Human and activity detection in ambient assisted living scenarios

By

MOHAMAD REZA SHAHABIAN ALASHTI

School of Physics, Engineering and Computer Science

UNIVERSITY OF HERTFORDSHIRE

A thesis submitted to the University of Hertfordshire
in partial fulfilment of the requirement of the degree of
DOCTOR OF PHILOSOPHY.

APRIL 2024

Abstract

Human activity recognition (HAR) is crucial in assistive technology and human-robot interaction (HRI) as it enables robots and assistive devices to understand and respond to the individual's movements and actions, facilitating personalized assistance for those with mobility challenges or disabilities. In the context of HAR for ambient assisted living (AAL) environments, integrating additional cameras with the robot's perspective has the potential for significant advancement of detection outcomes. However, considering the computational limits in robots, the caveat is that processing additional video streams presents challenges in terms of both computation complexity and data integration.

The primary goal of this research is to create an efficient multi-view skeleton-based HAR system that optimises accuracy without sacrificing efficiency. By leveraging the strengths of the skeleton-based models and incorporating diverse perspectives, this system aims to enhance overall performance in AAL scenarios. To support this goal, an open dataset for skeletal data in HAR is developed and utilised. For objective evaluation, this work considers computational needs and algorithmic efficiency in HAR methods, exploring the potential of multi-view systems to improve human-robot interaction.

This thesis is grounded on a thorough literature review including extensive dataset analysis. It explores pivotal research questions centred around the effectiveness of skeleton-based models in multi-view settings compared to image-based models. It explores the role of perspectives in multi-view HAR and investigates the optimal models for multi-view recognition. These inquiries lead to the development and evaluation of a novel lightweight and multi-view HAR architecture.

This thesis significantly contributes to the field by introducing a multi-view skeleton-based dataset, dataset analysis metrics to evaluate and compare different perspectives, and a novel lightweight HAR architecture. Performance analysis supports the importance of integrating robot vision with observations from additional cameras. These results reveal variations in performance based on different views. For instance, the results highlight how the robot's tracking of the human subject during action performance can lead to higher data collection quality in activities like climbing stairs up and down, while proximity to the subject may result in missed body joints. Conversely, other views offer a wider perspective of the scene, presenting unique advantages and challenges. In this study, integrating the additional view with the Robot-view resulted in an accuracy increase of up to 25%.

Notably, the proposed skeleton-based architecture exhibits improved efficiency compared to its image-based counterpart when applied to the same dataset, demonstrating a notable 15% improvement in HAR. Moreover, the comparison between image-based and skeleton-based methods reveals that the robot movement affects them differently. Unlike image-based methods, where the movement of the robot's camera can create confusion between the subject's movement and the camera's movement, the skeleton-based method is less affected by the robot's movement.

In addition, this work introduces various multi-view architectures for comparative analysis,

shedding light on different data combination methodologies. The proposed system achieves high accuracy (approximately 90%), with a minimal number of training parameters (0.6M), and demonstrates significantly lower computational demands (0.00106 Giga FLOPs)compared to well-known CNN and GCN models. For instance, the ResNet models with 11.2M and MobiNet with 2.2M parameters achieved 90.9% and 83.5% accuracy, respectively.

Given the pivotal role of HAR in diverse applications, the emphasis on its efficiency and effectiveness is crucial. This work not only addresses these concerns but also establishes a foundation for future research directions. The proposed skeleton-based architecture lays the groundwork for various applications such as ambient assisted living scenarios, offering a flexible platform for the development of efficient multi-person activity recognition, continual and real-time HAR, and utilisation in Human-Robot Interaction (HRI) studies.

DEDICATION AND ACKNOWLEDGEMENTS

I dedicate this PhD to my beloved parents, whose unwavering support and boundless love have been my guiding light throughout this journey. Their selflessness and encouragement have inspired me to reach for the stars and pursue my dreams without hesitation or expectation. To my cherished wife, thank you for being my steadfast companion and trusted friend, offering unwavering support and understanding every step of the way.

I extend my deepest gratitude to my principal supervisor, Prof. Farshid Amirabdillahian, whose wisdom, guidance, and unwavering belief in my abilities have been instrumental in shaping my academic journey. Your invaluable insights, encouragement, and patience have been truly appreciated, and I am profoundly grateful for your mentorship.

I would also like to express my heartfelt thanks to Catherin Menon and Patrick Holthaus for their invaluable support and encouragement throughout this process. Your dedication, expertise, and kindness have been invaluable, and I am sincerely grateful for your contributions to my success.

Finally, I wish to conclude with a timeless verse from the esteemed Hakim Omar Khayyam, a Persian philosopher, mathematician, astronomer, and poet from the medieval period.

"With them the Seed of Wisdom did I sow,
And with my own hand labour'd it to grow:
And this was all the Harvest that I reap'd –
I came like Water and like Wind I go."

یک چند به کودکی به استاد شدیم
یک چند به استاد خودشاد شدیم
پایان سخن شنو که ما را چه رسید
از خاک دمیدیم و برباد شدیم

Author's declaration

I declare that the work presented in this thesis was carried out in accordance with the requirements of the University of Hertfordshire Policies and Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific references in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: ..... DATE: 7th of June 2024

TABLE OF CONTENTS

List of Tables

Table **Page**

List of Figures

Figure **Page**

List of Acronyms

AAL Ambient Assisted Living i, 1–4, 6, 8, 57, 88, 91, 92

CNN Convolutional Neural Network 9, 21, 29, 57

DL Deep Learning 11

FLOPs floating-point operations per second 73

GNN Graph Neural network 21

HAR Human Activity Recognition i, ii, 1–12, 15, 20, 21, 28, 31, 40, 84, 88, 91, 92

HG high-level co-learning 77, 78

HRI Human-Robot Interaction ii, 1, 8

LW low-level fusion 77, 78, 80

MD mid-level co-learning 77, 78

MH combined MD and HG xiv, 77–79, 84

MI Mutual Information 44

MICM Mutual Information Correlation Matrix 7, 51, 53, 60, 89

Min-Max Min and Max distance 44

MV-HAR Multi-View Human Activity Recognition 25, 57

PD Pairwise Distance 41, 42, 44

R Robot view 77

RQ Research Question 40

SOTA state-of-the-art 13

ST-GCN Spatio-Temporal Graph Convolutional Network 21

CHAPTER 1

INTRODUCTION

A assistive technology has emerged as a promising approach to address the challenges faced by older adults in maintaining their independence and quality of life [2, 3]. Ambient assisted living AAL systems are often designed to be integrated into the daily environment of individuals, providing sensitive and responsive services [4]. These systems aim to support individuals in various aspects, including personal care, handling future interfaces, recognizing frailty and mobility, preventing accidents, and providing support in daily living [5, 6]. At the forefront of this transformation is the integration of robots into AAL systems, introducing a new level of support and assistance. Robots provide a chance to have an active presence in the scene as an embodiment of the artificial agent. Moreover, social interaction between robots and humans has been identified as one of the top ten challenging and progressing tasks in robotics [7]. This interdisciplinary area encompasses understanding and modelling human behaviour, as well as their companion pets [8] and using this understanding to improve the interaction with robots, as well as various technological methods and tools. The integration of robots into AAL systems opens up new possibilities for social interaction and assistance, enhancing the overall user experience and promoting a sense of companionship[9].

Within the evolving landscape of AAL technology, two key priorities come to the forefront: efficient Human activity recognitionHAR and seamless Human-robot interactionHRI [10]. The HAR serves as the foundation for perception in the AAL systems, enabling them to identify and promptly respond to users' actions, ensuring that assistance is provided precisely when and where it is needed [10]. This allows for an effective perception-action loop. On the other hand, HRI focuses on promoting intuitive and seamless communication between individuals and their robot companions, enhancing the overall user experience [11]. Therefore, HAR is not only effective in the AAL systems but also plays a crucial role in HRI, where it enables robots to recognize and understand human activities, facilitating more natural and intuitive interactions.

1.1 Human Activity Recognition

Human activity recognition is a complex task that involves capturing spontaneous human behavior in a chosen context such as activities of daily living at home or work. For a human,

the learning process begins in infancy and evolves through experiencing and understanding [5] via significant repetitions and significant feedback. For example, a study found that 12 to 19-month-old toddlers, on average, take 2368 steps and fall 17 times per hour and new walkers (those just starting to walk independently) were found to fall an average of 69 times per hour [12].

However, implementing this technique in robotic systems has been a challenge. Many methods have been developed for human activity recognition, including the use of sensors, technologies, atmospheres, conditions, activities, objects, and machine learning and representation methods. Moreover, the diversity and variations in human body movements and lifestyles make HAR even more challenging. For example, individuals may choose different ways to grasp an object based on their preferences or the task context.

Activity recognition research has focused on various aspects, including daily life activities, indoor/outdoor activities, gait recognition, and transitions between activities. Algorithms and sensor data utilization have enabled the recognition of specific activities and behaviours in different contexts [13]. Companion robots, which are designed to interact and assist humans in various tasks, can benefit greatly from HAR. By incorporating vision-based activity recognition methods, companion robots can understand and respond to human activities in real-time scenarios, providing timely and appropriate assistance [14]. For example, a companion robot equipped with vision-based activity recognition can recognize when a person is performing a specific task, such as cooking or cleaning, and offer assistance or perform complementary tasks. This enhances the overall user experience and promotes seamless interaction between humans and robots.

1.2 Vision-Based HAR

Vision-based HAR methods have gained significant attention in AAL research. These methods utilize cameras and computer vision algorithms to analyze and interpret human activities [15]. Vision-based activity recognition methods can be categorized into two general branches: image-based and skeleton-based approaches. image-based methods utilize colour information such as RGB or grayscale data to recognize human activities. These methods often involve computer vision techniques and machine learning algorithms to extract visual features and classify activities [16]. These methods can capture fine-grained details and subtle movements, but they may be sensitive to variations in lighting conditions, occlusions, and background clutter [17]. On the other hand, skeleton-based methods focus on capturing the spatial and temporal information of human skeletons [14]. These methods utilize techniques such as skeleton tracking or pose estimation to represent human activities based on the movement and configuration of skeletal joints. Both image-based and skeleton-based HAR techniques have been explored in the context of AAL, particularly in smart homes and healthcare facilities [18].

Skeleton-based methods are a subset of vision-based methods for HAR and can be categorized into two main approaches: RGB-D and 2D skeleton-based methods [19, 20]. The RGB-D method utilizes depth sensors to capture both colour (RGB) and depth information, enabling the extraction of 3D skeletal data [19]. On the other hand, the 2D skeleton-based method relies on pose estimation techniques to extract human skeletal information from 2D images or videos.

Pose estimation methods are deemed efficient due to their capability to accurately capture human body keypoints while remaining computationally effective [20]. Despite the complexity of some pose estimation models, ongoing research and development efforts have led to optimizations that enhance their computational efficiency [21]. These advancements enable pose estimation

methods to efficiently extract key body joint positions from images or videos, contributing to the overall effectiveness of skeleton-based human activity recognition systems by the ability to handle occlusions and variations in lighting conditions [22]. This efficiency is crucial as it allows for the representation of human activities in a lower-dimensional space, reducing the computational burden compared to analyzing raw image data. Moreover, these methods often prioritize real-time or near-real-time processing, ensuring timely and responsive activity recognition systems.

1.3 Motivation and Research Goals

In recent years, numerous investigations have been conducted to address the challenges and improve the performance of vision-based activity recognition methods. These include the development of large datasets for training and evaluation [23], the exploration of multi-modal sensor fusion techniques [24], and the utilization of deep learning architectures [25]. These advancements have contributed to the progress and effectiveness of vision-based HAR systems.

However, in the context of AAL and the need for seamless interaction between humans and robots, it often becomes necessary for robots to simultaneously execute multiple processes like navigation, obstacle avoidance, and object detection to achieve advanced perception capabilities and perform effectively. Therefore, this requirement poses limitations on computational resources, and there is a need to explore strategies to optimize and streamline HAR methods to mitigate the computational burden and enhance the overall efficiency of the system.

One critical aspect of this advanced perception is the integration of additional camera perspectives alongside the robot's view. These supplementary camera views expand the robot's sensory awareness, allowing it to perceive the environment from different angles. This is particularly valuable in recognizing complex human activities, as it provides a more comprehensive understanding of the user's actions and the surrounding context.

However, the incorporation of multiple camera views not only intensifies the computational demands on the system but also raises several practical concerns. Firstly, privacy becomes a paramount consideration, as the increased number of cameras may intrude upon users' privacy rights. Moreover, the cost associated with installing and maintaining multiple cameras needs to be carefully evaluated, as it may pose financial challenges. Additionally, the invasiveness of having multiple cameras in private spaces must be addressed to ensure user comfort and acceptance of the system.

Furthermore, the integration of additional camera views introduces complexities in data processing and storage, potentially exacerbating privacy and security risks. Therefore, while these additional perspectives enhance the system's capabilities, they also pose computational challenges and elevate concerns related to privacy, cost, and invasiveness.

This work focuses on addressing the computational cost of integrating extra camera views into the HAR system. Such optimizations are essential to maintain the overall efficiency of the AAL system, ensuring that it can provide timely and accurate assistance while balancing computational resources.

The effectiveness of a perception algorithm is often demonstrated via its detection accuracy, as well as its efficiency which is entailed by the number of training parameters and performance time of the algorithm. There are important factors that allow us to compare and contrast methods when designing and developing a HAR system. In this research, the overarching aim is to design and develop a multi-view skeleton-based human activity recognition system that evaluates algorithmic efficiency by comparing and contrasting state-of-the-art approaches. By leveraging

the advantages of skeleton-based models and incorporating multiple views, it is hypothesised that the proposed system will achieve improved accuracy, reduced training parameters, and faster performance.

The main goals of this research can be summarized as follows:

"To exhibit the development of a streamlined multi-view human activity recognition system utilizing skeleton-based tracking, incorporating strategies for optimization and efficiency to enhance accuracy without significantly increasing training parameters or performance time, thus improving the overall performance of the system for effective and efficient human-robot interaction in AAL environments."

To achieve this, a comprehensive review of the existing literature on HAR, skeleton-based models, and multi-view HAR was conducted. This review provided a solid foundation for the development and evaluation of the proposed multi-view skeleton-based HAR system. Additionally, the effectiveness and efficiency of the proposed system was validated through experimental studies using benchmark datasets.

1.3.1 Research questions

To validate the effectiveness and efficiency of the proposed system, an experimental study using benchmark datasets was performed. The following research questions have been chosen to address different aspects of the main research goal:

RQ1: **How effective is the use of skeleton-based human activity recognition?**

With the ongoing advancement of 2D pose estimation and skeleton extraction methods, special abilities such as detecting multiple humans at the same time, handling occlusions, and managing variations in lighting conditions have been achieved. Moreover, since skeleton data represents abstract human body key points, serves as input with higher information despite its reduced dimensionality. Then, the integration of skeletal models in multi-view settings reduces the number of HAR model parameters and facilitates the deployment of simplified HAR models, resulting in enhanced efficiency of the HAR pipeline for assistive scenarios and human-robot interaction.

Utilizing skeleton-based models allows for the extraction of spatial and temporal information from human skeletons in 2D images or videos, enabling effective recognition of human activities. In contrast, image-based human activity recognition methods, particularly those employing deep learning models, often entail a large number of parameters. Deploying such models in a multi-view context can significantly increase the computational load, posing challenges in resource-constrained environments like a robot in the AAL scenario.

RQ2: **What is the impact of different camera perspectives on the quality and richness of data captured for human activity recognition in multi-view scenarios?**

Understanding the information collected and delivered by each view is crucial for assessing its quality, as well as its spatial and temporal similarity. This comprehension plays a pivotal role in multi-view HAR, as different perspectives may offer complementary or redundant information about human activities. Therefore, this research seeks to explore how factors such as camera type and movement, especially when mounted on moving robots, influence the information captured by each view. By investigating these variables and comparing

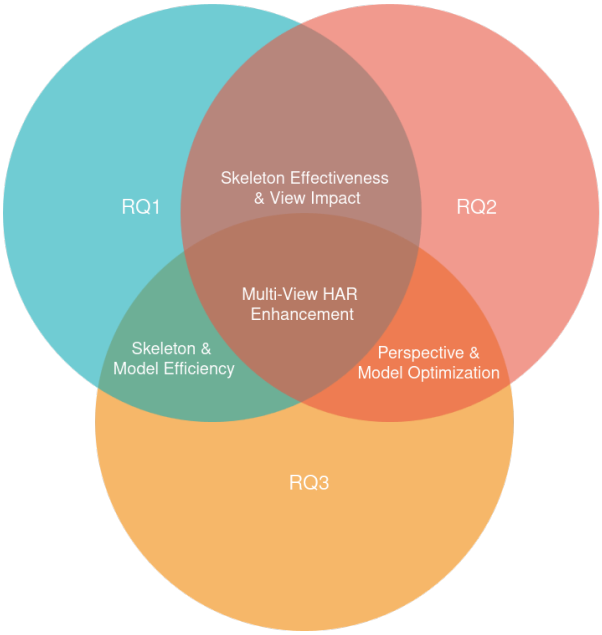


Figure 1.1: Venn diagram illustrating the research focus areas and their intersections in multi-view HAR

data from various viewpoints, the study aims to uncover the unique contributions of each perspective, thereby emphasizing the significance of multi-view settings in enhancing the effectiveness of human activity recognition systems.

RQ3: **What are the optimal models for multi-view human activity recognition, and how do different methods for data combination influence their performance?**

To improve the efficiency of the Human Activity Recognition (HAR) pipeline in various assistive scenarios and human-robot interaction, a comprehensive evaluation of different models at various stages becomes pivotal. This assessment aims to identify the optimal model choice based on criteria encompassing accuracy, learning parameters, and processing time. Simultaneously, to enhance the performance of multi-view HAR systems, the combination of information from multiple camera perspectives proves instrumental. This fusion resolves issues like occlusions, and disparate camera viewpoints fostering a more robust and comprehensive HAR system. Integrating the strengths of the 2D skeleton-based method with multi-view learning and fusion techniques holds promise for augmenting accuracy and reliability in human activity recognition. Specifically, this research endeavours to ascertain the superiority of high-level co-learning over low-level and mid-level data combination approaches within multi-view HAR systems, scrutinizing factors including recognition accuracy and computational efficiency.

This Figure 1.1 highlights the intersections between three key research questions (RQs) in the study of multi-view HAR systems:

Skeleton Effectiveness & View Impact: The overlap between RQ1 and RQ2 focuses on how the effectiveness of skeleton-based HAR is influenced by different camera perspectives.

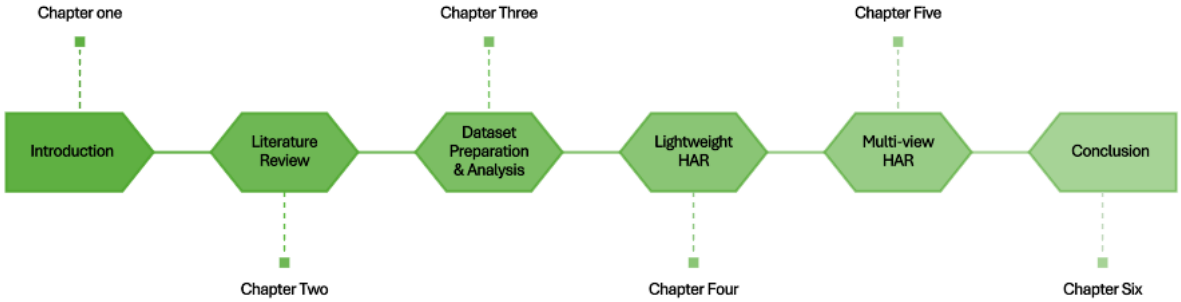


Figure 1.2: Chapters overview

Perspective & Model Optimization: The intersection of RQ2 and RQ3 explores the impact of various camera perspectives on the selection and optimization of models for multi-view HAR.

Skeleton & Model Efficiency: The area between RQ1 and RQ3 examines the relationship between the effectiveness of skeleton-based HAR methods and the identification and performance of optimal models.

Multi-View HAR Enhancement: The central intersection of RQ1, RQ2, and RQ3, representing the integrated approach of using skeleton-based tracking, diverse camera perspectives, and optimal model selection to comprehensively enhance multi-view HAR systems for effective human-robot interaction in Ambient Assisted Living (AAL) environments.

1.4 Thesis Outline

The thesis comprises of six chapters, visually presented in Figure 1.2.

1.4.1 Introduction

Chapter 1 initiates with the thesis narrative and introduces the research motivation. It introduces the research questions and hypotheses. Three primary questions have been outlined, and each subsequent chapter explores various dimensions of these inquiries.

1.4.2 Literature Review

Chapter 2 conducts a comprehensive review of the current literature surrounding HAR technologies, skeleton-based models, and multi-view HAR. This review serves as a robust foundation for addressing all three questions, Specifically:

Question one establishes the absence of a diverse, skeleton-based HAR dataset encompassing various views, notably the Robot-view, in the AAL applications.

In addressing question two, the focus is on highlighting the deficiency in multi-view analysis while concurrently developing advanced methodologies capable of presenting spatial and temporal information derived from various perspectives.

Finally, to address research question three, the chapter emphasizes the crucial need for evaluations to discern suitable models for multi-view HAR in AAL contexts.

1.4.3 Dataset Preparation and Analysis

In Chapter 3, the focus lies on addressing all research questions.

To address research question one (RQ1), a multi-view skeleton-based HAR dataset has been developed (RHM-HAR-SK) using an image-based dataset (RHM) as its foundation. This allows for a comparative assessment of the results obtained from skeleton-based models against other work based on RGB data within the RHM dataset (The RGB data refers to images or videos containing visual information captured by cameras, where each pixel in the image is represented by a combination of red, green, and blue color values).

Moreover, the qualitative analysis conducted in this chapter notes the variance in quality across different perspectives. Additionally, the spatial and temporal analyses conducted demonstrate the capability of the skeleton-based approach in extracting valuable human biomechanical data.

Concerning question two, the findings reveal the significant impact of perspectives on pose extraction across various camera types, including their attitudes and wider views. Additionally, the spatial and temporal analyses conducted on joint movements indicate both similarities and differences within different action classes.

