

Computer vision is a branch of Artificial Intelligence (AI) that teaches and allows computers to comprehend the visual environment. Computers can reliably recognise, categorise, and respond to things using digital images and deep learning models (Simplilearn, 2021). Computer vision models strive to mimic biological design in order for systems to accomplish meaningful tasks (Kumar et al., 2021).

Computer vision in AI is concerned with the creation of automated systems capable of interpreting visual input (such as images or motion pictures) in the same way that humans do. The goal of computer vision is to teach computers to analyse and comprehend pictures pixel by pixel (Simplilearn, 2023). This is the cornerstone of the field of computer vision. In terms of technology, computers will attempt to collect visual data, manage it, and analyse the results using sophisticated software programmes.

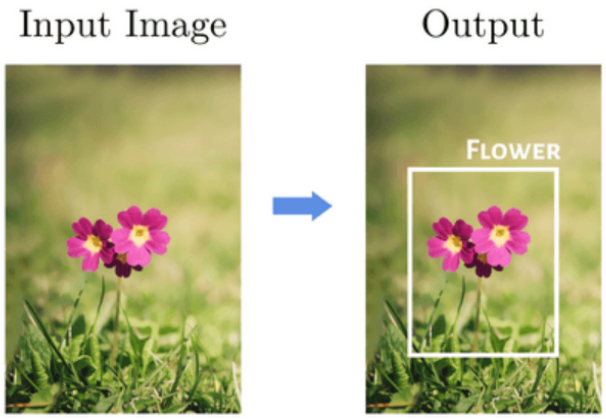


Figure 2.2: Example of Computer Vision (Recognising object from images)

2.2.1 Image Processing and Computer Vision

Image processing is used in the early stages of computer vision models . In Figure 2.3, we use image-processing techniques to improve brightness and contrast in order to see certain writing more clearly. As a result, the performance of the computer vision model that discovers and recognises words in text will improve. In short, an image-processing method enhances the attributes of the input image. Computer vision, on the other hand, will attempt to understand what is displayed in the image or video (Baeldung, 2020).

Figure 2.3: Example of image processing improves the performance of a computer vision model

2.3 Human Detection

Human detection is the task of detecting all instances of human beings present in an image, and it has been most extensively performed by examining all places in the image, at all feasible sizes, and comparing a tiny region at each position with known templates or patterns of humans (Davis, 2009). It is a task of computer vision systems for finding humans in video footage. Human detection is commonly regarded as the initial phase in a video surveillance pipeline, and it can feed into higher-level reasoning modules like action recognition and dynamic scene analysis.

Generally, the detection procedure is divided into two steps: object detection and object classification. Figure 2.4 elegantly outlines the two crucial steps involved (object detection and object classification), along with the corresponding techniques employed for each step. Background subtraction, optical flow, and spatio-temporal filtering might all be used for object detection. Background subtraction is a common object detection approach that seeks to find moving objects by comparing the current frame to a background frame on a pixel-by-pixel or block-by-block basis (Paul, 2013). According to Yuwono's research (2018), the integration of background subtraction with other detection methods yields superior results in terms of accuracy and processing efficiency. By incorporating background subtraction into the detection process, the combined method exhibits enhanced accuracy while minimising processing time.

Upon successful object detection, the detected object seamlessly progresses to the subsequent step, namely object classification. This pivotal stage entails the detection

and recognition of whether the objects are human beings, contingent upon their distinctive features. Within the realm of object classification, three remarkable techniques can be employed: shape-based method, motion-based method, and texture-based method.

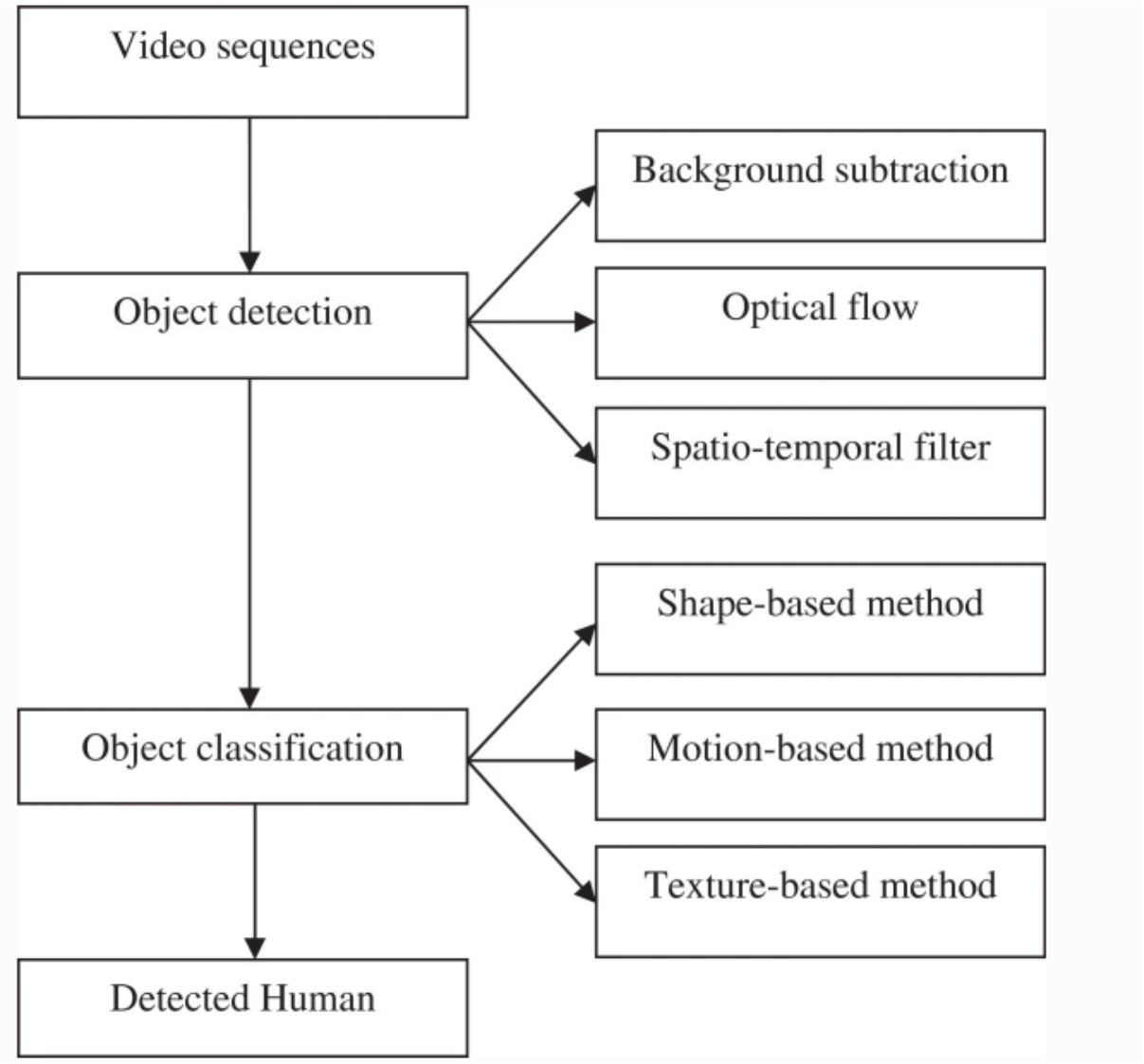


Figure 2.4: Flowchart of human detection

According to Paul (2013), occlusion is the main problem for the 3 object detection techniques. The term "occlusion" refers to a phenomenon in which an item of interest is partially blocked out. For example, when we want to detect a pedestrian crossing a road, but the pedestrian is occluded by a car. In this case, the pedestrian may not be detected and the car which is blocking the pedestrian will act as an occluder. However, Wang et al. (2009) has proposed a human detection method that can handle

partial occlusion, as well as a feature set that blends trilinear interpolated Histograms of Oriented Gradients (HOG) with Local Binary Pattern (LBP) in an integrated image framework.

The research done by Sumit et al. (2021) has discussed the pros and cons of different detection techniques. In the research, it stated that although Viola-Jones and Speeded Up Robust Features (SURF) can identify objects in real time and overcome filtering restrictions, they are still sensitive to light. Besides, Scale-Invariant Feature Transform (SIFT), Bag-of-Words (BoW), and Orthogonal Moments (OM) exist and give additional intriguing features like occlusion and clutter insensitivity, simplicity, and low-order element creation, but they are computationally costly. HOG is yet another approach with some interesting features such as invariance to photometric and geometric changes and illumination, but it suffers from the lack of spatially neighbouring pixel context.

The research suggests that the Deformable Part-based Model (DPM) approach is a feature extraction technique for human detection that outperforms other techniques. DPM's advantages include the ability to manage specific position variations, multiple views, and its application-free nature, which means it performs quite well in real-time applications. However, the DPM approach has several drawbacks. One disadvantage of DPM is that it is computationally costly since it relies on heuristic initialization during the training process to optimise the non-convex cost function. Despite this drawback, the DPM approach is thought to be somewhat superior to other available human detection techniques.

2.4 Human Activity Recognition (HAR)

Human activity recognition is crucial in interaction between people and interpersonal relationships. It is tough to extract since it provides information about a person's identity, personality, and psychological condition. One of the key objects of research in the scientific fields of computer vision and machine learning is the human ability to recognise another person's activity (Vrigkas et al, 2015). Recognising human activities from video sequences or still images is difficult due to issues such as background clutter, partial occlusion, and so on.

2.4.1 HAR methods

In Figure 2.5, the current HAR (Human Activity Recognition) method is depicted. The figure illustrates that HAR methods can be categorised into two types: Unimodal and Multimodal. Each category comprises multiple methods. In this subsection, we will delve into a comprehensive discussion of all these methods.

