Faculty of Natural and Mathematical Sciences
Department of Engineering/Information

King's College London
Strand Campus, London,
United Kingdom

7CCSMPRJ/7CCSMUIP

Individual Project Submission 2023

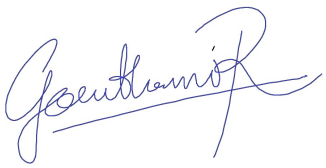
Name:	Gowthami Rasanayagam
Student Number:	k20124758
Degree Programme:	Data Science
Project Title:	FamtamAI: Fusion of AI-MoCap Technology in Augmenting the Human Motion
Supervisor:	Dr. Rita Borgo
Word Count:	11,135

RELEASE OF PROJECT

Following the submission of your project, the Department would like to make it publicly available via the library electronic resources. You will retain copyright of the project.

☑ I **agree** to the release of my project

☐ I **do not** agree to the release of my project

Signature:  **Date:** August 15, 2023

Department of Engineering/Information
King's College London
United Kingdom

7CCSMPRJ/7CCSMUIP Individual Project

FamtamAI: Fusion of AI-MoCap Technology in Augmenting the Human Motion

---

Name: **Gowthami Rasanayagam**
Student Number: k20124758
Course: Data Science

Supervisor: Dr. Rita Borgo

This dissertation is submitted for the degree of MSc in Data Science.

Acknowledgements

I am sincerely thankful to my supervisor, Dr. Rita Borgo, for her invaluable unwavering support, mentorship, and insightful feedback right through the journey of this project. Her domain expert knowledge and motivation have been instrumental in forging my research work.

I extend my deepest appreciation to parents, whose unconditional love, encouragement, and sacrifices have been a constant source of motivation. Your belief in me has been the driving force behind my achievements.

To my friends, thank you for your camaraderie, encouragement, and countless discussions that have enriched my ideas and perspective. Your presence has made this academic journey all the more enjoyable.

I am also thankful to the research communities that I have been a part of. The exchange of knowledge, experiences, and ideas within these communities has played a significant role in molding my research and enhancing my understanding of the subject.

To everyone who has contributed to my growth and success, your support has been invaluable, and I am deeply appreciative of the roles you've played in shaping this endeavor.

Abstract

The recognition of actions holds a pivotal role across a spectrum of applications. This research works puts forth a novel approach aimed at discerning human actions by harnessing the wealth of data derived from motion capture actions.

The goal of this research work is to test the hypothesis of general and global motion fingerprints for both activity classification and individual identification. We propose a comprehensive and self-sustainable system dubbed as **FamtamAI**: **F**usion of **A**i-**M**ocap **T**echnology in **A**ugmenting the human **M**otion. The proposed model is able to produce fingerprints regards to individual atomic motion and can leverage this knowledge to enhance security, ergonomic conditions, recommend improvements, and optimize performance in sports and other activities.

Our main contribution to the domain is,

- Proposed a novel idea of Global Motion Fingerprint for both activity classification and identity recognition.
- Proposed a model "Famtam AI" for recognizing the global motion fingerprint and tested it under the standard protocol.
- Contributed to open-source motion projects including bvh-python Github repository.

We tested our model with Berkerly MHAD [1] test dataset and achieved an accuracy of circa 98% and compared the results against various other existing researchers.

Keywords: AI, Famtam AI, Motion Fingerprint, Motion Capture, Motion Analysis, Identity Recognition, Activity Recognition

Nomenclature

$ACCAD$	Advanced Computing Center for Arts and Design
AGI	Artificial General Intelligence
AI	Artificial Intelligence
$BerkerlyMHAD$	Berkeley Multimodal Human Action Database
$BRNN$	Bi-Directional Recurrent Neural Networks
$CAGR$	Compound Annual Growth Rates
$CIloss$	Clip-based Incremental losses
CNN	Convolutional Neural Network
DAQ	Data Acquisition
DL	Deep Learning
$DTTB$	Directional Temporal Transformer Block
$FamtamAI$	Fusion of Ai-Mocap Technology in Augmenting the human Motion - Artificial Intelligence
FPS	Frames Per Second
GAT	Graph Attention Network
GCN	Graph Convolutional Network
GNN	Graph Neural Network
GRN	Graph Recurrent Networks
HAR	Human Motion Recognition
HCI	Human Machine Interaction
$IMCT$	Inertial MoCap Technology

IMU	Inertial Measurement Unit
kNN	K-Nearest Neighbors
$LSTM$	Long Short Term Memory
$McRBFN$	Meta-Cognitive Radial Basis Function Network
ML	Machine Learning
$MoCap$	Motion Capture
PBL	Projection Based Learning
RBF	Radial Basis Function Network
$RGNN$	Residual Graph Neural Network
RNN	Recurrent Neural Network
RPA	Robotic Process Automation
RPG	Role Playing Game
$SR-TSL$	Spatial Reasoning and Temporal Stack Learning
SRN	Spatial Reasoning Networks
SSL	Self Supervised Learning
STB	Spatial Transformer Blocks
$STST$	Spatial Temporal Specialized Transformers
SVM	Support Vector Machines
$TSLN$	Temporal Stack Learning Network

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The sentient Artificial General Intelligence (AGIs) [4],[5] we fear is not just already transforming multiple industries declaring "code red" for human jobs but also augmenting human capabilities in many unseen ways. AGI's are seamlessly integrated into our day-to-day life and play simple roles like revamping our languages to assisting us in moon shot missions by making intelligent and calculated moves.

Figure 1.1: Illustration of Artificial General Intelligence
This illustration is a part of our poster presentation at BCS Women Lovelace Colloquium 2023 (Refer to Appendix..1.3)

At the same time, traditional Motion Capture (MoCap) technologies have moved out from silver screens and started appearing in our day-to-day life with affordable wearable and vision-based Motion capture technologies. There has been a paradigm shift of interest observed in this domain by researchers for various applications using human motion recognition.

In this research work, we delve into this domain to study the gaps and propose an Artificial Intelligence (AI) system called **Famtam AI**: **F**usion of **A**i-**M**ocap **T**echnology in **A**ugmenting the human **M**otion to solve problem with respect to general global motion fingerprint that is generated by our system for all possible recognition tasks. The project name "Famtam AI" is coined in order to celebrate women in technology and AI space as the one among them.

1.1 Objectives and Aims

This section delineates the objective and aims behind the project **Famtam AI**: **F**usion of **A**i-**M**ocap **T**echnology in **A**ugmenting the human **M**otion. The objectives of the projects are provided in detail in Chapter 1.1.2.

1.1.1 Aim

The aim of this research work is to propose a comprehensive system dubbed as **FamtamAI** for identifying the general/global motion fingerprints (Refer to Chapter 1.7 for a detailed explanation of motion fingerprints) of individuals and leveraging this knowledge to enhance ergonomic conditions, recommend improvements, and optimize performance in sports and other activities.

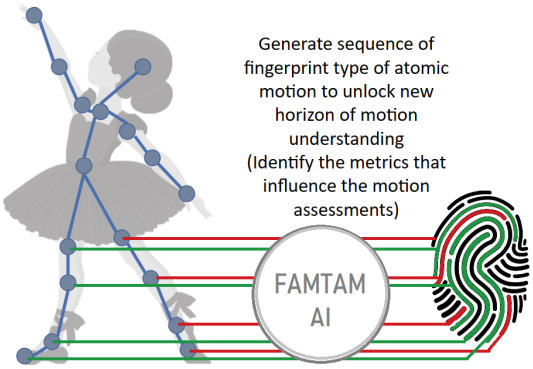


Figure 1.2: Illustration for an idea of Motion Fingerprint (Refer to Chapter 1.7)

1.1.2 Objectives

The objectives of the FamtamAI project are laid down as follows, and in order to formulate the objectives for the project, we have carefully considered the various ethical and social issues as well. Refer to Chapter 5 for additional information about the fair policies of the project.

- *Literature Review and Gap Analysis*: Conduct a thorough analysis of existing motion assessment solutions and related research to identify the current state of knowledge in motion fingerprint identification. Identify gaps and limitations in the current studies to lay the foundation for further investigation.

- *Developing Motion Assessment Solutions*: Based on the findings from the literature review, design and implement novel solutions for assessing motion fingerprints.

Explore various techniques for motion analysis, including gender recognition, individual recognition, action recognition, and potentially expanding to include recognition of emotions, personality traits, and intentionality, as well as bodies, faces, and biological motion.

- *Identifying Influential Metrics*: Determine the key metrics that significantly influence motion assessments. Investigate how these metrics contribute to the understanding of individual motion fingerprints and their potential impact on ergonomics, health, and sports performance.

- *Methodology Development*: Develop robust methodologies and approaches for calibrating motion performance and interpreting human motion data. These methodologies should be capable of dynamically recognizing and interpreting motion patterns, allowing for the recommendation of improvements and rehabilitation strategies.

- *Model Applicability Evaluation*: Evaluate the applicability and effectiveness of the developed model for identifying motion fingerprints in various real-world scenarios. Test the model's accuracy, efficiency, and adaptability to different individuals and motion contexts.

- *Additional Objectives*:
  - *Enhancing Ergonomic Conditions*: Utilize the motion fingerprint identification system to create better ergonomic and healthy conditions for individuals. Recommend personalized ergonomic adjustments based on the dynamic recognition and interpretation of their motion patterns.

  - *Optimizing Sports Performance*: Explore the application of motion fingerprints in sports to enhance athletes' performance. Develop calibration techniques that calibrate individual motion fingerprints to improve sports-related movements and optimize athletic capabilities.

By achieving these objectives, this research aims to contribute to the field of motion assessment and foster the implementation of personalized solutions to improve ergonomics, health, and sports performance based on individual motion fingerprints.

1.2 Background

Motion capture, also known as *MoCap* or Motion Tracking, is a cutting-edge technology that enables the accurate recording and analysis of human movements in various applications, spanning from entertainment and gaming industries to biomechanics, sports science, and healthcare [6]. This revolutionary technique involves capturing the intricate nuances of human motion and converting them into digital data, allowing for a detailed understanding of movement patterns and behaviors.

In recent years, motion capture has witnessed remarkable advancements, driven by the rapid evolution of sensor technology, machine learning algorithms, and computational power. These advancements have opened up exciting possibilities for researchers and professionals seeking to delve deeper into the intricacies of human movement and its applications across diverse fields.

The creative objectivity of this chapter is to equip you with an end-to-end knowledge and overview of motion capture technology, its historical development, underlying principles, and its prevalent use cases in both academic and industrial settings. Secondly, we will do a thorough literature review on machine learning approaches and researches around the domain to harvest valuable information and produce effectiveness on human motion planning. By understanding the foundations and the progress made in this field, we can lay the groundwork for our research and identify the existing gaps and challenges that motivate further investigation.

In the following sections, we will explore the key components and methodologies involved in AI-based motion capture systems and the wide-ranging applications that have demonstrated the transformative potential of this technology. Additionally, we will highlight some of the current limitations and open questions in the realm of motion capture, aiming to address these in our research to contribute to the advancement of this innovative field. Through this exploration, we seek to leverage the capabilities of motion capture to enhance our understanding of human motion, paving the way for novel solutions and advancements in areas such as animation, rehabilitation, sports performance, and human-computer interaction.

1.2.0.1 What is Motion Capture?

Motion capture, often abbreviated as *Mocap*, is a technique used in the to digitally capture and record the movements of subjects. It involves tracking the position and orientation of various points on the subject's body or objects in real time or during a performance. The captured data is then analyzed and used to create highly realistic and accurate computer-generated animations or visual effects.

The subject wears a specialized suit or markers on specific body parts. These markers or sensors emit signals that are tracked by cameras or other motion capture devices [7]. Alternatively, non-optical methods such as inertial sensors or magnetic systems can be used to capture the motion. The motion capture system collects the positional and orientation data from the markers or sensors. The captured data is then processed and translated into a digital format that can be used by computer software.

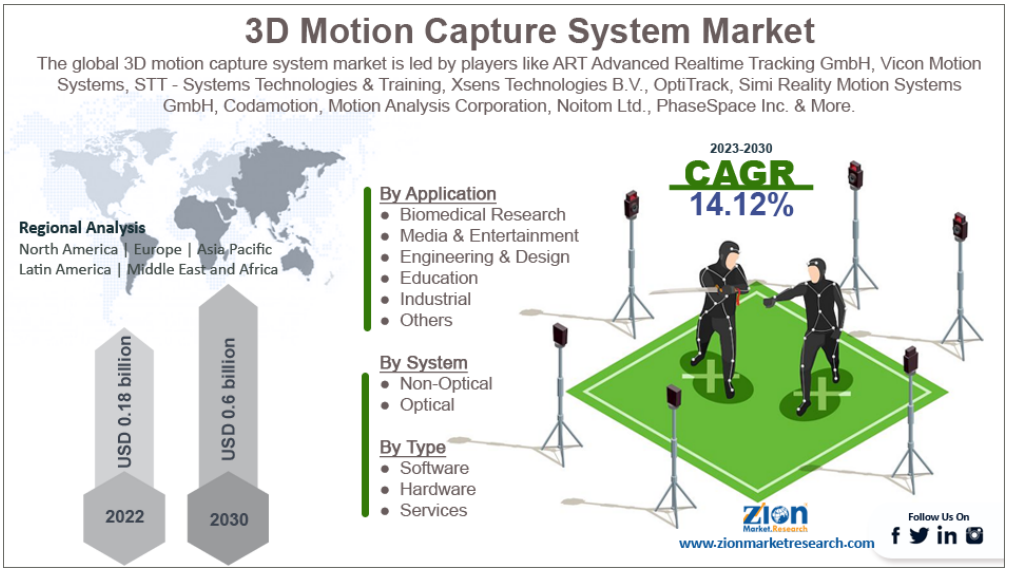


Figure 1.3: Motion Capture Tech Market Condition and Growth (Source: Zion Market Research)

Motion capture is widely used in various fields, including film, television, video games, virtual reality, augmented reality, and biomechanical research. According to a report by Zion Market Research, the global motion capture market was valued at around $180 million in 2022. The report projects that the market is expected to grow at 14.12% of CAGR - Compound Annual Growth Rates and will reach a value of a whopping $600 million by 2030, growing at a (CAGR) of 14.12% during the forecast period. Moreover, the application of motion capture in virtual reality and augmented reality experiences, as well as its use in biomechanical research for analyzing human movement and improving athletic performance, has further fueled the market's growth.

1.3 Sports and Performance

Motion capture technology has a long history of assisting Sports and performance evaluation and enhancement [8]. By capturing and analyzing the movements of subjects, it enables coaches, trainers, and sports scientists to gain insights into technique, biomechanics, and performance optimization. Here are some ways motion capture is used in sports:

1.3.0.1 Biomechanical Analysis

Motion capture allows for detailed analysis of a person's movements, providing precise data on joint angles, body positions, and forces exerted during different sports and athletic activities. This information helps identify pathological human motion sequences, inefficiencies, or flaws in presentation techniques, enabling biomechanical experts and coaches to make targeted corrections and improve performance [9].

1.3.0.2 Injury Prevention and Rehabilitation

Athletes employ a multitude of joint movements with numerous degrees of freedom during high-speed sporting maneuvers. Motion capture can aid in assessing movement patterns and identifying potential injury risks during this fast-paced motion sequences. By monitoring an athlete's movements and comparing them to biomechanical norms, motion capture can highlight deviations or imbalances that may increase the risk of injury. Additionally, it can be used during rehabilitation to track progress, assess movement quality, and guide the recovery process [10], [11].

1.3.0.3 Performance Optimization

Motion capture can assist in optimizing athletic performance by analyzing movements and identifying areas for improvement. It allows for comparisons between different techniques or training methods, helping athletes and coaches make informed decisions regarding training strategies, equipment selection, and performance-enhancing modifications as demonstrated in Wei et al. [12] for Basketball Resistance Training and Shooting Hit Rate.

1.3.0.4 Skill Development

By providing accurate real-time feedback, motion capture helps athletes refine their skills and technique. It can be used to measure key performance indicators, such as speed, agility, and power, and provide immediate feedback to athletes during training sessions. This enables athletes to make adjustments and enhance their performance in real-time.

1.3.0.5 Sports Research

Motion capture contributes to sports research by providing quantitative data on movement patterns, joint dynamics, and performance metrics. Researchers can utilize this

data to study biomechanics, understand the mechanics behind successful techniques, and develop evidence-based training protocols [13], [12].

Overall, AI-assisted building strategies and game theories for team sports like football fed with MoCap data of players have pushed the limits to dominate the games and open a new horizon of injury prevention while promoting healthy and efficient human motion. Motion capture technology plays a significant role in sports performance and enhancement by providing objective measurements, precise analysis, and feedback to athletes and coaches. It helps optimize training, prevent injuries, refine technique, and advance our understanding of human movement in sports.

1.4 Health

1.4.1 Work Life balance

Numerous research studies have established that the use of technology plays a pivotal role in shaping Work-Life Balance. These technologies possess the capability to either blur or enhance the demarcation between work and personal life. Particularly, mobile and motion analysis technologies facilitate and empower communication, presentation, and the traversing of temporal and spatial boundaries [14], [15].

1.4.1.1 Enhanced Productivity

AI-based motion capture technology has become a valuable tool in improving productivity in the workplace. By employing sophisticated algorithms and sensors, it accurately captures human movement, translating it into digital representations that can be used in various applications such as animation, gaming, and virtual reality. This technology enables employees to efficiently collaborate and contribute to projects regardless of their physical location, eliminating the constraints of traditional office environments. With the ability to seamlessly integrate remote workers, teams can work together in real-time, breaking down barriers of time and distance. Consequently, employees can allocate their time more effectively, resulting in increased productivity and a better work-life balance.

1.4.1.2 Remote Work Opportunities

The implementation of AI-based motion capture technology has facilitated the rise of remote work opportunities. Traditionally, individuals were required to commute to their workplaces, often enduring long hours in traffic or crowded public transportation. This daily grind not only consumed a significant amount of time but also contributed to stress and decreased overall job satisfaction. However, with motion capture technology, employees can now work remotely without compromising the quality of their work. Through wearable devices and advanced motion tracking systems, individuals can participate in meetings, engage in collaborative projects, and perform tasks from the comfort of their

own homes. The flexibility and freedom offered by this technology empower individuals to design their work schedules around personal commitments, leading to improved work-life balance.

1.4.1.3 Promotion of Employee Well-being

AI-based motion capture technology contributes to the promotion of employee well-being, a critical component of achieving work-life balance. Traditional office settings often necessitate long hours of sitting in front of a computer, which can lead to physical health issues such as musculoskeletal disorders and decreased mobility. However, with motion capture technology, employees can engage in more active work environments. The technology encourages movement and physical activity, as it tracks and translates gestures, postures, and body language into the digital realm. This promotes a healthier lifestyle and mitigates the negative effects of sedentary work, ultimately enhancing employee well-being. Furthermore, by allowing individuals to work remotely, this technology reduces the stress associated with commuting and offers the opportunity to create a comfortable work environment tailored to personal preferences.