Moreover, the temporal analysis coupled with mutual information analysis highlights the potential for multi-view models to improve classification accuracy by identifying distinct patterns, thus presenting opportunities to enhance the effectiveness of classification algorithms.

Regarding question three, the mutual information correlation matrix (MICM) analysis indicates significant disparities at the frame level among various views. This analysis shows the similarity of each frame in an action sample and enables us to compare it with the other synchronised views.By utilising the information theory and entropy concept this matrix can show the frames differences in information.The MICM shows frame similarity, highlighting redundant information and identifying similar frames in each view.

Moreover, the joints movement analysis results underscore the potential advantages of incorporating additional views to augment the overall performance of the system. Metrics such as Min-Max and Pairwise distance capture movement patterns across different views, highlighting how combining multiple perspectives can yield richer information.

However, the inclusion of supplementary views raises concerns regarding heightened computational demands and associated costs. Additionally, the research acknowledges the need for a more comprehensive exploration of different multi-view methods that show the impact of multi-view settings. This aspect will be examined and addressed in Chapter 5.

1.4.4 Lightweight HAR

In Chapter 4, a lightweight HAR model has been devised, presenting results that address all research questions.

Concerning question one, the findings demonstrate that leveraging the skeleton-based model allows for the transformation of input data into a 2D tensor format, enabling the utilization of lightweight classifier structures.

About question two, the low-level data fusion shows that incorporating additional views can potentially boost accuracy. While the performance improvement is not significant, it suggests a possibility for enhancement.

Regarding question three, the comparison between CNN-based and transformer-based models indicates that CNN models possess the capability to deploy lightweight HAR models with

reasonably high accuracy. Subsequently, the focus shifts towards CNN-based models in Chapter5.

1.4.5 Multi-view HAR

Chapter 5 focuses on crafting a multi-view HAR structure that encompasses various levels of data combination, providing results that address all three research questions.

Concerning research question one, the comparison of multi-view combinations between the skeleton and image-based methods on the same dataset demonstrates the superiority of the skeleton-based approach in effectively capturing and amalgamating data to deploy a more streamlined model.

Moving on to research question two, the results portray the potential for improvement by integrating various perspectives, notably enhancing the accuracy of the Robot-view within the multi-view setup.

Addressing research question Three, the findings reveal that high-level co-learning yields higher performance, and the combination of mid-level with high-level components exhibits even greater performance without an increase in model complexity. Moreover, the comparison among different CNN architectures indicates that employing a lightweight structure proves to be more efficient in the context of multi-view HAR when compared to complex CNN structures.

1.4.6 Discussions and Conclusion

In conclusion, this chapter provides the basis for exploring the potential of multi-view HAR for enhancing HRI within AAL environments. The growing importance of efficient and accurate HAR in AAL settings was identified, particularly for robots seeking to understand and assist older adults. Recognizing the limitations of existing methods, this research explores the potential of skeleton-based approaches due to their advantages in simplicity, computational efficiency, and robustness. We further highlight the significance of multi-view approaches, where incorporating information from different perspectives can improve recognition performance and thereby improve robot perception in AAL scenarios.

CHAPTER 2

LITERATURE REVIEW

Human activity recognition in ambient assistive scenarios is a crucial research area that aims to enhance the well-being and safety of individuals by detecting and understanding their activities in real-time [26]. This chapter aims to provide an in-depth understanding of human activity recognition, emphasizing its significance in leveraging various technologies. Specifically, considering the lake of multi-view models and datasets in the HAR domain, wherein a robot is involved in AAL scenarios, it focuses on vision-based methodologies, skeleton-based approaches, the multi-view paradigm, and the architecture of CNNs in the multi-view context. This analysis aims to highlight the critical role of these techniques in enhancing the efficacy and accuracy of human activity recognition systems in the AAL environment.

2.1 Human Activity Recognition in Assistive Robotics

Human activity recognition enables robots to understand and respond to human users' needs and activities, particularly in the context of ambient assisted living. However, existing comprehensive reviews of assisted living technology and HAR have highlighted a notable shortcoming: there is a lack of dedicated datasets that focus on skeleton-based and multi-view HAR specifically within assistive robotics contexts [] [27–29]. Thus, the development of a novel dataset specifically tailored for assistive robotics applications presents a valuable move in this evolving field. This initiative not only fills an evident void but also lays the foundation for more efficient and effective models, covering various facets ranging from data preparation to classification techniques and developing multi-view learning methodologies.

2.2 Level of Human Activities

The primary concept of human activity necessitates a comprehensive description. Human activity can be divided into five key components: gestures, actions, human-object interactions, human-human interactions, and group activities [30]. This perspective is further expanded by Herath et al [31], who describe human activity into six distinct categories: a human interaction, an

individual action, a human-object interaction, an action captured by an RGBD sensor, a group action, and a multi-view action.

Considering a gesture as a very primary act of human behaviour to convey meaning, Dang et al. define it as a position or subtle movement that communicates an idea [32]. In contrast, Herath et al. describe an action as the most basic interaction between a human and their environment that carries significance meaning[33]. However, meaning in human activity is defined as the relationship between the subject and the object, reflecting the understanding of subjective meaning in psychology, philosophy, and linguistics [34].

In the context of human-robot interaction, these definitions become particularly significant. When a human performs a gesture, the robot must accurately interpret this subtle movement to understand the intended message or command. This requires advanced recognition algorithms capable of detecting and interpreting the nuanced positions and movements that constitute a gesture. For example, a simple wave of the hand could indicate a greeting or a command for the robot to follow, depending on the context.

Similarly, understanding actions involves recognizing the broader interactions between the human and their environment. This means the robot must not only detect the action but also comprehend the significance behind it. For instance, when a person picks up an object, the robot needs to understand whether the person is preparing to use the object, passing it to someone else, or simply moving it out of the way. This understanding enables the robot to assist more effectively, such as by offering help, adjusting its behaviour, or planning subsequent actions.

The ability to discern gestures and actions is crucial for robots in AAL environments, where they need to interact seamlessly with humans and provide assistance. By accurately interpreting these basic forms of human behaviour, robots can offer timely and appropriate responses, enhancing the overall quality of interaction and support. Therefore, the integration of sophisticated gesture and action recognition systems is vital for developing intuitive and responsive human-robot interaction frameworks that can operate effectively in dynamic and diverse environments.

2.3 Action Recognition Methods

The implementation of the HAR module could involve a variety of applications and devices. For instance, smartwatches employ various sensors, while robots utilize different types of vision systems plus sensors. In a comprehensive survey[32], various applications have been mentioned, including smart appliances and homes, security and surveillance, self-driving cars, healthcare and elderly care, and entertainment.

More precisely, the input data type in the HAR specialisation may usefully be divided into two general categories: *vision-based*, like RGB, depth, stereo cameras and *sensor-based* systems, including sensors like global positioning systems, gyroscopes, accelerometers, and magnetometers. Plus, some multi-modal approaches [35] benefit from multiple sensors and cameras. This combination of diverse technologies and applications underscores the complexity and versatility of the HAR module. In this work, the focus is on vision-based technology and more specifically on the skeleton-based with 2D human pose estimation.

In the following subsections, some common features of HAR datasets have been explored, exploring deeper into the background work within vision-based technology. This exploration aims to discuss the various aspects and evolution of vision-based methods, providing a comprehensive understanding of these evolving techniques within the domain of HAR. The goal is to be familiar

with the diverse methodologies and innovations within vision-based technology, particularly emphasizing the complexities of skeleton-based approaches using 2D human pose estimation.

2.3.1 Datasets and Benchmarks

The realm of Human Activity Recognition (HAR) encounters challenges in both the generation of datasets and the practical application of their outcomes in real-world settings [36]. Particularly, methodologies such as crowd-sourcing and *DL* approaches have emerged as potent tools in HAR but demand vast quantities of data for effective training and validation [37]. This requirement poses a significant challenge, as acquiring, annotating, and curating large-scale datasets remains resource-intensive and time-consuming.

The significance of HAR datasets lies not only in their sheer volume but also in their characteristics and annotations. These datasets encompass a diverse array of features that provide vital insights into the underlying activities being recognized [38]. Characterizing HAR datasets often involves identifying several prominent features, such as:

- **Technology**
  Vision-based or Sensor-based or hybrid approaches.
- **Number of Sensors/Camera**
  The quantity and arrangement of sensors or cameras utilized for data collection, impacting the scope and coverage of the dataset.
- **Activity Diversity**
  The variety and quantity of activities represented within the dataset, crucial for training models to recognize diverse human movements and behaviours.
- **Subject Variability**
  The involvement of multiple subjects or individuals in data collection, contributing to the dataset's variability and generalization capability.
- **Attribute Count**
  The number and nature of features or attributes associated with each data instance, aiding in the characterization and differentiation of activities.
- **Instance Volume**
  The total count of instances or data samples within the dataset, providing the necessary scale for training and evaluating HAR models.
- **Sampling rate/Frames**
  The frequency of data capture or number of frames per second, influencing temporal analysis and model performance in recognizing activities.
- **Data Trimmed or Untrimmed**
  The distinction between datasets containing full sequences of activities versus partial fragments, affecting model training and testing scenarios.
- **Labeling/Annotation Quality**
  The presence and quality of annotations or labels associated with each instance, facilitating supervised learning and model comprehension.

- **Controlled/Uncontrolled Environments**
  Whether the dataset is collected in controlled settings or captures activities in real-world ('in-the-wild' settings and uncontrolled environments), impacting model adaptability.

- **Indoor/Outdoor Setting**
  The environmental context of data collection, influencing model adaptability to different contexts.

- **Scene Variety**
  The diversity and number of distinct scenes or contexts within the dataset, contribute to model robustness.

Addressing these characteristics in HAR, dataset quality plays a pivotal role in fostering the development and benchmarking of robust recognition models. To truly quantify and qualify a dataset, it's essential to consider various characteristics that directly impact its usability and effectiveness in real-world applications.

One key consideration is the dataset's technology, whether it adopts vision-based, sensor-based, or hybrid approaches. This choice significantly influences the dataset's applicability in real-world scenarios where certain technologies may be more prevalent or suitable.

The number and arrangement of sensors or cameras used for data collection also contribute to dataset quality. Datasets with a diverse range of sensors or cameras can capture a broader scope of human activities, enhancing their representativeness and utility.

Activity diversity is another crucial factor to assess in HAR datasets. A dataset that covers a wide range of activities relevant to real-world scenarios enables models to learn robust representations and generalize effectively across different contexts.

Subject variability, or the inclusion of data from multiple subjects or individuals, further enhances dataset quality by ensuring models can adapt to different users and behaviours encountered in real-world settings.

Attribute count, instance volume, sampling rate, and data trim are additional features that contribute to dataset quality. These factors influence the dataset's richness, scale, temporal analysis capabilities, and suitability for model training and evaluation.

However, while a dataset with more variables may better represent real-world scenarios, it also introduces greater complexity and difficulty during model training. Striking the right balance between dataset richness and complexity is crucial to developing HAR models that can seamlessly integrate into everyday environments.

From a technical standpoint, there are several variables to consider. For example, camera types can vary in terms of resolution, viewpoint (wall mount, ceiling, fish-eye, robot view, egocentric view), whether they are single or stereo, two-colour, RGB, and RGBD. Therefore, each research project addressing a specific area should select an appropriate dataset or create the data they require.

The Table 2.1 presents an overview of popular HAR datasets along with their properties provided by Bamorovat et al. [1]. They meticulously analyzed 42 existing HAR datasets, uncovering critical gaps in the current landscape of HAR datasets. Their comprehensive assessment highlighted the scarcity of datasets capturing dynamic perspectives from a robot's viewpoint, with only one dataset, LIRIS, exhibiting such characteristics. Moreover, they identified the absence of datasets providing top views, particularly from ceiling perspectives, which presents challenges for applications such as caregiving. Additionally, their analysis unveiled redundancy in multi-view

datasets, with only a few offering a robot-view perspective alongside motion, notably LIRIS and InHARD.

To address these deficiencies, Bamorovat et al. introduced the RHM HAR dataset, strategically designed to fill the identified gaps in existing datasets. This pioneering dataset aims to incorporate dynamic perspectives from a robot's viewpoint, provide top views, and mitigate redundancy issues found in multi-view datasets. Their work not only sheds light on the limitations of current HAR datasets but also lays the groundwork for future advancements in activity recognition research.

Generally, *Vision-based* activity recognition methods can be divided into two categories: *image-based* and *skeleton-based*. As a result, the machine learning techniques and developments for these methods also differ [38]. In the following sections, a review of relevant works is presented to discuss the distinctions between image-based (Sec. 2.3.2) and skeleton-based (Sec. 5.1.2) methods.

2.3.1.1 Generative and Non-generative Datasets

When it comes to data preparation techniques, generative and non-generative view-invariant HAR methods are the two primary dataset groups. As implied by the name, generative approaches produce their input data from one or more actual views [39], whereas non-generative approaches acquire their data from genuine input devices like sensors and cameras. For instance, [40] is a SOTA perspective-shifting approach that transforms an action into many views and is based on the angle representation in skeleton data. Their method proved reliable when dealing with incomplete data. Moreover, Generative Adversarial Networks (GAN) [41] and encoder-decoder CNN networks are popular for image-based approaches [42, 43]. However, there currently exists no non-generative skeleton-based HAR dataset including a robot view, and this work addresses this gap. Additionally, the presented dataset (Sec. 3) can provide sufficient data to create generative datasets in the future and can be adopted for the future development of assistive scenarios.

Dataset Name	Year	Video	An	Act	FV	En	Si	Mot	PoV	Modality	B	MV	AT	L	So	U	T	Acc
BON	2022	2.6K	2.6K	18	_	Di	UC	Dy	FP	RGB	Dy	No	No	No	C	Home	Tr	No
EPIC-KITCHENS-100	2021	700	90K	4053	_	I	UC	Di	FP	RGB	Dy	No	No	No	C	Kitchen	A	Link
HOMAGE	2021	5.7K	5.7K	75	2	I	UC	Di	FP/TP	12 Sensors	Dy	Yes	Yes	No	C	Home	A	Link
HA500	2021	10K	591K	500	_	Di	UC	St	TP	RGB	Dy	No	Yes	No	W	Diversity	A	Link
M-MiT	2021	1M	2M	292	_	Di	UC	St	TP	RGB	Dy	No	No	Yes	W	Diversity	A	Link
MovieNet	2020	1.1K	65K	80	_	Di	UC	St	TP	RGB	Dy	No	No	No	M	Diversity	A	Link
Multi-ViewPointOutdoor	2020	2.3K	503K	20	3	O	UC	Di	TP	RGB	Dy	Yes	No	No	YT	Sport	A	No
HVU	2020	572K	9M	3457	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
AViD	2020	80k	80K	887	_	Di	C	St	TP	RGB	St	No	No	No	W	Diversity	A	Link
LEMMA	2020	1.1K	0.9M	641	3	I	C	Di	FP/TP	RGB,D	Dy	Yes	Yes	No	C	Home	A	Link
InHARD	2020	4.8K	2M	14	3	I	C	S	TP	RGB,D	Dy	Yes	No	No	C	Industrial	A	Link
FineGym	2020	503	32.5K	15	_	I	UC	Di	TP	RGB	Dy	No	Yes	No	M	Sport	A	Link
Ava_Kinetic	2020	500	230K	80	_	Di	UC	St	TP	RGB	Dy	No	No	Yes	YT	Diversity	A	Link
Kinetic_700_2020	2020	648K	648K	700	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
Jester	2019	148K	5.3M	27	_	I	C	St	TP	RGB	Dy	No	Yes	No	C	Gesture	Tr	No
HACS	2019	504K	1.5M	200	_	Di	UC	St	TP	RGB	Dy	No	No	Yes	YT	Diversity	A	Link
Kinetic_700	2019	650K	650K	700	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
NTU RGB+D 120	2019	114K	8M	120	155	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Daily	A	Link
MiT	2019	1M	1M	339	_	Di	UC	Di	TP	RGB	Dy	No	No	No	W	Diversity	Tr	Link
20BN-sth_sth-V2	2018	220K	220K	174	_	I	UC	Di	FP	RGB	Dy	No	No	No	W	Diversity	A	No
Kinetic_600	2018	496K	496	600	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
Charades-Ego	2018	8K	68.5K	157	2	I	C	Di	FP/TP	RGB	Dy	Yes	Yes	Yes	C	Daily	A	Link
AVA	2017	430	197K	80	_	Di	UC	St	TP	RGB	Dy	No	Yes	Yes	M	Diversity	A	Link
SLAC	2017	520K	1.17M	200	_	Di	UC	Di	TP	RGB	Dy	No	No	Yes	YT	Diversity	A	No
MultiTHUMOS	2017	38.6K	38.6K	65	_	Di	UC	Di	TP	RGB	Dy	No	No	No	W	Diversity	A	Link
20BN-Sth_Sth	2017	100K	100K	174	_	I	UC	Dy	FP	RGB	Dy	No	Yes	No	W	Diversity	Tr	No
Kinetic_400	2017	300K	300K	400	_	Di	UC	St	TP	RGB	Dy	No	Yes	No	YT	Diversity	A	Link
M2I	2017	1784	1784	22	2	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Diversity	Tr	No
DALY	2016	8133	8133	10	_	Di	UC	St	TP	RGB	Dy	No	Yes	Yes	YT	Diversity	A	Link
YouTube-8M	2016	8.2M	8.2M	4800	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Diversity	A	Link
NTU RGB+D	2016	56K	56K	60	3	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Daily	Tr	Link
Charades	2016	10K	10K	157	2	I	UC	St	TP	RGB	Dy	No	No	Yes	YT	Daily	Tr	Link
UTD-MHAD	2015	861	861	27	5	I	C	St	TP	RGB,D	St	Yes	Yes	No	C	Daily	Tr	Link
ActivityNet	2015	23K	23K	203	_	Di	UC	St	TP	RGB	Dy	No	No	No	W	Diversity	A	Link
Sport-1M	2014	1M	1M	487	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Sport	A	Link
Berkeley MHAD	2013	660	660	11	12	I	C	St	TP	RGB,D	St	Yes	Yes	No	C	Diversity	Tr	Link
Multi-View 3D Events	2013	3.8K	383K	11	3	I	C	St	TP	RGB,D	Dy	Yes	Yes	No	C	Diversity	Tr	No
ASLAN	2012	10K	10K	432	_	Di	UC	St	TP	RGB	Dy	No	No	No	YT	Diversity	Tr	Link
UCF101	2012	13K	13K	101	_	Di	UC	Di	TP	RGB	Dy	No	Yes	No	YT	Diversity	Tr	Link
LIRIS	2012	828	828	10	2	I	C	Di	TP	RGB,D	Dy	Yes	Yes	Yes	C	Daily	Tr	Link
HMDB51	2011	6.8K	6.8K	51	_	Di	UC	Di	TP	RGB	Dy	No	No	No	YT	Daily	Tr	Link
UCF_ARG	2010	480*3	480*3	10	3	O	C	St	TP	RGB	Dy	Yes	Yes	Yes	C	Daily	Tr	Link

Table 2.1: OVERVIEW OF POPULAR HAR AND THEIR PROPERTIES.

PRESENTED IN DESCENDING ORDER OF YEAR STARTING FROM 2022 AND ENDING WITH 2010[1]. An: Number of Annotation, Act: Number of classes, FV: Number of Fixed Views, En: Environment Type (I: Indoor, O: Outdoor, Di: Diverse), Si: Situation (C: Controlled, UC: UnControlled), Mot: Camera motion capability (Dy: Dynamic, St: Static, Di: Diverse), PoV: Point of View (FP: First Person, TP: Third Person), B: Background (Dy: Dynamic, St: Static), Mu: Multi-View, At: Atomic, L: Localization, So: Source (C: Created, W: Web, M: Movie, YT: YouTube, U: Usage, T: data preparation type (Tr: Trimm, A: Annotation), Acc: Accessibility)

2.3.2 Image-based Methods

Image-based video representation techniques, particularly RGB-based methods, are among the most popular approaches in the field. This popularity stems from the widespread availability and compatibility of RGB data in various applications and devices, making it easily accessible for analysis and processing. RGB-based methods involve extracting features from image frames within a video stream, leveraging the rich color information captured by RGB cameras to characterize and analyze visual content effectively. Within this domain, two primary types of representation methods exist, the *hand-crafted* features and *deep features*, with the latter being more prevalent in current trends. Hand-crafted features are sometimes referred to as shallow learning.

One notable method within the realm of hand-crafted features is the dense trajectories technique, as popularized by Wang et al. in their work on dense trajectories [44]. This approach initiates by generating a set of trajectories based on the optical flow observed between frames. Subsequently, it employs feature extraction methods such as *SIFT, HAR, HOG*, among others, to enhance the trajectory data. These extracted features are then utilized to create a higher-level representation through feature encoding techniques.

The resultant model, formulated via this method, is commonly employed for action classification tasks. However, an advanced version of this technique known as an improved dense trajectory (iDT) [45] has been developed to further enhance precision and robustness in action recognition. Despite its superior performance, deploying iDT poses challenges due to its higher computational demands, making its practical implementation difficult.