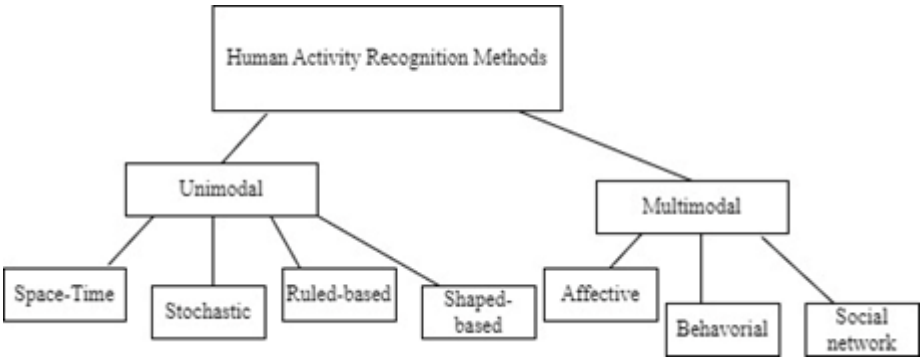


Figure 2.5: HAR methods

Unimodal

Unimodal human activity recognition methods are used for identifying human activities from data of a single modality. Most existing techniques depict human activities as a set of visual characteristics derived from video sequences or still images, and multiple classification algorithms are used to recognise the underlying activity label (Kong et al., 2014). For identifying human activities based on motion characteristics, unimodal techniques are acceptable. However, recognising the underlying class just through motion is a difficult challenge in and of itself (Vrigkas et al, 2015). The key issue is ensuring motion continuity along time when an activity occurs consistently or non-uniformly inside a video stream. In this category there are 4 methods which are Space-Time, Stochastic, Rule-based and Shape-based methods.

Space-Time Method

Space-time methods are concerned with identifying activities based on space-time properties or trajectory matching. They assume a 3D space-time volume activity consisting of the concatenation of 2D spaces in time (Wang, 2013). A set of space-time characteristics or trajectories taken from a video stream is used to describe an activity. These approaches, which are frequently sensitive to noise and occlusion, are not designed to recognise complicated activities (Beddiar et al,, 2020).

Stochastic Method

In recent years, various stochastic techniques such as hidden Markov models (HMMs) and hidden conditional random fields (HCRFs) have been used for human activity recognition. Several researchers have proposed different models and algorithms for this task, employing features like position, velocity, local descriptors, and motion information. Some approaches include modelling human behaviour as a stochastic sequence of actions, using hierarchical models with latent variables, and employing probabilistic models like Markov random fields (MRFs) and conditional random fields (CRFs) (Vrigkas et al, 2015).

There are some challenges and limitations associated with existing probabilistic methods for human activity recognition. One challenge is that these methods can be complex and computationally expensive, often requiring dynamic programming or HMMs with a varying number of parameters (Vrigkas et al, 2015). This complexity can make them impractical for real-time applications. Additionally, HMMs assume that features are conditionally independent, which may not hold true in many cases.

According to Vrigkas et al. (2015), there are trade-offs and considerations when selecting between HMMs and CRFs for different types of human activity recognition tasks. HMMs are better suited for simpler activities, while CRFs provide better generalisation but may be less practical for real-time applications due to their computational requirements.

Rule-based Method

Rule-based methods represent an activity using rules or sets of attributes that characterise an event to identify going on events. Each activity is seen as a set of primitive rules/attributes, allowing the development of a descriptive model for human activity recognition (Vrigkas at al., 2015). They pose certain challenges during the development of these rules and qualities, as well as during the analysis of extended video sequences and the detection of complicated behaviours (Beddiar et al,, 2020).

Shaped-based Method

This method recognises human actions relies on obtaining accurate human body parts from videos. The representation of the human body includes describing body parts in 2D space as rectangular patches and in 3D space as volumetric shapes. Several algorithms focus on action

recognition from still images or videos by considering scene appearance (Vrigkas et al., 2015). These methods include representing actions using histograms of pose primitives, combining actions and human poses together, and introducing robust representations of human pose and appearance. The availability of a large variety of low-cost pose estimation devices makes these approaches particularly practical (Beddiar et al,, 2020). The problem of human pose estimation remains challenging for real-time applications. It is affected by factors such as occlusions, background clutter, and variations in lighting and clothing. Dimensionality reduction, data association, and handling self-occlusions are ongoing research areas to address these challenges.

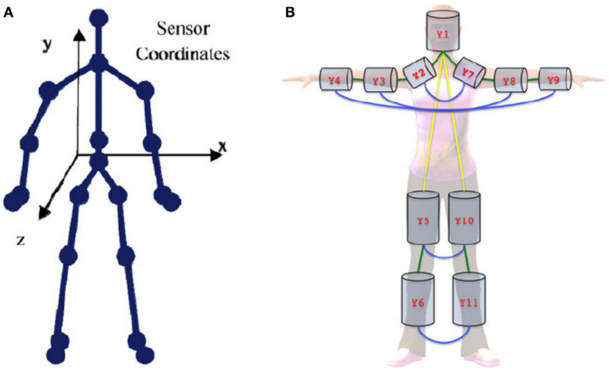


Figure 2.6: Human Body Representations. (A) 2D skeleton model (Theodorakopoulos et al., 2014) and (B) 3D pictorial structure representation (Belagiannis et al., 2014).

Multimodal

To recognise human activities, multimodal human activity recognition (HAR) systems employ data from many modalities (not only from image). Sensors with many modalities might be employed to overcome some of HAR's inherent problems, such as lighting fluctuations, clutter, occlusions, and backdrop variety (Yadav et al., 2021). Sensors such as an RGB-D camera, infrared sensors, thermal cameras, inertial sensors, and so on might be used. It will improve activity recognition performance by providing more and important information from various sources.

Affective Method

Affective computing, which focuses on understanding and modelling human emotions and affective states through various modalities such as facial expressions, gestures, physiological changes, and speech. Accurate annotation of affective data is a key

challenge in this field, as emotions can be expressed differently by different individuals and can be influenced by simultaneous activities and feelings (Vrigkas at al., 2015). Fusion techniques, regression models, and deep learning methods are utilised for combining modalities, continuous prediction, and feature extraction in affective computing. Classifying affective states is difficult due to the semantic gap between low-level features and high-level concepts like emotions. Automatic affective recognition systems are proposed to reduce the effort in selecting proper affective labels and to better assess human emotions.

Behavioural Method

Recognising human behaviours from video sequences is a difficult challenge for the computer vision community (Candamo et al., 2009). A behaviour recognition system may provide information about a person's personality and psychological condition, and its applications range from video surveillance to human-computer interaction. Behavioural approaches seek to identify behavioural characteristics such as emotions, mood, and so on. These approaches use complicated classification models to recognise complex human actions, making it challenging to specify emotional qualities (Beddiar et al,, 2020). Various methods have been proposed for behaviour recognition, including integration of audio-visual features, probabilistic approaches combining facial expressions and audio information, real-time emotion recognition systems, and human activity recognition using auditory information (Vrigkas at al., 2015). The selection of proper features for behaviour recognition is a trial-and-error process, and the combination of visual features with informative features related to human emotions and psychology is necessary.

Social Networking Network

Social interactions are an important part of daily life, and they involve adapting behaviour according to the group of people present. Human behaviour recognition involves understanding behaviours in single-person, multi-person, and object-person interactions. Researchers have studied social interactions in various contexts, including social events, sports, mice behaviour, and human group activities. Social interactions can be represented using graphs, contextual information, and multimodal features. Deep learning methods have had a significant impact on human activity recognition,

including emotion recognition, group activity recognition, event detection, and event recognition from images and videos (Vrigkas at al., 2015). Challenges in modelling social interactions include occlusions, interacting motion patterns, varying number of participants, and complex event recognition (Beddiar et al,, 2020). Contextual information alone may not capture the full understanding of activities and poses during social interactions. Recognizing social interactions with a dynamically changing number of participants is more complex and challenging.

2.4.2 Application

Sports

Human Activity Recognition (HAR) in sports involves using technology to detect and analyse players' movements and actions during training or matches. It uses sensors and algorithms to track players, recognize their actions, and provide statistical analysis (Host & Ivašić-Kos, 2022). HAR helps monitor player performance, compare different actions, evaluate team activities, and generate objective data for performance analysis and training optimization. It aids in understanding players' skills, tactics, and physical abilities, assisting coaches and athletes in making informed decisions and improving overall performance.

Video Surveillance

Human Activity Recognition (HAR) has emerged as a promising technology in human-computer interaction, computer vision, and mass surveillance. The application of it in public places such as railway stations, ATM machines, schools, colleges, and traffic signals enables cost-effective surveillance by detecting and recognising human activity from video footage (Urankar et al.,2020). HAR has the ability to increase security, monitor capabilities, and automate analysis, all of which contribute to safer public places. As HAR technology advances, it has the potential to revolutionise public safety by providing real-time surveillance, early identification of suspect behaviour, and effective deployment of security resources. Furthermore, the insights gathered from HAR data may help in the creation of proactive security policies for diverse public areas, as well as enhance evidence-based decision-making and resource optimisation.

Healthcare

In the context of smart healthcare, HAR is critical in monitoring and measuring patients' movements, assisting in the diagnosis and treatment of a variety of illnesses (Islam at al. 2023). It allows healthcare providers to analyse activity patterns, diagnose irregularities, and track progress throughout the rehabilitation process. HAR may assist with personalised treatment planning and enhance patient outcomes by giving objective data on patients' activity levels and movements (Ong & Bee, 2014). Furthermore, HAR has the ability to facilitate remote patient monitoring, allowing healthcare personnel to follow and intervene remotely based on real-time activity data, improving the accessibility and efficiency of healthcare services.