AI-based motion capture technology has emerged as a transformative force in achieving work-life balance. By enhancing productivity, enabling remote work opportunities, and promoting employee well-being, this technology paves the way for a more flexible and harmonious work environment. As it continues to evolve and be integrated into various industries, its potential to revolutionize work-life balance is undeniable. Embracing this technology can lead to a future where individuals can achieve professional success while maintaining a fulfilling personal life, ensuring a healthy and balanced lifestyle for all.

1.4.2 Nutrition and Hydration Management

1.4.2.1 Personalized Nutritional Recommendations

FamtamAI like AGIs may analyze an individual's movement patterns, energy expenditure, and physical characteristics to provide personalized nutritional recommendations. By accurately capturing body movements and tracking vital signs, the technology can gather valuable data on an individual's metabolic rate, hydration levels, and nutrient requirements. This data is then processed by AI algorithms, which generate tailored meal plans, portion sizes, and hydration schedules. By optimizing nutrient intake based on individual needs, professionals can fuel their bodies more effectively, leading to improved physical and mental performance. This personalized approach to nutrition management contributes to overall well-being and work-life balance by ensuring that individuals are adequately nourished to meet the demands of their work and personal lives.

1.4.2.2 Real-Time Hydration Monitoring

Motion capture technology integrated with AI can also revolutionize hydration management. Through wearable devices equipped with sensors, motion capture technology accurately monitors an individual's hydration levels in real time. By capturing and analyzing data such as body temperature, sweat rate, and electrolyte balance, AI algorithms can provide timely hydration reminders and recommendations. This technology helps professionals maintain optimal hydration throughout the day, reducing the risk of dehydration-related fatigue, headaches, and decreased cognitive function. By prioritizing hydration, individuals can sustain their energy levels, improve concentration, and enhance their overall well-being.

AI and motion capture technology have transformed nutrition and hydration management, fostering a healthier work-life balance for professionals. By providing personalized nutritional recommendations, real-time hydration monitoring, and promoting physical activity, these technologies empower individuals to optimize their well-being and performance. Embracing AI and motion capture technology in nutrition and hydration management enables professionals to sustain their energy levels, enhance their productivity, and prioritize their health while juggling work and personal commitments. As these technologies continue to advance, their potential to revolutionize work-life balance and improve overall quality of life is immense.

1.4.3 Posture correction

1.4.3.1 Preventing Sedentary Behavior

FamtamAI like intelligent systems can address the issue of sedentary behavior in the workplace. Prolonged sitting has been linked to various health problems, including musculoskeletal disorders and decreased productivity. By utilizing motion capture technology, professionals can receive real-time feedback and reminders to engage in physical activity or adopt a better posture. AI algorithms analyze body movements and detect periods of inactivity, prompting individuals to take breaks, stretch, or engage in light exercises. By encouraging movement throughout the workday, this technology promotes a healthier and more active lifestyle, positively impacting both physical and mental well-being. The incorporation of physical activity into the work routine contributes to a better work-life balance by ensuring professionals prioritize their health while meeting work responsibilities.

Detecting bad posture in real time helps mitigate musculoskeletal conditions which are the leading contributors to disability worldwide and supports assessing the ergonomic conditions in many different environmental settings. Personal motion inferences using FamtamAI enable us to track our nutrients and hydration levels and lead to healthy practices.

1.4.4 Gait Analysis

Gait analysis, the study of human walking patterns, plays a crucial role in various fields, including sports medicine, rehabilitation, and biomechanics. The integration of artificial intelligence (AI) and motion capture technology has brought significant advancements to gait analysis, enabling more accurate and comprehensive assessments. By combining the power of AI algorithms with motion capture systems, researchers and healthcare professionals can gain deeper insights into gait abnormalities, leading to improved diagnosis, treatment, and overall patient care [16].

1.4.5 Artificial Robotic Limbs

Prosthetics powered by AI trained on personal motion data seeds a new generation of enhanced and powerful bionic human beings that impersonate natural Gait patterns for amputees. This also creates an enhanced and powerful workforce with multi-actuators and sensors that can be effectively applied to labor incentive applications.

1.5 Security and Surveillance

Motion recognizable-based AI models may be deployed in security and surveillance tasks at extreme corner cases such as low-light conditions where regular vision-based monitoring fails. This application is one of our key factors for developing identity recognition based on motion fingerprints. The hypothesis is that every individual has unique motion sequences when it comes to doing an activity like walking, running, and moving.

1.6 Human Computer Interaction (HCI)

Human and computer/robot interaction has become part of vital technology since the introduction of the digital era. Simply a computer mouse can be considered as a motion tracker and used motion to interact with the machine. This legacy technology along with AI is becoming more powerful in Robotic Process Automation (RPA) and many other tasks.

1.6.1 Transport

Motion assessment can pave a path to perceive human behavior with dead reckoning by Autonomous navigation systems for safe and convenient operation. This allows human pedestrians and autonomous vehicles to co-exist and operate safely.

1.6.2 Entertainment

Role Playing Games (RPG) are stars of this century in mesmerizing humans with their own flavor of addiction and tieing them to chairs by devouring most of their leisure time.

These games are mostly embedded with motion-perceivable AIs that react to a certain state of the player.

1.7 Motion Fingerprints

Motion Fingerprint is a generic term that we have dubbed in order to recognize an encoding of the atomic motion pattern in our FamtamAI's hidden fingerprint layers. Atomic motion patterns are a subset of uniquely identifiable motion sequences that make human motion movements. For. e.g. Throwing a ball is a human motion movement that can be constructed with atomic motion sequences such as "Raising the right arm", "Taking a pace on the left foot", "Twisting the wrist", "Lean back on the trunk", and etc...

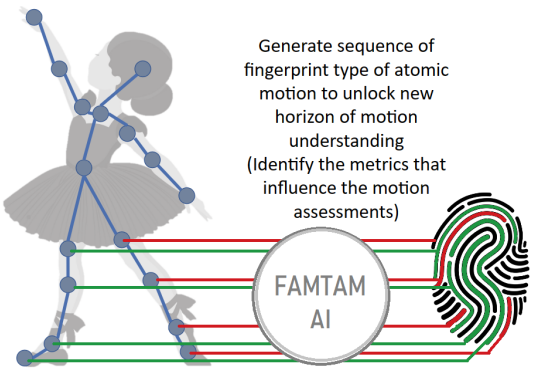


Figure 1.4: Illustration for an idea of Motion Fingerprint
This illustration is a part of our poster presentation at BCS Women Lovelace Colloquium 2023 (Refer to Appendix..1.3)

We create and are willing to test the hypothesis of using such motion fingerprints as a global fingerprint to not only recognize the action but also to identify individuals based on their atomic motion fingerprints. The idea behind this is that every person has an individual and unique motion sequence of activities. For e.g. The way "SAM" walk (consider atomic motions like his pace length, his maximum angle in knee joints, etc..) may significantly differ from how "MARRY" walk.

We believe that FamtamAI's competency in uniquely identifying individuals and recognizing their activity based on their movement fingerprints can open up a wide range of applications that do not just have to be limited to surveillance and security.

Chapter 2

Literature Survey

This chapter studies the current domain knowledge to identify the research gaps. We have included relevant studies and research within our lenses and attempt to bridge details from different areas to lay the foundation for our research.

2.1 Human Motion/Action Recognition

Human Motion/Action Recognition (HAR) endeavors to comprehend subject behaviors and categorize each actions with a label. Due to its diverse applications, HAR has garnered escalating attention within the domain of computer vision. To represent human actions effectively, various data modalities have been explored, encompassing RGB images, skeleton hierarchical structures, depth images/fields, infrared images, 3D point clouds, event streams, audio, inertial measurements, ultrasonic, radar, and WiFi signals. These modalities encode distinct and valuable information sources, offering specific advantages tailored to different application scenarios. As a result, numerous research endeavors have been undertaken to explore diverse approaches for HAR, leveraging the potential of these varied modalities [17].

Initially, the majority of research in Human Action Recognition (HAR) concentrated on utilizing RGB or grayscale videos as input, owing to their widespread usage and accessibility. However, in more recent times, there has been a notable surge in the adoption of alternative data modalities for HAR, including depth data, skeleton hierarchical structural data, infrared motion sequences, 3D point cloud data, event-based data stream, audio/ultrasound data, inertial data (acceleration, and angular velocity), radar and WiFi echo data. The paradigm of shift can be attributed to the advancement of precise and cost-effective sensor technologies, which have enabled the exploration of various modalities. Additionally, each modality offers distinct advantages in HAR, depending on the specific application scenarios, further driving the adoption of multi-modal approaches.

RGB video data stands as the prevailing data type in Human Action Recognition

(HAR), finding extensive usage in security, surveillance, and monitoring systems. On the other hand, Hierarchical structural data such as skeleton models, represents the path/trajectories of motion pattern and proves particularly efficient for HAR tasks not that does not involve any object reconstruction or scene context. 3D Point cloud and depth data have gained popularity in HAR applications related to robot navigation and self-driving, as they capture essential 3D structural and distance information. Additionally, infrared, audio/ultrasound, and radar data find utility in HAR within extreme environments such as in low light conditions where cameras become useless, while event streams excel in preserving the focal point movement of targets and eliminating visual redundancy, armed them as a well-suited technology for HAR applications.

During the previous era, extensive studies have focused on Human Motion Recognition and Analysing using single modalities. However, due to the distinct strengths and limitations of different data channels in HAR, there has been a growing interest in the fusion of many modalities and the exchange of information between them to enhance accuracy and robustness in recent times. Fusion, in this context, refers to combining the information from two or more channels to identify the respective actions effectively. For instance, ultrasound data can complement camera-based vision data channels, enabling the discrimination within motions like "throwing an object" and "placing an object". Apart from fusion, other methods have explored co-learning, where knowledge is transferred among multiple channels to bolster the robust performance of HAR applications.

Building upon the insights into various modalities provided above, our research emphasis lies in developing a comprehensive global motion fingerprint to recognize not only Human Actions but other information as well. Therefore, the subsequent section will delve into intricate methodologies for effectively recognizing human actions using skeleton modalities as this is the basic interface for most of the motion recording/actuation.

The next chapter will explore the advanced deep-learning (DL) methodologies for human motion recognition and analysis. These Machine Learning (ML) approaches can be divided into four main categories, such as K-Mean and K-Nearest Neighbors (kNN), Support Vector Machines (SVM), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Radial Basis Function Networks (RBF), Graph Neural Networks (GNN), and Transformers-based Neural Networks.

2.1.1 K-Mean and kNN - K Nearest Neighbors

K-means is a clustering algorithm used popularly in various machine learning and data analysis tasks [18]. Its objective is to group a set of data points into K - no. of clusters, where each data belongs to the group with the smallest distance to the centroid (mean). The algorithm iterates through and assigns data points to the nearest clusters and recalculates the centroids until convergence occurs. K-means is often regarded as a grouping of similar data points based on their features, making it a widely used technique for unsupervised clustering tasks.

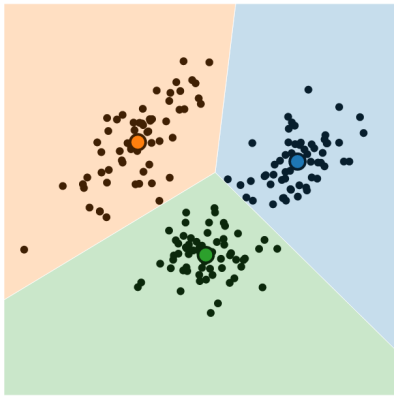


Figure 2.1: Illustration of clustering approach: K-Mean and kNN - K Nearest Neighbors

Unlike the idea of the K-Mean which is an unsupervised approach, K-Nearest Neighbors (kNN) is a supervised machine-learning algorithm used for classification and regression tasks [19]. In K-NN, for a given data point, the algorithm finds the K training data points that are closest to it based on a chosen distance parameter measure (often Euclidean distance). The majority class among the K neighbors is then designated to the query data point for the clustering application. In regression, the mean or weighted mean of the K nearest neighbors' values is provided as the predicted value. K-NN is simple and intuitive, as it relies on the idea that related data points tend to have relational outcomes.

I. Kapsouras and N. Nikolaidis [20] proposed an innovative adaptation of the K-means algorithm, designed to accommodate the cyclical nature of angle data channels of skeleton joints. Subsequently, each frame is allocated to distinct motion patterns, and histograms are created to capture the occurrence frequency of these motion patterns. The recognition of actions in the test data is accomplished using Nearest Neighbor and Support Vector Machine (SVM) classification techniques.

The method's efficiency and resilience are demonstrated through comprehensive experiments conducted over the Berkerly Multi-modal Human Action Detection dataset

(MHAD) [1]. The accuracy of the simplex method has been reported as 98.18%.

2.1.2 SVM - Support Vector Machines

As depicted in Figure 2.2, Support Vector Machines (SVM) represent a group of interconnected techniques used in supervised learning, suitable for tackling classification and regression applications. The SVM model constructs an optimal hyperplane with the widest achievable margin within a transformed input domain. This hyperplane effectively partitions the instances of different classes, while also maximizing the distance to the nearest properly separated support vectors. The parameters governing the properties of this solution hyperplane are obtained through the resolution of a quadratic programming optimization issue [21].

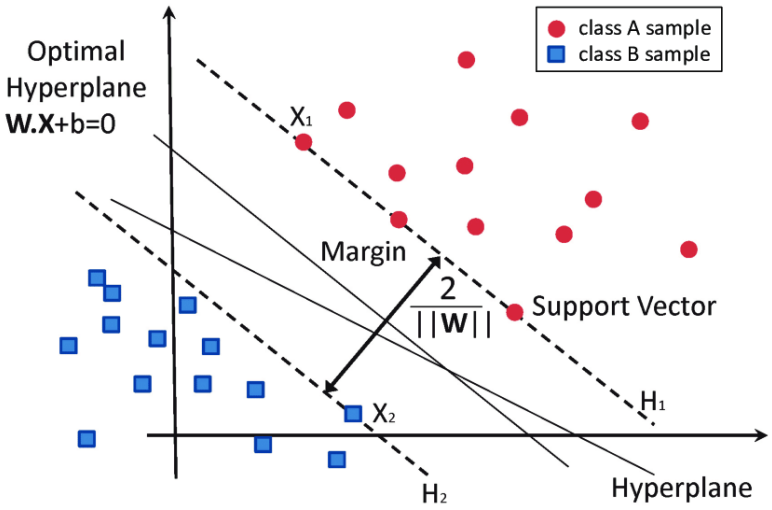


Figure 2.2: Support Vector Machine

Vantigodi et al. [22] proposed a method for a motion capture system that utilizes the 3D coordinates of the subject's skeleton structural joints to analyze the dynamic aspects of the motion. The characteristics used for classification involve the temporal variance of each individual joint in the skeleton, coupled with its variance weighted by time. This combination of features effectively captures temporal nuances and aids in distinguishing actions that might otherwise be perplexing, such as sitting down and standing up. These features can be swiftly extracted and are well-suited for real-time action recognition.

The effectiveness of the proposed approach is showcased through its application of both a correlation-based metric and Support Vector Machines (SVM) on the Multimodal Human Action Detection dataset (MHAD)[1]. The achieved recognition accuracy surpasses 95%, validating the robust performance of this SVM-based method.

2.1.3 RNN - Recurrent Neural Network

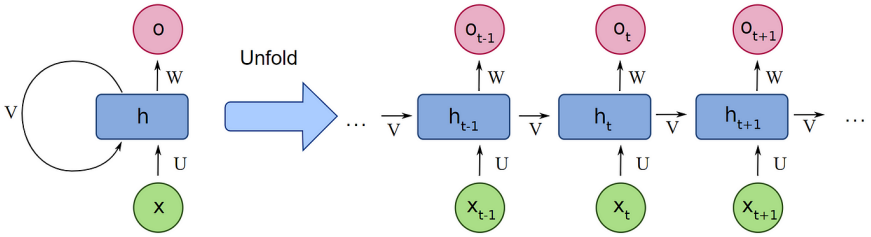


Figure 2.3: Illustration of RNN (source: Medium Article [2])

Recurrent Neural Networks are mainly used to perform analysis on time domain data due to their in-built recurring nature baked into their neurons. Early adoption of this technology introduced traditional vanilla RNNs. These are very simple recurrently connected neurons that take time-based input data and perform operations recursively while storing a hidden state every time that represents the previous data. However, these traditional vanilla RNNs are subject to all sorts of vanishing and exploding gradient problems [23]. This makes the network not an optimal solution for effectively learning the long temporal dependencies in the time domain data. Hence, various methods [24], have built and adopted the idea of RNNs and Long Short Term Memory (LSTM) neural networks to efficiently model the temporal context information within the human motion sequences.

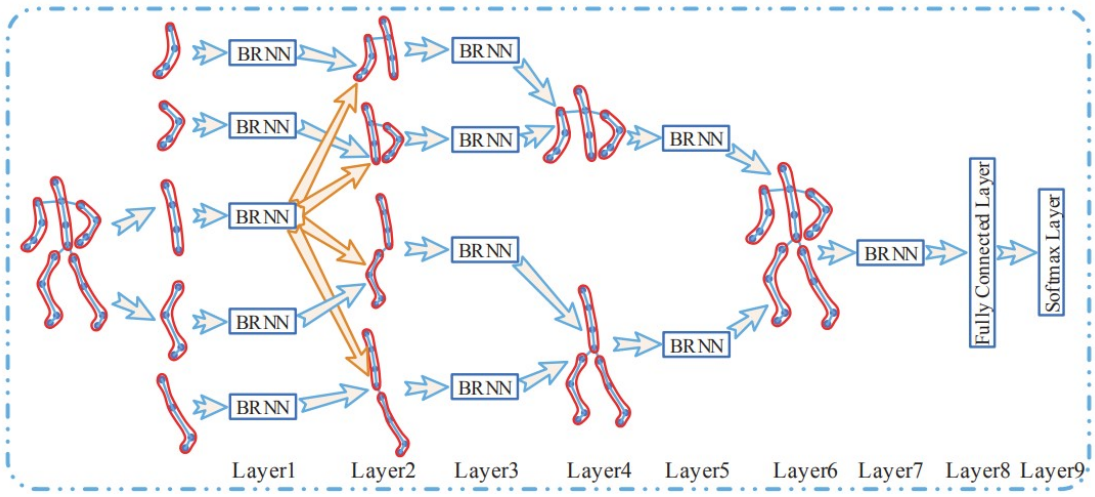


Figure 2.4: Du et al. [24] proposed hierarchical RNN is presented in the above illustration.

A conventional approach presented by Du et al. [24] introduced a complete hierarchical Recurrent Neural Network technology wherein they partitioned the subjects' skeleton structural data into 5 distinct body parts, instead of processing an entire structure as a single feature vector within each frame. Each of these body parts was subsequently individually inputted into multiple bidirectional RNNs. The resulting output representations from these RNNs were then hierarchically combined to produce comprehensive high-level representations of the action.