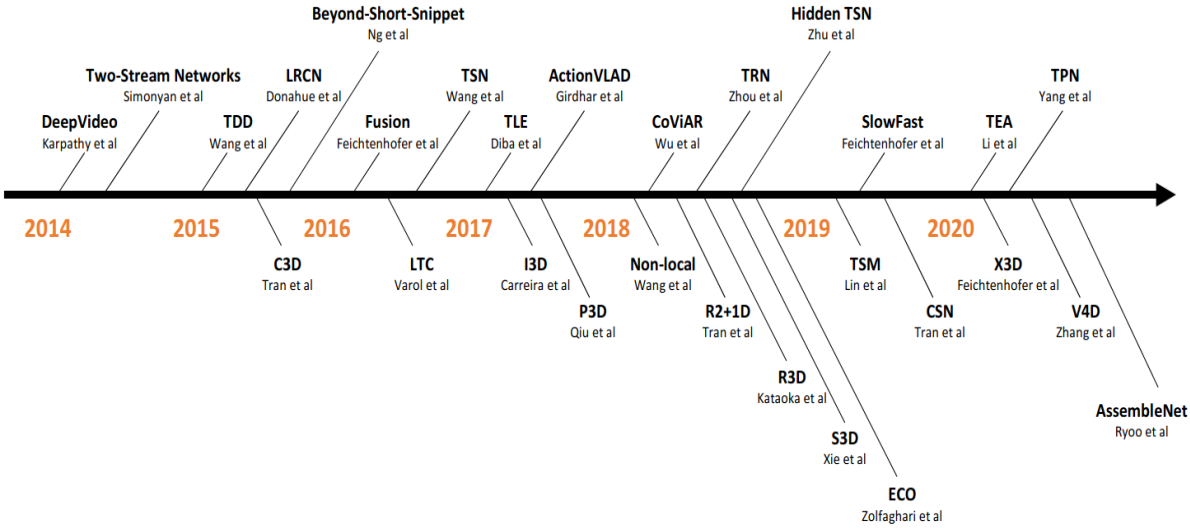


Figure 2.1: Chronological overview of representative HAR methods [46]

On the other hand, deep learning [47], offers a high level of abstraction that facilitates the unravelling of interpretative factors, simplifying generalisation and transfer. A comprehensive chronological overview of HAR is provided highlighting the evolution of representative learning techniques since 2014 (Figure 2.1) [46]. This overview reveals three main trends in deep learning based on RGB data for HAR tasks.

The **first** trend is the emergence of *Two-Stream Networks*These networks separate spatial and temporal information to capture dynamic changes in video data better. The notable models in this category include:

- **Trajectory-pooled Deep-convolutional Descriptor(TDD)** [48], combines the strengths of hand-crafted and deep-learned features by using deep architectures to create discriminative convolutional feature maps and then aggregating these through trajectory-constrained pooling. This method incorporates spatiotemporal and channel normalization to enhance robustness.

- **Long-term Recurrent Convolutional Networks (LRCN)** [49], This model combines CNNs with recurrent neural networks (RNNs). The convolutional layers extract spatial features, while the recurrent layers handle temporal dependencies, making it effective for sequential data analysis.

- **Fusion** [50], This approach integrates spatial and temporal streams at different stages of the network, allowing for a more comprehensive understanding of video content by merging the two types of information.

- **Temporal Segment Networks (TSN)** [51], TSN divides a video into segments and models long-range temporal structures by sampling frames from each segment. This helps capture diverse temporal information and improves the robustness of activity recognition.

The **second** trend involves the use of *3D convolutional kernels*, which extend traditional 2D convolutions into the temporal dimension, thereby capturing motion and spatial information simultaneously. Key models in this trend are::

- **Inflated 3D ConvNet(I3D)** [52], I3D inflates 2D convolutional filters and pooling operations into 3D, allowing it to learn spatiotemporal features directly from video frames.

- **ResNet 3D(R3D)** [53], This model applies 3D convolutions to capture motion information. It adapts the residual learning framework to the 3D domain, making it powerful for video analysis.

- **Separated 3D(S3D)** [54], S3D separates spatial and temporal convolutions, reducing the model's complexity while maintaining performance. This separation helps in balancing the trade-off between computational efficiency and recognition accuracy.

- **Non-local Networks**[55], These networks capture long-range dependencies within video data by incorporating non-local operations, which improve the ability to model global context in videos.

- **SlowFast Networks** [56], SlowFast employs two pathways operating at different frame rates—one slow pathway to capture detailed spatial semantics and a fast pathway to capture rapid motion, enhancing the model's ability to process varying temporal dynamics.

The **third** trend is the focus on computational efficiency, These models aim to reduce computational load without significantly compromising performance, which is crucial for real-time applications. Prominent examples include:

- **Hidden Two-Stream Network (Hidden TSN)** [57], This model employs a hidden two-stream architecture, optimizing the fusion of spatial and temporal features for efficient processing.

- **Temporal Shift Module (TSM)** [58], TSM introduces an innovative technique where a portion of the channel information is shifted along the temporal dimension, allowing for efficient temporal modelling without additional parameters.
- **X3D** [59], is a scaled-down version of I3D that reduces the number of parameters and computational requirements by progressively expanding the network along multiple dimensions.
- **Tiny Video Networks(TVN)** [60], focuses on achieving efficiency through a compact architecture designed specifically for low-resource environments, enabling fast and effective video processing.

These trends highlight the ongoing evolution and diversification of deep learning techniques in the field of RGB-based HAR, underscoring the importance of continued research and development in this area.

2.3.3 Pose Extraction for Activity Recognition

Since the pose extraction method is applied at an early-stage task in the HAR pipeline (as shown in Figure 2.4), it plays a vital role in skeleton-based HAR [61]. The accuracy in this section directly affects the rest of the procedure. Pose extraction typically relies on either 2-dimensional (RGB) or 3-dimensional (RGB-D) input data [62, 63]. While depth data in 3-dimensional approaches allows for better recognition results, they require special sensors that are sometimes costly or unsuitable for certain environments. Additionally, these sensors are limited in their ability to estimate distances beyond approximately 6 meters, making them less effective in larger or more open spaces.

Moreover, the storage size of such datasets increases drastically compared to RGB-based ones. Hence, publicly available datasets often provide 2-dimensional data only. To allow for later comparison to other datasets and approaches, this work relies on 2-dimensional data. Moreover, the simplicity, affordability and accessibility of RGB cameras allow us to apply a high-performance pose extraction method independent of specific hardware on a robot.

2.3.3.1 High-Resolution Network(HRNet)

The High-Resolution Network (HRNet) is an advanced architecture designed for effective pose estimation by employing three parallel sub-networks that operate at different resolutions: low, medium, and high [64]. This multi-resolution approach enables HRNet to capture features at various scales, which is essential for achieving high accuracy in pose estimation.

Low-resolution branch: Processes the entire input image, capturing global context and spatial relationships between body parts.

Medium-resolution branch: Operates on a downsampled version of the image, focusing on extracting intermediate-level features. Intermediate-level features refer to patterns or elements in an image that capture mid-range details and structures, sitting between low-level features (such as edges, textures, and colours) and high-level features (such as objects and faces). It serves as a bridge between the global context and fine details, capturing essential mid-level information that contributes to the model's understanding of the pose.

High-resolution branch: Processes the image at full resolution, preserving fine details crucial for precise keypoint localization. Maintaining high resolution, ensures that subtle and

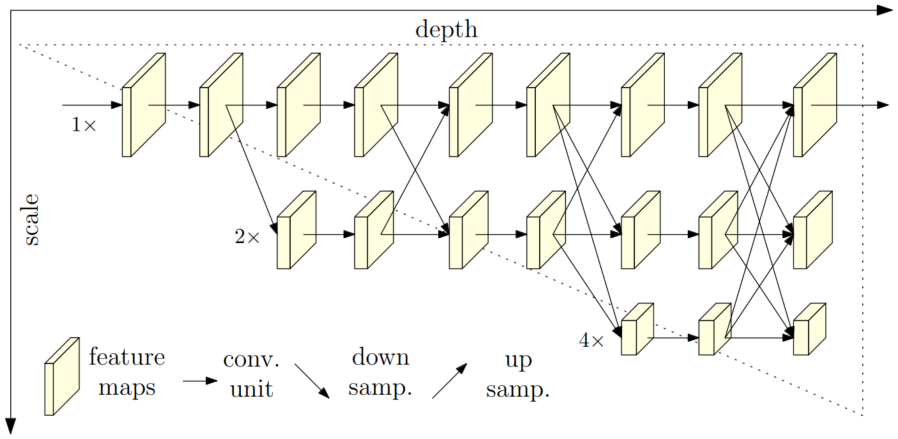


Figure 2.2: The architecture of HRNet
, which consist of three high-to-low resolution parallel subnetworks[64].

intricate details of the human pose are accurately captured, leading to more precise and reliable keypoint detection.

The effectiveness of HRNet in pose estimation stems from its multi-resolution architecture, which captures a wide range of visual information crucial for accurate keypoint localization. This approach not only enhances the understanding of the human body's structure but also ensures that finer details and broader spatial relationships are seamlessly integrated. By processing the input image through low, medium, and high-resolution branches, HRNet ensures a comprehensive analysis, making it a robust solution for pose estimation tasks.

Following the extraction of features at varying resolutions, the subsequent pivotal step involves effectively integrating these features to facilitate robust keypoint estimation. This integration process, known as feature fusion, combines the diverse information obtained from each resolution branch. Once fused, these features are inputted into the keypoint prediction stage. Subsequently, in the keypoint refinement stage, post-processing techniques are applied to further enhance the accuracy and reliability of keypoint localization.

Feature fusion: The extracted features from each sub-network are then efficiently fused through channel concatenation or other techniques. This process integrates the strengths of each branch, combining low-level details with high-level context.

Keypoint prediction: The fused features are fed into a final prediction stage, typically consisting of convolutional layers, to estimate the heatmaps for each keypoint. Each heatmap represents the probability of a specific keypoint existing at each pixel location in the image.

Keypoint refinement: Post-processing techniques like peak search and pose refinement algorithms are often applied to identify the most likely locations of the keypoints based on the generated heatmaps.

There are two general methods in two-dimensional pose estimators, *BottomUp* [63] and *TopDown* [65, 66]. The difference between the two is the sequence of finding poses and humans. The TopDown method first finds the Region of Interest (ROI), which is the human body, and then finds the poses. The provided dataset in this work (Chapter 3)also used the TopDown method. On the other hand, in the BottomUp approach, the poses needed to be found, and then by grouping them, the human skeleton data will be created.

HRNet Variants:

1. Top-Down Approach:

Functioning: This is the standard HRNet structure that has been used in this work. It follows a two-stage process:

Stage 1: Person Detection: An initial stage utilizes a pre-trained object detection model (often another CNN) to identify people in the image and generate bounding boxes around them.

Stage 2: Keypoint Estimation within Bounding Boxes: The features extracted by HRNet's sub-networks (low, medium, high resolution) focus only on the area within each bounding box. This allows for more focused and accurate keypoint prediction within each individual.

2. Bottom-Up Approach (HigherHRNet):

Functioning: This variant deviates from the top-down approach by directly predicting keypoints from the entire image:

Keypoint Heatmap Prediction: HRNet's sub-networks predict heatmaps for each keypoint across the entire image. Each heatmap represents the likelihood of a specific keypoint existing at each pixel location.

Pose Grouping: After generating heatmaps, a separate step groups the predicted keypoints into complete human poses. This involves techniques like associating nearby keypoints based on anatomical relationships and body structure constraints.

Key Differences between Top-Down and Bottom-Up HRNet:

Object Detection: Top-down HRNet relies on a pre-trained object detector, while bottom-up directly predicts keypoints without prior person detection.

Focus: Top-down focuses on keypoint estimation within pre-defined bounding boxes, while bottom-up predicts keypoints across the entire image and then groups them into poses.

Computational Cost: Bottom-up can be slightly more computationally expensive due to the additional grouping step. For instance, in a scenario with a high-resolution image (e.g., 1920x1080) and a large number of keypoints (e.g., 17 keypoints per person), the keypoint detection step alone may require processing millions of pixels and performing keypoint estimation on each pixel. Subsequently, the keypoint grouping step further increases computational demands, as it involves analyzing pairwise relationships between all detected keypoints to form coherent human poses.

2.3.3.2 YOLOv7

YOLOv7, released in 2023 by Wang et al. [67], is a state-of-the-art object detection model, recently being adapted for pose estimation tasks.

Architecture:

Backbone: Similar to most object detection models, YOLOv7 utilizes a convolutional neural network (CNN) as its backbone. This network extracts features from the input image, capturing relevant information for object and keypoint localization. YOLOv7 utilizes the EfficientNet series as its backbone, known for balancing accuracy and speed.

Neck: The extracted features are then passed through a "neck" network, responsible for combining information from different levels of the backbone. In YOLOv7, this neck utilizes Feature Pyramid Network (FPN) connections, allowing the model to integrate features at various resolutions, crucial for capturing objects of different sizes.

Head: Finally, the processed features are fed into separate head networks designed for specific tasks. For pose estimation, one head network predicts bounding boxes for each individual in the image. Another head network predicts heatmaps for each keypoint, indicating the likelihood of its presence at each pixel location within the predicted bounding box.

Key Features:

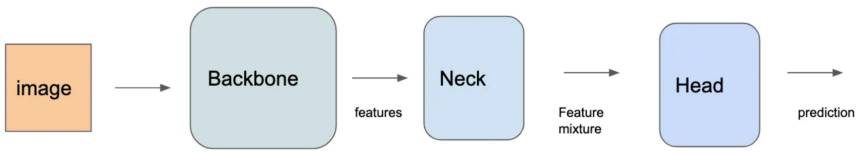


Figure 2.3: Yolov7 high level architecture

Feature	**HRNet (Top-Down/Bottom-Up)**	**YOLOv7**
Approach	Two-stage (Detection, Keypoint) / Single-stage	Single-stage
Object Detection	Pre-trained model (Top-Down) / No	Yes
Keypoint Prediction	Stage 2 (Top-Down) / Heatmaps (Bottom-Up)	Heatmaps
Output	Keypoints within bounding boxes (Top-Down) / Individual poses (Bottom-Up)	Bounding boxes and Keypoints
Strengths	High Accuracy (Top-Down), Handles Occlusions (Bottom-Up)	Fast Processing, Token-based detection, Focus mechanism
Weaknesses	Slower than YOLOv7 (Top-Down), Grouping complexity (Bottom-Up)	Potentially lower accuracy, Less suitable for high-resolution tasks

Table 2.2: Comparision between HRNet and Yolo7

Single-Stage Detection: Unlike HRNet's two-stage approach, YOLOv7 performs both bounding box and keypoint prediction in a single stage. This makes it computationally efficient and faster than HRNet, especially useful for real-time applications.

Token-Based Object Detection: YOLOv7 introduces a novel token-based approach for object detection. It divides the image into a grid of cells and predicts bounding boxes and keypoints relative to these cells. This method improves the model's ability to handle objects of different sizes and positions within the image.

Focus Mechanism: YOLOv7 employs a focus mechanism that focuses the model's attention on specific regions of the image containing objects with high confidence. This helps to improve the accuracy of object and keypoint predictions, especially for smaller or less prominent individuals.

2.3.4 Skeleton-based Methods

Skeleton data can be obtained through pose estimation algorithms applied to RGB videos [64] or depth maps [68]. On the other hand, this information can be gathered through motion capture systems. Although human pose estimation is affected by changes in viewpoint, motion capture systems remain unaffected by variations in view and lighting, ensuring the accuracy of skeleton data. Nonetheless, utilizing motion capture systems might not always be practical in numerous application scenarios.

Utilizing skeletal information in HAR presents several benefits. It offers details regarding body structure and pose, offering a straightforward yet comprehensive representation. Additionally, it remains unaffected by scale differences, providing stability. Furthermore, it exhibits

resilience against changes in clothing textures and backgrounds. Given the accessibility of precise and cost-effective depth sensors, the focus on skeleton-based HAR has notably increased within the research community.

While RGB methods offer high accuracy in HAR tasks, the simplicity, scalability and less noisy nature of skeleton-based procedures make them an attractive alternative. Less noise, in this context, means that when a human is detected using skeleton-based methods, the HAR model focuses solely on the specific areas corresponding to the human body joints and movements. This contrasts with RGB-based methods, where the presence of other objects and human movements in the background can introduce noise and potentially affect the model's input and performance.

Moreover, the skeleton-based models are less prone to environmental clutter and varying light conditions, allowing for a focus on the activity being conducted [69]. For instance, changes in lighting conditions in an indoor environment can alter colour intensity, potentially impacting the accuracy of RGB-based models. In contrast, skeleton-based HAR remains unaffected by such variations, as it focuses solely on body joints and movements, thereby maintaining consistent performance regardless of changes in ambient lighting.

Early research in skeleton-based HAR, similar to image-based methods, relied on shallow learning and hand-crafted features and relations between poses [70]. However, recent work has revolutionized skeleton-based activity recognition methods by applying deep learning methods for higher performance [71].

Various methods have emerged based on the spatial and temporal nature of human activity. *Sequence models* like Recurrent Neural Network (RNN) [72] or Long Short Term Memory (LSTM) consider the sequential nature of input skeleton data as time series. CNN based models [73, 74] have shown great results in spatial information compared to RNN models. GNN based models [75, 76] represent spatial and temporal information by the human skeleton's natural topological graph structure. The ST-GCN [75] is the first model in this category, which considers harmony in spatial and temporal dependencies and benefits from the CNN architecture. Recently, *transformer* models have been used in HAR tasks to achieve competitive outcomes. Some have modified GCN models [77, 78], while others [79] are purely transformer-based.

2.3.4.1 Spatial–Temporal Graph Convolutional Networks(ST-GCN)

This section discusses the pipeline of one of the prominent models in HAR known as ST-GCN, as introduced by Yan et al. [75]. Renowned for its innovative design, the integration of convolutional and graph models has garnered significant attention within skeleton-based models. Numerous studies have adopted and benchmarked their results against this model, emphasizing its significance within the field. In the following, exploring the pipeline and structure will gain insights into the overall workflow.

Diagram 2.5 offers an overview of the ST-GCN process, delineating three key components: *pose extraction* from the input video stream, the *graph and convolution network*, and the *classifier*. This illustration serves to show the various stages and their connection within the ST-GCN model.

Generally, the pipeline contains keypoints extractor, then skeleton data is created from body joints, which can be extracted by motion capture devices like Xtion or Microsoft Kinect and pose estimation algorithm. This frame output data type is a vector of coordinate joints. Then a sequence of joint data constructs a spatial-temporal graph. Each graph node is a body joint, and edges are the connectivity of joints of the body and their connection in time. Finally, the multi-

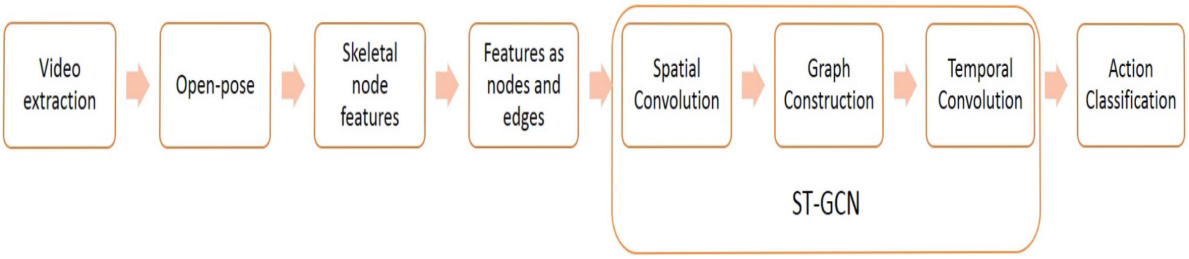


Figure 2.4: ST-GCN Model pipeline [75]

layer convolution can exploit high-level features. In the last section, a classifier is responsible for categorizing the activities.

Two datasets have been used in the ST-GCN model, the Kinetics-skeleton and NTU RGB+D. They have been chosen because the Kinetics-skeleton includes transformed skeleton data derived from the original dataset by the Openpose library, and NTU RGB+D data includes skeleton data initially from the dataset. Thus the skeleton extraction (keypoints) phase in the main process does not exist, yet there is a data preparation phase.

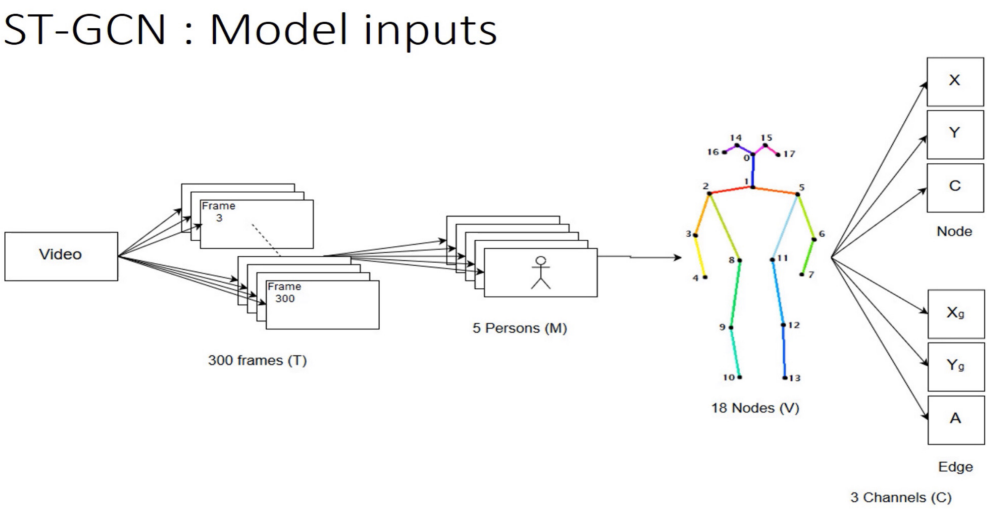


Figure 2.5: ST-GCN Model inputs [75]

In [75], the Spatial Temporal Graph is fed with incoming data from two datasets. The Kinetics dataset only includes raw video without skeletons, necessitating the extraction of skeletons using tools like Openpose. The resulting dataset is referred to as Kinetics-skeleton. On the other hand, the NTU RGB+D contains raw video and skeleton, in which the skeleton data is captured by RGB-D devices like Microsoft Kinect. Both skeleton datasets provide different data types, the Kinetics-skeleton's output is a *Json* file, while NTU RGB+D skeleton data is *.skeleton* format. ultimately, both data formats need to be transformed into an array to function as a database.

Figure 2.6 shows the spatial and temporal graph of the human skeleton. The blue nodes represent the body poses and the edges between each joint in each frame define the human body poses connections. Between each frame, every node has a link to the end of the sequence. All these joints create the graph model and feed the ST-GCN section.