2.5 Open Pose

OpenPose is an open-source library built in C++ with Python API support for real-time multi-person keypoint detection and multi-threading. The Carnegie Mellon University Perceptual Computing Lab created it, and it may be used to assess human postures in photos or movies. It uses a non-parametric representation to recognise key points on human bodies, hands, faces, and feet (a total of 135 keypoints) on single images (Kim et al., 2021).

OpenPose works by using a deep learning-based approach to detect keypoints on the human body. The input image is first passed through a convolutional neural network (CNN) called VGG net model to generate a set of confidence maps for each keypoint. To create the confidence map of given input, VGG net model uses the first 10 layers (Badave & Kuber, 2021). These confidence maps represent the probability of a keypoint being present at each pixel location in the image. The confidence maps are then processed to extract the locations of the keypoints.

The identified keypoints are parsed into Part Confidence Maps (PCM) and Part Affinity Fields (PAFs), with PCM obtaining the key points of the human body and PAFs predicting the direction of bone points. After obtaining the key point and the key point vector, the correlation between the two key points is calculated by integrating the point product between the two key point connecting vectors and the PAFs vectors of each

pixel on the two key point connecting lines (Bulat & Tzimiropoulos, 2016). Finally, the greedy analysis method encodes the main points of the human body.

2.5.1 Advantage

One of the advantages of OpenPose is that it is compatible with a variety of platforms, including Ubuntu, Windows, Mac OS X, and embedded systems (for example, Nvidia Tegra TX2). It also supports a variety of hardware including CUDA GPUs, OpenCL GPUs, and CPU-only devices. He/she (the user) may choose between images, video, webcam, and IP camera streaming as an input. He may also choose whether to show or store the results, activate or disable each detector (body, foot, face, and hand), enable pixel coordinate normalisation, regulate the number of GPUs used, skip frames for quicker processing, and so on (Martinez, 2019). OpenPose's inference time beats all state-of-the-art approaches while maintaining high-quality findings. The scientific community has previously utilised OpenPose for numerous vision and robotics applications, such as person re-identification.

When compared to Kinect v2, OpenPose may offer a greater number of joints from the face and foot, as well as more steady tracking capability when tracking is occluded or non-frontal (Kim et al., 2021). This means that OpenPose may be able to provide more precise and accurate monitoring of body motions even when the individual is not facing the camera or when parts of the body are concealed. As a result, OpenPose may be a preferable solution for applications requiring accurate tracking of body movements.

OpenPose employs a bottom-up multi-person posture estimation method (Jo & Kim, 2022). As a result, it is more robust to the amount of people. First, all key points in a particular image are recognised, and then they are categorised by human instances. Because it only detects key points once and does not redo pose estimates for each individual, this strategy is generally faster than the top down approach (Osokin, 2018).

2.5.2 Disadvantage

One of the disadvantages of OpenPose is that its low-resolution outputs limit the degree of information in keypoint estimations. This makes OpenPose less appropriate for precision-demanding applications like elite sports and medical examinations, which all rely on a high degree of precision in movement kinematics assessment. Furthermore, OpenPose is regarded as exceedingly inefficient since it spends 160 billion floating-point operations (GFLOPs) each inference. (Zhang et al., 2021).

It is unable to identify and estimate joints that are not visible to the camera. For example, with a leftward gait, the camera cannot accurately collect and estimate the right arm posture. Even when the joints are visible, OpenPose cannot always estimate them correctly, resulting in missing data. The missing data must be compensated by utilising acceptable data in this scenario (Abe et al., 2021).

When the ground truth example contains non-typical postures and upside down instances, OpenPose has difficulty predicting pose. Due to the overlapping PAFs that cause the greedy multi-person parsing to fail in extremely packed photos with individuals overlapping, the technique tends to mix annotations from different persons while missing others (GeeksforGeeksb, n.d.).

2.6 MediaPipe

MediaPipe is an open-source perception pipeline framework. It was created to solve the issues of developing apps that observe their surroundings. Developers may use MediaPipe to create prototypes by merging existing perceptual components, then progress them to polished cross-platform apps while measuring system performance and resource consumption on target platforms. These characteristics allow developers to concentrate on algorithm or model development while using MediaPipe as a platform for iteratively developing their application with results that are repeatable across several devices and platforms (Lugaresi et al., 2019). MediaPipe has been utilised in a range of applications, including human posture detection and recognition in real time. Researchers in one study utilised MediaPipe Holistic, which contains position, face, and hand landmark detection models, to parse frames collected from

real-time device feed using OpenCV. The obtained landmarks were exported as coordinates to a CSV file and used to train a bespoke multi-class classification model to identify and categorise custom body language postures (Singh et al., 2022).

2.6.1 Advantage

The speed of MediaPipe is one of its advantages. It achieves its speed through the use of GPU acceleration and multi-threading. Such development procedures are often challenging, but MediaPipe takes over and accomplishes them for you as long as you adhere to appropriate graph-making practices. Multi-threading and GPU acceleration enable modern phones to run away with frames, frequently at FPS levels too high to see with the naked eye. Another benefit of MediaPipe is its versatility and reuse. Because MediaPipe uses graphs, subgraphs, and calculators, the work of one project may simply be translated to the work of another. When combined with side packages, you may actually fine-tune the characteristics of each calculator to suit various tasks. MediaPipe currently has a wealth of "example calculators" that you may freely use, such as multi-platform renderers, multi-platform TensorFlow Lite, and pre-made neural networks (Alavi. A ,2019).

2.7 Deep Learning for Human Activity Recognition

Deep learning is a sort of machine learning that uses numerous layers of processing to extract progressively higher level characteristics from input. It is based on artificial neural networks and enables systems to cluster data and produce extremely accurate predictions (IBM, n.d.). Deep learning is a cutting-edge method that excels in autonomously discerning both spatial and temporal complexities within raw sensor data, removing the need for time-consuming human feature engineering. This enables the algorithm to recognise intricate patterns that human-crafted features could have missed. The result is remarkable: a significant improvement in the precision and durability of human activity recognition, especially when dealing with multidimensional and diversified activities that frequently display nuanced changes. Deep learning emerges as a potent technique that enhances the accuracy and

dependability of activity recognition systems by smoothly responding to the particular problems given by such activities (Zhang et. al.).

Deep learning offers several advantages for recognising human activities, but it also has certain drawbacks. Deep learning requires a significant quantity of labelled data to train the models, which is one of the obstacles. For some activities, this data may be limited, noisy, or uneven, making it challenging to build accurate and robust models. Another issue is that deep learning models are frequently computationally costly and need high-end gear and software. This can limit their deployment and scalability, particularly in limited-resource contexts. Finally, deep learning models are sometimes difficult to comprehend and explain, which might pose ethical and societal concerns for human activity recognition applications. It may be difficult, for example, to understand why a model produced a specific forecast or to guarantee that the model is not biassed against certain groups of individuals (Ullah et. al., 2021).

2.7.1 CNN-LTSM network

Several research on the usage of CNN-LSTM (Convolutional Neural Network-Long Short-Term Memory) architecture for human activity recognition (HAR) have been conducted. One such study suggests a convolutional neural network-long short-term memory network (CNN-LSTM)-based holistic deep learning-based activity detection architecture. This CNN-LSTM technique not only increases the predictability of human actions from raw data, but it also decreases model complexity while removing the requirement for significant feature engineering. The CNN-LSTM network is dense in both space and time. On the iSPL dataset, an internal dataset, the suggested model achieves 99% accuracy, while on the UCI HAR public dataset, it achieves 92% accuracy (Mutegeki & Han, 2020).

2.7.2 LTSM network

Long Short-Term Memory (LSTM) has had a profound impact on the disciplines of machine learning and neurocomputing because of its ability to solve the exploding gradient problem, which arises frequently while training recurrent or very deep neural networks. The architecture of LSTM will be shown in Figure 2.7.With over 4 billion LSTM-based translations performed daily as of 2017 (Van Houdt et al., 2020), it has

been widely adopted by tech giants like Google for speech recognition and machine translations, Amazon for enhancing Alexa's functions, and Facebook for translations. The game industry has also made use of LSTM's learning capabilities; AlphaStar, an AI created by Google's Deepmind, is capable of playing Starcraft II (Van Houdt et al., 2020). In addition, LSTM has developed into state-of-the-art machine learning methods and is applied in a number of fields, such as health monitoring, supply and demand forecasting, and finance (Ab Kader et al., 2022). Various strategies, including hybrid, ensemble, and hyperparameter optimization, have been employed to improve the long short-term memory (LSTM) methodology for time series analysis and forecasting (Ab Kader et al., 2022).