The entire skeleton hierarchical input data structure is partitioned into 5 distinct body parts and fed into 5 Bi-Directional Recurrent Neural Networks (BRNNs). The inputs of the second layer are a fused product of previous layers which extract the representation from its inputs. A dense layer with a soft-max classification layer provides classification and makes a prediction about the temporal input data.

The performance achieved by this innovative approach is nothing short of exceptional. It achieves an accuracy rate of circa 100% when put against testing on the widely embraced Berkeley Multimodal Human Action Database (MHAD [1]).

2.1.4 CNN - Convolution Neural Network

CNNs have demonstrated remarkable achievements in 2D image analysis as illustrated in Figure 2.5, mainly attributed to their exceptional ability to learn spatial features [25]. However, when it comes to hierarchical data structures (skeleton data) as input to recognize the motion sequences, effectively modeling spatiotemporal information becomes a challenge. Despite this, numerous advanced approaches have emerged to address this issue. Some methods utilize temporal convolutions on skeleton data, while others transform skeleton sequences into pseudo-images, subsequently processed by standard CNNs to tackle HAR tasks.

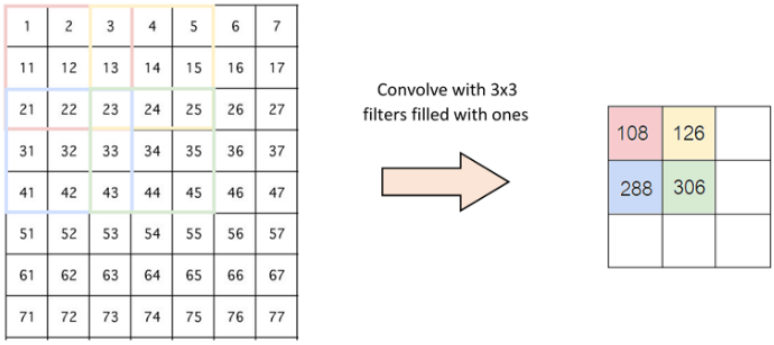


Figure 2.5: Illustration of CNN (source: Medium Article [3])

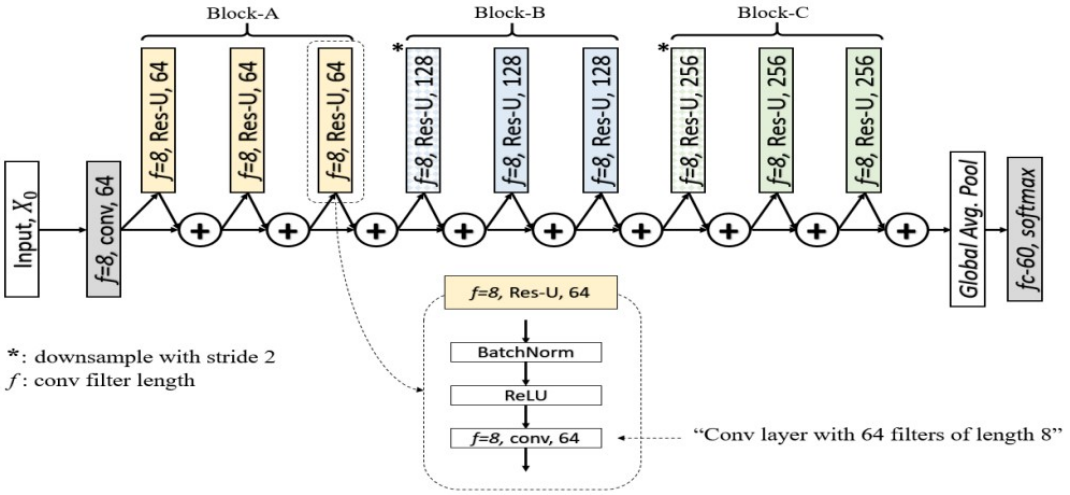


Figure 2.6: Soo et al. [26] proposed Res-TCN model architecture.

Soo et al. [26], propose the utilization of Temporal Convolutional Neural Networks

(TCN) as a new class of models for 3D human action recognition, offering interpretability when using 3D skeletons as input. By redesigning TCN with interpretability in mind, the model allows the explicit learning of readily interpretable spatiotemporal representations for action recognition. The resulting model, Res-TCN Figure 2.6, the target is to model a spatiotemporal solution that is more easily understood, explained, and interpreted, contributing to the advancement in the domain of 3D human motion analysis and recognitions.

Li et al. [27] introduce a novel complete framework for convolutional co-occurrence feature learning. The co-occurrence features are acquired using a layered approach, aggregating contextual information at different levels. Initially, point-level information of each skeletal structural joint is independently embedded into words, followed by the assembly of semantic representations in both spatial and temporal domains. To enhance joint co-occurrence features, a global spatial aggregation scheme is introduced, outperforming local aggregation methods. Additionally, raw hierarchical structural data coordinates and their time domain gradients were integrated using a two-stream paradigm.

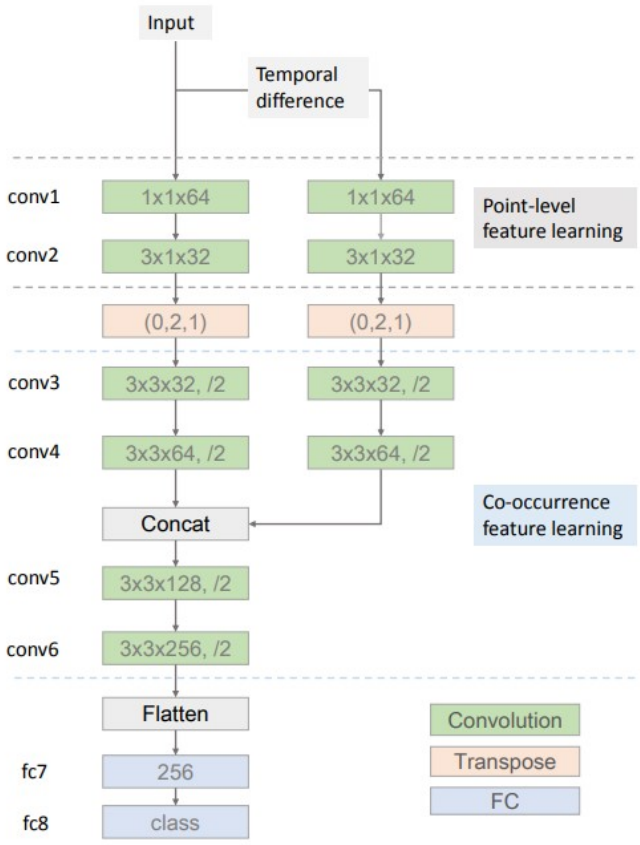


Figure 2.7: Li et al. [27] proposed Hierarchical Co-occurence Network architecture.

2.1.5 RBF - Radial Basis Function Networks

A Radial Basis Function (RBF) neural network [28] is a specific type of neural network architecture characterized by its feed-forward structure. In this network, there exists a single hidden layer composed of a set of functions that are linearly independent of each other. These functions collectively establish a basis within a multi-dimensional function space.

At the core of the method proposed by Vantigodi et al. [29] lies the representation of the skeletal configuration in each frame of an action sequence as a vector spanning 129 dimensions. In this vector domain, every element represents a distinct 3D angle originating from its joint's connection with a fixed point on the skeleton structure. To encapsulate the visual essence of the video, a histogram is meticulously constructed over a "codebook" that emerges from a combination of all motion sequences. The approach is further enhanced by incorporating the time domain variance (temporal differences) inherent in the structural joints, effectively introducing an auxiliary feature into the recognition process.

The crux of action categorization rests on the application of the "Meta-Cognitive Radial Basis Function Network (McRBFN)", complemented by its "Projection Based Learning (PBL)" algorithm following the approach of Nelson and Naren's models [30]. The results garnered by this pioneering approach are nothing short of remarkable; it achieves an astounding accuracy rate surpassing their previous proposed approach [22] of 95% to 97% when tested against the widely embraced Berkeley Multimodal Human Action Database (MHAD [1]).

2.1.6 GNN/GCN - Graph Neural/Convolution Network

Numerous learning tasks necessitate the handling of graph data, which encompasses intricate relationships among components. The demand for a model capable of learning from graph inputs arises in various contexts, including motion analysis using skeleton structural datasets where structural information is inherently present, there's a burgeoning need for models that can analyze these structures.

Graph Neural Networks (GNNs) are neural models engineered to capture graph dependencies by facilitating the exchange of messages between graph nodes. Over the past decade, GNN variants like Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and Graph Recurrent Networks (GRN) have exposed stellar performance across various domains and tasks [31].

Graph-based learning is used heavily to analyze hierarchical and structural data models and has gained significant attention recently due to the attention power of graph networks. Skeleton hierarchical data inherently takes the form of a graph-based structure. Therefore, the representation of the hierarchical structural data solely in a vectorized form of the sequence when processed by Recurrent Neural Networks or 2D / 3D matrix processed by Convolutional Neural Networks may not be able to fully capture the intricate spatiotemporal relationships between body joints. Consequently, graph-based topological representations are expected to be the best candidate for expressing the complexity of hierarchical structural data. In response, many different approaches based on these Graph-based Neural Networks (GNN) and Graph-based Convolutional Networks (GCN) have been introduced and experimented to treat the skeleton data as graphs with nodes and edges.

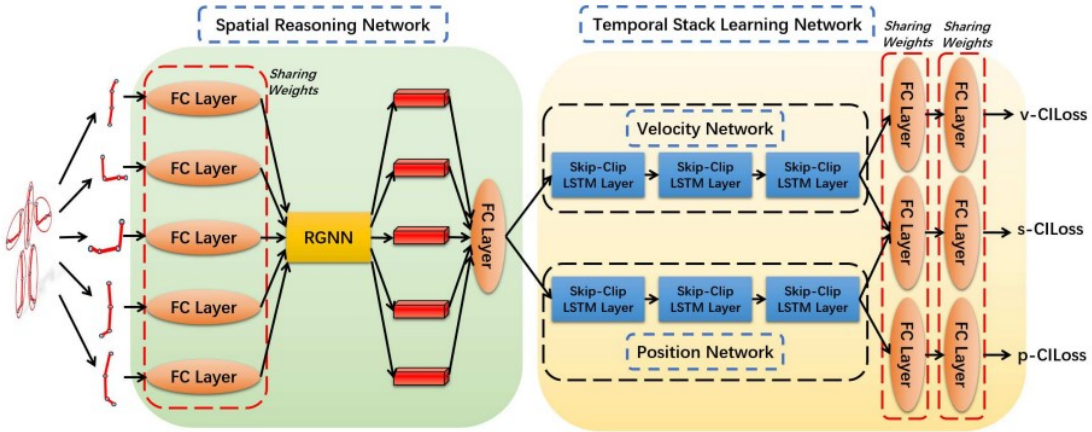


Figure 2.8: (Si et al. [32] proposed SRN and TSLN model

Si et al. [32] present an innovative approach called "Spatial Reasoning and Temporal Stack Learning (SR-TSL)" for hierarchical structural-based motion analysis and recognition. This approach comprises two interconnected components: the "Spatial Reasoning Network (SRN)" and the "Temporal Stack Learning Network (TSLN)". The SRNs utilize a residual graph neural network to study the higher dimensional spatial hierarchical information within individual frames. On the other hand, the TSLN employs a fusion of many skip-clip LSTMs to model very intricate time-domain information of hierarchical data sequences. The authors also propose a "clip-based incremental loss" to effectively optimize the proposed model during training.

Figure 2.8 shows the end-to-end pipeline of the Si et al. proposed model which has SRN and TSLN networks buttered together. In the SRN, a Residual Graph Neural Network (RGNN) is utilized to analyze the high-dimensional spatial hierarchical and structural information between the different body parts. The TSLN which is a temporal stack learner may learn the highly oriented time domain aspects of the hierarchical structural data in sequence.

2.1.7 Transformers

The Transformer [33], a cutting-edge deep learning model, has emerged as a prominent force in the machine learning domain, boasting impressive capabilities and holding great promise. It is usually structured with an encoding network and a decoding network. The encoding network employs multiple "self-attention blocks" to embed the input data sequence effectively. Similarly, the decoding networks adopt a comparable network structure, with an addition of an encoder-decoder attention mechanism in every block. This attention mechanism-based model empowers the Transformer architecture to excel in tasks involving long-term memory and dependency learning, the fusion of multi-modal data, and processing multiple tasks.

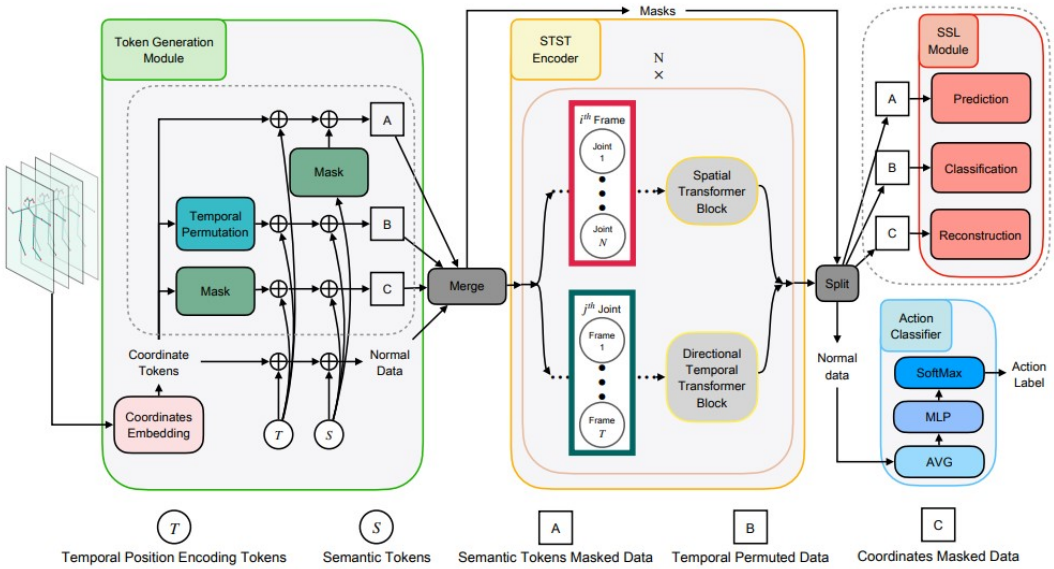


Figure 2.9: Zhang et al. [34] proposed STST-Encoder

The authors in Zhang et al. [34], introduced two distinct components, the Spatial Transformer Block (STB) and the Directional Temporal Transformer Block (DTTB), to effectively model sequences of hierarchical structural data in the space and time axeses, respectively. The need of capturing the trajectory of the entire motion body in the time domain was a driving factor behind this design. Unfortunately, in practice, the extracted motion data often contains temporal and spatial dimensional noises, which can compromise the identification capabilities of the models due to occlusion, sensor limitations, or raw video quality. To address this challenge and adapt to imperfect information, the authors proposed a multi-task self-supervised learning approach. By introducing noisy data in various situations, they sought to enhance the model's potential. By fusing these designed components and the self-supervised learning strategy, they have introduced the

"Spatial-Temporal Specialized Transformer (STST)" to tackle all specific demands of hierarchical structural data modeling in an effective way.

As shown in Figure 2.9, the initial block called "Token Generation Module" will generate three different tokens for structural data sequences. The Encoder of the Spatial-Temporal Specialized Transformer (STST) learns the pattern of motion sequences. The Self Supervised Learning (SSL) Module is composed of three different self-supervised learning tasks in order to boost the STST-Encoder. The action-recognizing module "Action Classifier" receives the processed structural data sequence representation from the previous blocks and predicts the label for each action.

Chapter 3

Methodology

The Chapter discusses the foundational idea behind our research theme and walks the reader through each and every design consideration to arrive at a plausible solution.

3.1 Objectives, Specification and Design

In the realm of motion capture technology, the significance of multi-channel time series datasets, particularly Motion Capture (MoCap) systems, lies in their ability to document human body movements. Prior research has predominantly focused on addressing biases arising from sensor placement, sensor types, preprocessing methods, feature extraction, and label irregularities in Human Activity Recognition (HAR) models. However, a recent and noteworthy area of investigation revolves around the potential for recognizing individual identity and soft biometrics from these activity recordings. Nevertheless, the impact of subject-specific characteristics on classifier models remains relatively unexplored.

The present study endeavors to curate training data by leveraging heterogeneity measures (HM) - a statistical concept that pertains to the non-uniformity of qualities within a given set. Our hypothesis is rooted in the belief that by incorporating variations in physical characteristics of subjects, we can establish a comprehensive motion fingerprint capable of recognizing individuals and their atomic motions using motion sequence data.

The initial phase involves identifying and assessing strategies for developing Motion Assessment Solutions based on insights garnered from a thorough literature review. We aim to present novel solutions for evaluating motion fingerprints as a significant contribution to the field. The scope of exploration spans diverse motion analysis techniques, encompassing individual recognition, action recognition, and the potential extension into recognizing emotions, personality traits, intentionality, as well as bodies, faces, and biological motion. This entails identifying Influential Metrics within motion fingerprints that significantly impact motion assessments. Furthermore, we will investigate how these

fingerprints enhance our understanding of individual motion patterns and their potential implications on ergonomics, health, and sports performance.

Ultimately, we will rigorously evaluate the applicability and efficacy of the proposed model in identifying motion fingerprints across various real-world scenarios, utilizing publicly available datasets. The assessment will encompass testing the accuracy, efficiency, and adaptability of the model to different individuals and motion contexts. By doing so, this research aims to advance the domain of motion capture technology and shed light on its far-reaching implications across diverse fields.

In this chapter, we delve into the methodology and data preparation process employed in conducting the research to develop a motion fingerprint as discussed in Chapter 1.7. A comprehensive understanding of the methodology employed is crucial for ensuring the validity, reliability, and generalizability of the study's findings. This chapter outlines the systematic approach taken to address the research objectives and sheds light on the steps undertaken to gather, analyze, and interpret the data. Additionally, it highlights the strategies used to ensure data quality, as well as the ethical considerations observed throughout the research process.

3.2 Methodology and Implementation

This section provides an overview of the selected research design and justifies its suitability for addressing the research objectives. It describes the nature of the study, the data collection methods employed, and the rationale behind their selection. The section also discusses any potential limitations and how they were mitigated to enhance the validity of the findings.

3.2.1 Domain

A detailed study is conducted to review this emerging domain of literature and identify the gaps in existing research works and presented in Chapter 2. This step of the research work is a key milestone in developing motion capture methods that are economical and practically relevant to many stakeholders such as clinical practicians, performance coaches, and motion researchers around the world.

We understood that there are no global means of understanding atomic motions exists in any of the previous literature according to our knowledge and research. Therefore, we decided to attack the gap in generating a global fingerprint for atomic motions and identity recognition.