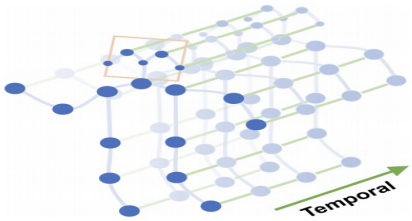


Figure 2.6: Spatial and temporal graph of human key-joints [75]

Network Architecture: Introducing the ST-GCN network architecture provides a deeper understanding of one of the prominent approaches in skeleton-based models for human activity recognition. Figure 2.7 illustrates that the model includes nine layers of ST-GCN units called spatial-temporal graph convolution operators. Sixty-four channels of output for the first three layers, 128 channels for the following three and 256 channels for the final three layers. Nine temporal kernel sizes have been applied in layers. The Resnet mechanism is used for each ST-GCN unit. The random dropout with 0.5 Probability has been chosen to overcome the overfitting. The stride of all layers is equal to one except for layers four and seven, which are two. After the shared convolutional layers, a global pooling is applied, which feeds the softmax classifier.

2.3.4.2 Skeleton-based HAR Leader Board Analysis

This section explores three well-known skeleton-based human activity recognition datasets, each distinguished by unique characteristics. The examination centres on employing the same human activity recognition models across various datasets, highlighting the significance of dataset characteristics. The investigation of skeleton-based action recognition reveals that NTU-RGB+D [80], NTU-RGB+D 120 [81], and Kinetics-skeleton [82] datasets are trending nowadays. Table 2.3 illustrates these datasets' top-ranked skeletal model performances. The rank number, model's accuracy, and year of publication have been provided to show the diversity of ML models and their sometimes varying behaviour in different datasets.

The *PoseC3D* [83] method has the highest accuracy in two datasets (Kinetics-skeleton and NTU-RGB+D) and stands at rank nine in one other. In addition, a different variation of PoseC3D, RGB + Pose, has ranked five in kinetics skeleton and first and second rank in two others.

Considering the available number of Skeleton-based HAR models, NTU RGB+D has the highest with 85, followed by NTU RGB+D 120 with 38, and then Kinetics-skeleton with 18. In Table 2.3, the top ten models in terms of accuracy in almost all datasets have been considered. Kinetics-skeleton is the base dataset for sorting the model ranks. Given that not all models are applied in all three datasets, comparable results are not always available. The total number of models is 21.

The range of accuracy is not the same in all datasets. The highest performance in the Kinetics-skeleton dataset belongs to PoseC3D (w.HRNet 2D skeleton) with 47.7%, following that by almost 9% difference, the 2s-AGCN+TEM [84] model accuracy is 38.36%. Ironically, the rest of the model's accuracy was distributed in 1%, from 38.4% for DualHead-Net [85], the 3rd rank, to

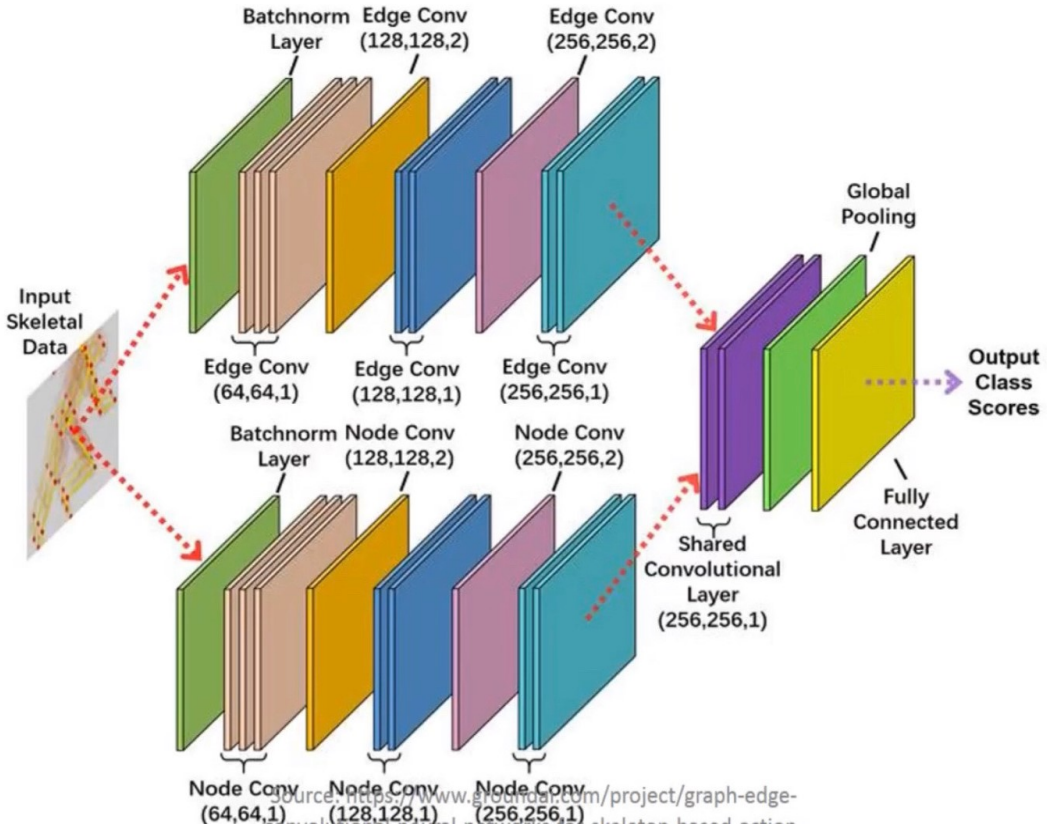


Figure 2.7: ST-GCN Model architecture [75]

37.4% for ST-TR-agcn [77], the 10th rank. However, the total accuracy range of ranks in two other datasets is like a uniform distribution. Low difference, 0.7% for NTU RGB+D and 3%, for NTU RGB+D 120. However, based on this evidence, it is unreliable to say that a method is superior by considering its rank in just one dataset. For example, the EfficientGCN-B4 [86] model stands in the third stage on the leaderboard for NTU RGB+D 120 dataset, but its rank in NTU RGB+D is 22. Likewise, PoseC3D (w. HRNet 2D skeleton), which has outstanding results in the Kinetics-skeleton dataset, and the highest accuracy in NTU RGB+D, stands in stage nine in the leaderboard for NTU RGB+D 120 dataset. However, another variant of PoseC3D (RGB + Pose) has conspicuous accuracy in the NTU RGB+D 120 (95.3%) dataset and high accuracy in two others.

On the other hand, models like CTR-GCN [87], Skeletal GNN [88], 2s-AGCN+TEM [84], DualHead-Net, AngNet-JA + BA + JBA + VJBA [89], MS-G3D [76], CGCN [90], has reasonable accuracy because they stand in top rank in two or all datasets, respectively.

To summarise, although the number of skeleton-based human activity recognition methods and their variation is increasing, there is still room for improvements for models to be applied in challenging datasets such as Kinetics-skeleton. The comparison reveals that dataset details have a direct effect on the ML model accuracy. For example, kinetic-skeleton data is collected from YouTube videos and includes an uncontrolled environment, and the NTU RGB+D videos were captured in a controlled environment. The top accuracy for the kinetic is almost 50% lower than others. Besides, this review illustrates that the same model may not perform as well in a

Table 2.3: Results Of Skeleton-Based HAR Leader Board In Three Datasets

Model	Kinetics-skeleton	NTU-RGB+D	NTU-RGB+D120
PoseC3D(Pose)	1 , 47.7% , 2021	1, 97.1%, 2021	9, 86.9%, 2021
PoseC3D(P+RGB)	5 , 38% , 2021	2, 97.0%, 2021	1, 95.3%, 2021
CTR-GCN	NA	3, 96.8%, 2021	2, 89.9%, 2021
EfficientGCN-B4	NA	22, 95.7%, 2021	3, 88.3%, 2021
Skeletal GNN	NA	4, 96.7%, 2021	7, 87.5%, 2021
PA-ResGCN-B19	NA	17, 96%, 2021	8, 87.3%, 2020
Ensemble-top5	NA	NA	9, 87.22%, 2020
2s-AGCN+TEM	2 , 38.6%, 2020	NA	NA
4s Shift-GCN	NA	6, 96.5%, 2020	13, 85.9%, 2020
DualHead-Net	3, 38.4%, 2021	5, 96.6%, 2021	4, 88.2%, 2021
AngNet-JA	NA	7, 96.4%,2021	6, 88.2%, 2021
DSTA-Net	NA	8, 96.4% 2020	11, 86.6%, 2020
Sym-GNN	NA	9, 96.4%, 2019	NA
MS-G3D	4, 38%, 2020	NA	NA
Dynamic GCN	6, 37.9%, 2020	13, 96%, 2020	NA
MS-AAGCN	7, 37.8%, 2019	11, 96.2%, 2019	NA
CGCN	8, 37.5%, 2020	10, 96.4%, 2020	NA
JB-AAGCN	9, 37.4%, 2019	15, 96%, 2019	NA
ST-TR-agcn	10, 37.4, 2020	12, 96.1%, 2020	17, 82.7%, 2020

Three values in datasets' row define the *Rank* , *Accuracy*, and *Year* of publication respectively.

different dataset.

On the other hand, developing a comprehensive and real-world activity recognition is demanding, particularly given the nature of some *Deep-Learning* (DL) approaches, which require extensive data and significant processing power e.g. CPU and GPU nodes. This results in a lack of comprehensive benchmarks [91] for evaluating the performance of activity recognition algorithms. One approach to solve this problem is dataset specialization, in which elements such as theme, activity, task, and subject adhere to specific criteria. For example, in This work, the aim is to apply HAR in the AAL context using a skeleton-based and multi-view dataset.

However, in the MV-HAR systems, a lightweight machine learning approach is essential for providing real-time and resource-constrained applications like robots. A low computational cost, fewer training parameters, and an efficient algorithm enable the system to be more practical for long-term deployment in assistive living scenarios. These systems can perform effectively in resousce constrains Hardware since they demands less resourse like memomor and process. However, focusing on the number of training parameters of the existing skeleton-based models shows that many methods are not computationally effective. For example, considering some single-view high-accuracy models in the Table 2.3, *PoseC3D* [83] in different variation has 2m to 8m parameters and 2s-AGCN+TEM [84] has 6.94m parameters. Expanding the comparison to the multi-view could indicate models with significantly more parameters.

2.4 Convolutional Neural Networks

Convolutional neural networks (CNNs) have demonstrated remarkable capabilities in feature representation [92], particularly in interpreting data that may not be easily interpretable by the human mind [93]. Through multiple layers of convolutional and pooling operations, CNNs can capture intricate patterns and relationships within the data, enabling them to recognize complex visual patterns and make accurate predictions. CNNs have been successfully applied in various domains, including image recognition, speech recognition, and natural language processing [93–95]. These models are designed to capture local dependencies and scale invariance, making them well-suited for tasks involving feature representation and classification [96].

LeNet, initially introduced by LeCun et al. in the early 1990s [97], is a relatively simple CNN model with a limited number of training parameters [98]. Originally developed for handwritten digit recognition, it proves suitable for tasks with constrained computational resources [99]. M-Lenet is a modified version of LeNet specifically designed for multi-view human activity recognition, which is described in detail in Chapter 4.

MobileNet [100], SqueezeNet [101], and MnasNet [102] all share the common goal of being lightweight and efficient neural network architectures, each designed with a specific purpose in mind. MobileNet, tailored for mobile and embedded devices, focuses on computational efficiency through depth-wise separable convolutions, making it well-suited for real-time applications on resource-constrained platforms. SqueezeNet excels in minimizing model size without compromising accuracy, making it ideal for resource-constrained scenarios. With comparable accuracy to AlexNet, SqueezeNet drastically reduces model complexity by utilizing 50 times fewer parameters and maintaining a remarkably small size [101]. MnasNet, short for Mobile Neural Architecture Search, leverages neural architecture search techniques to strike a balance between accuracy and efficiency, making it a suitable choice for mobile and edge devices where optimized architecture is vital.

Residual Network [103](ResNet) and DenseNet [104] are complex and highly influential convolutional neural network architectures, both devised to excel in demanding tasks. ResNet's primary innovation of residual connections combats the vanishing gradient problem in deep networks, enabling the training of exceptionally deep models, while DenseNet fosters feature reuse by densely connecting layers, promoting efficient training of very deep networks. ResNet was created to facilitate the training of deep networks for image classification and computer vision, while DenseNet's core objective was to enhance feature propagation and reuse, making it invaluable for image classification and computer vision applications, marking both as groundbreaking in deep learning.

2.4.0.1 Graph-based Convolutional Neural Networks

There are numerous variations of CNN-based models that have been utilized in the field of Human Activity Recognition (HAR). The chosen CNN models were contrasted with Graph-based Convolutional Networks (GCN), which is one of the emerging trends in the development of HAR. Table 2.4 depicts some of these models, focusing on their complexity by comparing the number of training parameters and FLOPs.

Among these models, ST-GCN (spatiotemporal graph convolutional networks) [75] stands as a foundational architecture specifically designed for this task. It harnesses the potential of graph convolutional networks to comprehend the intricate dynamics of human skeletons. However, ST-GCN's computational complexity, measured in terms of FLOPs, is relatively high,

	ST-GCN	CoST-GCN	AGCN	CoAGCN	S-TR	CoS-TR	DGNN	AS-GCN	AGC-LSTM
FLOPs(G)	16.73	0.27	18.69	0.3	16.14	0.22	126.8	27	54
Params(M)	3.14	3.14	3.47	3.47	3.09	3.09	-	-	-

Table 2.4: A comparative Analysis of graph-based convolutional networks in human activity recognition.

featuring computational complexity (Giga FLOPs) and model size (Million Parameters)

making it resource-intensive. While this architecture excels in predictive accuracy, its significant computational demands might pose challenges for real-time applications.

To address the computational burden inherent in online inference, researchers have proposed the use of continual inference networks [105], introducing optimized versions of GCN models. CoST-GCN, CoAGCN, and CoS-TR represent these evolved models. Notably, CoST-GCN offers a remarkable 109× reduction in time complexity, demonstrating its suitability for real-time online inference scenarios. CoAGCN and CoS-TR further underline the efficiency of continual inference networks, with reduced computational requirements, making them well-suited for applications where minimizing processing overhead is essential.

2.5 Multi-modality and Multi-view

The terms "multi-view" and "multi-modality" are related concepts in the field of deep learning, but they refer to different aspects of data representation and processing. The multi-view concept, as discussed in [106], involves utilizing multiple perspectives or representations of the same data to enhance the learning process. In traditional machine learning approaches, a single view of the data is used, which may limit the model's ability to capture the underlying patterns and relationships. Multi-view learning covers various aspects including representation learning [107, 108], feature selection [107, 109], and fusion methods [110–112].

On the other hand, multi-modality refers to the integration of information from different modalities or types of data to improve the learning process and capture complementary information [113]. In this context, modalities can refer to different types of data, such as images, text, audio, or sensor data, that provide different perspectives or aspects of the same phenomenon [114]. Multi-modality approaches aim to leverage the strengths of each modality and combine them to obtain a more comprehensive and informative representation of the data [115]. This can be achieved by designing architectures that can process and fuse information from multiple modalities, such as using parallel streams or fusion mechanisms [113].

The importance of the multi-view concept lies in its ability to provide a more comprehensive understanding and representation of complex and high-dimensional data [116]. By considering different views, the model can capture different aspects and nuances of the data, leading to improved accuracy [106] and generalization [117, 118]. The applications of the multi-view concept are wide-ranging. In computer vision, for example, multi-view learning can be applied to object recognition tasks, where different views of an object (e.g., different angles or lighting conditions) can be leveraged to improve recognition accuracy [107]. In natural language processing, multi-view learning can be used for sentiment analysis, where different views of textual data (e.g., word embeddings, syntactic structures) can provide a more comprehensive understanding of sentiment [108]. In this work, the robotic perspective is augmented by supplementary cameras observing the same human subject engaged in indoor activities.

2.5.1 Multi-view Convolutional Neural Networks

CNNs excel in domains like image classification [92, 119]. However, single-view CNN architectures may not fully utilize the multi-view information available in the target data. To address this limitation, multi-view architectures aim to integrate information from different views of the same data to obtain more discriminative and comprehensive representations [120].

There are two main types of multi-view CNN architectures: the one-view-one-net and the multi-view-one-net mechanisms [106]. In the one-view-one-net mechanism, each view is processed by a separate CNN, and the outputs are combined to obtain the final representation [106]. On the other hand, the multi-view-one-net mechanism models multiple feature sets together and aims to learn a common representation that captures the multi-view information [106].

Several studies have explored the application of multi-view CNNs in different domains. For example, in 3D shape recognition, multi-view CNNs have been used to model multiple views of 3D shapes and achieve better recognition performance [120]. In multivariate electroencephalography (EEG), multi-view CNNs have been employed to integrate information from multiple electrodes and improve classification accuracy [121]. Additionally, multi-view CNNs have been applied to tasks such as multi-feature aggregation [122] and lung nodule classification [123].

The design of multi-view CNN architectures involves various parameter optimization techniques [120]. For instance, the use of attention mechanisms, such as SoftPool attention, has been proposed to enhance the feature extraction and classification process [121]. Another approach is the fusion of spatial and temporal networks at different layers, which has been shown to improve performance while reducing the number of parameters [50].

Despite the strides made in multi-view CNNs, there remain critical gaps in the existing research, prompting further exploration. Firstly, there is no current comparison of various multi-view HAR approaches employing CNN methodologies to assess their performance. Furthermore,introducing multiple perspectives simultaneously significantly increases the overall complexity of the models, an area that is relatively under-explored in existing literature.. Notably, extending the concept of multi-view CNNs into the realm of activity recognition may compound this complexity. Furthermore, the research landscape has seen limited investigation into the optimal utilization of CNN models in the context of MV-HAR.

Hence, the present work seeks to bridge these gaps by presenting a comprehensive comparison of CNN methods, while investigating the effects of incorporating additional views on the model's efficiency. Ultimately, this work aims to provide a streamlined pipeline that addresses these challenges head-on.

2.5.2 Multi-view HAR Efficiency

The HAR pipeline encompasses multiple stages, as detailed in Section 2.3.4.1, using the ST-GCN pipeline as an illustration. These stages typically involve data collection, preprocessing, feature extraction, representation, classification, and activity recognition. Each stage within this pipeline significantly influences the overall efficiency and effectiveness of the model. This holds, particularly in the context of multi-view pipelines, where similar concerns are present. Therefore, to attain a highly efficient model, it becomes imperative to carefully consider and address all these aspects across the entire HAR process.

In the data collection phase, various sensors, including vision-based sensors such as cameras, are used to capture data related to human activities [5]. Vision-based HAR, specifically skeleton-based HAR, utilizes 2D pose estimation techniques to extract skeletal information from the

captured data. These techniques analyze the spatial relationships between body joints to recognize activities [124]. Once the data is collected, it undergoes preprocessing to remove noise, filter out irrelevant information, and normalize the data for further analysis [125].

In skeleton-based HAR, features are extracted from the 2D pose estimation results, such as joint angles, joint velocities, or spatial relationships between joints. These features capture the essential characteristics of human activities and serve as input for activity recognition algorithms. Activity recognition algorithms are trained on labelled data, where human activities are manually annotated, to learn the patterns and characteristics of different activities [124]. In skeleton-based HAR, these algorithms utilize the extracted features from the 2D pose estimation to classify the data into specific activities [124].

The HAR pipeline can be streamlined and made more efficient by incorporating various strategies. One approach is to leverage deep learning architectures, such as CNNs, which have shown superior performance in HAR tasks [126]. These models can automatically learn hierarchical representations from raw sensor data, eliminating the need for manual feature engineering and reducing computational complexity.

Another strategy is to explore multi-view fusion techniques, which combine data from multiple sensors or camera views to enhance recognition accuracy [127]. By fusing information from different views at different stages, such as the last convolutional layer or class prediction layer, the pipeline can be optimized for improved performance.

Context-aware approaches can also be employed to improve efficiency and accuracy in HAR [128]. By incorporating contextual information, such as the environment or additional sensor modalities, the system can better understand and recognize human activities in real-world scenarios.

Efficient feature representations, such as dynamic imaging or view-invariant shape dynamics, can be utilized to reduce computational complexity and storage requirements while capturing the essential characteristics of human activities [127] [129].

Ensemble methods, such as classifier ensembles, can enhance the performance of the HAR system by combining the predictions of multiple classifiers [130, 131]. This approach benefits from diverse perspectives and can improve overall accuracy, especially in complex and diverse activity recognition tasks.

Optimizing data collection and preprocessing techniques is crucial for a streamlined HAR pipeline [132, 133]. Carefully selecting appropriate sensors, optimizing sensor placement by analysing data quality, and applying effective noise reduction and data normalization methods can ensure high-quality and well-preprocessed data, leading to more efficient subsequent stages of the pipeline.

Through an extensive analysis of the Multi-view human activity recognition (MV-HAR) pipeline, it's apparent that each stage significantly impacts the model's efficiency, especially within multi-view pipelines. As we progress through subsequent chapters, the focus will revolve around revisiting and refining specific segments of this pipeline. By exploring deeper into data collection, preprocessing, feature extraction, classification, and activity recognition phases, the aim is to optimize individual components. This strategic approach intends to amalgamate diverse insights, ultimately crafting a more efficient and effective MV-HAR pipeline capable of superior activity recognition in multi-view scenarios.

CHAPTER 3

DATA PREPARATION AND ANALYSIS

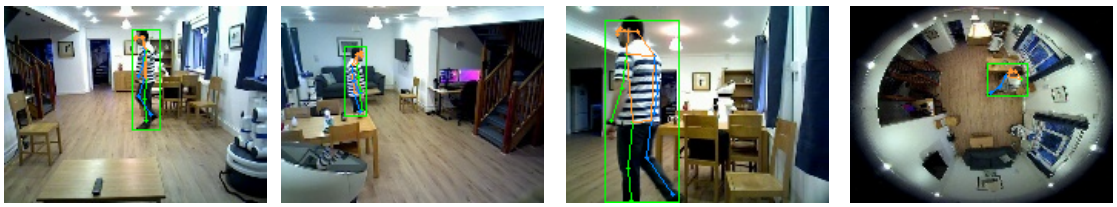
This chapter presents two main contributions to human activity recognition (HAR) in ambient assisted living scenarios. Firstly, a novel dataset, RHM-HAR-SK, comprising human skeleton data produced using an existing RGB dataset [1] is presented. The dataset contains extracted skeletons of human activities from four different perspectives, including a Robot-view, which makes it particularly valuable for HRI AAL scenarios. It aims to provide a rich information source to train and test the performance of human activity recognition approaches. Moreover, the dataset allows for detection algorithms to rely on low-dimensional skeleton data instead of videos. It therefore reduces computing resources, which are otherwise computationally expensive considering the multiple parallel video streams. The availability of both RGB and skeleton data within the same dataset facilitates the opportunity to evaluate and compare the performance of various classification methods.