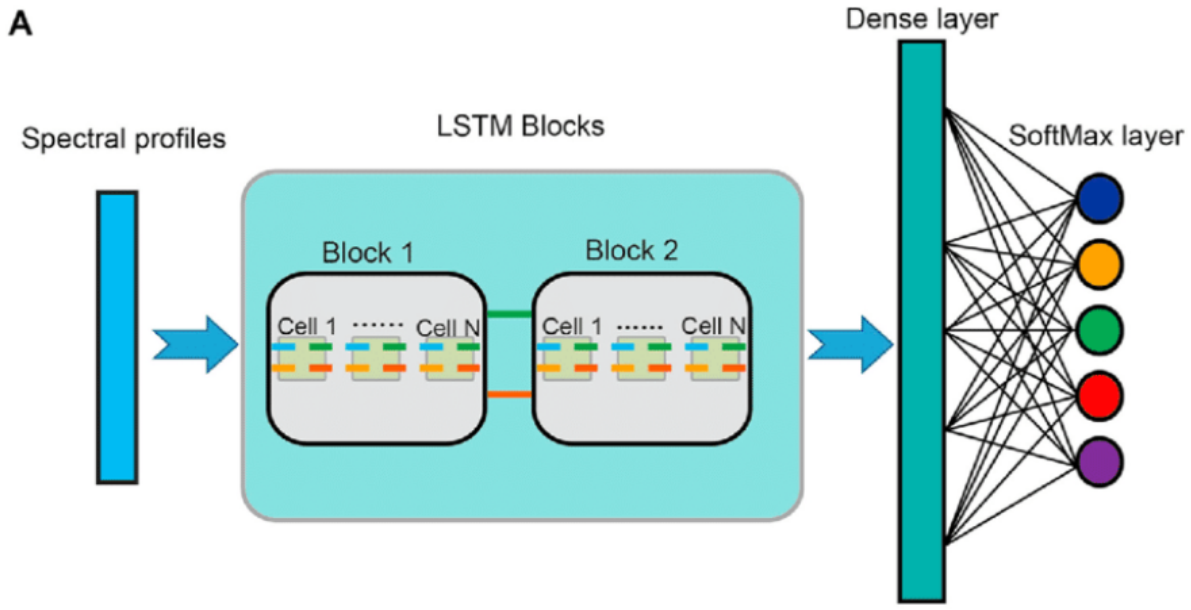


Figure 2.7: Architecture of LSTM model (Kang et al., 2021)

2.8 Comparison of tools

Table 2.1: Comparison of MediaPipe, OpenPipe, PaddleDetection, ActionAI and OpenCV

Tools	**MediaPipe**	**OpenPose**	**PaddleDetection**	**ActionAI**	**OpenCV**
Year Founded	2019	2016	2020	2018	1999
Founding Company	Google	Carnegie Mellon University	PaddlePaddle	ActionAI	Intel
Licence Pricing	Open Source	Open Source	Open Source	Open Source	Open Source
Supported Features	• Real-time 2D/3D pose estimation • Hand tracking	• Real-time multi-person keypoint detection	• Object detection • Instance segmentation	• Action recognition • Gesture recognition	• Image and video processing • Object

	• Face detection • Object detection • Augmented reality • Audio and video processing • Gesture recognition	• Body, face, and hand pose estimation • Pose tracking • Human action recognition	• Semantic segmentation • Face recognition • Human detection and tracking • OCR • Image classification	• Behaviour analysis • Human pose estimation • Object detection and tracking • Event detection	detection and tracking • Feature extraction • Camera calibration • Machine learning algorithms • GUI development
Limitations	• Limited support for non-Google platforms • Requires deep learning expertise for customization	• GPU-intensive, requires powerful hardware • Limited to human pose estimation and related tasks	• Requires PaddlePaddle framework knowledge • Limited pre-trained models compared to other frameworks	• Proprietary license, pricing details need to be obtained from ActionAI • Limited information available publicly	• Limited support for deep learning compared to specialized frameworks • Performance may be lower for complex tasks
Common Application	Augmented reality, gesture recognition,	Human action recognition, sports	Object detection, image classification,	Video surveillance, behaviour analysis,	Image/video processing, robotics,

	real-time image processing	analysis, animation	document analysis	activity recognition	augmented reality

Chapter 3

Methodology

3 Methodology

This chapter will provide a description of the dataset that will be used and the methodology that will be implemented in this project.

3.1 Dataset

UCF50 is an action recognition data set containing 50 action categories made up of real YouTube videos. This data set is a supplement to the YouTube Action data collection (UCF11), which contains 11 action types. The vast majority of accessible action recognition data sets are not realistic and have been produced by actors. The major goal of this dataset is to give the computer vision community an action recognition data set composed of realistic films acquired from YouTube. Due to enormous changes in camera motion, item look and posture, object scale, perspective, complex backdrop, lighting conditions, and so on, the dataset is quite difficult. The videos are divided into 25 groups for each of the 50 categories, with each group containing more than four action clips. The video clips in the same group may have certain similarities, such as the same person, a similar setting, a similar point of view, and so on.

The action included in this dataset are:

Baseball Pitch, Basketball Shooting, Bench Press, Biking, Biking, Billiards Shot,Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, Playing Guitar, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Lunges, Military Parade, Mixing Batter, Nunchucks, Playing Piano, Pizza Tossing, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Skateboarding, Skiing, Skijet, Soccer Juggling, Swing, Playing Tabla, TaiChi, Tennis Swing, Trampoline Jumping, Playing Violin, Volleyball Spiking, Walking with a dog, and Yo Yo.

Few Sample of the dataset will be shown as below:

Figure 3.1: Sample video of BaseballPitch from UCF50

Figure 3.2: Sample video of Basketball from UCF50

Figure 3.3: Sample video of BenchPress from UCF50

Figure 3.4: Sample video of Biking from UCF50

Figure 3.5: Sample video of Billiards from UCF50

3.2 Algorithm Used

3.2.1 Pose Extraction

In pose extraction, the selected tool used is MediaPipe. MediaPipe is an open-source platform that enables developers to create complex pipelines for processing perceptual data such as video and audio streams. It includes pre-trained models for identifying human positions, faces, and hands, as well as a variety of tools for developing machine learning applications. Data may be processed in real-time via MediaPipe. This means it can swiftly analyse video streams and offer findings, making it suited for applications like public security and gaming. MediaPipe also offers a set of pre-trained models for detecting human positions, faces, and hands. These models may be integrated to give a comprehensive knowledge of human activity that includes body motions, face expressions, and hand gestures. The MediaPipe posture estimation approach is very good for recognising human activities. To categorise motions, it can identify key body areas and analyse posture. Body posture markers are produced by the model in both image coordinates and 3-dimensional world coordinates (Google

Developers, 2021). This data may be utilised to predict human body position and emotions as well as comprehend the relationship between different body components. In conclusion, MediaPipe offers a variety of tools and pre-trained models that make it well-suited for human activity recognition. The ability to handle data in real-time and deliver precise findings makes it a valuable tool for developers in this industry. Its full grasp of human behaviour is provided by its holistic approach to interpreting human activity, which includes body motions, face expressions, and hand gestures.

3.2.2 Deep Learning Model

The algorithm chosen in this project as the deep learning model is the LSTM network. Since LSTM can automatically learn complex features from raw accelerometer signals and distinguish between common human activities, it is especially well-suited for Human Activity Recognition (HAR) applications (Nabriya, 2021). The LSTM model exhibits higher activity detection capabilities and robustness, as it can extract activity features in an adaptive manner (Xia et al., 2020). It also optimises efficiency in real-time activity recognition by doing away with the requirement for extensive feature engineering overhead and data pre-processing (Choudhury & Soni, 2023). In comparison to standard machine learning techniques, deep learning is more efficient because it can be used more simply and does not require domain knowledge (Zakaria Benhaili et al., 2022). Moreover, the LSTM model has the capacity to eliminate superfluous data, offering superior comprehensibility and harmonising with common sense (Zheng et al., 2019).

3.3 Project Flow

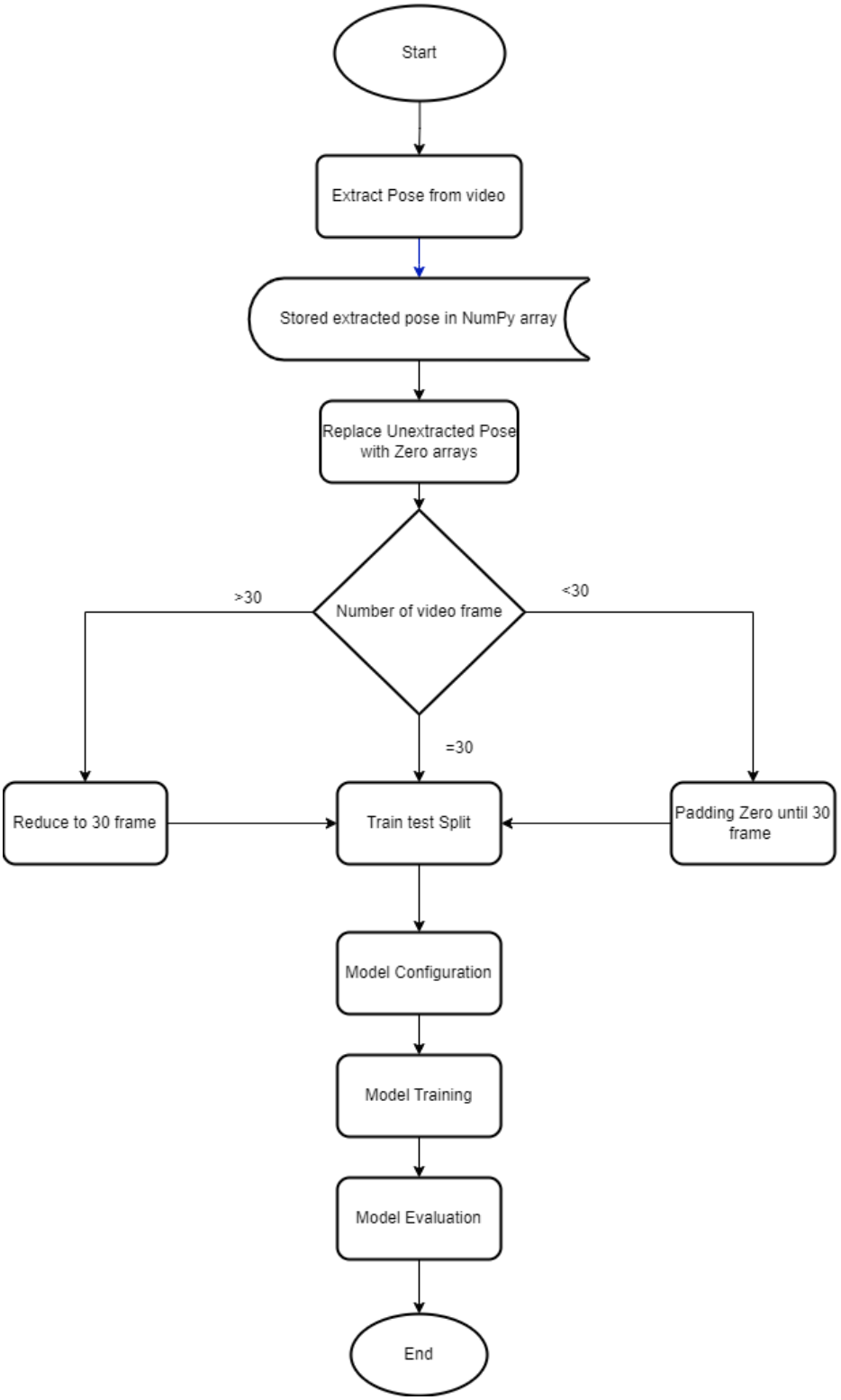


Figure 3.1: Flowchart of project flow.