3.2.2 Motivation

The motivation behind this research is to create better ergonomic and healthy conditions by dynamically recognizing and interpreting a human's movements. This is enabled by correctly understanding the motion sequences and comparing them against the standard.

3.2.3 Strategy

In order to understand the motion sequences, we analyze the motion sequences of individuals in atomic parts. Atomic action is an action defined by a very small time frame defined motion sequences. Machine learning methods such as LSTMs or Temporal Convolutional Neural Networks can be employed to predict qualitative and quantitative outcomes and can act as an encoder machine to generate a global fingerprint for such atomic motions.

3.3 DATA

This section provides insights into the data sets gathered and considered for the research. It also explains the preparation strategy of the data and discusses the limitation of the datasets.

Motion data depicts the kinematic information of moving objects, which can be represented using a sequence of poses. In our research, we focus only on human motion activities, where the human poses are usually informed with positions and rotations of the body joints, linked as a hierarchical bone structure as shown in Figure 3.1.

3.3.1 Data Formats

Motion capture technologies produce data in a 2D representation on a sub-domain that takes place in a 3D world along with temporal directions. There are many popular file formats such as FilmBox(*FBX* [35]), *C3D* [36], Biovision Hierarchy (*BVH* [37]), OpenSim *TRC* and *MOT* file, etc. to store such motion data along with additional metadata including volumetric information of models such as mesh, skin, and muscles.

FBX, FilmBox [35], is one of the famous three-dimensional motion data formats that has been developed originally by Kaydara for MotionBuilder. *FBX* format usually consists of mesh/grid information, texture, material, and hierarchical structural motion frame data. These properties mark them well-suitable to be used in compute-based animation and gaming applications. Many applications use this *FBX* format mainly because of their robustness. However, there exist additional bottlenecks before blindly choosing the data format:

- *FBX* format versions are typically not backward compatible with older versions of software.
- *FBX* format data may be quite large in size and may consist of information that is redundant for users making them not the best choice for portable applications.
- Redunctant information for specific applications as not every computer program is equipped with support for all of the features of *FBX* format. For example, some software may not be able to render certain animations and bake exact materials/skin properly and hence breaks the entire pipeline.

In our context, we have no use case for rendering any material as such and hence we have no such extreme requirements to use rich file formats such as *FBX* or *C3D*.

The *C3D* [36] format stands as a widely used open-domain file format within Biomechanical analysis, Animating 3D models, and Gait Analysis fields. This format has been employed since the mid-1980s to capture synchronized 3D and analog data. *C3D* files

serve as a comprehensive standard, encompassing all the requisites for reading, showcasing, and scrutinizing 3D motion capture data, accompanied by supplementary analog data sourced from electromyography, force plates, and other inertial sensors. Notably distinct from other 3D formats, the *C3D* standard has endured the test of time and remains unchanged despite new product version releases. As a result, data saved in the *C3D* format retains its readability over extended periods. This format compiles 3D coordinate and numeric data from diverse measurement trials, along with the gamut of parameters that characterize the data, all within a solitary file. Consequently, the necessity of attaching motion-related data with additional notes and testing particulars is eliminated, as these often tend to get detached from the data during its circulation, making accessibility seamless.

However, such rich formats as *FBX* or *C3D* are a heavy lift for our application. Therefore, we have chosen to go with an efficient way to retrieve information to represent human motion in terms of the Biovision Hierarchy (*BVH*) file format [37]. *BVH* is a simple avatar animation portable file format developed by a motion capture service company called "Biovision". An *BVH* file consists of textual content in ASCII format, with its initial segment detailing specifications for the primary pose of a human skeleton. The subsequent portion encompasses a chronological series of diverse specifications, each characterizing subsequent poses. Commencing with the keyword HIERARCHY, the initial portion of a *BVH* file designates the Hips joint as the ROOT joint or ROOT node, signifying that it has no parent joint. This pivotal juncture establishes the foundation for a hierarchical arrangement of parent and child joints.

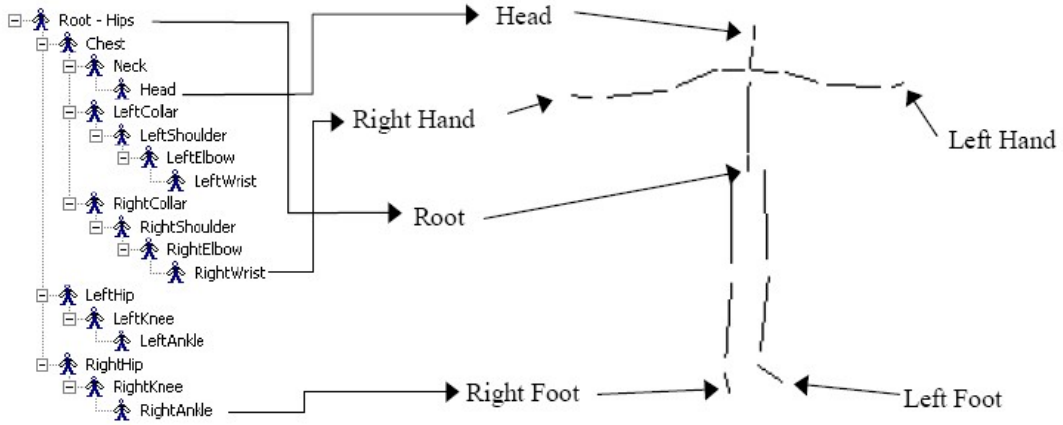


Figure 3.1: *BVH* skeletal structure (Source: www.cs.cityu.edu.hk)

The initial segment, known as the ROOT section, establishes the precise spatial coordinates of the Hips joint within a three-dimensional space. Subsequently, positioned

beneath this ROOT section, are the JOINT sections. Each JOINT section contains vital information that outlines the specific location of a skeletal joint with respect to its parent joint. By capturing these relative positional specifications between parent and child joints, it becomes possible to deduce the length of the "bone" that bridges these interconnected elements. In cases where a joint lacks a corresponding child joint, it forms a link with an End Site section. This organizational structure ensures a comprehensive representation of the skeletal hierarchy within the *BVH* file.

In the *BVH* file, both the ROOT section and the various JOINT sections play a role in defining the CHANNELS. These CHANNELS are essential for managing the temporal sequence of translation and/or rotational coordinates presented in the latter segment of the file.

```
HIERARCHY
ROOT Hips
{
  OFFSET 13.6774 106.653 -34.4427
  CHANNELS 6 Xposition Yposition Zposition Zrotation Xrotation Yrotation
  JOINT ToSpine
  {
    OFFSET -2.69724 7.43032 -0.144315
    CHANNELS 3 Zrotation Xrotation Yrotation
    JOINT Spine
    {
      OFFSET -0.0310711 10.7595 1.96963
      CHANNELS 3 Zrotation Xrotation Yrotation
      JOINT Spine1
      {
        OFFSET 19.9056 3.91189 0.764692
        CHANNELS 3 Zrotation Xrotation Yrotation
        JOINT Neck
        {
          OFFSET 25.9749 7.03908 -0.130764
          CHANNELS 3 Zrotation Xrotation Yrotation
          JOINT Head
          {
            OFFSET 9.52751 0.295786 -0.907742
            CHANNELS 3 Zrotation Xrotation Yrotation
            End Site
            {
              OFFSET 16.4037 0.713936 2.7358
            }
          }
        }
        JOINT LeftShoulder
        {
          OFFSET 17.7449 4.33886 11.7777
          CHANNELS 3 Zrotation Xrotation Yrotation
```

Figure 3.2: *BVH* file header - Skeleton Hierarchy

The latter portion of a *BVH* file commences with the keyword MOTION, succeeding which there is a set of details provided. These details encompass the count of FRAMES as the first item, followed by the sampling rate per second mentioned after the keyword FRAME TIME as the second component. Lastly, there is an indication of the number of lines, corresponding to the FRAMES count, which are furnished with translation and/or rotation coordinates. These coordinates are processed in alignment with the CHANNELS specifications outlined in the initial section of the file.

```
MOTION
Frames: 187
Frame Time:     0.0333333
13.6653 106.685 -34.2172 -5.63641 -71.8545 -95.0052 -70.1915 0 88.8779 62.1586 69.2427 22.2372 5.96825 19.0255 6.37525 -150.291 21.1193
176.098 -163.439 60.5096 -174.866 2.55503 -2.37731 -32.3426 -104.152 50.6314 -90.632 -17.1186 51.6158 12.0422 10.0607 -0.703492 -6.10608
0.646576 -2.09883 44.7788 -85.5555 -35.9695 91.5527 -6.5549 -17.0566 7.55529 15.5878 0.028334 -0.15186 -58.3355 -0.973316 95.6992 2.53043
17.3374 -0.586912 4.36456 -3.16644 -0.747004 -1.97672 -4.51903 2.86016 -59.9066 5.81827 83.4352 -1.5491 17.0043 2.13418 -4.42889 -3.55316
0.822805 -0.0273048 -0.871444 -1.07255
13.6558 106.685 -34.1923 -5.33004 -71.8075 -94.7879 -70.1915 0 88.8779 61.6647 69.1881 22.4401 5.96476 18.9544 6.63873 -150.243 21.132
176.351 -163.965 60.546 -174.303 2.56451 -2.3418 -32.3762 -104.357 50.6798 -90.4701 -17.0503 51.6706 12.0574 9.8348 -0.670281 -5.91275
0.648379 -2.10412 44.7943 -85.6025 -35.9407 91.4572 -6.50378 -16.6606 7.69196 15.8044 -0.0272584 -0.540942 -58.2947 -1.11338 95.6324
2.56058 17.3294 -0.589208 4.36552 -3.0575 -0.743324 -2.05604 -4.71274 2.99267 -59.9123 5.7451 83.457 -1.58736 16.9287
2.15541 -4.54213 -3.57447 0.833658 -0.0336935 -0.985649 -1.1899
13.6444 106.662 -34.1839 -5.14114 -71.8291 -94.6807 -70.1915 0 88.8779 62.085 69.2124 21.7681 5.63167 18.8114 6.91391 -150.068 21.0416
176.452 -163.392 60.4091 -174.634 2.61429 -2.16969 -32.584 -103.98 50.454 -90.5512 -17.126 51.279 11.952 9.46835 -0.65526 -6.02126
0.628342 -2.0419 44.7046 -84.5226 -36.1212 92.0003 -6.61624 -14.3614 8.20695 16.8382 -0.279099 -2.33467 -58.2406 -1.18212 95.6529 2.55021
17.485 -0.674181 4.19501 -2.96761 -0.724124 -2.15306 -4.94598 3.20737 -59.9984 5.61693 83.4138 -1.52028 16.8367 2.30906 -5.07006 -3.53977
0.880786 -0.0636369 -1.22449 -1.31501
13.656 106.696 -34.1903 -4.81748 -71.8409 -94.2098 -70.1915 0 88.8779 62.3745 69.1783 21.6117 5.68774 18.747 6.93164 -150.3 21.1659
176.106 -161.648 60.4274 -176.058 2.62838 -2.14294 -32.7043 -103.754 50.3806 -90.7786 -16.9137 50.1518 11.6633 9.53264 -0.61324 -5.48578
0.629508 -2.04549 44.7113 -84.105 -36.0702 92.1762 -6.59764 -13.9723 8.37312 16.8665 -0.257928 -2.17151 -58.2263 -1.09168 95.6286 2.60954
17.502 -0.88676 4.1438 -2.98443 -0.719846 -2.04965 -4.644 3.39078 -59.9524 5.73153 83.3708 -1.46257 16.8566 1.88739 -4.96938 -3.20194
0.862125 -0.175012 -1.6681 -1.10251
13.6667 106.654 -34.1811 -4.93931 -71.8299 -94.4227 -70.1915 0 88.8779 62.1863 69.171 21.8288 5.81083 18.7989 6.77913 -150.375 21.1812
175.896 -160.762 60.3818 -176.764 2.63055 -2.11441 -32.6531 -103.848 50.4667 -90.6369 -16.8571 50.1093 11.653 9.48986 -0.619298 -5.58752
0.640837 -2.08589 44.64 -83.6004 -36.1539 92.5646 -6.75752 -13.1259 8.43563 17.1377 -0.258437 -2.14532 -58.2087 -1.1645 95.6347 2.58104
17.5247 -0.679457 4.43633 -3.04942 -0.749775 -2.1136 -4.85192 3.10897 -60.0476 5.65686 83.4309 -1.5207 16.7485 2.18539 -5.0786 -3.6401
0.88442 -0.0316408 -0.967264 -1.17889
13.6801 106.642 -34.1335 -5.11809 -71.634 -94.4949 -70.1915 0 88.8779 61.4268 69.014 22.9695 6.0843 18.8586 6.51014 -150.402 21.2219
175.862 -160.752 60.4214 -176.843 2.63494 -2.16313 -32.8539 -103.706 50.2366 -90.7783 -16.9446 50.2644 11.6912 9.21232 -0.616324 -5.76726
0.639059 -2.08676 44.4822 -84.1287 -36.2484 92.2163 -6.54707 -14.3605 8.35472 16.4011 -0.255097 -2.20468 -57.897 -1.03586 95.6672 2.60768
17.8091 -0.733175 4.20504 -3.18136 -0.732661 -2.08719 -4.76497 3.23556 -59.9601 5.81739 83.5705 -1.53608 16.5893 2.14525 -5.03144 -3.64172
0.880124 -0.00384874 -0.873908 -1.25487
```

Figure 3.3: *BVH* file Frames - Motion Sequences

In addition, the availability and abundance of datasets plays a significant role in selecting the file formats. In our dataset exploration for the application, we found that the *BVH* format is widely available and most of the other formats can be converted to the *BVH* formats using standard software stacks.

Despite the availability of alternatives such as *AMC*, which shares similarities with *BVH* but follows a different skeletal format, and *TXT*, involving comma or space delimited files that directly extract translation data (x, y, z) from Vicon Blade's c3d data, these options either lack the same level of widespread usage as *BVH* or do not provide equivalent benefits such as available to read by an open source software.

3.3.2 Motion Capture Technologies

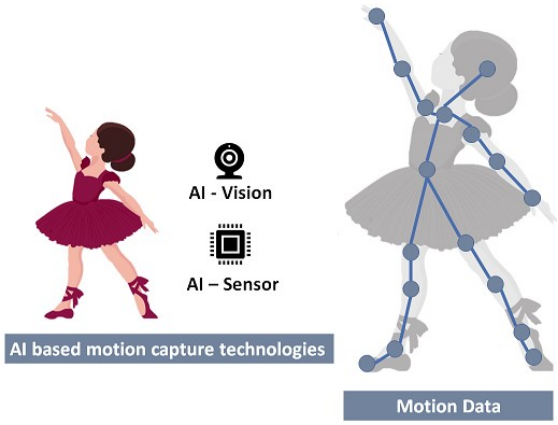


Figure 3.4: Motion Capture Technologies

The motion capture data can be generated using multiple technologies.

- Motion Sensor based data capture (IMU) or Inertial Mocap technology (IMCT)

Figure 3.5: IMU-based motion capture suit (Source: Xsens)

- Optical motion capture data

Figure 3.6: Vicon optical marker-based motion capture technology

- Retroreflective markers using IR camera (Vicon, Qualysys, Optitrack, etc)
- Depth Cameras (Kinect)
- Markerless Vision-based tracking using regular cameras

These motion data are very expensive to generate as they require proper laboratory setups, volunteers to participate in motion trials, and sophisticated software stacks to build motion sequences. In addition, guaranteeing the integrity of the data with privacy concerns as well as labeling the same is beyond the budget of the research. Given that we are constrained by time and resources, we have decided to pick a public dataset for our research purpose allowing us to concentrate more on model architecture and fingerprint generations.

3.3.3 Dataset

3.3.3.1 Berkeley Multimodal Human Action Database (MHAD)

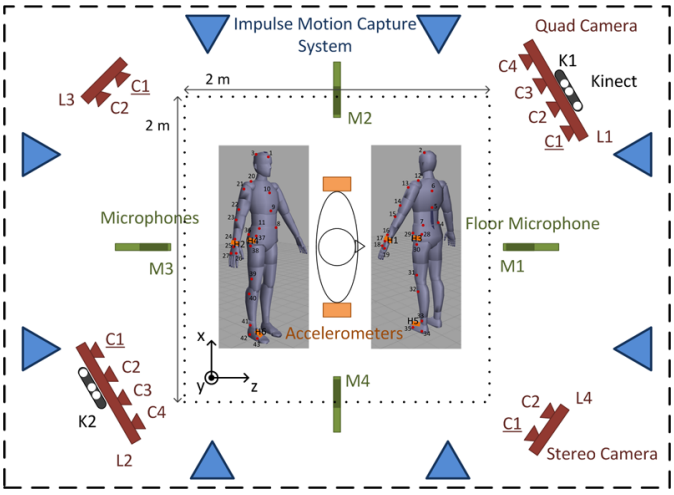


Figure 3.7: Diagram of the data acquisition system (Source: Berkeley MHAD [38])

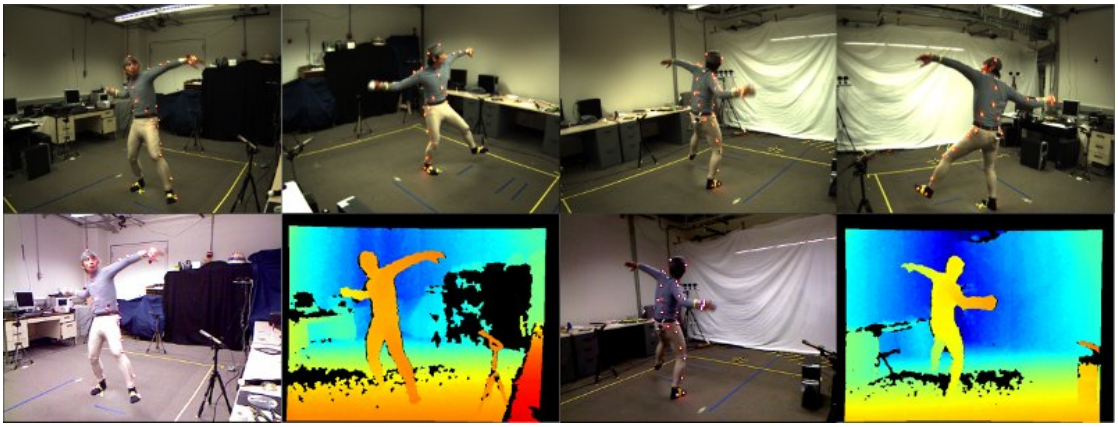


Figure 3.8: The act of throwing is demonstrated through the reference camera in each camera cluster, along with both of the Kinect cameras. (Source: Berkeley MHAD [38])

The Berkeley Multimodal Human Action Database (MHAD) [38] comprises 11 distinct actions executed by 5 female and 7 male subjects, with ages ranging from 23 to 30, except for one elderly participant. Figure 3.7 explore the data acquisition system (DAQ). Every motion sequence was performed by all participants for 5 repetitions, resulting in approximately 660 action sequences, amounting to a total recording time of approximately 82 minutes. Moreover, for each subject, a T-pose has been captured, which

facilitates skeleton extraction. Furthermore, the database encompasses background information, featuring scenarios both with and without the inclusion of the chair used in certain activities.