As a second major contribution, this chapter demonstrates the potential of using additional camera perspectives to enhance an assistive robot's activity recognition pipeline. For that, it relies on a measurement of the information content provided by the different views by analysing the number of missed frames and missed poses. Additionally, this exploration involved comparing two pose extraction methods to assess their influence on the analysis. Furthermore, spatial and temporal analyses from multiple viewpoints are conducted, exploring human skeleton movement and the information it conveys across various activities.

These comprehensive analyses aim to enhance the understanding of the importance of skeleton-based data, and the effects of incorporating additional perspectives. Subsequently, these insights contribute significantly to addressing and potentially resolving the research questions.

3.1 RHM-HAR-SK Dataset

This section provides information about the *RHM-HAR-SK* dataset that is created on top of the extended version of RHM [1] RGB data, a multi-view human activity dataset. It includes a *single person*, *trimmed video* from *four* independent cameras, two wall-mounted cameras (Front-view and Back-view), one mobile robot camera (Robot-view), and one ceiling fish-eye camera (Omni-view). Cameras were used to cover the whole area resembling an ordinary living



(a) Walking Back-view. (b) Walking Front-view (c) Walking Robot-view. (d) Walking Omni-view.

Figure 3.1: Synchronized skeleton output from different views of the "walking" action.

room, and the videos from different views overlap. This dataset captures fourteen daily indoor activities *[walking, bending, sitting down, standing up, cleaning, reaching, drinking, opening can, closing can, carrying object, lifting object, putting down object, stairs climbing up, stairs climbing down]* in a typical living room of a British home. The conspicuous feature is a *mobile robot* camera synchronized with three other cameras. It enables us to explore the added value of mobile observations in HRI in the context of social and assistive robotics.

In all video clips, the frame size is 640×480 pixels except the Omni-view with 512×486. As shown in Figure 3.1 the bounding box size varies in different frames. The variation is based on the distance of the detected human to the camera, the camera type and position, the subject's body dimension, and the number of detected poses. Looking at Figure 3.1 shows a living room which is equipped with three fixed cameras and one mobile robot with a camera. In Figure 3.1a, the right side features the Fetch robot capturing the video. Figure 3.1c displays the Robot-view, while positioned on the top right of the shelf, to the right side of the image, you can locate the Front-view camera (Figure 3.1b). Contrarily, the Back-view camera is situated atop the sofa. Additionally, Figure 3.1d, depicting a fish-eye view, is positioned in the middle of the living room ceiling, offering an encompassing perspective of the environment.

The dataset comprises a total of 6,701 synchronized videos with each activity sample captured simultaneously from four different camera perspectives. This results in a cumulative total of 26,801 videos across all viewpoints (as illustrated in Figure 3.2). In other words, each sample activity in every activity class is recorded from four perspectives at the same time. The distribution of videos for each action class varies, spanning from 407 to 700 videos across different perspectives. Notably, each video clip in the dataset has a duration ranging from 1 to 5 seconds, capturing essential moments of the activities conducted. Given a video capture rate of 30 frames per second, the video samples contain approximately 30 to 150 frames.

The *HRNet* [64] and YOLOv7 [67] have been used to extract poses from videos. These models have been trained over the *COCO* keypoint detection dataset [134], and the *MPII* Human pose dataset [135]. The pros and cons of these two pose estimation methods have been described in Chapter 2.

The skeleton dataset retains all characteristics of the RGB dataset (RHM), such as the number of samples, video length, views, and action numbers. However, a notable distinction lies in the data generated from each frame, where a body skeleton with 17 keypoints is extracted. The total number of video frames varies across different video streams and activities, aligning with the video length in the RHM dataset. Each pose encompasses both X and Y positions in the 2D scene. Initially, the extracted poses are stored in a *JSON* file, which is subsequently transformed into a *Tensor* file for input into the machine learning training model.

All actions from different views are combined in a single five-dimensional tensor:
$T = \{n, c, f, p, s\}$, where

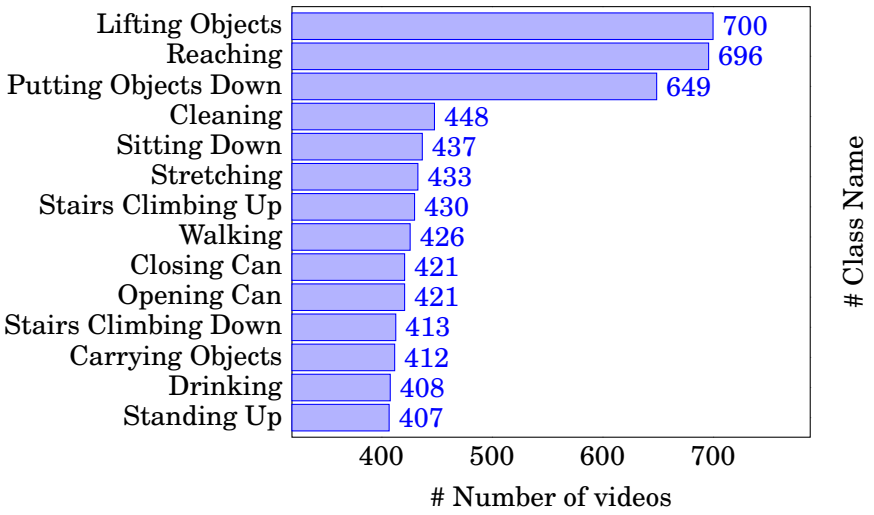


Figure 3.2: RHM [1] Videos number in each class-view

The dataset comprises a total of 6,701 synchronized videos across all camera views, with each viewpoint contributing to a cumulative sum of 26,801 videos (as illustrated in Figure 2). The distribution of videos for each action class varies, spanning from 407 to 700 videos across different perspectives. Notably, each video clip in the dataset has a duration ranging from 1 to 5 seconds, capturing essential moments of the activities conducted. Given a video capture rate of 30 frames per second, the video samples contain approximately 30 to 150 frames. The figure displays a comprehensive list of all the classes included in the dataset, along with the corresponding number of video samples available for each class in every view.

- $n \in \{\mathbb{N}_0 | n < 6701\}$ denotes the *sample number*.

  Note: videos are synchronized, meaning each sample across the four videos from a different camera. Some of the videos are filled with zero (0) values. These refer to a video clip with missing poses;

- $c \in \{\mathbb{N}_0 | c < 4\}$ identifies one of the four *camera views*;

- $f \in \{\mathbb{N}_0 | f < 34\}$ refers to the frame number. Because the nature of the matrix does not support different dimensions, to unify it, 34 frames were randomly selected and sorted as the original sequence.

- $p \in \{\mathbb{N}_0 | p < 17\}$ denotes the *number of extracted poses* up to a maximum of 17 identifiable poses (c.f. Table . 3.1);

- $s \in \{\mathbb{R} | s < 3\}$ combines the relative x and y position plus the *score* of this pose are in this section. The confidence score depicts the reliability level of the extracted pose.

- $l \in \{\mathbb{N} | l < 14\}$ is an individual tensor L with the same dimension of sample number, which shows the class labels for the actions.

3.1.1 The Input Data Size and Sampling

One of the most challenging parts of the HAR task is the video frame sampling. Every video is labelled as a single activity, and the video length is different based on action type and situation.

Then, for the ML models, this variation means having a dynamic input size. Consequently, all parameters in the model should be modified based on the input size. Designing this dynamic model is a significant structural challenge in AI modelling, which is still an open area for improvement. Similarly, the skeleton-based methods need to use fixed-size input data. However, sampling or other reduction-based methods could lose valuable data from a video stream. In this work, *ordered random sampling* method has been used, which fixes the number of frames like 34, 64 and 128, which were selected randomly from entire frames.

A 2D image (Figure 3.3) visualizes the spatial-temporal data. It shows the results of transforming 20 video streams of skeleton data from walking action in robot view to 2D images. The spatial information which is extracted from each video frame is transformed into a single row, a dimension vector with 17 elements. Each element of this row can show the relevant body pose information. They could be X, Y, or the results of a specific function like the Mean square. The X value of all 17 positions is shown in Figure 3.3, the information of these experiments is depicted with a grayscale image to give a better understanding. It emphasizes the variations in length and the change patterns, displaying them graphically in a two-dimensional grayscale image.

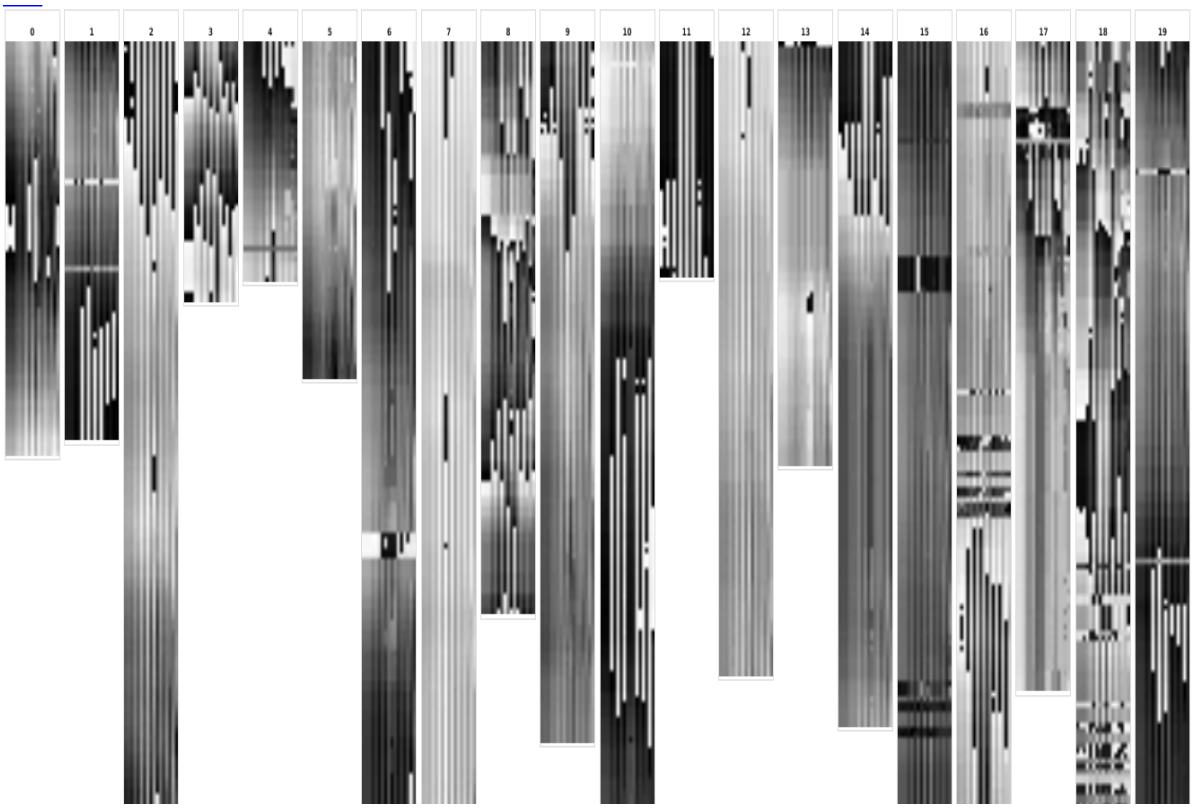
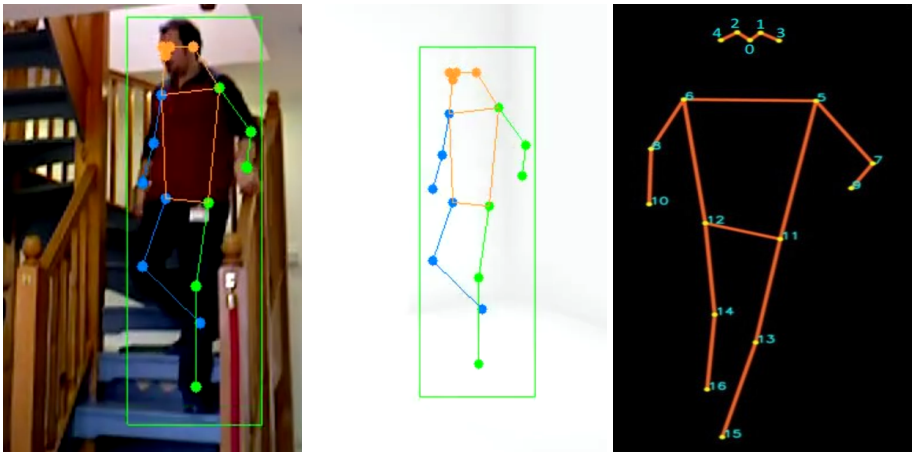


Figure 3.3: The two-dimensional representation of x position from 20 videos with different lengths in Robot-view from walking action.

Figure 3.4 displays a real frame capturing a human engaged in stair climbing down the action, along with the extracted body poses and skeleton, as depicted in Figure 3.4a. Additionally, Figure 3.4b showcases the individual human skeleton data devoid of RGB data. Each pose is represented by a unique index number, as demonstrated in Figure 3.4c, with corresponding nomenclature provided in Table 3.1.

Table 3.1: Table of keypoints index

Index	Keypoint	**Index**	Keypoint
0	Nose		
1	Left eye	2	Right eye
3	Left ear	4	Right ear
5	Left shoulder	6	Right shoulder
7	Left elbow	8	Right elbow
9	Left wrist	10	Right wrist
11	Left hip	12	Right hip
13	Left knee	14	Right knee
15	Left ankle	16	Right ankle



(a) (b) (c)

Figure 3.4: 3.4a shows a subject performing the "stair climbing down" action with skeleton overlay. 3.4b shows only the skeleton of the same action, and 3.4c another skeleton with index.

3.2 Quantitative and Qualitative Analysis

This section focuses on the *quantity* and *quality* of the extracted skeleton and its poses from the RHM-HAR-SK dataset. Two general terms are considered to describe the quality of extracted skeleton from RGB images, the number of *missed frames* and the number of *missed poses*. The primary objective of the analysis is to provide an improved comprehension of the effectiveness of different camera views, and pose extraction models, in human detection and pose extraction quality.

3.2.1 Missed Frames

The RGB frames on which the pose extraction methods could not find any human skeleton is considered a *missed frame*. In cases where the human body is only partially detected but still present in the frame, it is not considered a missed frame. In the RHM-HAR-SK dataset, 14 actions have been captured from four synchronized camera views. The number of frames in all views is the same, but it's different from action to action. Figure 3.5 illustrates the total count of missed frames across four perspectives and 14 actions, each analyzed using the HRNet pose extraction

method. On the other hand, Figure 3.6 displays the outcomes obtained through YOLOv7 pose extraction. A comparison of these figures indicates a nearly identical representation of results, showcasing only minor discrepancies between the two methods in which YOLOv7 has slightly fewer missed frames.

For both figures, the black bars show the total frame number distribution (RGB image frame) in the dataset and each activity individually. The orange bar shows the statistics of Omni-view's camera missed frames, which illustrates that the majority of actions missed the frames, higher than 45%. To be specific in Figure 3.5, the *walking* and *carrying object* actions by 29.6% and 36.5% have the lower missed frames in the Omni-view, respectively. At the same time, these actions have higher frames error in the Robot-view with 0.9% for walking and 1.3% for carrying objects, which is negligible.

Excluding the Omni-view, the highest missed frames belong to the Front-view in *stairs climbing up/down* with 13.3% and 9.4%. Following that, the Back-view has the same pattern in stairs climbing actions by 10.2% and 4.7%.

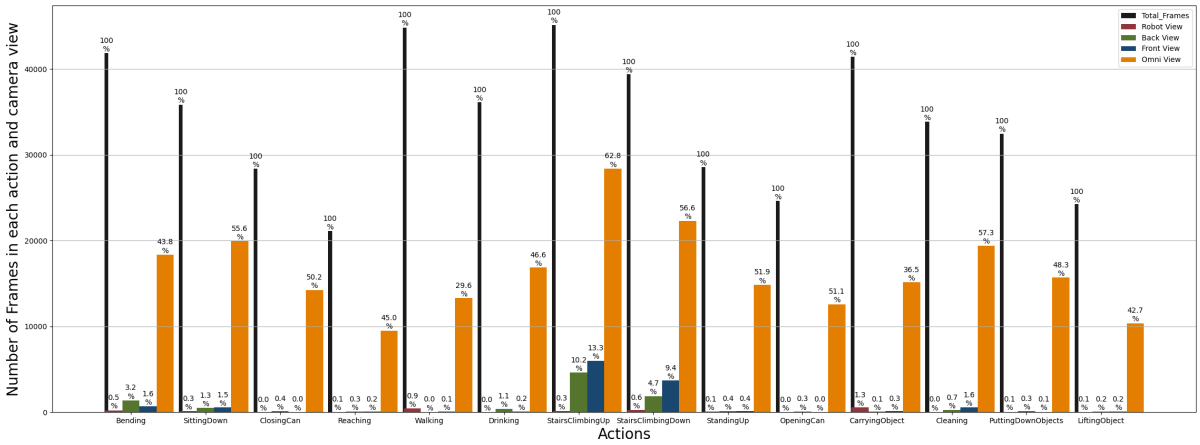


Figure 3.5: Missed frames Across all actions grouped by the view from method

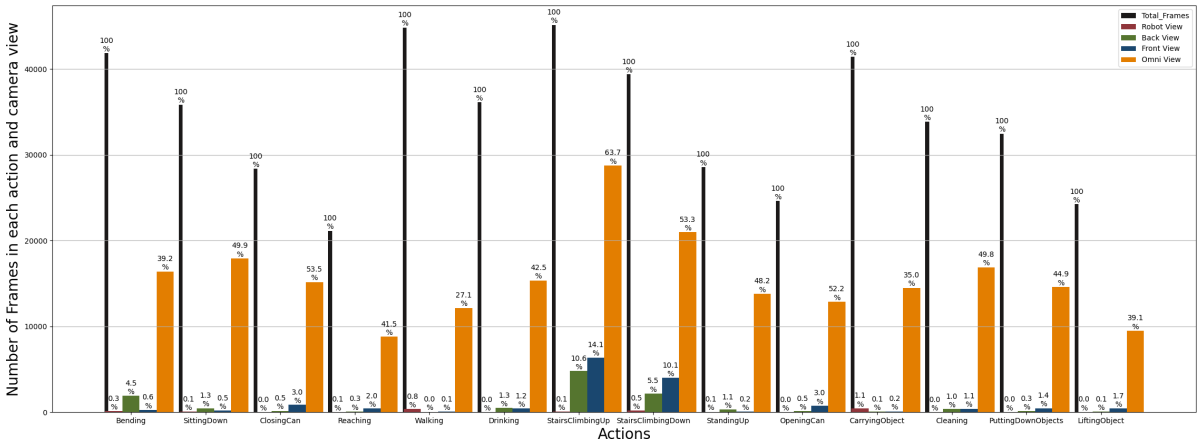


Figure 3.6: Missed frames Across all actions grouped by view From YOLOv7 method

The statistics on missed frames indicate variations in quality across different camera views. The Omni-view is not a reliable source for body pose extraction, while the robot view consistently

exhibits the lowest number of missed frames in both and YOLOv7 pose extraction methods. A comparison between the missed frame results of both pose extraction methods reveals that the YOLOv7 method generally has fewer missed frames. To further analyze, a box-plot comparison of two pose estimation methods is shown in Figure 3.7 and illustrates that there isn't a significant difference between the two methods regarding missed frame percentages.

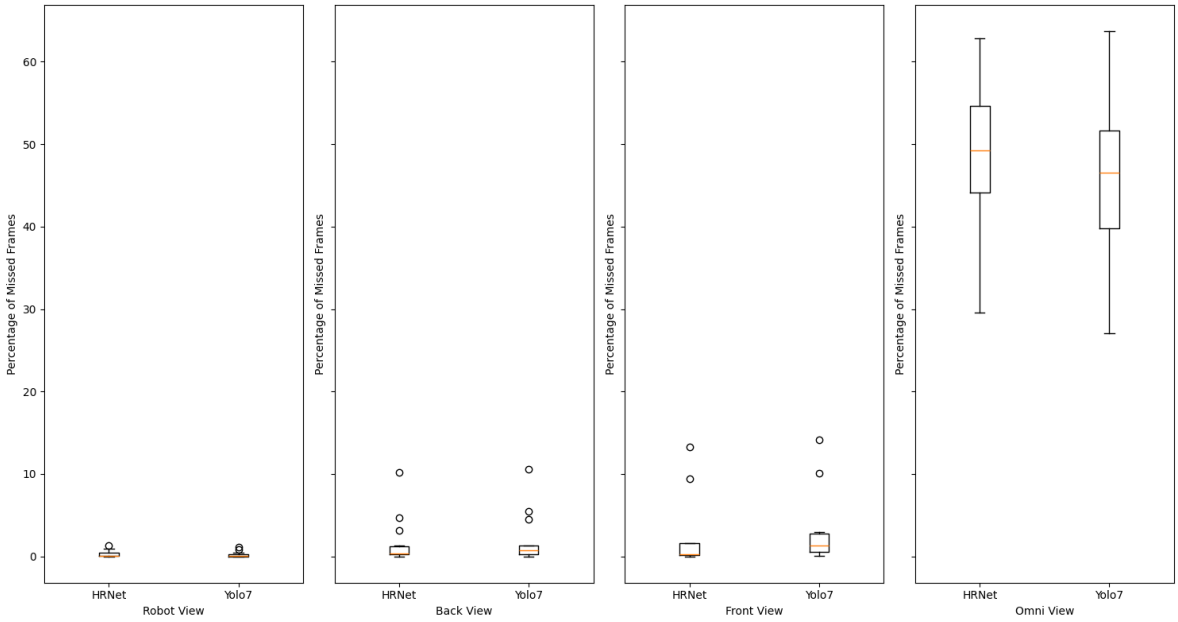


Figure 3.7: The boxplot comparison of and Yolo
Illustrates the performance in all actions on each view

3.2.2 Missed Poses

There are three parameters for each pose, X and Y values in 2D space and the *confidence* score. The confidence value refers to how much the extracted position is accurate. This value is between *0 to 1*, and in this work, the values less than *0.5* are considered as *missed poses*. For example, if the left leg's confidence value is 0.8 and the right leg's is 0.4, then the right is considered as missed poses or missed keypoint.