Figure 3.2 shows that a multi-step technique is used to improve the model's accuracy and prediction capabilities while refining the training process for a deep learning model using the UCF50 dataset, which consists of videos in.avi format. The first step is to extract critical human pose information from the videos, as straight use of the video data may not produce ideal results. This collected posture data, together with their related labels, is organised and saved in the NumPy array..

Following pose extraction, a careful data cleaning operation is carried out to remove any noisy or inconsistent data. Videos lacking extracted pose data are supplemented with empty placeholders. To ensure consistency across videos with varying frame rates, we standardise their sequences to 30 frames. This resampling process normalises the data, allowing for more accurate comparisons and analysis. This essential stage guarantees that the dataset retains a high degree of quality and conforms to the specified format. Now that the dataset has been improved, the attention switches to setting and fine-tuning the deep learning model. This stage entails carefully experimenting with various model parameter combinations to find the ideal configuration that best matches with the dataset's subtleties.

The model is ready for training after determining the most effective parameter settings. This stage comprises using the precisely produced posture data to train the model, allowing it to recognise detailed patterns within the dataset. Performance metrics are thoroughly maintained and assessed during this procedure. The fundamental assessment parameter for assessing the model's efficacy is accuracy, which provides a complete measure of its predictive power. In essence, this complete technique consists of a series of refined steps, beginning with the extraction of salient pose information and data purification and ending with a thorough assessment of the model's correctness. By painstakingly carrying out each step, the final deep learning model is primed to outperform the hard UCF50 dataset.

3.4 Implementation

Figure 3.2: Overview of implementation flow.

Figure 3.3 shows the flow of implementation. Within the scope of project implementation, users have the option of uploading a video or seamlessly utilising their device's built-in camera for activity recognition. The video footage, whether shot in real time or uploaded, is meticulously preprocessed. This preliminary stage serves as the foundation for the future analysis. The recognition of the actions represented in the video is the climax of this process. This identified action is then subtly placed onto the film, effectively labelling and boosting its contextual significance.

Chapter 4

Results and Discussion

4 Results and Discussion

This chapter will discuss the outcomes of our project and engage in an in-depth analysis to comprehend the reasons behind the obtained results. This section allows us to evaluate the significance and implications of our findings, facilitating a comprehensive understanding of the project's outcomes.

4.1 Test Plan

4.1.1 Test Cases

In accordance with established testing procedures, ten activities were randomly selected from the dataset's inventory of fifty for system evaluation.The test data will comprise a combination of real-time activity detection from webcam feeds and carefully selected Youtube videos.

Table 4.1: Test Cases

ID	**Activity**	**Method**
1	Push Ups	Real Time
2	Lunges	Real Time
3	Jumping Jack	Real Time
4	Breast and Stroke	Video
5	Playing Piano	Video
6	Taichi	Video
7	Clean and Jerk	Video
8	Golf Swing	Video
9	Fencing	Video
10	Basketball	Video

4.1.2 Test Environment

Table 4.2: Hardware Environment

Hardware	**Description**
CPU	Intel I7 - 13700HX
GPU	RTX 4060 8GB
RAM	16 GB
HDD	-
SSD	1 TB
Monitoring	-
Keyboard	-
Modem	-

Table 4.3: Software Environment

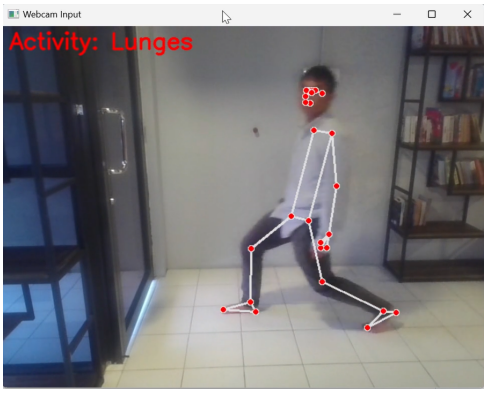
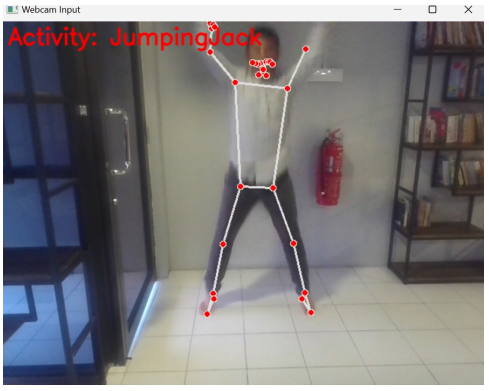
Software	**Description**
Operating System	Windows
IDE	PyCharm

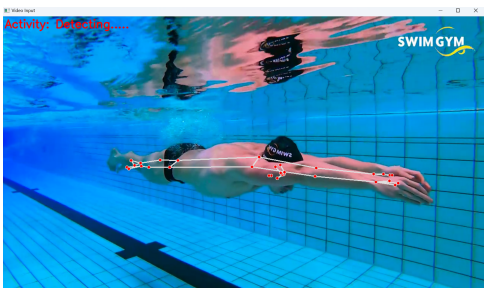
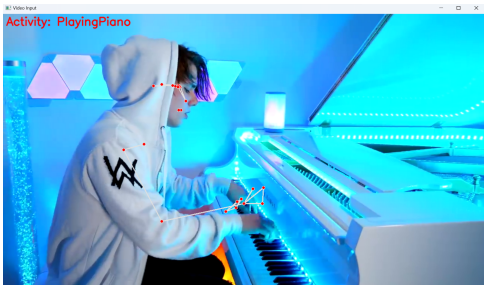
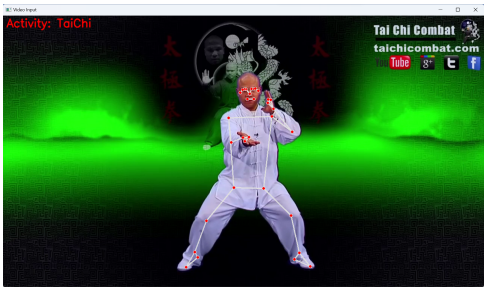
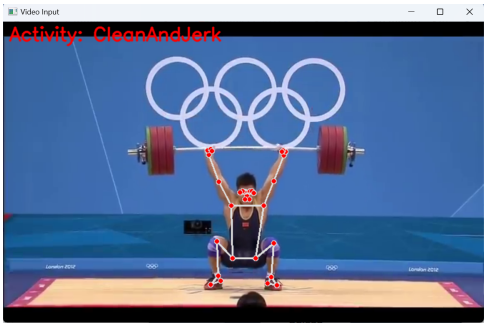
The successful operation of the system necessitates not only compatible hardware and software environments but also the specification of the uploaded video format. Fortunately, the system readily accepts a variety of formats, including .avi and .mp4 files, offering users a high degree of flexibility.

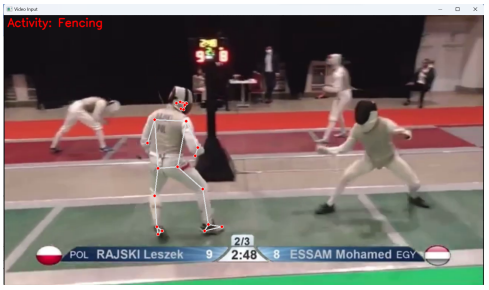
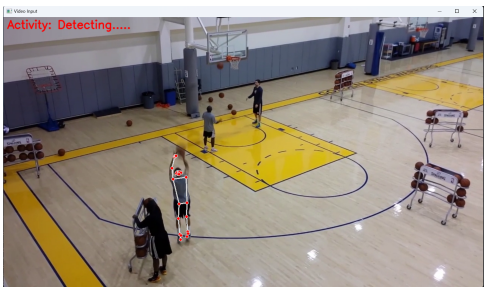
4.2 Results

4.2.1 Test Cases Result

Table 4.4: Result of Test Cases

ID	Activity	Results	Pass/Fail
1	Push Ups	Real Time 	Pass
2	Lunges	Real Time 	Pass
3	Jumping Jack	Real Time 	Pass

4	Breast and Stroke		Fail
5	Playing Piano		Pass
6	Taichi		Pass
7	Clean and Jerk		Pass
8	Golf Swing		Pass

9	Fencing		Pass
10	Basketball		Fail

Based on the data presented in Table 4.4, 80% (8/10) of the randomly selected test cases were successful. Nevertheless, it is important to note that this limited sample does not provide sufficient evidence to draw conclusions regarding the overall system performance.