In Figure 4.2, all the action classes { Jump, Jumping Jacks, Bend, Punch, Wave Two Hands, Wave One Hand, Clap, Throw, Sit Down/Stand Up, Sit Down, and Stand up} available in the Berkeley MHAD dataset showcases concurrent representations of the recorded actions alongside the corresponding point clouds extracted from the depth data of the Kinect sensors. Figure 3.8 shows detailed picture of a single action class "Throwing".

Additional information about the fair usage of databases and licenses shall be found in Chapter 5.0.1.

3.3.3.2 ACCAD: Advanced Computing Center for the Arts and Design

ACCAD [39] database is an Open Motion Project by the Advanced Computing Center for the Arts and Design of The Ohio State University and is licensed under a Creative Commons Attribution 3.0 Unported License.

The data is generated using 12 Vicon system "T40"s infrared cameras with 12.5mm and 24mm interchangeable lenses and retroreflective markers of size from 3mm to 24mm making the data more accurate as it is the current gold standard of the domain. The data is presented in various formats including ".bvh", ".c3d", ".amc"/".asf", ".tvd", and ".txt". It has a databank with various records tagged under different subject names,

- Female 1 (81 motion sequences)
- Male 1 (69 motion sequences)
- Male 2 (149 motion sequences)
- etc...

doing different motions/action sequences such as standing, walking, running, etc...

Additional information about the fair usage of databases and licenses shall be found in Chapter 5.0.1.

3.3.3.3 Other Datasets

We have also provided links to other useful datasets although we haven't used them in our research due to various reasons such as the use of markers in different locations making it difficult to merge with Berkerly MHAD [1] or ACCAD [39] data.

- *AMASS [40]*: **A**rchive of **M**otion capture **A**s **S**urface **S**hapes, serves as a comprehensive and diverse compilation of human motion data. It accomplishes the unification of 15 distinct optical/retroreflective marker-based motion capture datasets by presenting them within a harmonized framework and parameterization, thus creating a unified representation.

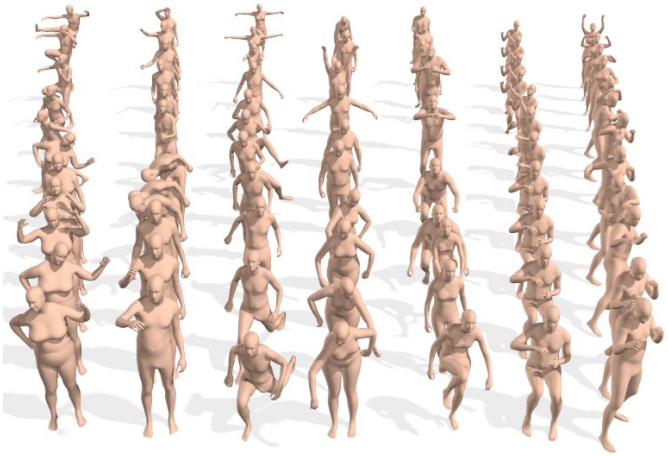


Figure 3.9: AMASS Dataset (Source: AMASS [40])

- *KiMoRe [41]*: **Ki**nematic assessment of **Mo**vement and clinical scores for remote monitoring of physical **Re**habilitation using Kinect cameras.

- *Physical Rehabilitation Movements Data Set (UI-PRMD) [42]*: Ten individuals in good health executed ten repetitions of various physical therapy motions. These motions were captured using a Vicon optical tracker and a Microsoft Kinect sensor. The data is structured to encompass the positions and angles of joints spanning the entire body.

  The database comprises the following ten movements: 1. Hurdle Step, 2. Standing Shoulder Internal-External Rotation, 3. Side Lunge, 4. Standing Shoulder Abduction, 5. Standing Active Straight Leg Raise, 6. Deep Squat, 7. Inline Lunge, 8. Sit To Stand, 9. Standing Shoulder Extension, and 10. Standing Shoulder Scaption.

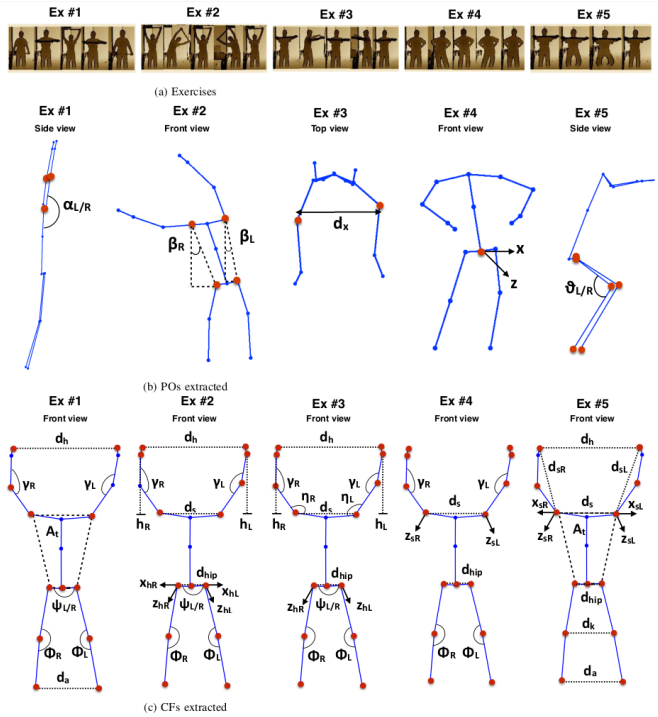


Figure 3.10: KiMoRe Dataset (Source: KiMoRe [41])

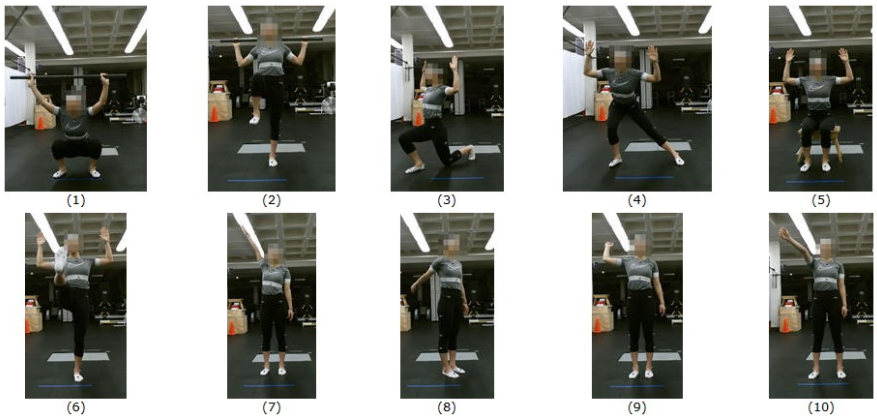


Figure 3.11: UI-PRMD Dataset (Source: UI-PRMD [42])

- *IntelliRehabDS (IRDS) [43]*: This database consists of three-dimensional hierarchical structural joint positions of 15 subjects out of which, 14 are healthy persons. Subjects perform a few simple motions and the data was collected by a Kinect depth camera. Each action is labeled with its activity type along with a label showing if the movement was executed correctly or not. The labeling of the activity class was done by two different annotators independently.

- *MSR Action3D dataset* [44],[45]: The dataset was recorded utilizing a depth sensor such as Kinect and encompasses 20 distinct actions executed by 10 individuals, with each action performed two or three times. Overall, there are 557 valid action sequences, with each sequence comprising 20 skeleton joints per frame.

- *UTKinect-Action dataset* [46]: This dataset was obtained using a lone stationary Kinect device and comprises 10 distinct actions carried out by 10 diverse subjects. Each subject performed each action twice, resulting in a total of 199 action sequences. The dataset provides the 3D coordinates of 20 joints. Due to variations in viewpoint and substantial intra-class differences, this dataset is considered challenging.

Figure 3.12: UTKinect-Action3D Dataset (Source: [46])

- *NTU RGB+D dataset* [47]: This dataset was recorded using three Microsoft Kinect v2 cameras and encompasses a total of 60 action classes, categorized into three main groups: 40 daily actions, 9 health-related actions, and 11 mutual actions. Each sequence in the dataset includes the 3D coordinates of 25 skeleton joints. Notably, the dataset presents significant challenges due to its substantial intra-class variations and viewpoint differences.

- *Northwestern-UCLA dataset* The Northwestern-UCLA dataset was acquired in real-time using three Microsoft Kinect v1 cameras. It comprises 1494 sequences, encompassing 10 distinct action categories. Each action is performed by ten subjects, ranging from one to six repetitions. The dataset offers a diverse range of viewpoints, adding to its variability and richness.

Figure 3.13: NTU RGB+D Dataset (Source: [47])

- *UWA3DII dataset* The data collection for this dataset involved four Microsoft Kinect v1 cameras, capturing a comprehensive set of 30 human actions, each performed four times by ten different subjects. Notably, each action was recorded from four distinct perspectives: front view, left side view, right side view, and top view. The dataset presents significant challenges due to the diversity in viewpoints, occurrences of self-occlusion, and the high similarity observed among certain actions.

3.3.4 Data Preparation

The dataset for building the initial strategy and architecture is obtained from ACCAD (Chapter 3.3.3.2) in BVH (Biovision Hierarchy) format and we prepared the data to be loaded as a standardized numpy array into our input pipeline for the respective AI model.

The bvh skeleton structure has 22 joints including the root joint (Hips). For the root joint (Hips), there are 6 channels exist ($X_{Position}$, $Y_{Position}$, $Z_{Position}$, $X_{Rotation}$, $Y_{Rotation}$, $Z_{Rotation}$). The rest of the 21 joints have only 3 channels ($X_{Rotation}$, $Y_{Rotation}$, $Z_{Rotation}$).

Similarly, for the test procedure in Chapter 4, the dataset is obtained from Berkeley Multimodal Human Action Database (MHAD) (Chapter 3.3.3.1) in BVH (Biovision Hierarchy) format and prepared. The bvh skeleton structure of Berkerly MHAD has 30 joints including the root joint (Hips). For the root joint (Hips), there are 6 channels exist ($X_{Position}$, $Y_{Position}$, $Z_{Position}$, $X_{Rotation}$, $Y_{Rotation}$, $Z_{Rotation}$). The rest of the 29 joints have only 3 channels ($X_{Rotation}$, $Y_{Rotation}$, $Z_{Rotation}$).

Refer to Appendix .1.1, for additional information on the functions.

3.3.4.1 Load and Standardize Data

Data in bvh format is first read using bvh-python API [48] and converted into ($FeatureVector \times No.of.Frames$), where $FeatureVector$ refers to channels for individual joints. In total, the feature vector size becomes 69 ($6 \times 1 + 3 \times 21$). Note that the feature vector size for the test dataset (Berkeley MHAD [1]) is 93 ($6 \times 1 + 3 \times 29$) and the model details are expressed in Chapter 4.

The data for $X_{Position}$, $Y_{Position}$, $Z_{Position}$ channel features are in "cm" while $X_{Rotation}$, $Y_{Rotation}$, $Z_{Rotation}$ channel features are in "degrees". Therefore, to standardize we first removed the initial offset from each frame making the first frame all zeros which make the rest of the frame as only the vector of changes. Additionally, for the rotation channels ($X_{Rotation}$, $Y_{Rotation}$, $Z_{Rotation}$), we also convert them into radians using the formula ($Val\_in\_Degree \times PI/180$).

Different records are in different sizes due to differences in $No.of.Frames$, hence, we first determine a frame size (30 Frames) to make the records into finite-size records. For. eg. $Male1\_A1\_Stand.bvh$ has 187 frames, hence this has been separated into 6 X 30 Frames samples. The rest of the 7 frames didn't satisfy the threshold of data availability of more than 70%($7/30 < 70\%$). In case where more than 70% of data is available, we pad the initial time frames with zeros.

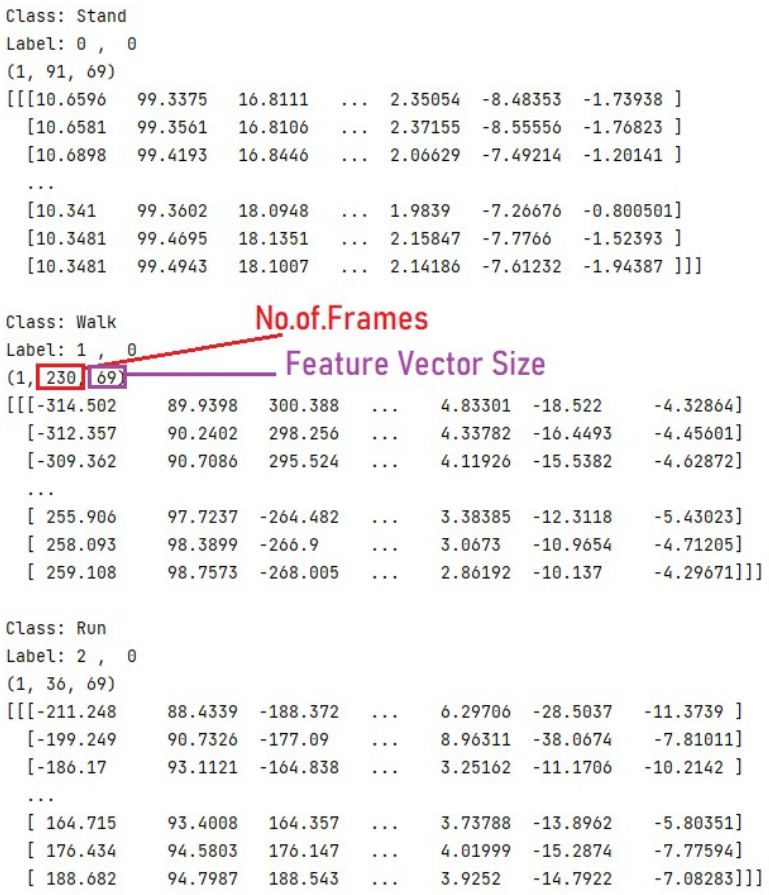


Figure 3.14: BVH file format read and converted as a numpy array

3.3.4.2 Data Labels

For each data (69×30), we assign two labels based on the file name namely the *Actionclasslabel* and the *Identityclasslabel*. For example, $Male1\_A1\_Stand.bvh$ is with *Stand*, *Actionclasslabel* and $Male1$, *Identityclasslabel*.

Using the labels we assign a one-hot vector for each *Actionclass* of records.

- *Stand*: [1, 0, 0, 0, 0]
- $Walk$: [0, 1, 0, 0, 0]
- *Run*: [0, 0, 1, 0, 0]
- *Skipping*: [0, 0, 0, 1, 0]
- $Crouch$: [0, 0, 0, 0, 1]

Similarly, the same file is tagged to subject $Male1$. Hence another *Identityclass* one-hot vector is assigned to each subject.

- $Female1$: [1, 0, 0]
- $Male1$: [0, 1, 0]
- $Male2$: [0, 0, 1]

3.3.5 Software and Tools

Python is used to implement the entire data preparation pipeline and Python numpy library is used as a default tool to prepare the data along with other default python environments.

We have used additional tools such as *Blender*3.5 [49] software to carve out data from motion sequences that are tagged with multiple atomic actions. For instance, there are records of motion files that are tagged under both $Walk$ and *Run*. Such files contain the first few frames performing walking actions and for the rest of the frames, the subject was transitioning to running action. These files are imported into *Blender* [49] and trimmed and saved as new files for atomic actions. The process helps us generate additional data for the training and testing of the models.

Additional information about the fair usage of software and license shall be found in Chapter 5.0.1.

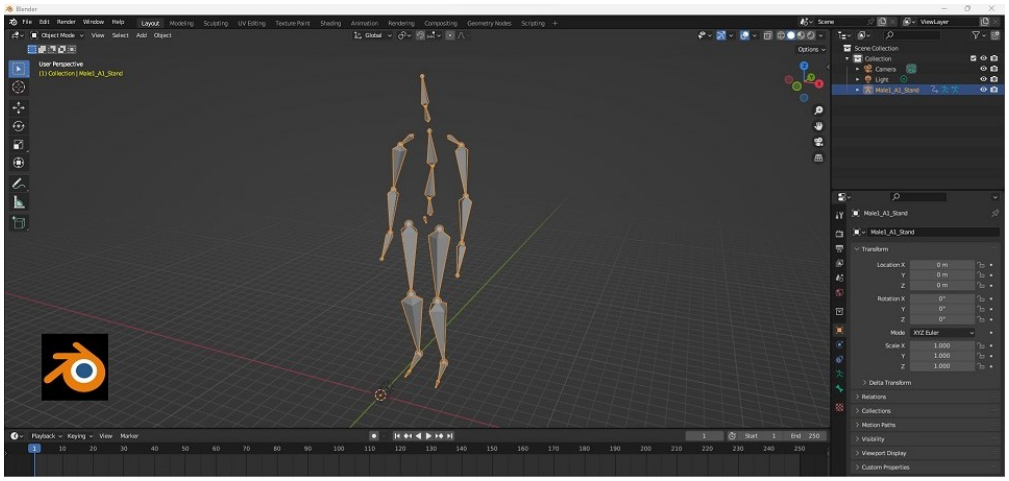


Figure 3.15: Blender 3.5 Software

3.4 Model Architecture

This section describes the design and development of the Artificial Neural Network model that is used to develop the motion fingerprint embeddings.

3.4.1 Model: Single I

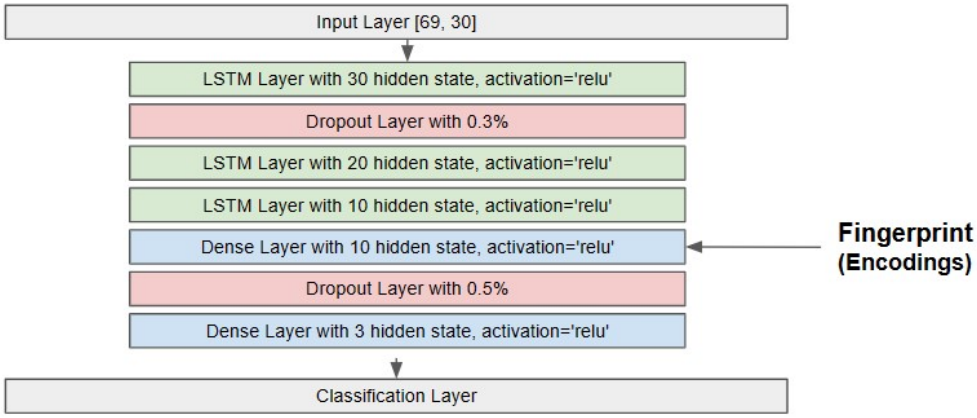


Figure 3.16: Model Single-I Architechture

The Model Single-I is the initial model developed to test the idea of generating a motion fingerprint to classify a specific atomic motion. This model is the first stage in our model population and will be augmented with more layers and classification properties to support the overall goal of the project.