Figure 3.8 illustrates the total number of all actions' missed poses from three views, and 17 poses separately for the HRNet method and Figure 3.9 show for the YOLOv7 method. The total number of each pose in all activities is almost the same, around 500000. The red, green, and blue bars show the Robot, Back, and Front view cameras' missed poses.

Overall, in Figure 3.8 (the Method) the Back-view has the lowest confidence among the other views (highest number of missed poses) in all poses, and the Front-view and Robot-view have the highest confidence, which changes in different joints. For the Robot-view, the highest number of missed poses belong to the lower body, with more than 50% in ankle joints and around 31% in knee joints, these results for the YOLOv7 method are almost 5% less. However, The outcomes from the YOLOv7 pose extraction method exhibit variations, particularly in the occurrence of missed poses across five specific joints located in the head region (including the nose, left/right

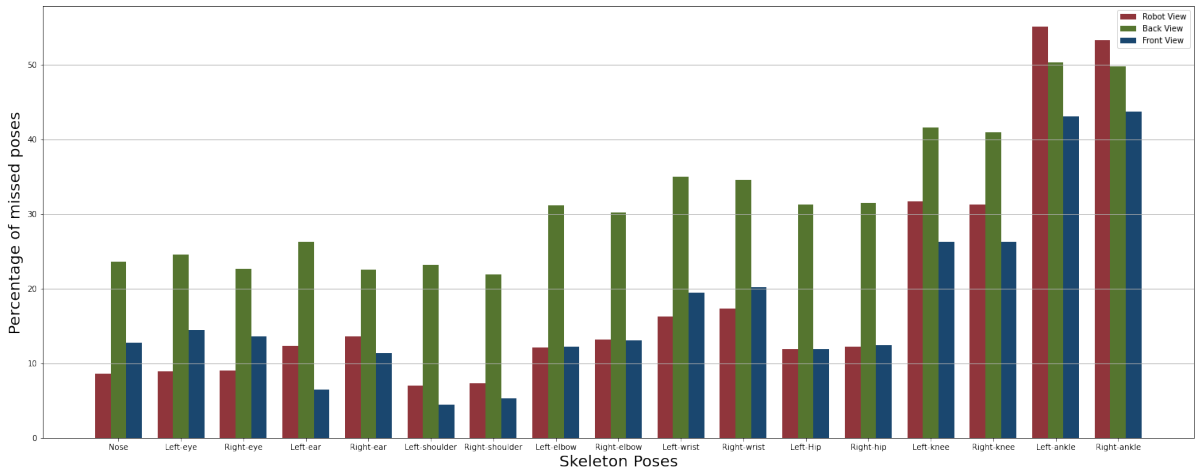


Figure 3.8: Average percentage of frames in all actions with missed skeleton poses across various views extracted via .

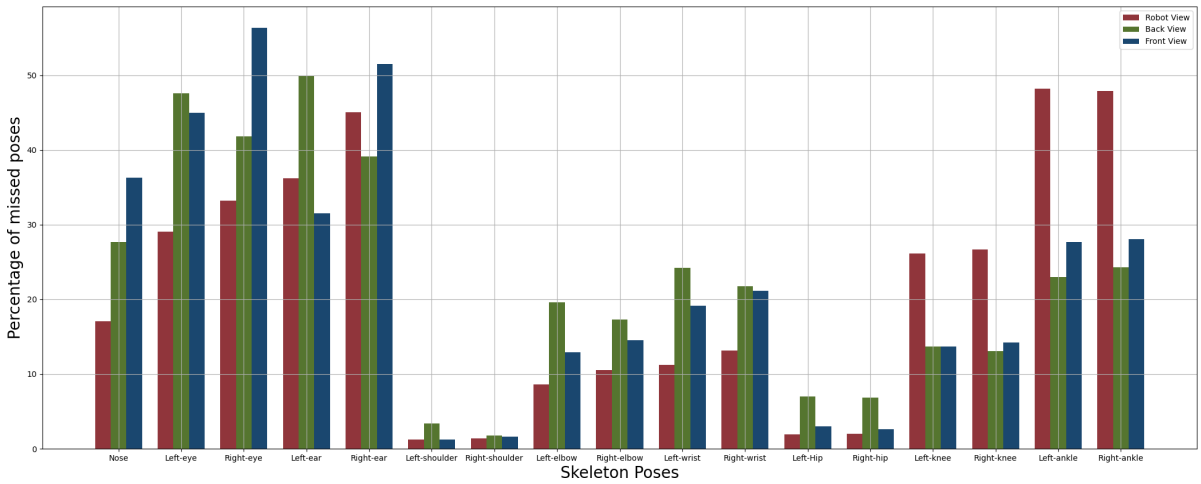


Figure 3.9: Average percentage of frames in all actions with missed skeleton poses across various views extracted via YOLOv7.

ears, and eyes). These discrepancies fluctuate within the range of 20% to 50% across different viewpoints.

The statistical data concerning missed poses for various activities can be found in Appendix 7. Notably, the left and right shoulders consistently demonstrate fewer missed poses across almost all actions compared to other body joints in both methods. Specifically, they exhibit less than 8% missed poses for Robot and Front views in the HRNet method and less than 3% for all views in the YOLOv7 method. Following that, the left and right shoulders consistently show the lowest occurrence of missed poses across both pose extraction methods. At the same time, the right and left elbow and wrist consistently demonstrate higher occurrences of missed poses in both methods. Although the pattern remains similar between the methods, the values in the YOLOv7 method tend to be lower.

Except for stairs climbing up and down actions all other action has a similar pattern, for instance, Figure 3.10 and 3.11 illustrate the walking action statistics, on the other hand, the

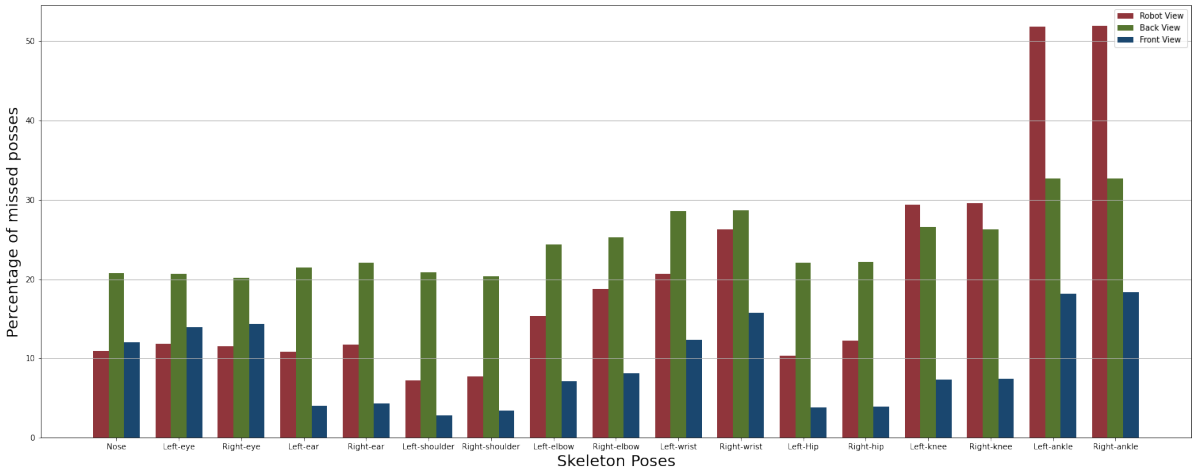


Figure 3.10: Percentage of frames with missed skeleton poses of "walking actions" across various views extracted via HRNet.

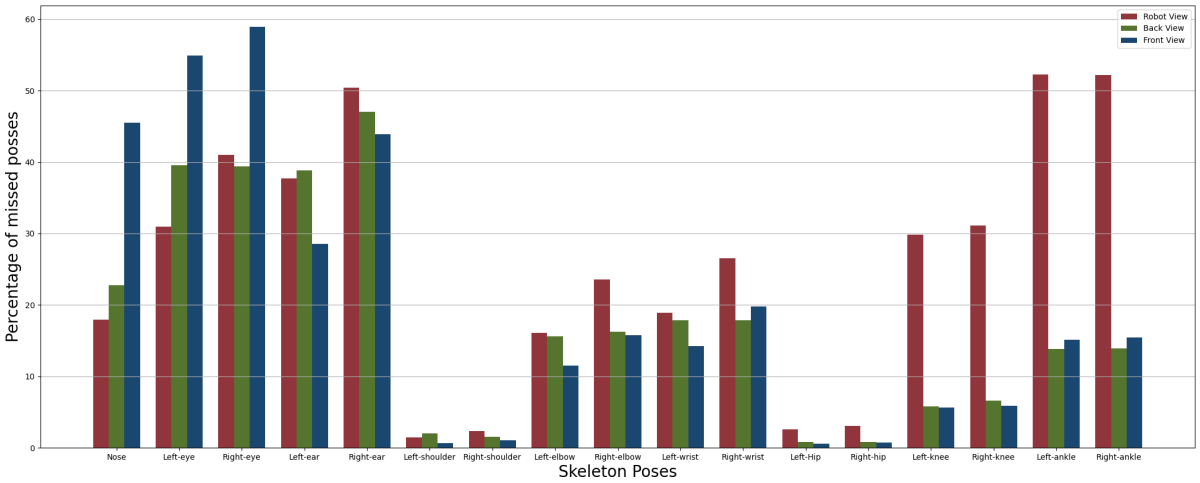


Figure 3.11: Percentage of frames with missed skeleton poses of "walking actions" across various views extracted via YOLOv7.

statistics in stairs climbing up (Figure 3.12 and 3.13) and down are slightly different from all other actions. Robot camera-view shows superior results in these actions with very low missed poses.

Chapter 4 has compared the results of two different pose extraction methods on the classification results. The results show that the YOLOv7 model has an improvement in the accuracy.

3.3 Spatial and Temporal Analysis

In this section, the spatial and temporal variations of the human skeleton are thoroughly analyzed within the RHM-HAR-SK dataset, which includes human skeletal data analysed in the previous section. The primary objective is to examine and compare different perspectives to evaluate the potential benefits of incorporating multiple views, particularly to enhance a single view, such as the Robot-view, in HAR. Through spatial analysis, differences in joint movements among

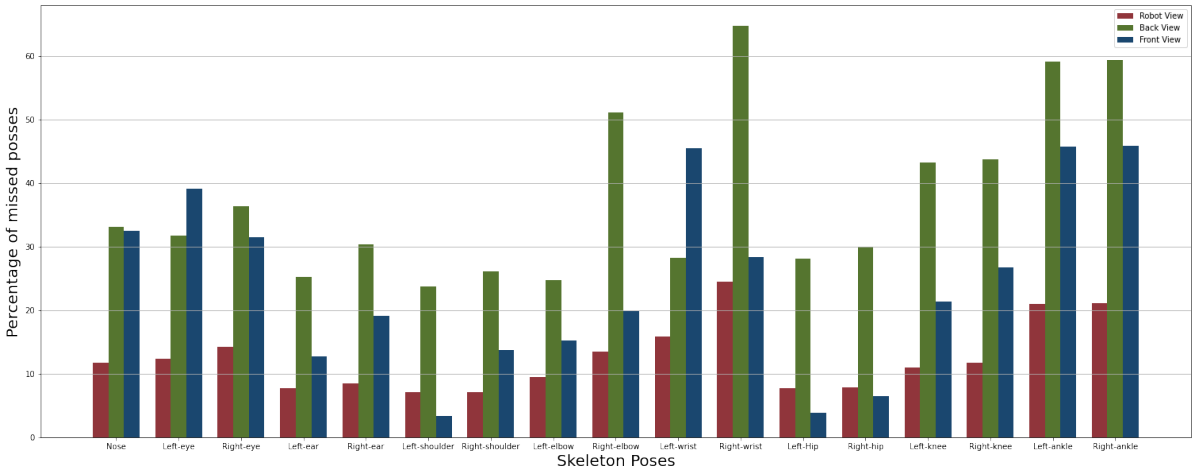


Figure 3.12: Percentage of frames with missed skeleton poses of the "stairs climbing up" actions across various views extracted via HRNet.

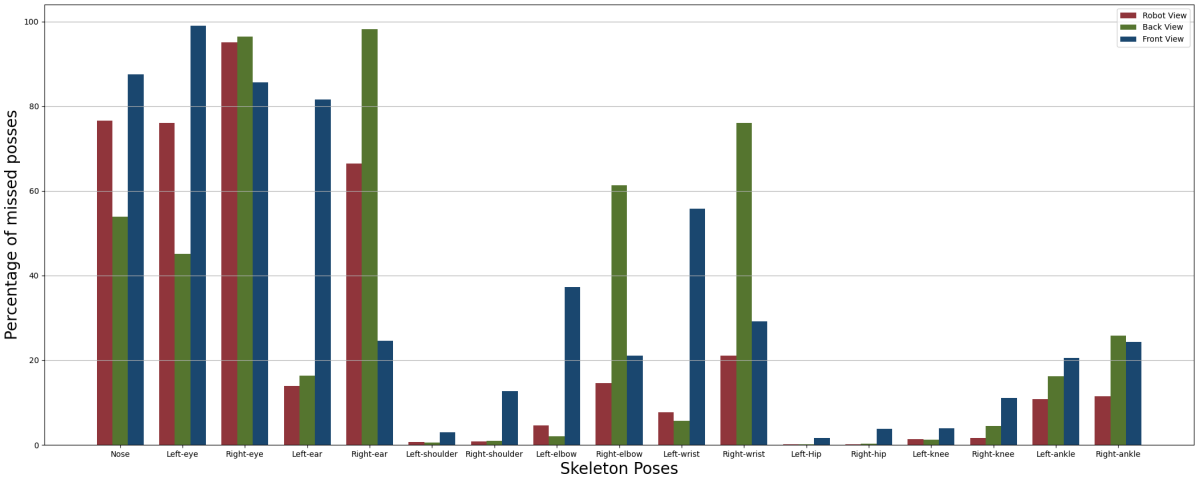


Figure 3.13: Percentage of frames with missed skeleton poses of the "stairs climbing up" across various views extracted via YOLOv7.

different activities and views are discerned.

Moreover, the temporal analysis explores each human skeleton portrayed in separate frames. By incorporating the concept of Mutual Information from information theory, the spatial information captured in each activity from various camera perspectives is assessed. This examination evaluates the amount of information conveyed during the occurrence of an action in synchronized views. The analysis demonstrates the potential to enhance overall detection accuracy in HAR by effectively combining multiple views.

This understanding guides us to answer research questions, specifically helping us to answer RQ two and indirectly related to RQ one and three.

3.3.1 Analytical Methods for Joint Movement Representation

To commence the analysis, the input data is transformed into a distance representation, encompassing the x and y values of each joint. These values denote the pixel positions in a 2D image.

Two approaches are introduced for measuring the Euclidean distance between joints: Pairwise Distance and Min-Max Distance. The mathematical formula (Formula 3.1) for calculating the distance between two points in a 2D space is called the Euclidean distance formula. It is derived from the Pythagorean theorem. Given two points $P1(x1, y1)$ and $P2(x2, y2)$, the Euclidean distance (d) between them can be calculated using the following formula:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{3.1}$$

Here, $(x2 - x1)$ represents the difference in the x-coordinates of the two points, and $(y2 - y1)$ represents the difference in the y-coordinates. Squaring these differences, summing them, and taking the square root gives you the Euclidean distance between the two points.

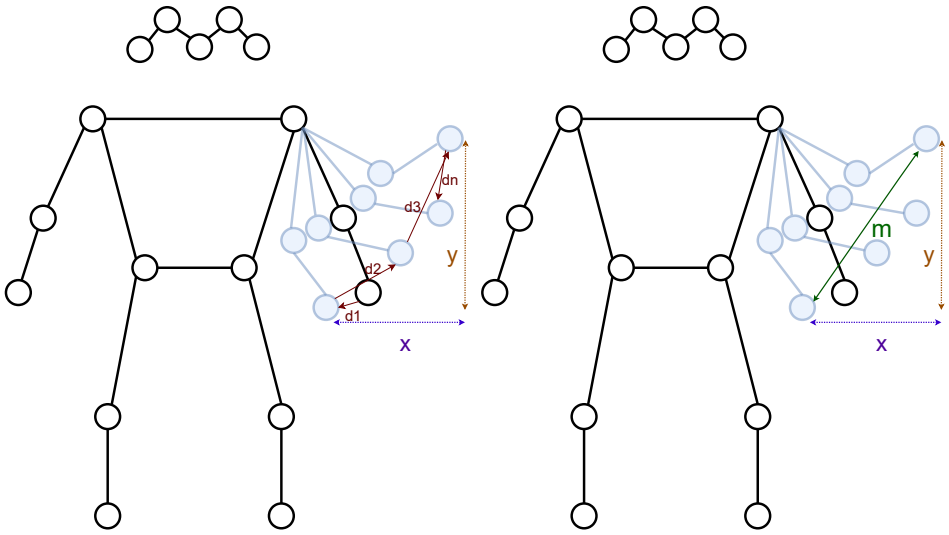


Figure 3.14: Movement visualization,
d on the left side shows the pairwise distance and m on the right side refers to the distance between the minimum and maximum position of a joint during acting.

3.3.1.1 Pairwise Distance

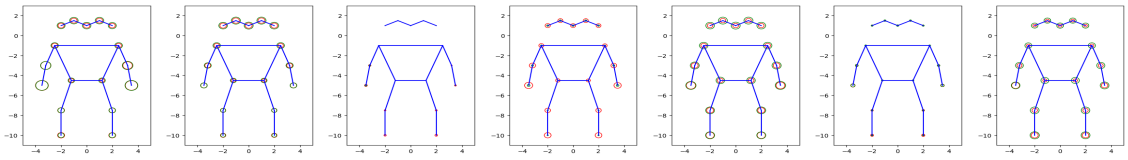
In the context of analyzing joint movements in human activity recognition, the first method employed is referred to as PD. This approach involves measuring the Euclidean distance between joints within consecutive frames of the captured video data. For instance, the distance between the left hand in frames one and two, followed by the distance between frames two and three, and so forth is shown in Figure 3.14. Each pairwise distance is represented by variables $d1$, $d2$, up to dn. By considering the positional variations between corresponding joints across frames, a comprehensive understanding of the movement patterns can be derived. The visual representation facilitates the identification of significant changes in joint positions and enables the comparison of joint movements across different actions and camera views. Ultimately, the PD method provides valuable insights into the temporal dynamics and spatial relationships of joint movements.

In summary, while PD diagram primarily showcases joints' position spatially, the fact that the average value (the red circles' diameter in Figure 3.15) is derived from sequences of movements over time infers a temporal aspect indirectly incorporated into the representation.

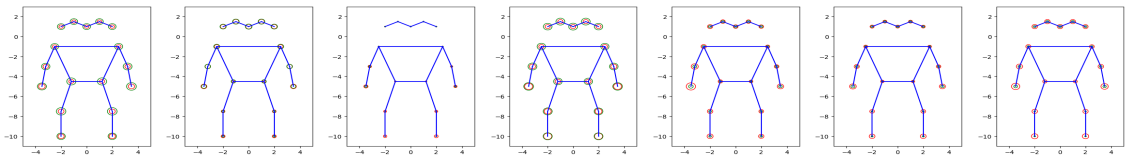
In Figure 3.15, the representations of pairwise distances across various activities (red circles), as observed from robot camera views, vividly showcase the diversity in joint movements among different actions. Normalization is performed using the maximum and minimum values observed across all joints within all actions in one view. This normalization process facilitates a robust comparison among joints across all actions, enabling a more comprehensive assessment of their variations.

This visualization distinctly highlights that certain actions exhibit a notably wider range of movement compared to others. Furthermore, Figure 3.16, 3.17, 3.18 presents the PD diagrams from three additional camera perspectives, revealing variation in the patterns observed in the robot view across all other camera angles. For example, the bending action in the Robot-view illustrates a different pattern compared to similar views (Front-view and Back-view). The head, leg and arm movements magnitude are contrasting in them. Observing the other action classes proves the fact that the view-invariant gives different features.

This visualization distinctly highlights that certain actions exhibit a notably wider range of movement compared to others. Furthermore, Figures 3.16, 3.17, and 3.18 present the PD diagrams from three additional camera perspectives, revealing variation in the patterns observed in the robot view across all other camera angles. For example, the bending action in the Robot-view illustrates a different pattern compared to the Front-view and Back-view, where bending appears similar. The magnitude of head, leg, and hand movements contrasts significantly among these views. Observing the other action classes further supports the fact that different perspectives provide unique features, underscoring the importance of view invariance in human activity recognition.



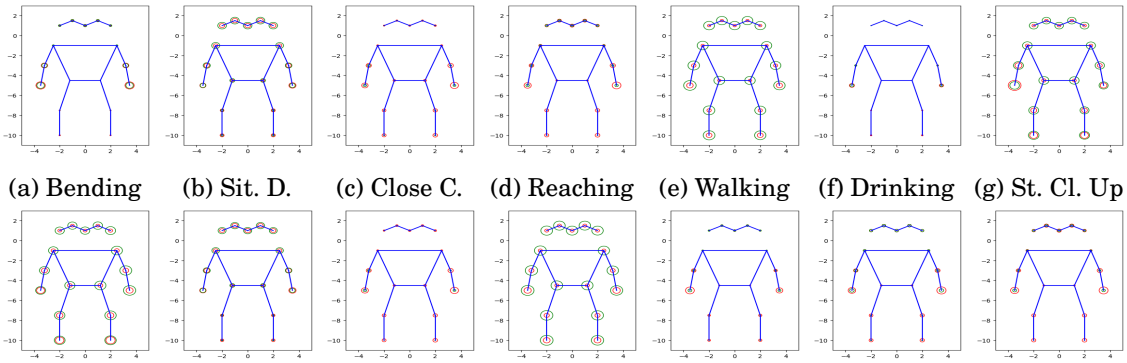
(a) Bending (b) Sit. D. (c) Close C. (d) Reaching (e) Walking (f) Drinking (g) St. Cl. Up



(h) St. Cl. D. (i) Stand. Up (j) Open. Can (k) Carr. Obj. (l) cleaning (m) P. D. Obj. (n) Lift. Obj.