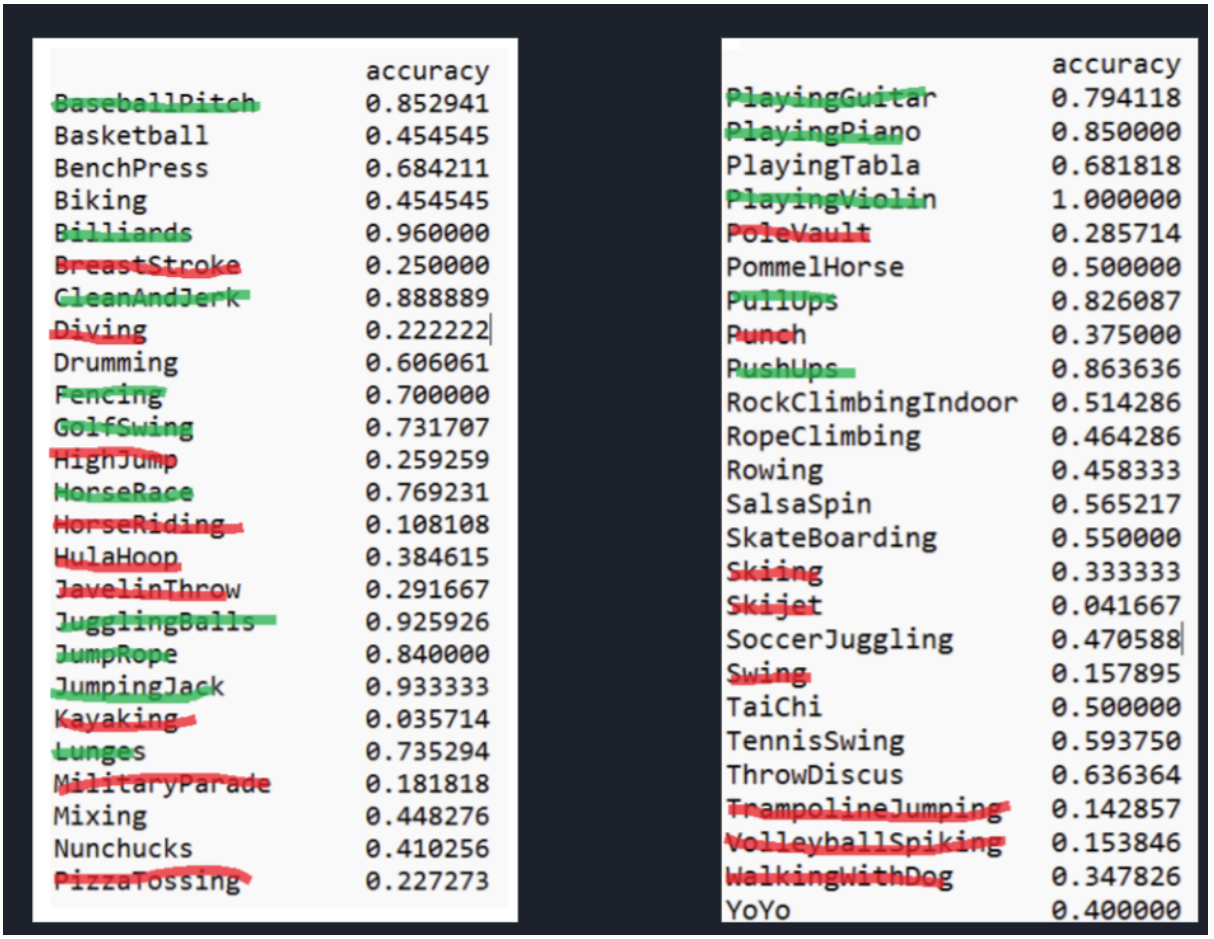
4.2.2 Model Performance

```
Epoch 30/50
84/84 [==============================] - 16s 190ms/step - loss: 0.5477 - accuracy: 0.8477 - val_loss: 2.1886 - val_accuracy: 0.4996
Epoch 31/50
84/84 [==============================] - 15s 184ms/step - loss: 0.5123 - accuracy: 0.8643 - val_loss: 2.2620 - val_accuracy: 0.5153
42/42 [==============================] - 3s 62ms/step - loss: 2.2620 - accuracy: 0.5153
Test Accuracy: 0.515332818031311
```

Figure 4.1: Performance of the LSTM model.

While the preliminary evaluation through the test cases offered valuable initial data points, a comprehensive assessment of the system's capabilities necessitates analysing its performance across the full spectrum of 50 activities. Figure 4.1, showcasing the performance of the LSTM model, with its 51.53% overall accuracy in recognizing all 50 activities, the model demonstrates a moderate potential for real-world applications.

4.2.3 Accuracy of Each Activities



	accuracy
BaseballPitch	0.852941
Basketball	0.454545
BenchPress	0.684211
Biking	0.454545
Billiards	0.960000
BreastStroke	0.250000
CleanAndJerk	0.888889
Diving	0.222222
Drumming	0.606061
Fencing	0.700000
GolfSwing	0.731707
HighJump	0.259259
HorseRace	0.769231
HorseRiding	0.108108
HulaHoop	0.384615
JavelinThrow	0.291667
JugglingBalls	0.925926
JumpRope	0.840000
JumpingJack	0.933333
Kayaking	0.035714
Lunges	0.735294
MilitaryParade	0.181818
Mixing	0.448276
Nunchucks	0.410256
PizzaTossing	0.227273

	accuracy
PlayingGuitar	0.794118
PlayingPiano	0.850000
PlayingTabla	0.681818
PlayingViolin	1.000000
PoleVault	0.285714
PommelHorse	0.500000
PullUps	0.826087
Punch	0.375000
PushUps	0.863636
RockClimbingIndoor	0.514286
RopeClimbing	0.464286
Rowing	0.458333
SalsaSpin	0.565217
SkateBoarding	0.550000
Skiing	0.333333
Skijet	0.041667
SoccerJuggling	0.470588
Swing	0.157895
TaiChi	0.500000
TennisSwing	0.593750
ThrowDiscus	0.636364
TrampolineJumping	0.142857
VolleyballSpiking	0.153846
WalkingWithDog	0.347826
YoYo	0.400000

Figure 4.2: Performance of the LSTM model on each activity.

Figure 4.2 presents a detailed evaluation of the model's accuracy for each individual activity within the dataset of 50. The performance varies considerably, with Skijet at 4.17% and PlayingViolin at 100% representing the respective extremes. Of particular interest is the identification of 15 activities (highlighted in green) that meet the target accuracy of 70% outlined in Objective 1, contrasting with 17 activities (highlighted in red) whose performance falls below 40% and requires further investigation.

4.3 Discussions

Further investigation revealed two significant limitations in the current system: MediaPipe's inability to accurately detect occluded individuals and multi-person interactions.

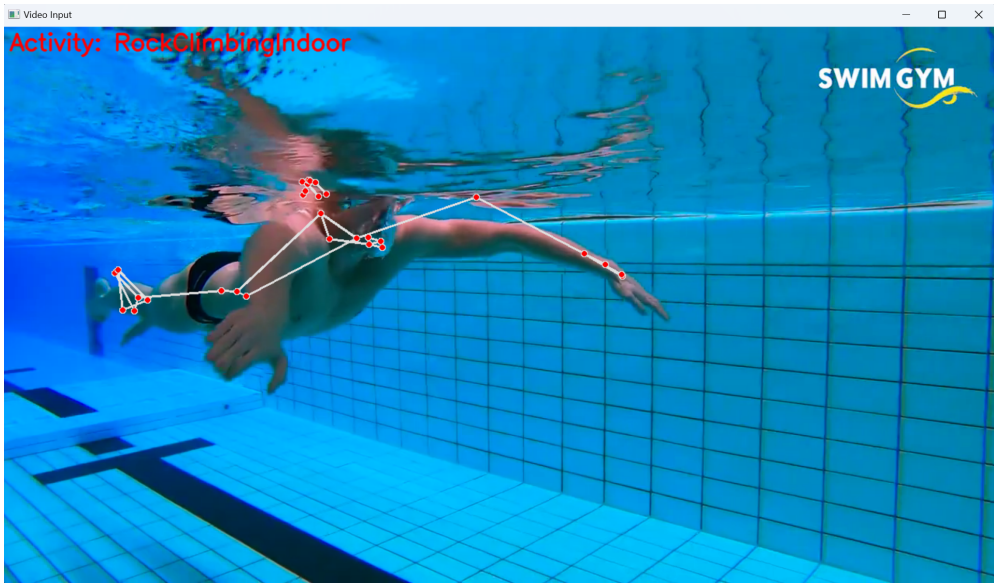


Figure 4.3: The swimmer is occluded by his own shadow.

Figure 4.4: The golfer is occluded by the tiger.

Figure 4.5: This basketball player is occluded by the basketball rack.

Figures 4.3, 4.4, and 4.5 illustrate the significant impact of occlusion on the system's performance. MediaPipe's inability to accurately estimate pose in occluded scenarios leads to the extraction of incorrect pose data. This data, if used for model training, has the potential to mislead the model and negatively affect its learning process.