The model is a sequential model consisting of 8 Layers (including the input layer) and made with LSTM cells. Input Layer is with the size of 69: $FeatureVector$ size $\times 30$: standard frame size. LSTM layers are used to capture the temporal relationship between frames in the input data and learn the key feature. Dropout layers are used to confirm that the dataset is not overfitting and is immune to any missing/noisy data. The final classification dense layer has only 3 nodes respectively for representing the classes: $Stand$, $Walk$, and Run.

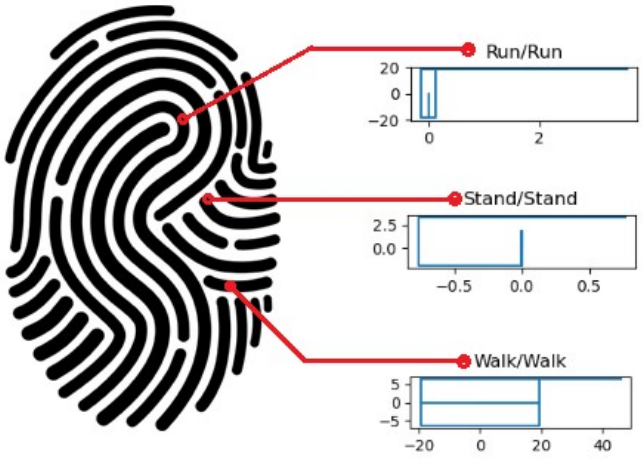


Figure 3.17: Model Single-I Motion Fingerprint

The hidden dense layer at level 6 of the proposed neural network builds and generates the embedding that is dubbed as motion fingerprints. The training caused the layer to learn unique patterns for each atomic motion/action class as shown in Figure 3.17. Since the Model Single-I only concentrates on 3 atomic motion class recognition, the respective fingerprints on the layer build unique and easily identifiable patterns for each atomic action.

3.4.2 Model: Single II

Similarly to the Model Single-I, we designed a new Model Singe-II to just identify the subject based on the atomic actions. The inspiration for the same is that just like motion sequences have unique patterns to identify the class of action, we can also assume that it has an embedded motion sequence that is unique to each individual.

The next iteration of the models will have more classes and deep architecture to recognize additional motion sequences and identity classification together.

3.4.3 Model: Multi I

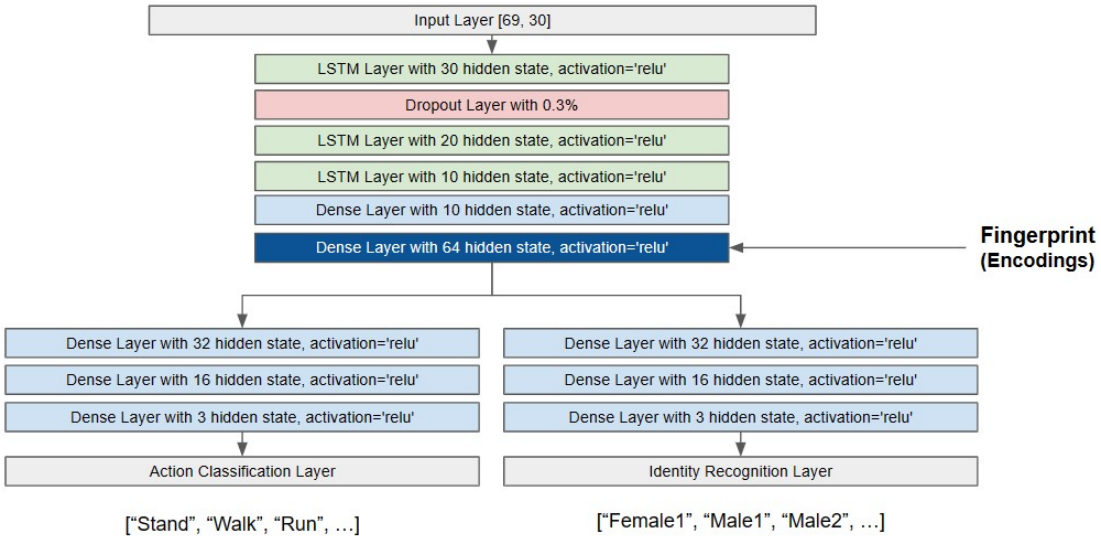


Figure 3.18: Model Multi-I Architechture

The Multi-I, model is the first fusion model of Single-I and Single-II to achieve the goal of generating a global motion fingerprint which is common for both individual identification and atomic action class recognition.

The model has 2 classification net branches from the fingerprint encoding layer and each net has its own 3 layer fully connected network as shown in Figure 3.18. The model Multi-I is also a sequential model consisting of 13 Layers (including the input layer) and made with LSTM cells very similar to Single models. Input Layer is with the size of 69: $FeatureVector$ size $\times 30$: standard frame size. LSTM layers are used to capture the temporal relationship between frames in the input data and learn the key feature. Dropout layers are used to confirm that the dataset is not overfitting and is immune to any missing/noisy data. At the fingerprint encoding dense layer (Layer 7), the net is branching out to two different classification dense layer-based networks consisting of 3 layers to each side of the net. Both branches have only 3 classification nodes respectively for representing the classes: $Stand$, $Walk$, and Run in the action classification Layer and $Female1$, $Male1$, and $Male2$ in the Identity Recognition Layer.

Figure 3.19 shows the generated encodings of motion fingerprint by the encoder part of the network. The model generated fingerprints for each motion sequence: $Stand$, $Walk$, and Run and for Identity recognition of $Female1$, $Male1$, and $Male2$. It has been observed that the fingerprints show a very distinguishable pattern and warrant the use of them as unique encodings to recognize different classes.

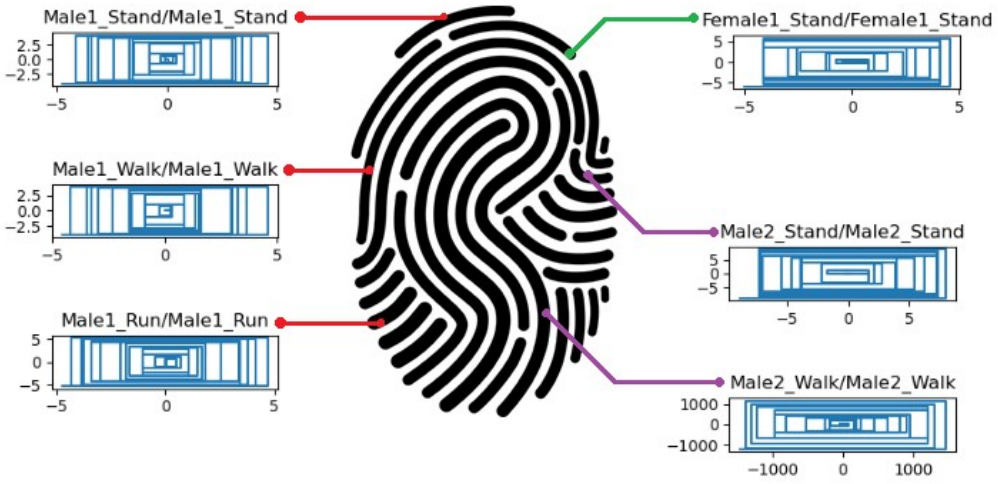


Figure 3.19: Model Multi-I Fingerprints

3.4.4 Model: Multi II

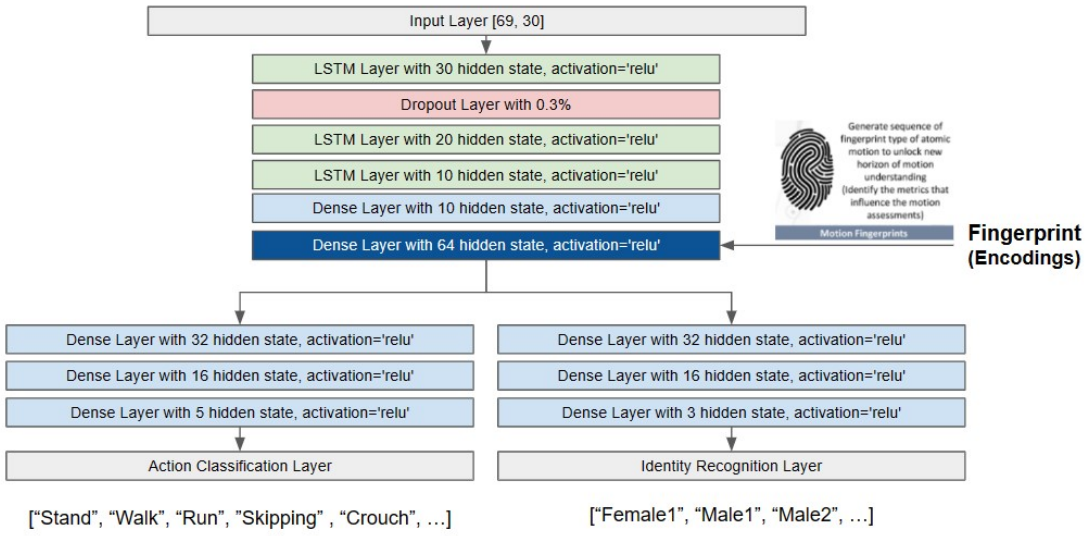


Figure 3.20: Model Multi-II Architechture

Model Multi-II is an extended version of Multi-I with more action classes in the Action Classification Layer as shown in Figure 3.20. This is the final blueprint architecture of our **Famtam AI** and will be evaluated for its performance in later chapters with tweaks in its size of input and classification layers and count of hidden layers.

Famtam AI testing and performance evaluation is provided in Chapter 4 and it explains detailed information about the tweaks made to create the final production model for testing.

3.4.5 Fingerprint Encoder/Decoder Model Training

All of the above models, including the final production **Famtam AI** model were initially trained using the standardized ACCAD data as inputs under supervised machine learning techniques. ADAM optimizer is used to train the model, and we have chosen the categorical cross entropy as the loss function to be optimized in the process. For the training, the mini-batch size is set to 4, and maximum epochs are set to 250. The best model based on the training and validation accuracy is chooses to be the final model.

Chapter 4

Results, Analysis and Evaluation

4.0.1 Model: Berkerly MHAD

We have used an 11-action (11 classes in action classification layer) and 5-subject classification model (5 classes in the Identity classification layer) of the same architecture (Chapter 3.4) that we proposed to evaluate the performance using the test data. We will refer to this 11-action, 5-subject classification model as the proposed **Famtam AI**: **F**usion of **A**i-**M**ocap **T**echnology in **A**ugmenting the human **M**otion model or simply as a "model" / "solution" from here onwards.

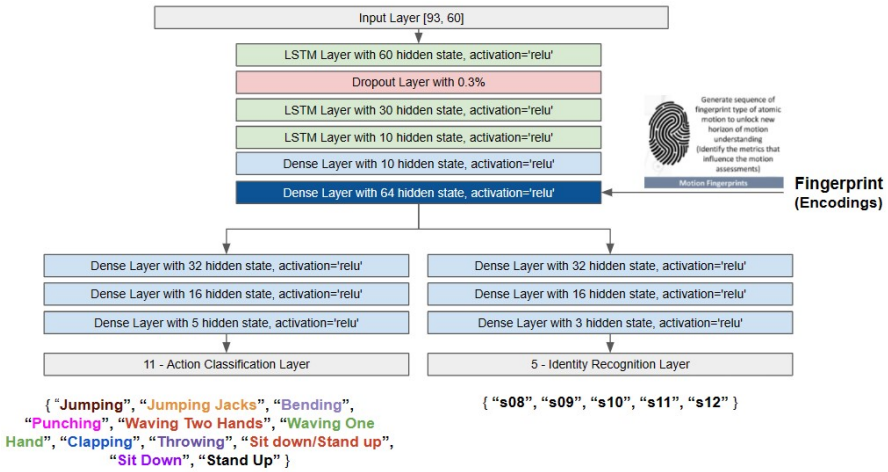


Figure 4.1: BerkerlyMHAD Model: 11-action, 5-subject classification Model

The model has an input size of 93 skeletal joints and 60 frames as compared to models with 69, and 30 skeletal, and frame respectively in Chapter 3.4. The changes are made to accommodate the Berkerly MHAD test data structure as its skeleton joint count is higher (93) and frame rate is 60 frames per second (FPS) while ACCAD (Chapter 3.3.3.2) is only 24 FPS.

4.1 Experiment Setup

The experimental setup aims to evaluate our proposed model using a well-known data set specifically in order to compare the results across the other research works in the domain. This section outlines the data testing process, and evaluation metrics used for assessing the system's performance.

The proposed solution is evaluated over the test dataset based on performance metrics such as accuracy, sensitivity/recall, specificity, and positive predictivity/precision.

$$Accuracy(ACC) = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1.1}$$

$$Recall/Sensitivity(SEN) = \frac{TP}{TP + FN} \tag{4.1.2}$$

$$Specificity(SPE) = \frac{TN}{FP + TN} \tag{4.1.3}$$

$$Precision(PPR) = \frac{TP}{TP + FP} \tag{4.1.4}$$

$$F1 = \frac{2 \times SEN \times PPR}{SEN + PPR} \tag{4.1.5}$$

where, TP = No. of True Positives, FP = No. of False Positives, TN = No. of True Negatives, and FN = No. of False Negatives.

For each test case, we have also provided the detailed performance of the model in various modalities such as using a confusion matrix and tables. The next section will explain the different phases of the model training experiment for comparing the model performance with other existing research works.

The experiment on the Berkerly MHAD test data (Chapter 4.1.1) starts with action recognition evaluation as this has to be performed in a controlled manner such that it can be compared against other research works. Then the same model is evaluated for a combined classification tasks (action and identity) with an extra phase of training which we explain in Chapters 4.1.2 and 4.1.3.

4.1.1 Test Data

Berkerly MHAD [38] dataset is used to evaluate the model performance as it has been widely used in the past and consistent across data collection and records. The data presented here is of homogeneous records (in terms of actions that are separated and labeled) which allow us to perform evaluation and compare against multiple other previous research works. Additional details of the dataset is presented in Chapter 3.3.3.1.

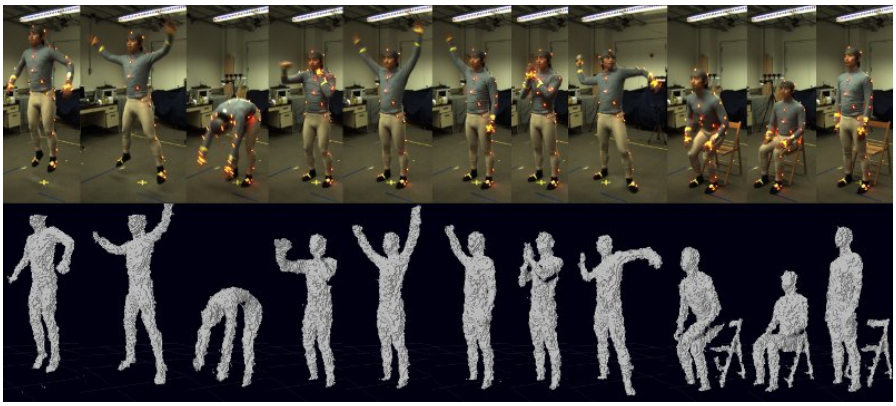


Figure 4.2: Actions recorded in the Berkeley MHAD [38]) along with their corresponding 3D point data obtained using Kinect stations. The figure shows the actions in the database from right to left: Stand Up, Sit Down, Sit Down and Stand Up, Throw, Clap, Wave One Hand, Wave Two Hands, Punch, Bend, Jumping Jacks, Jump.

The dataset consists of 671 motion files named in the format of $skl_sX_aY_rZ.bvh$. "skl" represents that the dataset is a skeletal one, meaning the data has the skeleton hierarchy (Chapter 3.3.1) and its channels for motion presentation. "sX" represents the $s$ubject number, i.e. "s03" means its belongs to subject number 3. There are 12 subjects present in the dataset with 56 records for each including a T-pose record (T-pose records are used to identify the skeleton structure clearly or to initialize the frames in different contexts). "aY" represents the $a$ction class, i.e. "a05" means it belongs to the action class of 5. There are 11 action classes defined in the database such as Jump and etc... as shown in Figure 4.2. Finally, the rZ represents the $r$epetition count of records belonging to the same subject and action class. There are 5 repetitions performed for each sequence of actions. Note this is not the repetition of motion/action within the record itself. The files are downloaded in (BVH) file format [37] for our evaluation, however, the Berkerly MHAD dataset allows you to download files in many different formats.

4.1.2 Action Recognition Evaluation

We strictly follow the similar experimental protocol recommended by [1] on the Berkerly MHAD database. The 384 sequences of the initial seven subjects are used for training whilst the next 275 sequences of the other five subjects are used for the testing procedure. We compare our proposed **Famtam AI** model with Vantigodi et al. proposed Support Vector Machines (SVM) [22], Ofli et al. proposed Sequence of the most informative joints (SMIJ) [38], Vantigodi et al. proposed Meta-Cognitive Radial Basis Function (McRBFN) networks [29], and Kapsouras and Nikolaidis K-Mean based k-Nearest Neighbour and SVM architecture [20] and Du et al. proposed Bi-Directional RNN architecture.

The Action class experimental results are presented in Table 4.1. This is not a part or summary of overall model performance which is provided in its full form in Table 4.3 which has used 70% of test data for training the Identity of each subject. The overall action class recognition accuracy of the proposed **Famtam AI** model is 98.68% with 92.73% sensitivity. The F1 score of the proposed method is 92.73% and it is very well on par with other previous approaches meticulously designed only for activity classification while ours target both the activity and identity recognition.

The table explains that the activities such as $Bending$, $Punching$, and $WavingTwoHands$ achieve the highest accuracy and sensitivity among all others. This is mainly due to the very crisp and clear differentiation in the motion sequences for the activities compared to other activities such as *sitdown/standup*, *sitdown*, and *standup* which share similar motion patterns between them.

Action Class	**TP**	**FP**	**TN**	**FN**	**ACC**	**SEN**	**SPE**	**PPR**	**F1**
01 - jumping	23	2	248	2	98.55%	92.00%	99.20%	92.00%	92.00%
02 - jumping jacks	22	1	249	3	98.55%	88.00%	99.60%	95.65%	91.67%
03 - bending	25	0	250	0	100%	100%	100%	100%	100%
04 - punching	25	0	250	0	100%	100%	100%	100%	100%
05 - waving two hands	25	0	250	0	100%	100%	100%	100%	100%
06 - waving one hand	25	2	248	0	99.27%	100%	99.20%	92.59%	96.15%
07 - clapping	25	1	249	0	99.64%	100%	99.60%	96.15%	98.04%
08 - throwing	24	2	248	1	98.91%	96.00%	99.20%	92.31%	94.12%
09 - sit down/stand up	21	5	245	4	96.73%	84.00%	98.00%	80.77%	82.35%
10 - sit down	19	3	247	6	96.73%	76.00%	98.80%	86.36%	80.85%
11 - stand up	21	4	246	4	97.09%	84.00%	98.40%	84.00%	84.00%
Overall	255	20	2730	20	98.68%	92.73%	99.27%	92.73%	92.73%

Table 4.1: Berkerly MHAD - Action Class Recognition Performance)

4.1.3 Identity Recognition Evaluation

The **Famtam AI** model was trained with 70% of data for identity recognition with a very low learning rate ($\epsilon = 0.00001$) and the rest of the 30% of data is kept for testing the model. Note that this model training is only to evaluate identity recognition or a combination of both action and identity recognition. This phase has not been a part of training in Chapter 4.1.2.