Figure 3.15: Normalized pair distance (red circle) and Min-Max (green circle) presentation of all samples in categorized by action in Robot-view

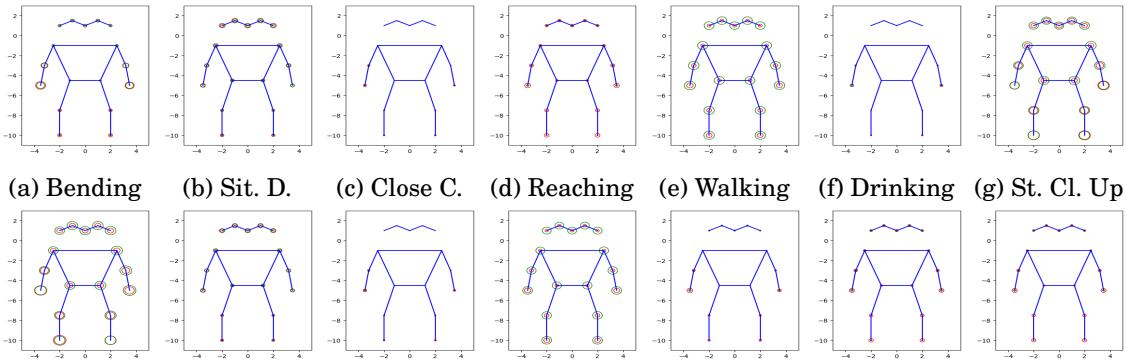
Joints with larger average pair distances indicate more significant movements and can be considered key contributors to the action's execution. On the other hand, joints with smaller average pair distances suggest relatively limited movement and can be considered less influential in the overall action performance. This differentiation between major and minor joint movements is crucial for understanding the dynamics of various activities, especially when analyzing data from different synchronized views. For instance, certain key joints may exhibit more pronounced



(a) Bending (b) Sit. D. (c) Close C. (d) Reaching (e) Walking (f) Drinking (g) St. Cl. Up

(h) St. Cl. D. (i) Stand. Up (j) Open. Can (k) Carr. Obj. (l) cleaning (m) P. D. Obj. (n) Lift. Obj.

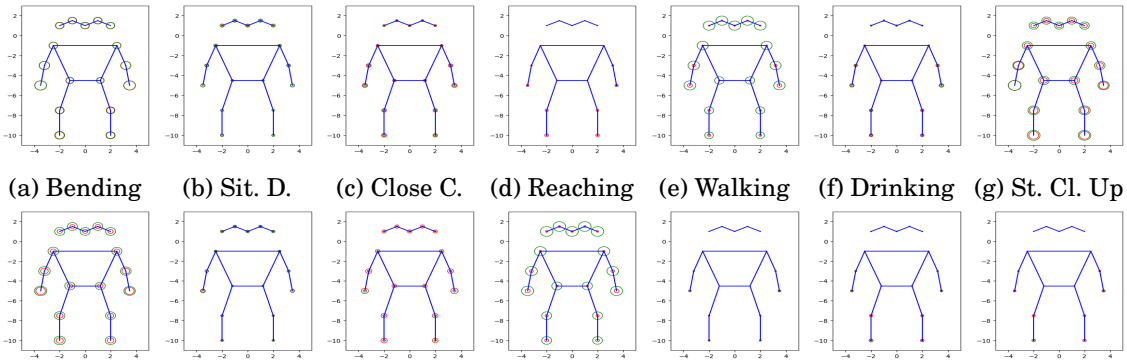
Figure 3.16: Normalized pair distance (red circle) and Min-Max (green circle) presentation of all samples categorized by action in Back-view



(a) Bending (b) Sit. D. (c) Close C. (d) Reaching (e) Walking (f) Drinking (g) St. Cl. Up

(h) St. Cl. D. (i) Stand. Up (j) Open. Can (k) Carr. Obj. (l) cleaning (m) P. D. Obj. (n) Lift. Obj.

Figure 3.17: Normalized pair distance (red circle) and Min-Max (green circle) presentation of all samples categorized by action in Front-view



(a) Bending (b) Sit. D. (c) Close C. (d) Reaching (e) Walking (f) Drinking (g) St. Cl. Up

(h) St. Cl. D. (i) Stand. Up (j) Open. Can (k) Carr. Obj. (l) cleaning (m) P. D. Obj. (n) Lift. Obj.

Figure 3.18: Normalized pair distance (Red circle) and Min-Max (Green Circle) presentation of all samples categorized by action in Omni-view

movements in one view while appearing less active in another, highlighting the importance of multiple perspectives for comprehensive human activity recognition.

3.3.1.2 Min-Max Distance

The second method employed for spatial analysis is the Min-Max. complements the PD approach by focusing on the minimum and maximum positions of each joint within each action separately. While the pair distance method provides insights into the overall pairwise distance travelled by a joint throughout an action, the Min-Max distance method offers a distinct perspective by emphasizing the extremities of joint positions. In Figure 3.14 m the green arrow shows this distance's overall movements.

In the Min-Max distance method, the minimum and maximum positions of each joint are identified within each sample of every action. By capturing the range of motion exhibited by each joint, this method provides a more localized understanding of the joint's movement patterns during specific instances within an action. For example, in Figure 3.14 the hand movement started from $d1$ move and ended with dn move, and the m shows the distance between maximum and minimum points that hand join passed.

Figure 3.15 displays the Min-Max diagram (the green circles) representing all actions observed from the robot perspective. The showcased values have undergone normalization by dividing all samples across all actions by the difference between the maximum and minimum values. Additionally, three supplementary views demonstrating similar patterns to the robot view are shown in Figure 3.16, 3.17, and 3.18 by green circles. These visualizations underscore the similarities and differences in joint movement patterns across multiple perspectives, highlighting the importance of multi-view analysis in enhancing the understanding and recognition of complex human activities.

3.3.1.3 Discussion

During the analysis, the results obtained from the examination of joint movement using the PD and Min-Max methods have revealed interesting insights worth discussing. The observed similarity in views and actions suggests a notable correlation. When actions exhibit similarity within a particular view, it signals potential vulnerability in the classification model's ability to differentiate between these actions. However, if these actions do not demonstrate a consistent pattern across different views, it implies an opportunity for the multi-view model to enhance the classification process.

For instance, action classes such as 'cleaning,' 'lifting objects,' 'putting down objects,' and 'lifting objects' showcase high similarity within samples captured by each respective view. Intriguingly, these actions display low similarity when observed from different camera perspectives. This indicates that while diverse patterns emerge for the same action across different views, a shared pattern exists within samples from the same camera view. This observation holds the potential for leveraging multi-view analysis to enrich the classification model, capitalizing on the distinctive patterns present within specific views to augment its discriminatory capability across various actions.

3.3.2 Temporal Analysis

In temporal analysis, a comprehensive examination of the information encapsulated within each frame's depiction of the human skeleton is conducted. To quantify the richness of information conveyed in these frames, this work employs the concept of MI, derived from information theory. This analytical approach allows us to assess and compare the amount of information carried by each frame concerning others within an action sample.

A distinctive aspect of the analysis lies in the multiple viewpoints captured simultaneously from the same subject. This concurrent acquisition of data from diverse camera angles enhances the effectiveness of the comparisons. By considering these multiple views, a more holistic perspective can be gained on the information content present in each frame in an action. This comparison across various viewpoints contributes to a more nuanced understanding of the information shared. It enables us to discern any variations or similarities present in the depiction of human skeletal movements.

3.3.2.1 Theory of entropy and mutual information

Information Theory is a branch of applied mathematics and electrical engineering involving the quantification of information. Originally introduced by Claude Shannon in 1948, it was primarily designed to find fundamental limits on signal processing operations such as compressing data and reliably storing and communicating data [136].

Key concepts in information theory include:

Entropy: This is a measure of the uncertainty, randomness, or disorder of a set of data, and it is defined as the average amount of information needed to describe the variable. In the context of information theory, entropy is used to quantify the amount of information in a source, often in units such as bits. The higher the entropy, the more unpredictable (and thus informative) the data is. Mathematically, for a discrete random variable X with possible values $\{x_1, x_2, ..., x_n\}$ and probability mass function $P(X)$, the entropy $H(X)$ is defined as:

$$H(X) = -\sum_i P(x_i) \cdot \log_2 P(x_i) \text{ for all } i \tag{3.2}$$

Mutual Information: In information theory, mutual information is a measure of the amount of information that is shared between two random variables. It quantifies the degree of dependence or correlation between the variables. If the variables are independent, their mutual information is 0.

In simple terms, if you have two variables, mutual information measures how much knowing the value of one of these variables reduces uncertainty about the value of the other variable. If the mutual information is high, then knowing the value of one variable tells you a lot about the other. If it is low, then the variables are independent.

Formally, let us consider two random variables X and Y. Mutual information between X and Y, denoted as $I(X;Y)$, measures the reduction in uncertainty of one variable when the other variable is known. It is defined as the average amount of information that X and Y share:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3.3}$$

where $H(X)$ and $H(Y)$ represent the entropies of X and Y, respectively, and $H(X|Y)$ and $H(Y|X)$ are the conditional entropies given the other variable.

The mutual information between X and Y is zero if and only if X and Y are independent. In this case, knowing the value of one variable provides no information about the other. As mutual information increases, the variables become more dependent or correlated.

Mutual information has several important properties:

- **Non-negativity**: Mutual information is always non-negative, meaning it is zero or positive.

- **Symmetry**: $I(X;Y) = I(Y;X)$, which implies that the order of the variables does not affect the mutual information.
- **Maximum value**: The mutual information between two variables is maximized when they are perfectly correlated or dependent.

Mutual information is widely used in various applications, including feature selection [137, 138], clustering [139, 140], and image registration [141, 142]. It provides a powerful tool for measuring and quantifying the relationships between random variables, helping to analyze and understand complex systems.

3.3.2.2 Human skeleton Entropy

To utilize the concept of entropy to show human skeleton information in a video stream, the entropy of the skeleton data in each frame needs to be computed. Entropy can measure the uncertainty or randomness in the distribution of joint positions or movements within the skeleton. Here is a step-by-step approach to applying entropy to the skeleton information:

Skeleton Representation: Each human skeleton in the video stream is typically represented as a collection of joints or keypoints. These joints include x and y values, which show the relevant pixel position in the 2D picture. This involves the pose extraction method by YOLOv7.

Data Preprocessing: The human skeleton is comprised of 17 distinct joints, with the spatial coordinates of each joint corresponding to the pixel positions within a two-dimensional image. The video stream consists of frames, each with dimensions of 640 by 480 pixels. For instance, within an image frame, the left shoulder joint is characterized by $x = 305$ and $y = 202$. To represent the skeletal information effectively, a single-dimensional array or tensor of size 34 is employed, where the first 17 elements denote the x-coordinates of the joints, and the subsequent 17 elements represent the y-coordinates. This organization ensures the coherent combination of all skeleton data within a compact data structure.

Normalization: Optionally, the joint position distribution can be normalized to bring all values within a common scale. X and Y coordinates are normalized separately, default division for X is 480 and 640 for Y. At the end, all coordinates in an array of 34 are normalized between zero and one. In the example figures the normalization was not applied to show the coordinates values.

Joint Position Binning: We divided the range of joint positions (both x and y dimensions) into several bins. The number of bins depends on the desired level of granularity in the probability estimation. For example, you can choose to have 10 bins, each covering a specific range of joint positions.

Calculate Histogram: For each frame in the video sample or entire video, the joint position bin to which each joint belongs (based on its x and y coordinates) was identified. Count the number of joint positions falling into each bin. This count represents the frequency of occurrence of joint positions in that bin for the given frames. Figure 3.19 displays a histogram depicting joints in a video sample, comprising 34 bins. It visualizes the frequency distribution of different x and y coordinates throughout the video, emphasizing all movements that occurred in a specific action. This histogram can be used to compute probabilities in the next step.

Joint Position Distribution: For each video sample, the skeleton data in each frame was analyzed to obtain the distribution of joint positions. This has been done by using the statistical measure, calculating the histogram of joint positions across all frames. This could also be done by computing other statistical measures such as mean, variance, or covariance of joint positions.

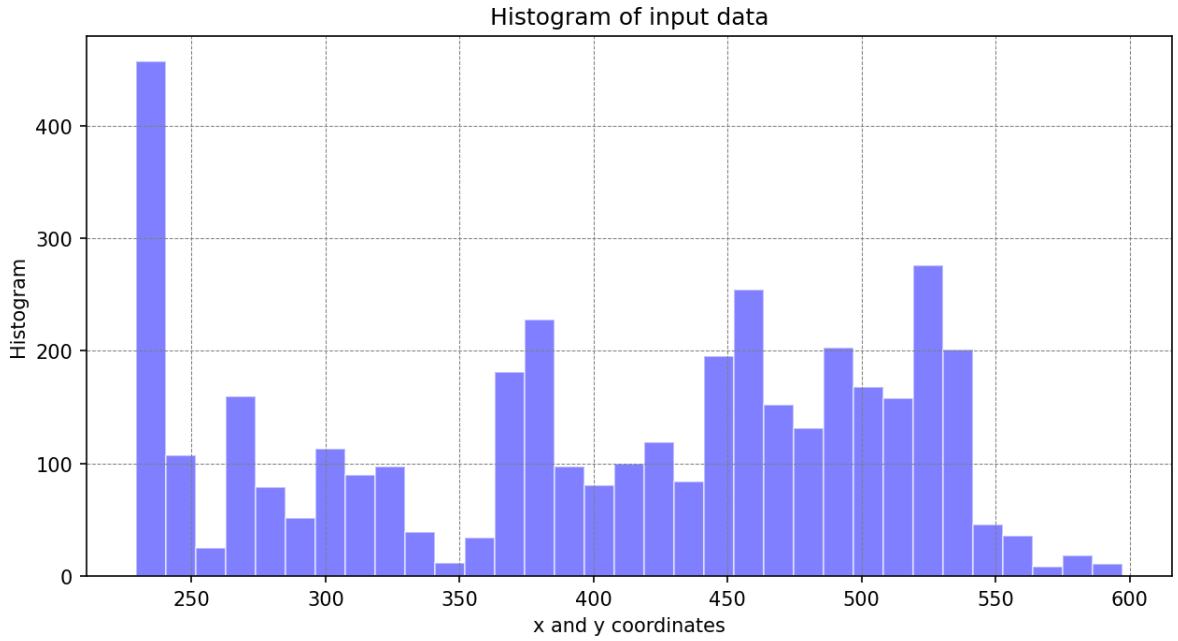


Figure 3.19: Histogram showcasing 34 bins representing joints within a video sample. The Y-axis represents the frequency of occurrences, while the X-axis displays the bins, which, in this context, represent coordinate values.

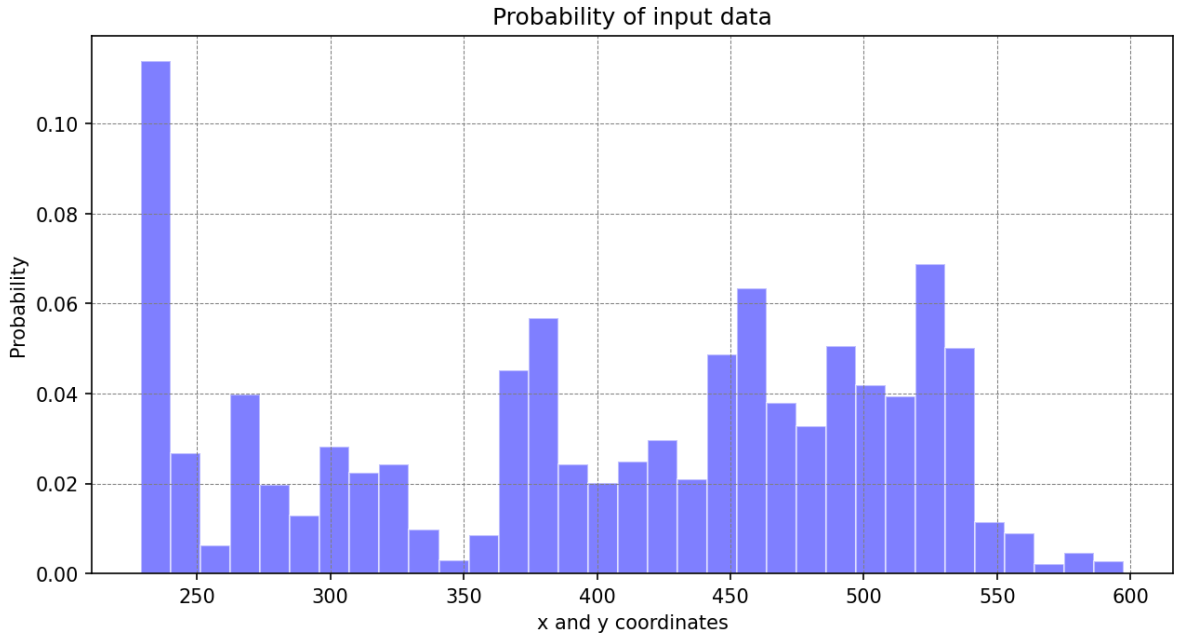


Figure 3.20: The probability distribution of joint coordinates within a video sample. It is derived by dividing the histogram (Figure 3.19) by the total count.

These measures provide insights into the spatial variability of the skeleton. The histogram of an input video sample provides a visual representation of the joints' position distribution across

all frames in the video. In the context of human activity recognition using skeleton data, the histogram illustrates the frequency of occurrence of different joint positions, both in the x and y dimensions, throughout the entire video sample.

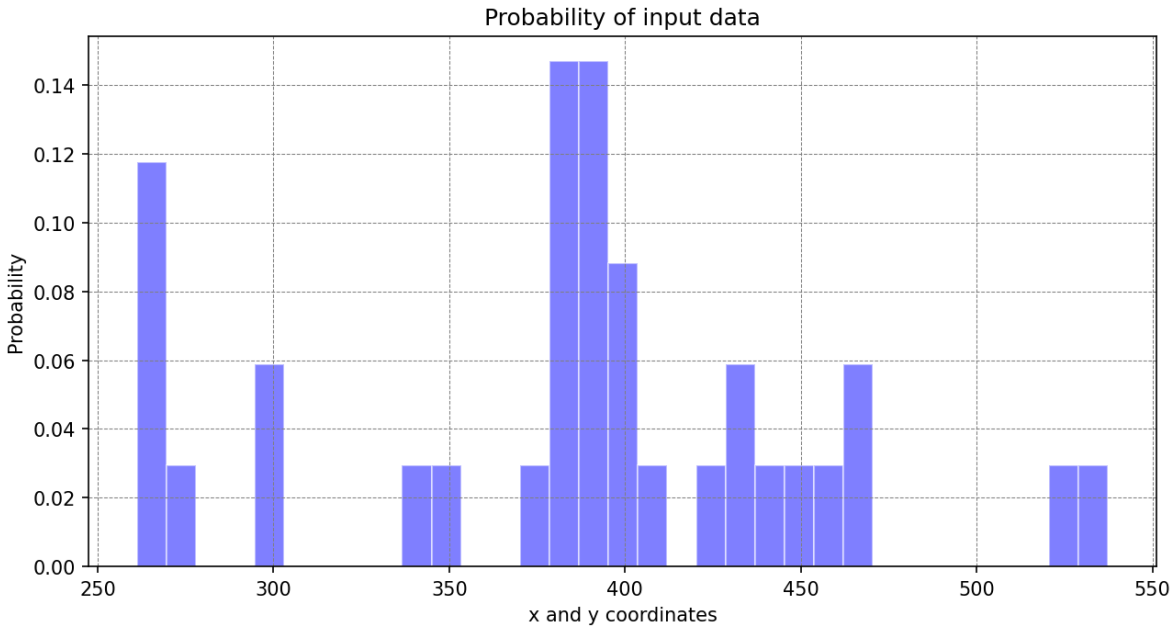


Figure 3.21: Probability distribution of frame one a human skeleton within video frames.

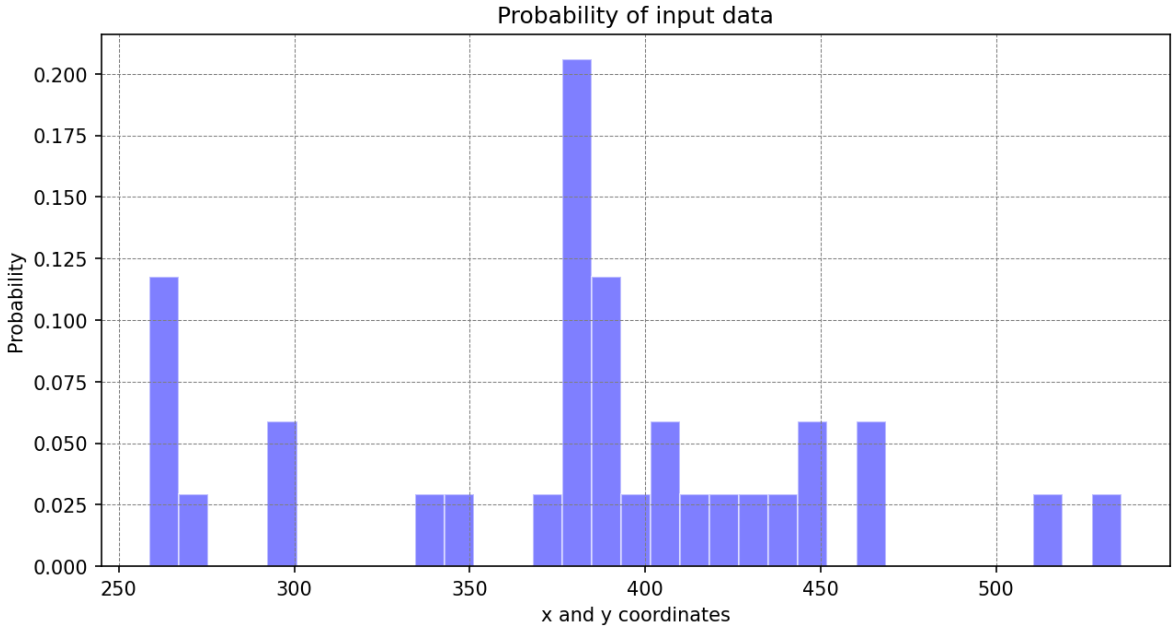


Figure 3.22: Probability distribution of frame two a human skeleton within video frames.