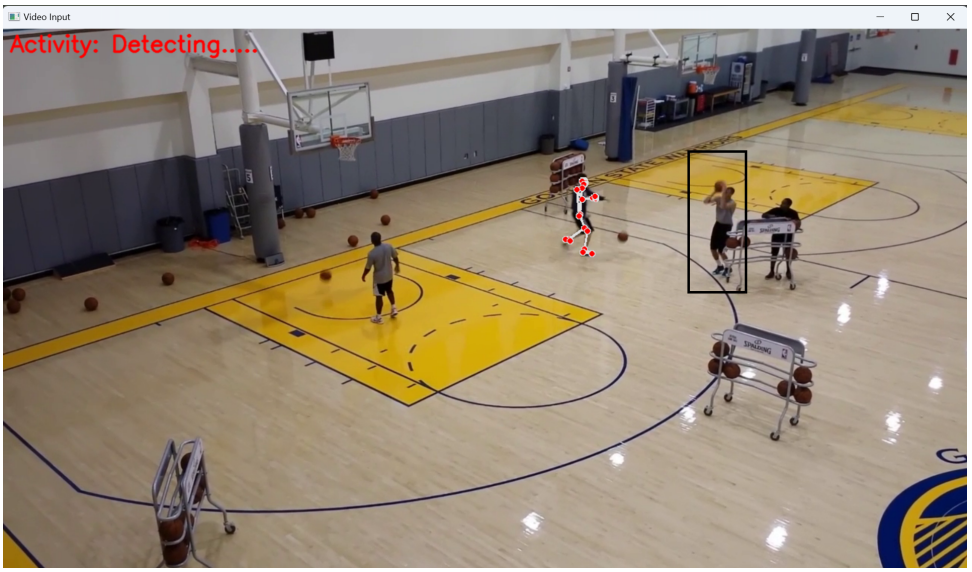


Figure 4.6: This basketball player is not targeted.

Figure 4.7: No person is detected when there are too many people.

The analysis presented in Figures 4.6 and 4.7 highlights a key challenge: MediaPipe's limitations in handling multi-person scenarios. When faced with a large number of individuals, MediaPipe's current architecture fails to perform multi-person detection, opting instead to randomly select a single individual. This random selection process introduces a significant risk of misidentifying the target person. Additionally, Figure 4.7 demonstrates that MediaPipe completely shuts down pose detection in crowded environments, further hindering its effectiveness.

label	Successful	Unsuccessful	total number of data
BaseballPitch	104	46	150
Basketball	63	74	137
BenchPress	33	127	160
Biking	45	100	145
Billiards	64	86	150
BreastStroke	4	97	101
CleanAndJerk	60	52	112
Diving	16	137	153
Drumming	33	128	161
Fencing	65	46	111
GolfSwing	124	18	142
HighJump	67	56	123
HorseRace	8	119	127
HorseRiding	47	150	197
HulaHoop	99	26	125
JavelinThrow	43	74	117
JugglingBalls	87	35	122
JumpRope	141	7	148
JumpingJack	96	27	123
Kayaking	27	130	157
Lunges	85	56	141
MilitaryParade	12	115	127
Mixing	11	130	141

label	Successful	Unsuccessful	total number of data
Nunchucks	112	38	150
PizzaTossing	45	69	114
PlayingGuitar	113	47	160
PlayingPiano	70	35	105
PlayingTabla	86	38	124
PlayingViolin	39	61	100
PoleVault	34	126	160
PommelHorse	40	83	123
PullUps	73	47	120
Punch	66	94	160
PushUps	77	29	106
RockClimbingIndoor	53	95	148
RopeClimbing	48	82	130
Rowing	18	119	137
SalsaSpin	97	36	133
SkateBoarding	47	73	120
Skiing	38	106	144
Skijet	16	84	100
SoccerJuggling	120	36	156
Swing	37	100	137
TaiChi	94	6	100
TennisSwing	112	55	167
ThrowDiscus	93	38	131
TrampolineJumping	46	73	119
VolleyballSpiking	57	59	116
WalkingWithDog	64	59	123
YoYo	89	39	128
Total	3118	3563	6681

Figure 4.8: Number of successful extracted pose from each action for training data

Analysis of Figures 4.2 and 4.8 reveals a critical limitation of the pose estimation tool, MediaPipe. Occlusion and multi-person scenarios significantly impact its effectiveness, often resulting in inaccurate pose extraction during training. This is evident in the red-highlighted actions in Figure 4.8, where less than 30 extracted poses per action lead to poor system recognition. Furthermore, feeding such unreliable data into the model can potentially introduce biases and hinder its ability to accurately recognize other actions.

Furthermore, the system's performance may be compromised by data outside its training set. Different individuals can exhibit variations in behaviour even for the same actions. Consequently, actions with different behaviour which is not included in the training data may be misidentified by the system. This is particularly evident for some activities in Figure 4.8, where limited training data hinders accurate recognition.

While the overall performance of the human activity recognition system remains unsatisfactory, as evidenced by both the test case results and Figure 4.2, it demonstrates some success in recognizing the 15 activities highlighted in Figure 4.2. This can be attributed to the system's inherent limitations: its single-person detection capability and susceptibility to occlusion. The training videos for these 15 activities likely featured only one individual and minimal occlusion, allowing for accurate pose extraction and consequently, effective recognition during testing.

Chapter 5

Conclusion

5 Conclusion

This chapter will draw the overall conclusions based on our project's outcomes and discuss the findings in relation to the initial objectives. This concluding section provides a summary of the entire thesis and highlights the key contributions and implications of our research.

5.1 Achievement

The developed Human Activity Recognition (HAR) system showcases its potential by recognizing human activities from both real-time webcam input and user-uploaded videos with an overall accuracy of 51.53%. While achieving the first objective of a 70% accuracy threshold proved elusive, the system excels in fulfilling our third objective, delivering real-time action recognition. Furthermore, all test case videos, except for "CleanAndJerk," met the minimum 720p resolution requirement and were successfully processed by the system, thus achieving the second objective. In conclusion, while the first objective remains unfulfilled, the HAR system demonstrably accomplishes two out of the three initial goals.

5.2 Contribution

The HAR system developed in this project demonstrates substantial promise despite not reaching the initial 70% accuracy goal. With an overall accuracy of 51.53%, it showcases its potential by recognizing human activities from both live webcam input and uploaded videos, achieving near-real-time performance that shines in applications like healthcare monitoring and sports analysis. Furthermore, the system successfully processed nearly all test videos, exceeding the minimum resolution requirement and solidifying its robustness. While the quest for higher accuracy continues, this project stands tall as a significant leap forward in real-time HAR technology, offering a multi-source approach that expands its potential impact.The research extends the frontiers of HAR capabilities by tackling the issues connected with the video data quality, algorithm selection, and real-time processing. This not only improves the accuracy and efficacy of existing HAR systems, but also prepares the way for future advancements and applications in a variety of fields.

5.3 Limitation

While achieving two out of three objectives is commendable, the system's current capabilities are not without limitations. Firstly, its recognition prowess remains confined to specific actions, resulting in an overall accuracy below our initial target. This limited scope can be attributed to the system's detection method, which falters when faced with challenges like occlusion and multi-person scenarios. Occlusion presents a significant challenge for the system, often hindering its ability to identify the target individual. Similarly, multi-person scenarios pose difficulties, as the crowded environment can lead to either a complete lack of person detection or the inadvertent overlooking of the targeted person. Furthermore, the diverse behavioural nuances exhibited by individuals performing the same action can lead to overfitting on the training data, potentially hindering generalizability to unseen situations. To overcome these limitations, an analysis of solutions to address these limitations will be presented in the next part.

5.4 Future Work

While the current system's accuracy warrants further progress, future improvements can capitalise on its potential. Enhancing accuracy can be addressed through strategic data curation. Prioritising training data free from occlusion and multi-person scenarios is crucial, as these situations often result in inaccurate or absent pose extraction. Feeding the model with such misleading data can negatively impact its overall recognition capabilities. Moving forward, incorporating diverse training videos showcasing various behavioural nuances for the same actions can significantly improve the system's ability to generalise and recognize actions more accurately. Last but not least, future research may focus on the integration of novel pose estimation approaches to improve handling of occlusion and multi-person situations.

Chapter 6

Appendix

6 Appendix

6.1 Acknowledgements and Publications

Paper Title: *The Science of Video: A Review of Human Activity Recognition Techniques*

Authors: Koh Jian Yong, Ts. Dr. Tan Chi Wee

Conference Title: *International Conference on Digital Transformation and Applications (ICDXA) 2024*

Publication Status: Accepted

References

Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., & Ilic, S. (2014). 3D pictorial structures for multiple human pose estimation. *IEEE Conference on Computer Vision and Pattern Recognition (pp. 1669-1676)*. https://doi.org/10.1109/CVPR.2014.216

Choudhury, N. A., & Soni, B. (2023). An Efficient and Lightweight Deep Learning Model for Human Activity Recognition on Raw Sensor Data in Uncontrolled Environment. *IEEE Sensors Journal, 23(20)*, 25579-25586. https://doi.org/10.1109/JSEN.2023.3312478

Davis, J.W., Sharma, V., Tyagi, A., Keck, M. (2009). Human Detection and Tracking. *In: Li, S.Z., Jain, A. (eds) Encyclopedia of Biometrics. Springer*, Boston, MA. https://doi.org/10.1007/978-0-387-73003-5_35

Google Developers. (2021). Pose Landmarker. MediaPipe. https://developers.google.com/mediapipe/solutions/vision/pose_landmarker

H. Badave & M. Kuber (2021), "Evaluation of Person Recognition Accuracy based on OpenPose parameters," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2021, pp. 635-640, doi: 10.1109/ICICCS51141.2021.9432108.

Host, K., & Ivašić-Kos, M. (2022). An overview of Human Action Recognition in sports based on Computer Vision. *Heliyon*, 8(6), e09633. https://doi.org/10.1016/j.heliyon.2022.e09633

Huang, C., Li, J., & Gao, G. (2023). Review of Quaternion-Based Color Image Processing Methods. *Mathematics, 11(9), 2056. MDPI AG*. http://dx.doi.org/10.3390/math11092056

IBM. (n.d.). Deep learning. IBM. https://www.ibm.com/topics/deep-learning

Islam, M. M., Nooruddin, S., Karray, F., & Muhammad, G. (2023). Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things. Information Fusion, 94, 17-31. https://doi.org/10.1016/j.inffus.2023.01.015

Jo, B., & Kim, S. (2022). Comparative analysis of OpenPose, PoseNet, and MoveNet models for pose estimation in mobile devices. *Traitement du Signal*, 39(1), 119-124.