The Identity recognition experimental results are presented in Table 4.2. This is a summary of overall **Famtam AI** model performance which is provided in its full version in Table 4.3. The overall action class recognition accuracy of the proposed **Famtam AI** model is 94.04% with 85.09% sensitivity. The F1 score of the proposed method is 85.09%.

Subject	**TP**	**FP**	**TN**	**FN**	**ACC**	**SEN**	**SPE**	**PPR**	**F1**
s08	**45**	8	212	10	93.45%	81.82%	96.36%	84.91%	83.33%
s09	46	7	213	9	94.18%	83.64%	96.82%	86.79%	85.19%
s10	46	9	211	9	93.45%	83.64%	95.91%	83.64%	83.64%
s11	51	10	210	4	94.91%	92.73%	95.45%	83.61%	87.93%
s12	46	7	213	9	94.18%	83.64%	96.82%	86.79%	85.19%
Overall	234	41	1059	41	94.04%	85.09%	96.27%	85.09%	85.09%

Table 4.2: Berkerly MHAD - Identity Recognition Performance)

The proposed **Famtam AI** model has performed exceptionally well in identifying the subjects that it has been trained with while also accommodating room for action classification. This is proof that our hypothesis of using a motion fingerprint for identity recognition is significant.

4.1.4 Combined Evaluation

We have followed a very similar approach as explained and experimented in Chapter 4.1.2. We have used the **Famtam AI** model, trained using the protocol proposed by [1] as starting point for this experiment. The model was then trained with another 70% of data for identity recognition with a very low learning rate ($\epsilon = 0.00001$) and the rest of the 30% of data is kept for testing the model as previously explained in Chapter 4.1.3. Note that this model training is an extra phase of training only to evaluate identity recognition or a combination of both action and identity recognition. This phase has not been the part of training in Chapter 4.1.2.

Table 4.3 provides the full view of each test record performance while Figure 4.3 expresses the Confusion Matrix plot for both subject and action recognition.

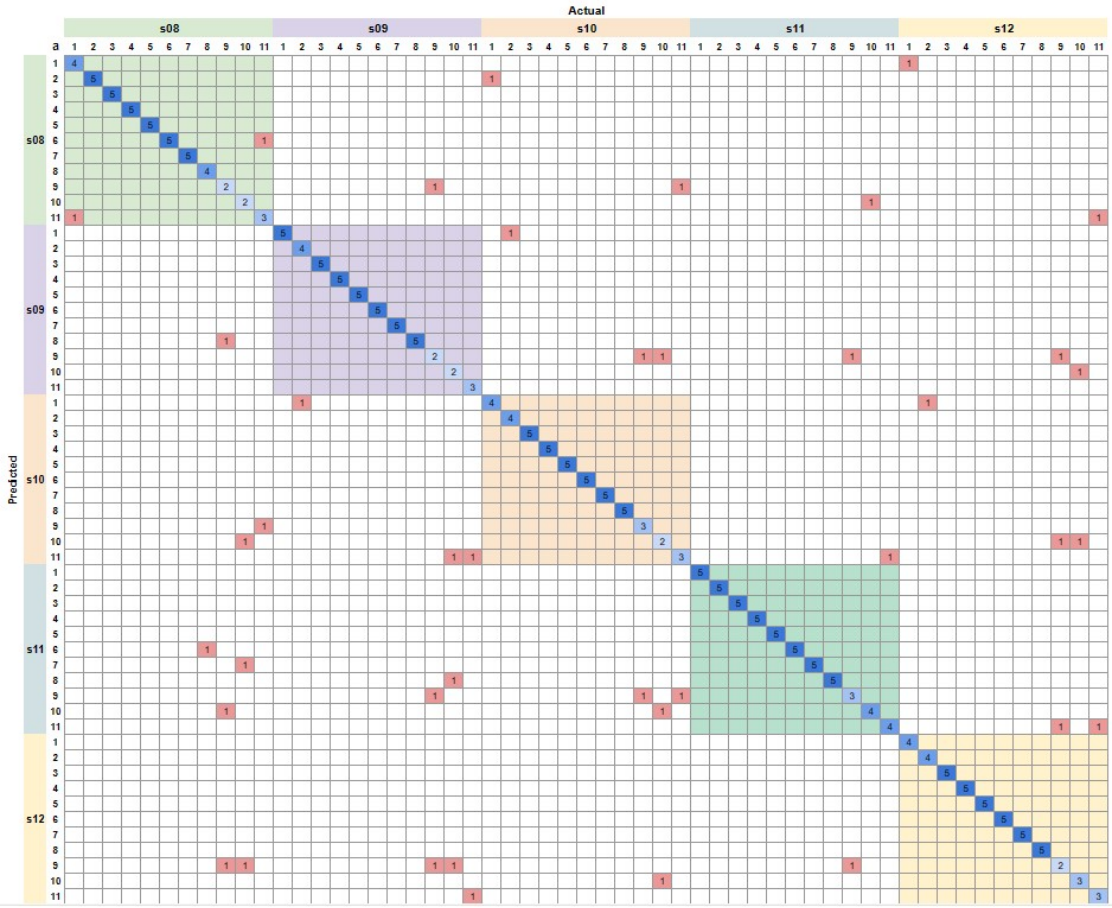


Figure 4.3: Confusion Matrix for the Overall test results. sX refers to the Xth subject in the test and the rest of the 1-11 is for action classes respectively. Refer to Chapter Chapter 3.3.3.1 and Chapter 4.1.1 for additional details on classes.

The test on the proposed **Famtam AI** model achieved an overall accuracy of 99.46% with the sensitivity of 85.09%. This serves as initial concrete evidence to prove our hypothesis about using a global motion fingerprint for action-based subject recognition tasks.

Table 4.3: Berkerly MHAD [1] Overall Test Results
(Action and Identity Recognition)

Subject	Action	TP	FP	TN	FN	ACC	SEN	SPE	PPR	F1
s08	1	4	1	269	1	99.3%	80.0%	99.6%	80.0%	80.0%
	2	5	1	269	0	99.6%	100.0%	99.6%	83.3%	90.9%
	3	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	4	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	5	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	6	5	1	269	0	99.6%	100.0%	99.6%	83.3%	90.9%
	7	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	8	4	0	270	1	99.6%	80.0%	100.0%	100.0%	88.9%
	9	2	2	268	3	98.2%	40.0%	99.3%	50.0%	44.4%
	10	2	1	269	3	98.5%	40.0%	99.6%	66.7%	50.0%
	11	3	2	268	2	98.5%	60.0%	99.3%	60.0%	60.0%
	1	5	1	269	0	99.6%	100.0%	99.6%	83.3%	90.9%
	2	4	0	270	1	99.6%	80.0%	100.0%	100.0%	88.9%
	3	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	4	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	5	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%

Table 4.3: Berkerly MHAD [1] Overall Test Results
(Action and Identity Recognition)

Subject	Action	TP	FP	TN	FN	ACC	SEN	SPE	PPR	F1
s09	6	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	7	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	8	5	1	269	0	99.6%	100.0%	99.6%	83.3%	90.9%
	9	2	4	266	3	97.5%	40.0%	98.5%	33.3%	36.4%
	10	2	1	269	3	98.5%	40.0%	99.6%	66.7%	50.0%
	11	3	0	270	2	99.3%	60.0%	100.0%	100.0%	75.0%
s10	1	4	2	268	1	98.9%	80.0%	99.3%	66.7%	72.7%
	2	4	0	270	1	99.6%	80.0%	100.0%	100.0%	88.9%
	3	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	4	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	5	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	6	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	7	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	8	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	9	3	1	269	2	98.9%	60.0%	99.6%	75.0%	66.7%
	10	2	3	267	3	97.8%	40.0%	98.9%	40.0%	40.0%
	11	3	3	267	2	98.2%	60.0%	98.9%	50.0%	54.5%
	1	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	2	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%

Table 4.3: Berkerly MHAD [1] Overall Test Results
(Action and Identity Recognition)

Subject	**Action**	**TP**	**FP**	**TN**	**FN**	**ACC**	**SEN**	**SPE**	**PPR**	**F1**
	3	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	4	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	5	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
s11	6	5	1	269	0	99.6%	100.0%	99.6%	83.3%	90.9%
	7	5	1	269	0	99.6%	100.0%	99.6%	83.3%	90.9%
	8	5	1	269	0	99.6%	100.0%	99.6%	83.3%	90.9%
	9	3	3	267	2	98.2%	60.0%	98.9%	50.0%	54.5%
	10	4	2	268	1	98.9%	80.0%	99.3%	66.7%	72.7%
	11	4	2	268	1	98.9%	80.0%	99.3%	66.7%	72.7%
	1	4	0	270	1	99.6%	80.0%	100.0%	100.0%	88.9%
	2	4	0	270	1	99.6%	80.0%	100.0%	100.0%	88.9%
	3	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	4	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	5	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
s12	6	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	7	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	8	5	0	270	0	100.0%	100.0%	100.0%	100.0%	100.0%
	9	2	5	265	3	97.1%	40.0%	98.1%	28.6%	33.3%
	10	3	1	269	2	98.9%	60.0%	99.6%	75.0%	66.7%

Table 4.3: Berkerly MHAD [1] Overall Test Results
(Action and Identity Recognition)

Subject	**Action**	**TP**	**FP**	**TN**	**FN**	**ACC**	**SEN**	**SPE**	**PPR**	**F1**
	11	3	1	269	2	98.9%	60.0%	99.6%	75.0%	66.7%
Overall		234	41	14809	41	99.46%	85.09%	99.72%	85.09%	85.09%

4.2 Comparison of Results and Analysis

We conducted our experiment using the proposed training protocol in Chapter 4.1.2 and compared it against many existing approaches that we have analyzed in Chapter 2.

Table 4.4 provides a comprehensive outlook of the comparison with existing approaches where enough informations are available on Berkerly Multi-modal Human Action Detection (MHAD) [1] test data.

Reference	**Author**	**Model**	**Accuracy**
[22]	Vantigodi et al.	Support Vector Machines (SVM)	95
[38]	Ofli et al.	Sequence of the most informative joints (SMIJ)	95.37
[29]	Vantigodi et al.	Meta-Cognitive Radial Basis Function (McRBFN) Network	97.58
[20]	Kapsouras and Nikolaidis	K-Mean with k-Nearest Neighbour and SVM	98.18
[24]	Du et al.	Bi-Directional RNN	100
Proposed **Famtam AI** model		LSTMs	98.68

Table 4.4: Comparison of our proposed model with existing research works using Berkerly Multi-modal Human Action Detection dataset (MHAD) [1] dataset.

As detailed in the above Table 4.4, we achieved a very high accuracy of 98.68% (Table 4.1) when compare to other models with Vantigodi et al. "SVM" [22], Ofli et al. "SMIJ" [38], Vantigodi et al. "McRBFN" [29], and Kapsouras and Nikolaidis[20] kNN which achieved an accuracy of 95%, 95.37%, 97.58% and 98.18% respectively. Du et al. proposed Bi-Directional RNN top the score with almost 100% accuracy in the test making our future work to lean towards experimenting with Bi-Direction LSTMs.

The highest False Positive occurred in the class "Sit Down/Stand Up", and this result is already anticipated as "Sit Down/Stand Up" class and "Sit Down", and "Stand Up" classes share very similar atomic motion fingerprints. The distinct motions such as "Bending", "Punching", and "Waving Two Hands" classes achieved very high accuracy,

sensitivity, and specificity of 100%. The power of the Famtam AI model in action recognition is tested positive by these results and is promisable for future application.

4.3 Discussion

The performance of the proposed **Famtam AI** model on the Berkerly MHAD dataset is a stellar performance achieving an overall accuracy of more than 99% according to the test results shown in Table 4.3. This has proved our hypothesis about proposing global motion fingerprints in action and identity class recognition dimensions. Given the person's action sequences, it is proven that there exist unique motion atomic patterns that specifically belong to each individual which can be further utilized to identify them in another instance.

Although, the overall result seems to be promising, the specific Identity recognition evaluation with 70% test data in the training batch for identifying a subject achieved around 94% accuracy only. On the other hand, in practice, the availability of such motion sequences and identity information for people in order to use this kind of technology as a substitute for conventional identity recognition technologies such as face recognition is not only hard but also not very economically sound.

The associated risk of data collection and privacy issues serves the top in limiting the usage of these kinds of identity recognition. However, in the increasing trend of wearable technology, most smartwatches, and wearables are already armed with many different sensors to provide different data modalities which can be used to fuse motion recognition. Therefore, the future seems to be bright for this approach and need more test and evaluation to come to arrive at a concrete conclusion.

Chapter 5

Legal, Social, Ethical and Professional Issues

This research project, the "FamtamAI" has given detailed attention on every stage of data collection, preparation, and training, while paramount importance has been shown to comply with the "Code of Conduct" and "Code of Good Practice" set out by the British Computer Society (BCS) [50].

5.0.1 Legal:

The data used to implement the model is open source and great effort has been shown to ensure the data is in the open public domain and an open license which allows it to be used for the project.

- Berkerly MHAD dataset [1]:
  The Berkeley MHAD database which has been used to test the model and compare it with existing research work is released under BSD-2 [51] license
- ACCAD dataset [39]:
  The ACCAD database which has been extensively used in the preliminary studies and model building phase in our research work is a project by The Ohio State University and licensed under the Creative Common Attribution 3.0 Unported License [52].
- bvh-python Github repo [48]:
  The bvh-python master Github repo is published under MIT License [53] for use without any restriction.
- Blender software:
  We have used blender animation software in various stages of the project to ensure data compatibility and fairness before feeding into our model training pipelines. The software is distributed under the GNU General Public License (GPL, or "free

software") [54] and it is free to use and share.

The rule "have due regard for public health, privacy, security and wellbeing of others and the environment" has been abided by here.

5.0.2 Social

There have been no other social issues identified with this project roadmap. In this project, there has been no other involvement of human subjects as we have used only the publicly available dataset for the studies.

Further studies may require ongoing engagement with society and social standards as they may involve human trials and therefore more concern should be given in near future works.

5.0.3 Ethical

The public domain data has been ethically acknowledged as the original source and credits have been provided to the researchers. Data integrity has been ensured by avoiding misrepresentations and modifications.

Chapter 6 Bibliography cites all such sources for our research work and acknowledges the scientific community's hard work. In any case, due to human errors, if we have missed any citation to the source, we would like to express that this is not intentional and would like to take immediate action upon request and will make necessary changes.

5.0.4 Professional

Integrity has been followed during the design, implementation, and report of the project. Open-source libraries and tools have been stated and the images, and design have been cited in the report explicitly. As a result of this, the code of "have due regard for the legitimate rights of Third Parties" have been complied with.

Chapter 6

Conclusion and FutureWorks

AI's are seamlessly integrated into our everyday lifestyle and play trivial but vital roles like revamping our languages to assisting us in moon shot missions by making intelligent and calculated moves. Motion Capture technologies have moved out from their status quo of being an unaffordable military-grade technology to an Over-The-Counter technology with affordable wearable and vision-based Motion capture technologies that have been unlocked by the power of AI.

In this research work, we have worked in order to test the hypothesis of using a global motion fingerprint (Chapter 1.7) for multi-tasking such as action class recognition and identity recognition and etc... In order to test the states hypothesis we launched this research project called "**Famtam AI**" and have designed a neural network architecture that can simultaneously predict the action class as well as identify the subject doing that action by recognizing their motion sequences. The idea behind the hypothesis is that every person has their own unique pattern of motion when it's come to doing an activity.

Preliminary studies and experimental results have proven that our hypothesis about using the global motion fingerprint to identify a person is valid and paved a new foundation to open up a whole new horizon of novel research. We believe that the future work of this continued research may disclose many key important metrics and knowledge in the same domain.

We strongly believe that "**Famtam AI**" project has contributed to the research domain in many different ways by Proposing a novel idea of Global Motion Fingerprint for both activity classification and identity recognition, Proposing a working neural network model "Famtam AI" for recognizing the global motion fingerprint and tested it under the standard protocol, and Contributed to open-source motion projects such as bvh-python Github repositories.

AI is a fascinating tool and Motion Capturing is a mesmerising technology. At the "Famtam AI" project we have tried to fuse both for the betterment of mankind. The

future of this work may involve the collection and creation of data specific to the project's objectivity and building better robust models tested under more generalized conditions. We believe that we are armed to make positive changes to society and eco system.

Bibliography

[1] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53–60, 2013.

[2] P. Borges, "Deep learning: Recurrent neural networks." `https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010`, Oct 2018. Accessed: 2023-7-15.

[3] Prabhu, "Understanding of convolutional neural network (cnn) — deep learning." `https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148`, Mar 2018. Accessed: 2023-7-15.

[4] OpenAI, "Openai charter." `https://openai.com/charter`, Apr 2018. Accessed: 2023-7-18.

[5] T. N. Y. T. Morozov, Evgeny, "The true threat of artificial intelligence." `https://www.nytimes.com/2023/06/30/opinion/artificial-intelligence-danger.html`, Jun 2023. Accessed: 2023-7-18.

[6] A. S. Reuter and M. Schindler, "Motion capture systems and their use in educational research: Insights from a systematic literature review," *Education Sciences*, vol. 13, no. 2, 2023.

[7] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports medicine-open*, vol. 4, no. 1, pp. 1–15, 2018.

[8] S. Barris and C. Button, "A review of vision-based motion analysis in sport," *Sports Medicine*, vol. 38, pp. 1025–1043, 2008.

[9] L. Mündermann, S. Corazza, and T. P. Andriacchi, "The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications," *Journal of neuroengineering and rehabilitation*, vol. 3, no. 1, pp. 1–11, 2006.

[10] J. Alderson, "A markerless motion capture technique for sport performance analysis and injury prevention: Toward a 'big data', machine learning future," *Journal of Science and Medicine in Sport*, vol. 19, p. e79, 2015.

[11] A. C. Alarcón-Aldana, M. Callejas-Cuervo, and A. P. L. Bo, "Upper limb physical rehabilitation using serious videogames and motion capture systems: A systematic review," *Sensors*, vol. 20, no. 21, p. 5989, 2020.

[12] W. Wei, Z. Qin, B. Yan, and Q. Wang, "Application effect of motion capture technology in basketball resistance training and shooting hit rate in immersive virtual reality environment," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[13] O. Mirabella, A. Raucea, F. Fisichella, and L. Gentile, "A motion capture system for sport training and rehabilitation," in *2011 4th International Conference on Human System Interactions, HSI 2011*, pp. 52–59, 2011.

[14] S. Holden and R. Y. Sunindijo, "Technology, long work hours, and stress worsen work-life balance in the construction industry," *International Journal of Integrated Engineering*, vol. 10, no. 2, 2018.

[15] R. Fleck, A. L. Cox, and R. A. Robison, "Balancing boundaries: Using multiple devices to manage work-life balance," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3985–3988, 2015.

[16] T. Cloete and C. Scheffer, "Benchmarking of a full-body inertial motion capture system for clinical gait analysis," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4579–4582, 2008.

[17] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[18] X. Jin and J. Han, *K-Means Clustering*, pp. 563–564. Boston, MA: Springer US, 2010.

[19] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "Knn model-based approach in classification," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (R. Meersman, Z. Tari, and D. C. Schmidt, eds.), (Berlin, Heidelberg), pp. 986–996, Springer Berlin Heidelberg, 2003.

[20] I. Kapsouras and N. Nikolaidis, "Action recognition on motion capture data using a dynemes and forward differences representation," *Journal of Visual Communication and Image Representation*, vol. 25, pp. 1432–1445, aug 2014.

[21] A. M. Deris, A. M. Zain, and R. Sallehuddin, "Overview of support vector machine in modeling machining performances," *Procedia Engineering*, vol. 24, pp. 308–312, 2011. International Conference on Advances in Engineering 2011.

[22] S. Vantigodi and R. V. Babu, "Real-time human action recognition from motion capture data," *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pp. 1–4, 2013.

[23] N. K. Manaswi, *RNN and LSTM*, pp. 115–126. Berkeley, CA: Apress, 2018.

[24] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1110–1118, 2015.

[25] Y. R., N. M., and D. R. et al., "Convolutional neural networks: an overview and application in radiology," vol. 9, p. 611–629, 2018.

[26] T. Soo Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 20–28, 2017.

[27] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *arXiv preprint arXiv:1804.06055*, 2018.

[28] P. Strumiłło and W. Kamiński, "Radial basis function neural networks: Theory and applications," in *Neural Networks and Soft Computing* (L. Rutkowski and J. Kacprzyk, eds.), (Heidelberg), pp. 107–119, Physica-Verlag HD, 2003.

[29] S. Vantigodi and V. B. Radhakrishnan, "Action recognition from motion capture data using meta-cognitive rbf network classifier," *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)*, pp. 1–6, 2014.

[30] L. Narens and T. O. Nelson, "Metamemory: A theoretical framework and new findings," 1992.

[31] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[32] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 103–118, 2018.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[34] Y. Zhang, B. Wu, W. Li, L. Duan, and C. Gan, "Stst: Spatial-temporal specialized transformer for skeleton-based action recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3229–3237, 2021.

[35] Autodesk, ".FBX - The biomechanics standard file format." https://www.autodesk.com/products/fbx/overview, Jun 2021. Accessed 15-Jul-2023.

[36] I. Motion Lab Systems, "C3D.ORG - The biomechanics standard file format — c3d.org." https://www.c3d.org/, Jul 2023. Accessed 15-Jul-2023.

[37] Biovision, "BVH motion capture data." https://www.cs.cityu.edu.hk/~howard/Teaching/CS4185-5185-2007-SemA/Group12/BVH.html, Jul 2023. Accessed: 2023-7-15.

[38] F. Ofli, R. A. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," *2013 IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 53–60, 2013.

[39] ACCAD, "Motion Lab: Mocap system and data — advanced computing center for the arts and design (osu.edu)." https://accad.osu.edu/research/motion-lab/mocap-system-and-data, Jul 2023. Accessed: 2023-7-15.

[40] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "Amass: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

[41] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriu, L. Romeo, and F. Verdini, "The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 7, pp. 1436–1448, 2019.

[42] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, 2018.

[43] A. Miron, N. Sadawi, C. Grosan, W. Ismail, and H. Hussain, "IntelliRehabDS - A dataset of physical rehabilitation movements," Mar. 2021. GitHub repository for basic exploratory data analysis: https://github.com/alina- miron/intellirehabds.

[44] W. Li, "MSR Action 3D." https://sites.google.com/view/wanqingli/data-sets/msr-action3d, Jul 2023. Accessed: 2023-7-22.

[45] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pp. 9–14, IEEE, 2010.

[46] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 20–27, IEEE, 2012.

[47] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[48] 20tab srl, "BVHPythonMaster: Python module for parsing bvh (biovision hierarchical data) mocap files." https://github.com/20tab/bvh-python, Jul 2019. Accessed: 2023-7-15.

[49] Blender, "Free and open source software licensed as gnu gpl, owned by its contributors." https://www.blender.org/, Jul 2023. Accessed: 2023-7-15.

[50] T. C. I. f. I. BCS, "Code of conduct for bcs members." https://www.bcs.org/media/2211/bcs-code-of-conduct.pdf, Jun 2022. Accessed: 2023-7-15.

[51] open source initiative, "The 2-clause bsd license." https://opensource.org/license/bsd-2-clause/, Jun 2022. Accessed: 2023-7-15.

[52] C. C. CORPORATION, "Creative commons attribution 3.0 unported." https://creativecommons.org/licenses/by/3.0/legalcode. Accessed: 2023-7-18.

[53] MIT, "Mit license." https://github.com/20tab/bvh-python/blob/master/LICENSE. Accessed: 2023-7-18.

[54] I. Free Software Foundation, "Gnu general public license." https://www.blender.org/about/license/, 1991. Accessed: 2023-7-18.

.1 Appendix

.1.1 Python Code and Contribution

We have used the bvh-python-master repository [37] to read our data in '.bvh' format. We have also contributed to the branch by adding a special utility function to "*get channel names*" of skeletal joints (Refer to Figure 1).

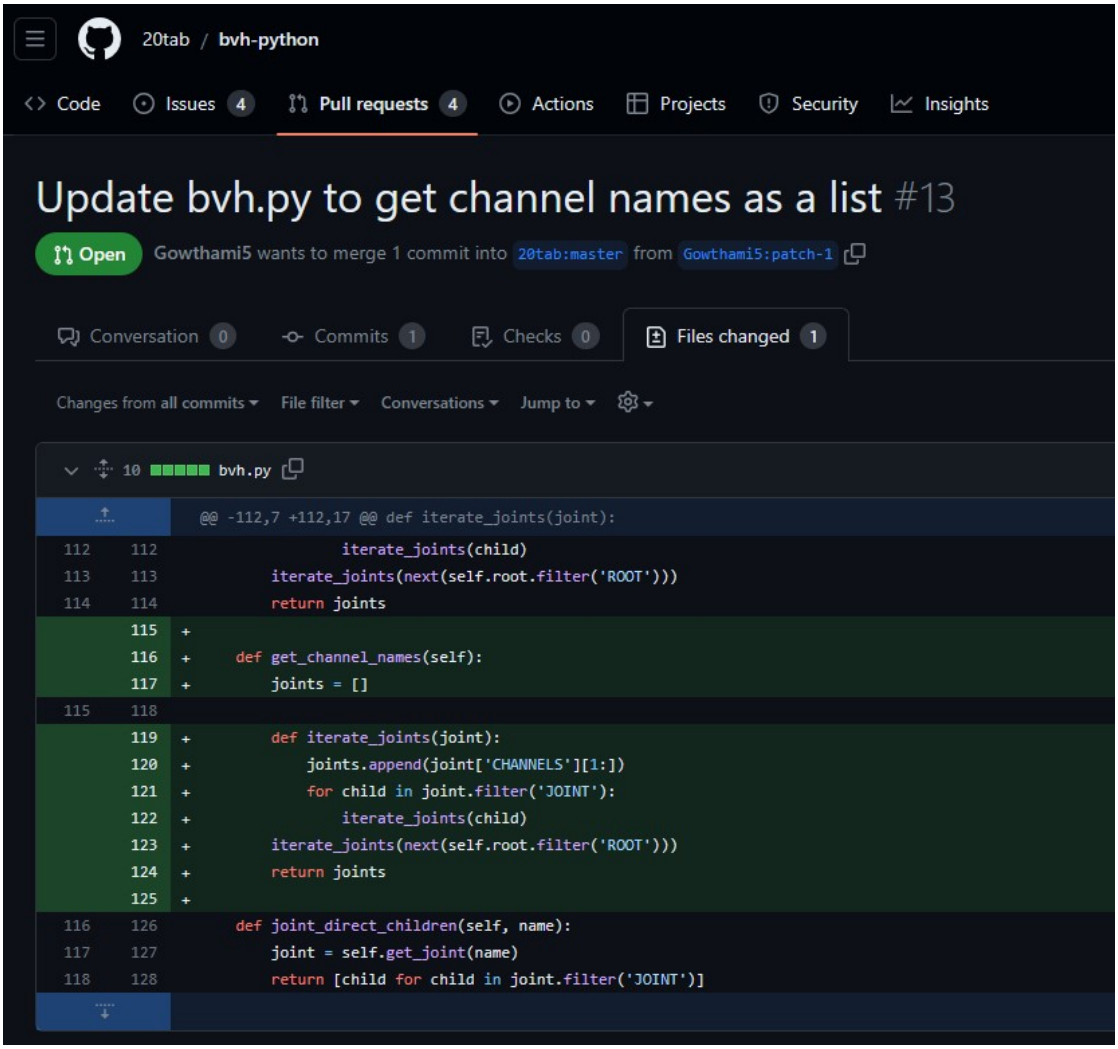


Figure 1: Contribution to BVH-python-master Github repo.

The single file read function, The function to process the mocap data from a single file read and the function to generate the models are provided below.

.1.2 Function: getBVHnumpy

```
def getBVHnumpy(filepath):
    mocap = readBVH(filepath)
    Frames, Frame_Title = processFrames(mocap)
    # print(Frame_Title)
    # print(Frames)
    return Frames, Frame_Title

def readBVH(file_name):
    with open(file_name) as f:
        # utility function from bvh-python-master repo
        mocap = Bvh(f.read())
    return mocap

def processFrames(mocap):

    nFrames = len(mocap.frames)
    nSize = len(mocap.frames[0])
    # print('Input Frame size:', nFrames, nSize)
    jointNames = mocap.get_joints_names()
    #print('Joint Names:', jointNames)
    channelNames = mocap.get_channel_names()
    #print('Channel Names:', channelNames)
    Frame_Title = ["_".join([jointNames[0], a]) \
                   for a in channelNames[0][:3]]
    Frame_Title.extend(["_".join([b,a]) \
                   for b in jointNames \
                   for list in channelNames for a in list[3:]])
    #print('RotFrame_Title:', Frame_Title)

    Frames = np.zeros((nSize, nFrames))
    i_cols, i_rows = 0,0
    for cols in mocap.frames:
        i_rows = 0
        for rows in cols:
            Frames[i_rows, i_cols] = float(rows)
            i_rows = i_rows+1
        i_cols = i_cols + 1

    return Frames, Frame_Title
```

.1.2.1 Function: Pad frames for equal size input

```
def padNumpyFrame(Frame, lenTs, no_of_channels, subsamplingFactor):

    lenFrame = int(Frame.shape[1]/subsamplingFactor)

    if (lenFrame % lenTs)/lenTs >= 0.8:
        shape = (int(lenFrame/(lenTs)) + 1, lenTs, no_of_channels)
    else:
        shape = (int(lenFrame / (lenTs)), lenTs, no_of_channels)

    retNpArr = np.zeros(shape)

    # print(Frame.shape[1])
    for i in range(shape[0]):
        adjLenFrame = lenFrame * subsamplingFactor
        adjlenTS = lenTs * subsamplingFactor
        if lenFrame <= lenTs:
            retNpArr[i,-lenFrame:,:] = Frame[:, \
                    -adjLenFrame::subsamplingFactor,:]
        else:
            retNpArr[i,:,:] = Frame[:,i*adjlenTS: \
                    (i+1)*adjlenTS:subsamplingFactor,:]
            lenFrame = lenFrame - lenTs

    return retNpArr
```

.1.2.2 Function: Standardise frames with degrees and distance vector

```
def standardizeFrames(Frames):

    nSample, nFrames, nSize = Frames.shape

    FramesX = np.zeros((nSample, nFrames, nSize))
    for ii_samp in range(nSample):
        for ii_cols in range(nFrames):
            for ii_rows in range(nSize):
                if ii_rows < 3:
                    FramesX[ii_samp, ii_cols, ii_rows] = \
                    Frames[ii_samp, ii_cols, ii_rows] \
                    - Frames[ii_samp, 0, ii_rows]
                else:
                    FramesX[ii_samp, ii_cols, ii_rows] = \
                    (Frames[ii_samp, ii_cols, ii_rows] \
                    - Frames[ii_samp, 0, ii_rows]) * math.pi / 180

    FramesX = np.array(FramesX)

    return FramesX
```

.1.2.3 Function: model

```
def model(lenTs, lenFs, No_Act_class = 11, No_ID_class = 5):
    # Define model layers.
    input_layer = keras.Input(shape=(lenTs, lenFs,))
    input_lstm = LSTM(30, activation='relu', \
                      return_sequences=True, \
                      input_shape=(lenTs, lenFs))(input_layer)
    first_dropout = Dropout(0.3)(input_lstm)
    second_lstm = LSTM(20, activation='relu', \
                      return_sequences=True)(first_dropout)
    third_lstm = LSTM(10, activation='relu')(second_lstm)
    pre_fingerprint_dense1 = Dense(10, activation='relu')(third_lstm)
    fingerprint_layer = Dense(64, activation='relu', \
                      name='fingerprint')(pre_fingerprint_dense1)
    fingerprint_dropout = Dropout(0.3)(fingerprint_layer)

    class_dense1 = Dense(32, activation='relu')(fingerprint_dropout)
    class_dense2 = Dense(16, activation='relu')(class_dense1)
    class_output = Dense(No_Act_class, activation='softmax', \
                      name='class_output')(class_dense2)

    identity_dense1 = Dense(32,activation='relu')(fingerprint_layer)
    identity_dense2 = Dense(16, activation='relu')(identity_dense1)
    identity_output = Dense(No_ID_class,activation='softmax', \
                      name='identity_output')(identity_dense2)
    # Define the model with the input layer
    # and a list of output layers
    model = keras.Model(inputs=input_layer, \
                      outputs=[class_output, identity_output])
    # Specify the optimizer, and compile the model with
    # loss functions for both outputs
    model.compile(optimizer='adam', \
                loss={'class_output': 'categorical_crossentropy', \
                'identity_output': 'categorical_crossentropy'}, \
                  metrics={'class_output': 'accuracy', \
                           'identity_output': 'accuracy'})
    return model
```

.1.3 BCS Women Lovelace Colloquium 2023 - Poster

The poster is available on the next page.

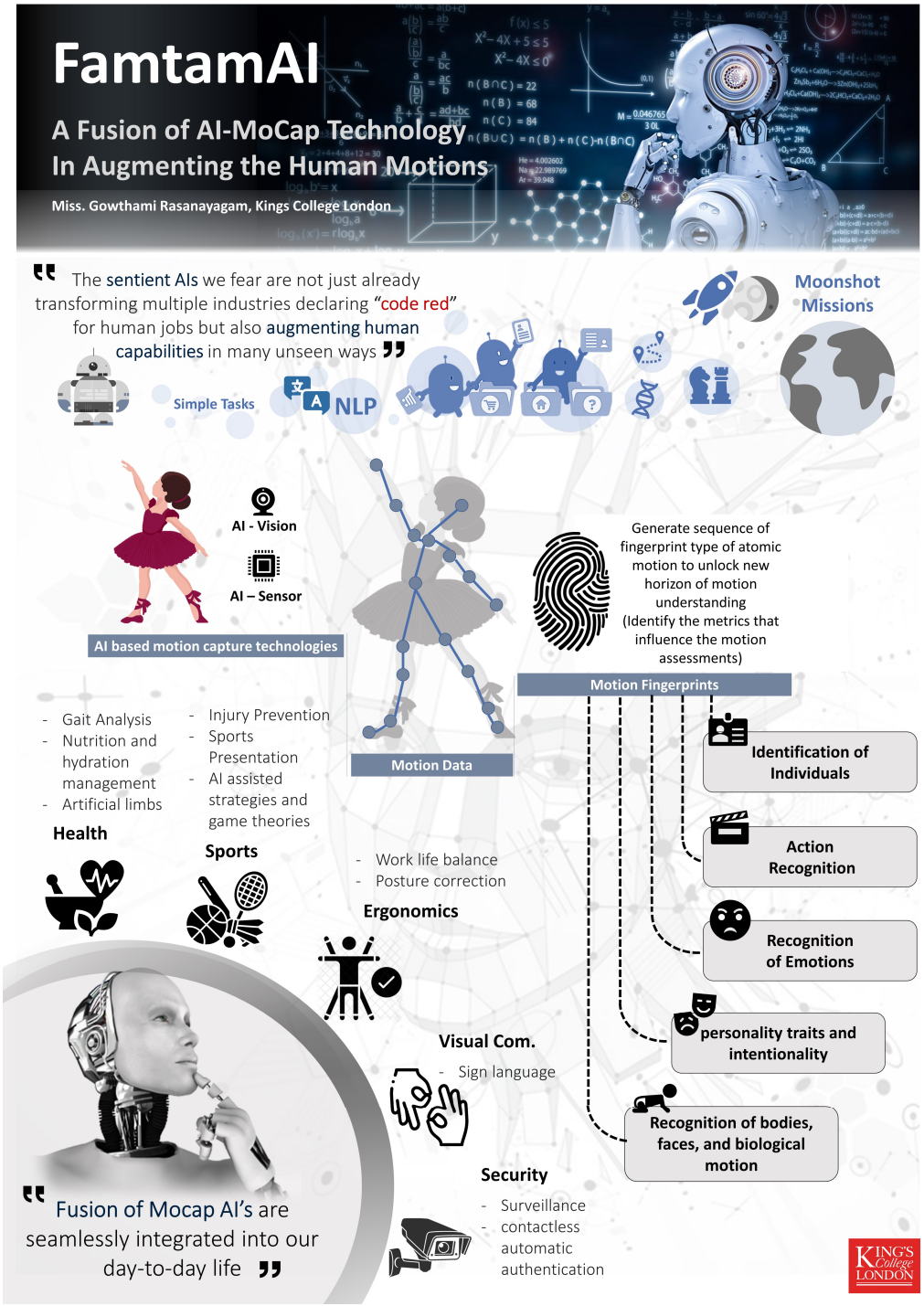


Figure 2: BCSWomen Lovelace Colloquium 2023
The poster was presented in the BCSWomen Lovelace Colloquium 2023 organized by the BCSWomen, The Chartered Institute for IT. The poster theme focused on the fusion of AI-MoCap technology in augmenting human motion.