K. Abe, K. -I. Tabei, K. Matsuura, K. Kobayashi and T. Ohkubo (2021), "OpenPose-based Gait Analysis System For Parkinson's Disease Patients From Arm Swing Data," *2021 International Conference on Advanced Mechatronic Systems (ICAMechS)*, Tokyo, Japan, 2021, pp. 61-65, doi: 10.1109/ICAMechS54019.2021.9661562.

Kang, R., Park, B., Ouyang, Q., & Ren, N. (2021). Rapid identification of foodborne bacteria with hyperspectral microscopic imaging and artificial intelligent classification algorithms. Food Control, 130, 108379. https://doi.org/10.1016/j.foodcont.2021.108379

Kamthe, U. M. & Patil, C. G. 2018, Suspicious Activity Recognition in Video Surveillance System, *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1-6.

Kim, W., Sung, J., Saakes, D., Huang, C., & Xiong, S. (2021). Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose). *International Journal of Industrial Ergonomics*, 84, 103164. doi: https://doi.org/10.1016/j.ergon.2021.103164.

L. Zhai, Y. Wang, S. Cui and Y. Zhou, "A Comprehensive Review of Deep Learning-Based Real-World Image Restoration," in *IEEE Access*, vol. 11, pp. 21049-21067, 2023, doi: 10.1109/ACCESS.2023.3250616.

Lara, O. D., & Labrador, M. A. (2013). Human activity recognition: A review. Sensors and Actuators A: Physical, 231, 169-178. https://doi.org/10.1016/j.sna.2013.10.029

Leemets, A., & Vajakas, T. (n.d.). Introduction to image processing. Retrieved June 19, 2023, from https://sisu.ut.ee/imageprocessing/book/1

Liu, Z. (2022). Literature review on image restoration. *Journal of Physics: Conference Series,* 2386(1), 012041. https://doi.org/10.1088/1742-6596/2386/1/012041

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M. (2019). MediaPipe: A framework for building perception pipelines. arXiv.org. https://arxiv.org/abs/1906.08172

Martınez, G. H. (2019). OpenPose: Whole-body pose estimation (Doctoral dissertation, Carnegie Mellon University).

Mostafa, K., & Hegazy, T. (2021). Review of image-based analysis and applications in construction. *Automation in Construction*, 122, 103516.

Mutegeki, R., & Han, D. S. (2020). A CNN-LSTM approach to human activity recognition. In 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (pp. 362-366). IEEE. https://doi.org/10.1109/ICAIIC48513.2020.9065078

Mutneja, Vikram. (2015). Methods of Image Edge Detection: A Review. Journal of Electrical & Electronic Systems. 04. 10.4172/2332-0796.1000150.

Nabriya, P. (2021). Implementing LSTM for Human Activity Recognition. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/implementing-lstm-for-human-activity-recognition-using-smartphone-accelerometer-data/

Nur, Umi Kalsom Yusof, Mohd, & Husain, N. (2022). A Review of Long Short-Term Memory Approach for Time Series Analysis and Forecasting. Lecture Notes in Networks and Systems, 12–21. https://doi.org/10.1007/978-3-031-20429-6_2

C. A., Ong and T. L., Bee. (2014) "Human activity recognition: A review." 2014 *IEEE international conference on control system, computing and engineering (ICCSCE 2014)*. IEEE, 2014.

Osokin, D. (2018). Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. arXiv preprint arXiv:1811.12004.

Paul, M., Haque, S.M.E. & Chakraborty, S (2013). Human detection in surveillance videos and its applications - a review. *EURASIP J. Adv. Signal Process*. 2013, 176. https://doi.org/10.1186/1687-6180-2013-176

Preksha, P., & Ankit, T. (2021). A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications. *The Artificial Intelligence Review*, 54(3), 2259-2322. doi:https://doi.org/10.1007/s10462-020-09904-8

Qi, Y., Yang, Z., Sun, W., Lou, M., Lian, J., Zhao, W., Deng, X., & Ma, Y. (2022). A comprehensive overview of image enhancement techniques. *Archives of Computational Methods in Engineering*, 29(2), 583-607. https://doi.org/10.1007/s11831-021-09587-6

S. Kumar et al. (2021), "Human-Inspired Camera: A Novel Camera System for Computer Vision," *2021 18th International SoC Design Conference (ISOCC), Jeju Island, Korea, Republic of*, 2021, pp. 29-30, doi: 10.1109/ISOCC53507.2021.9613914.

S. S. Sumit, D. R. A. Rambli and S. Mirjalili. (2021). Vision-Based Human Detection Techniques: A Descriptive Review. IEEE Access, 9, 42724-42761. https://doi.org/10.1109/ACCESS.2021.3063028

Sharma V., Gupta M., Pandey A. K., Mishra D. & Kumar A. (2022) A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets, *Applied Artificial Intelligence*, 36:1, DOI: 10.1080/08839514.2022.2093705

Siddiqi, M. H., Ali, R., Rana, M. S., Hong, E. K., Kim, E. S., & Lee, S. (2014). Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis. *Sensors (Basel, Switzerland)*, 14(4), 6370–6392. https://doi.org/10.3390/s140406370

Simplilearn. (2023). Computer vision: What it is and why it matters. Retrieved June 19, 2023, from https://www.simplilearn.com/computer-vision-article

Singh, A.K., Kumbhare, V.A., Arthi, K. (2022). Real-Time Human Pose Detection and Recognition Using MediaPipe. Soft Computing and Signal Processing. ICSCSP 2021. Advances in Intelligent Systems and Computing, vol 1413. Springer, Singapore. https://doi.org/10.1007/978-981-16-7088-6_12

Theodorakopoulos, I., Kastaniotis, D., Economou, G., & Fotopoulos, S. (2014). Pose-based human action recognition via sparse representation in dissimilarity space. *Journal of Visual Communication and Image Representation, 25(1), 12-23*. https://doi.org/10.1016/j.jvcir.2013.03.008

Ullah, H. A., Letchmunan, S., Zia, M. S., Butt, U. M., & Hassan, F. H. (2021). Analysis of deep neural networks for human activity recognition in videos—A systematic literature review. IEEE Access, 9, 126366-126387. https://doi.org/10.1109/ACCESS.2021.3110610

Urankar, S. S., K, S., Bhat, S., Kumar, R., J, M., & A S, D. K. (2020). Human Activity Recognition in Video Surveillance – A Survey. International Journal for Research in Applied Science & Engineering

Technology (IJRASET), 8(2). Retrieved from https://www.academia.edu/42100276/Human_Activity_Recognition_in_Video_Surveillance_A_Survey

Van Houdt, G., Mosquera, C., & Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review, 53(8)*. https://doi.org/10.1007/s10462-020-09838-1

Verma, P.K., Singh, N.P., Yadav, D. (2020). Image Enhancement: A Review. In: Hu, YC., Tiwari, S., Trivedi, M., Mishra, K. (eds) Ambient Communications and Computer Systems. *Advances in Intelligent Systems and Computing*, vol 1097. Springer, Singapore. https://doi.org/10.1007/978-981-15-1518-7_29

Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2015). A Review of Human Activity Recognition Methods. *Frontiers in Robotics and AI*, 2. https://doi.org/10.3389/frobt.2015.00028

W. S. Yuwono, D. Wisaksono Sudiharto and C. W. Wijiutomo (2018), "Design and Implementation of Human Detection Feature on Surveillance Embedded IP Camera," *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, Indonesia, 2018, pp. 42-47, doi: 10.1109/SIET.2018.8693180.

W. Zhou, X. Du and S. Wang (2021), "Techniques for Image Segmentation Based on Edge Detection," *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, Fuzhou, China, 2021, pp. 400-403, doi: 10.1109/CEI52496.2021.9574569.

Wang, H., Kläser, A., Schmid, C. et al. (2013). Dense Trajectories and Motion Boundary Descriptors for Action Recognition. *Int J Comput Vis 103*, 60–79 . https://doi.org/10.1007/s11263-012-0594-8

X. Wang, T. X. Han and S. Yan (2009), "An HOG-LBP human detector with partial occlusion handling," *2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan*, 2009, pp. 32-39, doi: 10.1109/ICCV.2009.5459207.

Xia, K., Huang, J., & Wang, H. (2020). LSTM-CNN Architecture for Human Activity Recognition. IEEE Access, 8, 56855–56866. https://doi.org/10.1109/access.2020.2982225

Y. Kong, Y. Jia and Y. Fu (2014), "Interactive Phrases: Semantic Descriptions For Human Interaction Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1775-1788, Sept. 2014, doi: 10.1109/TPAMI.2014.2303090.

Yadav, S. K., Tiwari, K., Pandey, H. M., & Akbar, S. A. (2021). A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223, 106970. doi: https://doi.org/10.1016/j.knosys.2021.106970

Zakaria Benhaili, Ihsane Kabbaj, Youssef Balouki, & Lahcen Moumoun. (2022). Human Activity Recognition Using Stacked LSTM. Lecture Notes in Networks and Systems, 33–42. https://doi.org/10.1007/978-3-030-91738-8_4

Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. Sensors, 22(4), 1476. https://doi.org/10.3390/s22041476

Zhang, Y., Zhang, J., & Zhang, Y. (2021). A novel human action recognition method based on multi-view deep learning. *Applied Intelligence*, 51(3), 1559-1572.

Zheng, Z., Shi, L., Wang, C., Sun, L., & Pan, G. (2019). LSTM with Uniqueness Attention for Human Activity Recognition. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-030-30508-6_40