Probability Distribution: Based on the joint position distribution, construct a probability distribution function that represents the likelihood of joint positions occurring within the entire

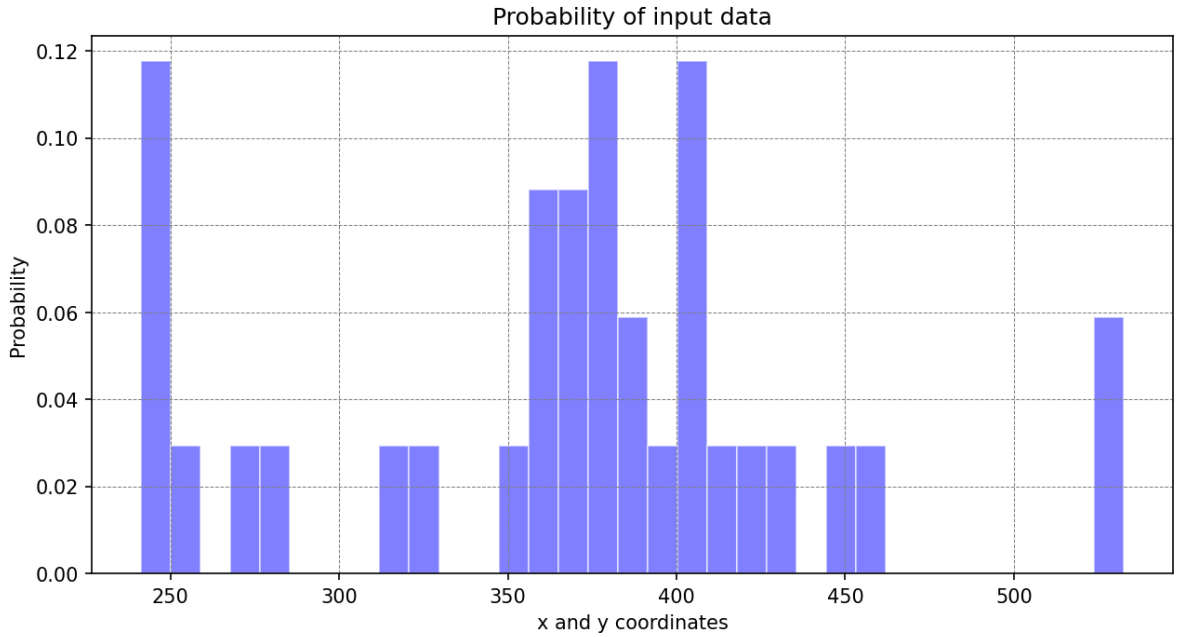


Figure 3.23: Probability distribution of frame 10 a human skeleton within video frames.

video sample or each frame. To estimate the probability of each joint position bin for a specific frame, divide the frequency of each bin by the total number of joints in that frame. This step normalizes the frequency counts to probabilities. Figure 3.21, 3.22, and 3.23 illustrates the joint position distribution of three skeletons, frame one, two and ten of a video sample that its probability distribution shown in Figure 3.20.

Entropy: Compute the entropy of the probability distribution. Entropy quantifies the amount of information or uncertainty in the distribution. We use the Shannon entropy formula:

$$H = -\sum P(x) \cdot \log_2(P(x)) \tag{3.4}$$

Where:

- H is the entropy value.
- $P(x)$ is the probability of a joint position bin in the distribution.

The entropy values for three frames are presented below:

- Frame one: 3.852928
- Frame Two: 3.803595
- Frame Ten: 3.984234

3.3.2.3 Joint Entropy

To calculate the joint entropy from the conditional probability, considering the frame-level joint distribution the following formula is used:

$$H(X,Y) = -\sum\sum(P(x,y)\cdot\log_2(P(x,y)) \tag{3.5}$$

Where:
$H(X,Y)$ is the joint entropy.
$P(x,y)$ is the conditional probability of joint position for a specific frame (Formula 3.7).
Negation:take the negation of the summed value to obtain the joint entropy.

Conditional Probability: To calculate the probability of a frame given the joint positions probability of the entire video, follow these steps:

- Calculate Frame Probability: First, the probability distribution of joint positions for each frame using the histogram approach mentioned earlier was calculated. Let's denote this frame probability as $P(frame)$. Each $P(frame)$ will represent the joint position distribution for a single frame.

- Calculate Video Probability: Next, calculate the joint position probability distribution for the entire video by considering all frames together. Let's denote this video probability as $P(video)$. $P(video)$ will represent the joint position distribution for the entire video sample.

- Calculate joint position probability: To calculate $P(frame \cap video)$, the element-wise product of $P(frame)$ and $P(video)$ has been taken, as both are probabilities, and their product gives the joint probability of their occurrence. This is also known as element-wise multiplication.

$$P(frame \cap video) = P(frame) \times P(video) \tag{3.6}$$

- Calculate Conditional Probability: To find the probability of a specific frame given the joint positions probability of the entire video, the conditional probability needs to be calculated, denoted as $P(frame|video)$. This can be calculated using the formula:

$$P(frame|video) = \frac{P(frame \cap video)}{P(video)} \tag{3.7}$$

where $P(frame \cap video)$ is the joint probability of the frame occurring together with the joint probability of the entire video, and $P(video)$ is the joint probability of the entire video.

3.3.2.4 Mutual Information (MI)

Mutual information (MI) is an information-theoretic measure that quantifies the amount of information shared between two random variables. In the context of this study on frame sampling for human activity recognition (HAR) using skeleton data, mutual information has been applied to analyze the joint positions probability distributions for each frame and the entire video sample. The goal is to explore the relationship between the joint positions in individual frames and the joint positions distribution of the entire video.

Mutual information is calculated using the following formula:

$$MI(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.8}$$

Where:

- $MI(X;Y)$ represents the mutual information between random variables X and Y.
- $H(X)$ is the entropy of random variable X.
- $H(Y)$ is the entropy of random variable Y.
- $H(X,Y)$ is the joint entropy of random variables X and Y.

3.3.2.5 Mutual Information Correlation Matrix (MICM)

MICM is a valuable representation of the relationships between pairs of frames in the video sample. It quantifies the mutual information values between different frames, providing insights into the redundancy or similarity of joint positions between frames.

MICM is populated with the calculated mutual information values between all pairs of frames in the video sample. MICM is a square matrix of size $N \times N$ (where N is the number of frames in the video), where MICM$[i,j]$ represents the mutual information between frame i and frame j.

Each entry (i,j) in the MICM represents the mutual information value between frame i and frame j. The value in this entry indicates the amount of shared information or similarity between the joint positions of these two frames. Higher mutual information values imply that the joint positions in the corresponding frames are more similar or redundant, while lower values suggest that the frames exhibit different joint position patterns.

Visualizing the MICM can provide an intuitive representation of the relationships between frames. The heatmap plots were used to visualize the mutual information values, where higher values are depicted using distinct colours, making it easier to identify similar frames. Further analysis, such as clustering frames based on mutual information values or exploring patterns in the MICM, can offer deeper insights into the structure and dynamics of the video sample, aiding in comprehensive activity recognition.

Frames with high mutual information values in the MICM are considered similar in terms of their joint positions. These frames are likely to capture similar human movements or activities. Such similarity can arise from repetitive actions or periods of inactivity during the video sample. Identifying and grouping similar frames can be valuable in streamlining the processing and analysis of video data, as it allows the system to focus on informative frames rather than redundant ones.

On the other hand, frames with low mutual information values in the MICM indicate dissimilarity in their joint positions. These frames represent instances of diverse human movement patterns or transitions between different activities. Uncovering such diverse frames can provide a comprehensive understanding of the range of activities and movements captured in the video sample.

In Figure 3.24, an insightful analysis is presented, shedding light on the MICM in the context of observing a single action across four distinct camera views. This metric serves as a measure of the variation in information content between frames within each view, revealing the distinctiveness of information conveyed by frames captured from different angles.

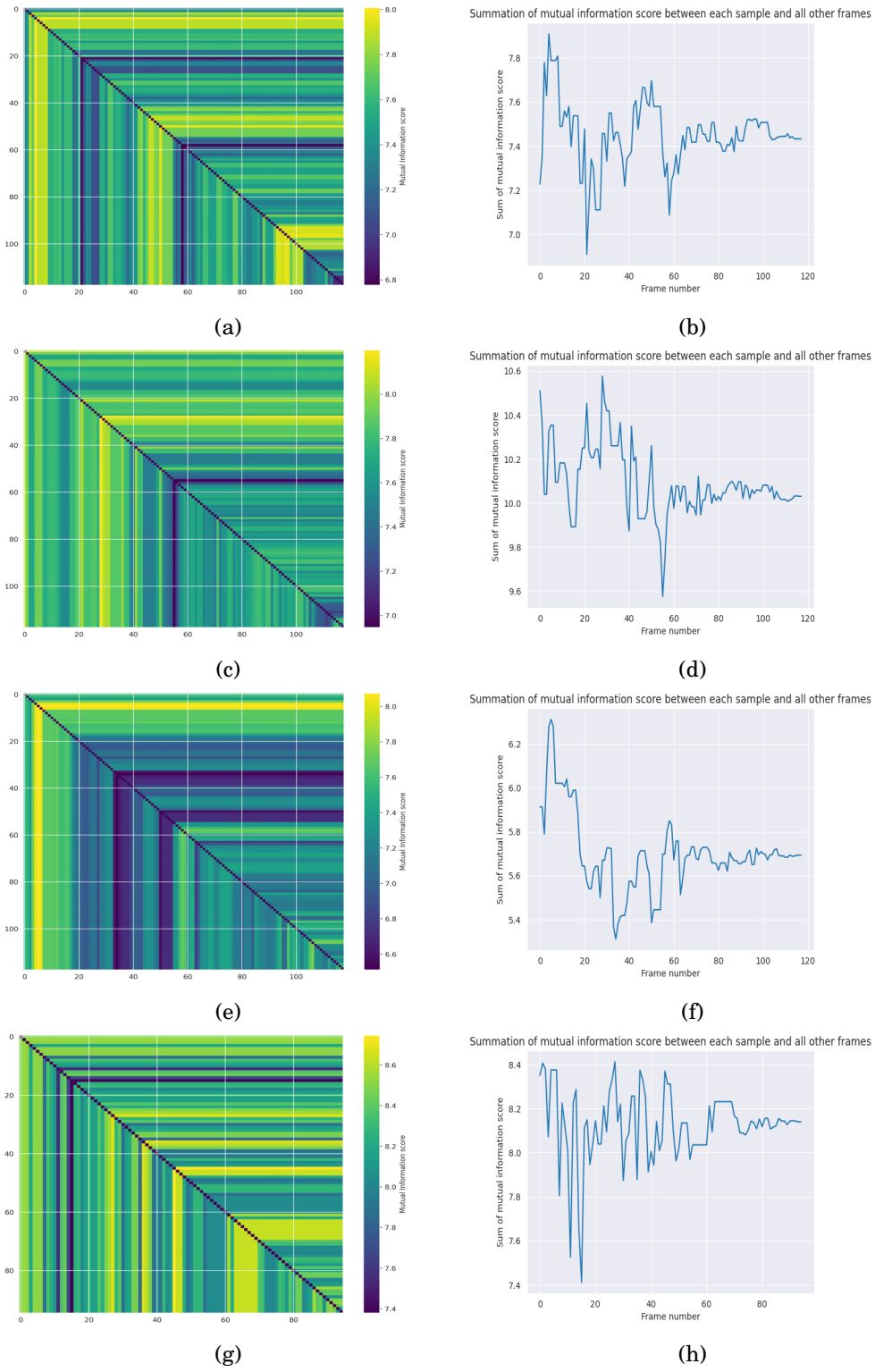


Figure 3.24: The illustration of the MICM (left side) and sum of mutual information (right side) of an action in four views

The left-hand side of the figure illustrates the MICM associated with performing a specific action within each view. This metric quantifies the variability in information content among frames within individual camera perspectives. The distinct visualizations for each view emphasize how frames captured from different angles exhibit diverse information content.

The right-hand side of the figure presents the cumulative MICM, visually represented through a graph. This cumulative MICM graph aggregates the individual information changes across the different camera views. It effectively summarizes the collective discrepancy in information content across all views.

The results in this analysis underscore that each frame in the same action captured from multiple perspectives can convey different information. This discrepancy is starkly evident when comparing frames captured from various viewpoints, emphasizing that frames within a video distinctly communicate varying information. Moreover, the cross-view analysis elucidates that frames captured from different angles offer unique perspectives, contributing diverse sets of information to the overall video content.

This insight is pivotal in understanding the significance of considering multiple perspectives within visual data analysis. The figure vividly illustrates that each frame carries distinct information within a video, and this discrepancy is further magnified across different camera views, emphasizing the need to account for and interpret the multi-faceted nature of visual information.

The MICM can be instrumental in developing an effective frame sampling strategy for human activity recognition. By analyzing the MICM, you can identify frames with high mutual information values, which represent informative and representative frames. These frames carry essential joint position patterns that contribute significantly to activity recognition and analysis. In contrast, frames with low mutual information values may not offer much additional information, making them less critical for the recognition process. This insight can guide the development of frame sampling techniques that prioritize informative frames while reducing the computational overhead associated with redundant frames.

3.4 Conclusion and Addressing the Research Questions through Analysis

The missed frames statistics reveal that different camera views have different qualities and show that an Omni-view is an unreliable source for body pose extraction from both HRNet and YOLOv7 pose extraction methods. Besides, Comparing the results of missed frames in both pose extraction methods shows that the YOLOv7 method has slightly lower missed frames. However, it has been noted that the Omni-view delivers good information in actions with long-range movements like walking and carrying objects. In Figure 3.7 the comparison of these two methods shows the similarity clearly.

These statistics also reveal that the number of missed frames is correlated with the action type. Actions like *stair climbing up and down*, *bending*, *sitting down*, and *cleaning* that need more vertical and horizontal courses have more missed value in two fixed wall mount cameras compared to the Robot-view. This is because the robot head follows the human, whereas the wall-mounted cameras do not. At the same time, the robot view has moderately more missed frames due to being too close to the human or being within a cluttered environment. These manifest mainly in actions *carrying objects*, *walking*. Considering the results of both missed frames and missed poses in Robot-view, it is deduced that being close to humans when they are moving around quickly or for long distances can decrease recognition quality due to partially

observable or not observable joints. The reason is that this view has a closer view of the human and the scene, causing missing the lower body joints.

The previous discourse might led to the proposition that utilizing a wide-angle camera in a robot could potentially facilitate the research; nevertheless, comparable cameras were employed to circumvent any further technical examination, which may be explored in a subsequent investigation.

The missed poses' statistics in stairs climbing actions prove that the robot's camera movement and ability to follow the human results in fewer errors. The human has vertical movement in this action, which can be followed by a robot camera that other fixed cameras might miss. For example, the Front-view, which has the fewest missed poses on all actions on average in the HRNet method, has the higher number of missed poses in stairs climbing up and down actions(Figure 3.12).

Comparing two wall-mounted cameras with the same technical feature emphasizes the effectiveness of the viewpoint. The missed pose statistics index in Figure 3.8 shows that the Front-view has better results regarding HRNet pose extraction quality. On the other hand, the Back-view, which is also a wall-mount camera with the same technical features, results in the most missed poses in almost all actions with HRNet. The only difference between these two wall-mount cameras is the altitude and view side. Reviewing the videos from these camera views in different activities suggests that the higher attitude and broader view in wall-mounted cameras can decrease the missed poses.

However, the results obtained for YOLOv7, as depicted in Figure 3.9, demonstrating the average missed poses, do not exhibit a similar pattern compared to HRNet. Specifically, concerning lower body joints, the Back-view demonstrates the highest quality, followed by the Front-view and the Robot-view, which exhibit comparatively lower quality. Among all joints analyzed in both methods, the left and right shoulders consistently exhibit the highest quality, although the YOLOv7 method shows superior quality in these joints overall. These findings indicate that YOLOv7 exhibits higher quality in terms of missed poses, suggesting a potential reflection on classification accuracy.

Overall, the results of the qualitative analysis show that the camera perspective, pose extraction method, activity type, and joints are highly significant in the quality of pose extraction. Theoretically, combining a Robot-view camera and two other cameras can enhance skeleton extraction. The integration of an extra camera may incur substantial expenses both in terms of computational resources and monetary cost, yet this concern has been subject to further discussion in Chapter 5 which addresses the efficient MV-HAR model.

The joint movement analysis reveals variations in human joint movements observed across different camera views, highlighting both similarities and differences in movement patterns within action classes. While the PD and Min-Max methods uncover similarities in actions within specific camera views, potentially complicating the classification model's discrimination, discrepancies across views offer an opportunity for a multi-view model to enhance classification by capitalizing on distinct patterns within each view.

The temporal analysis using Mutual Information emphasizes the diverse information conveyed by frames from different camera perspectives. It distinctly highlights the unique information captured within frames from varying viewpoints, emphasizing the significance of considering multiple perspectives in visual data analysis. This insight underscores that each frame in a video offers distinct information, accentuating the importance of interpreting the multi-faceted nature of visual data, especially across different camera views.

RQ1: How effective is the use of skeleton-based human activity recognition? This chapter provides valuable insights into the efficacy of skeleton extraction methods across diverse camera

viewpoints. It highlights the notable variability in quality and reliability observed among different camera perspectives. Additionally, the examination of human biomechanics demonstrates the skeleton's capacity to capture nuanced spatial and temporal information.

RQ2: What is the impact of different camera perspectives on the quality and richness of data captured for human activity recognition in multi-view scenarios? The significance of perspective in multi-view Human Activity Recognition (HAR) becomes evident from the analysis of body pose extraction quality across various camera types and perspectives. This examination underscores the critical role played by perspective, particularly evident in the comparison among different wall-mounted cameras sharing similar technical features. It suggests that altitude variations and a wider field of views could significantly influence the quality of pose extraction. This observation is supported by discussions on missed poses across diverse views, indicating distinct performance levels regarding pose extraction quality.

Moreover, the analysis of human biomechanics through temporal and spatial examinations provides crucial insights into the influence of different perspectives. While the temporal analysis reveals clear differences in perspectives, showcasing unique variations in the observed activities, the spatial analysis unearths high-level similarities in activities across different perspectives. Additionally, the study of joint movement patterns across various camera views underscores both similarities and differences within action classes. Despite uncovering similarities in actions within specific camera views via the PD and Min-Max methods, which might complicate discrimination in classification models, the discrepancies across views present an opportunity for a multi-view model. Such a model could capitalize on these distinct patterns within each view to enhance classification accuracy.

Furthermore, the Mutual Information-based temporal analysis emphasizes the diverse and unique information encapsulated within frames from different camera perspectives. This underscores the paramount importance of considering multiple perspectives in visual data analysis, highlighting that each frame in a video offers exclusive information. Understanding this multifaceted nature of visual data becomes crucial, especially when interpreting content across various camera views for comprehensive insights into human activity recognition.

RQ3: What are the optimal models for multi-view human activity recognition, and how do different methods for data combination influence their performance? The different data combination approaches in multi-view Human Activity Recognition (HAR) highlight the potential benefits of incorporating additional views to enhance overall performance. The spatial and temporal analyses underscore that certain actions exhibit similarities within individual views, suggesting that combining data from multiple views could significantly enhance performance. Additionally, the analysis using MICM analysis vividly 3.24 illustrates substantial differences at the frame level across different views. These findings emphasize the intricate variations in information conveyed by frames from diverse perspectives.

While the analysis briefly touches upon the advantages of combining views, it also acknowledges associated concerns regarding increased computational demands and costs. However, the discussion lacks an in-depth exploration into the specific advantages and disadvantages of merging data from different viewpoints in detail, leaving room for a more comprehensive analysis of the potential trade-offs and benefits associated with different data combination strategies. A more comprehensive examination in Chapter 5 could explore deeper into how combining data from multiple views impacts the robustness, accuracy, and computational complexity of the HAR model.

CHAPTER 4

LIGHTWEIGHT HAR

In this chapter, a classic LeNet [97] classification model is modified, termed as M-LeNet for the HAR task. In contrast, the Vision Transformer [143] (ViT) for the classification task is used, and the results are compared between both models in single-view and multi-view setups. LeNet is a widely used and relatively simplistic CNN in image classification tasks. In contrast, ViT is a more advanced transformer-based model that has gained popularity in recent years owing to its ability to process images without relying on traditional convolutional layers. Both classifiers are chosen due to their comparable number of training parameters, and FLOPs, thus allowing for a fair comparison between the two distinctive models. Besides, several parts of the HAR pipeline like input spatial-temporal data transformation, data sampling, and representation and classification methods have been modified.

4.1 Lightweight MV-HAR pipeline

The process of recognizing human activity via a skeleton-based multi-view approach typically encompasses the acquisition or loading of video data, the extraction of joint information, and the generation of skeleton data. Subsequently, a machine learning algorithm is employed to classify the recorded actions. The utilization of a lightweight pipeline in this context allows for the integration of cameras in robotic and AAL environment, enabling their effective operation in a variety of scenarios. As depicted in Figure 4.1, the proposed methodology for MV-HAR emphasizes the central concept of leveraging multiple camera viewpoints to enhance the recognition of activities via a lightweight pipeline. Since the data has been combined at the feature level, the combination is called low-level fusion. The pipeline started with data collection from different views, followed by pose extraction and preprocessing. Then, the prepared tensor file feeds the training model.

4.1.1 Input Data

In Chapter 3 the RHM-HAR-SK [144] dataset has been introduced which has been created by this research on the top of a RGB dataset (RHM) [1]. A robot-view camera, two wall-mounted

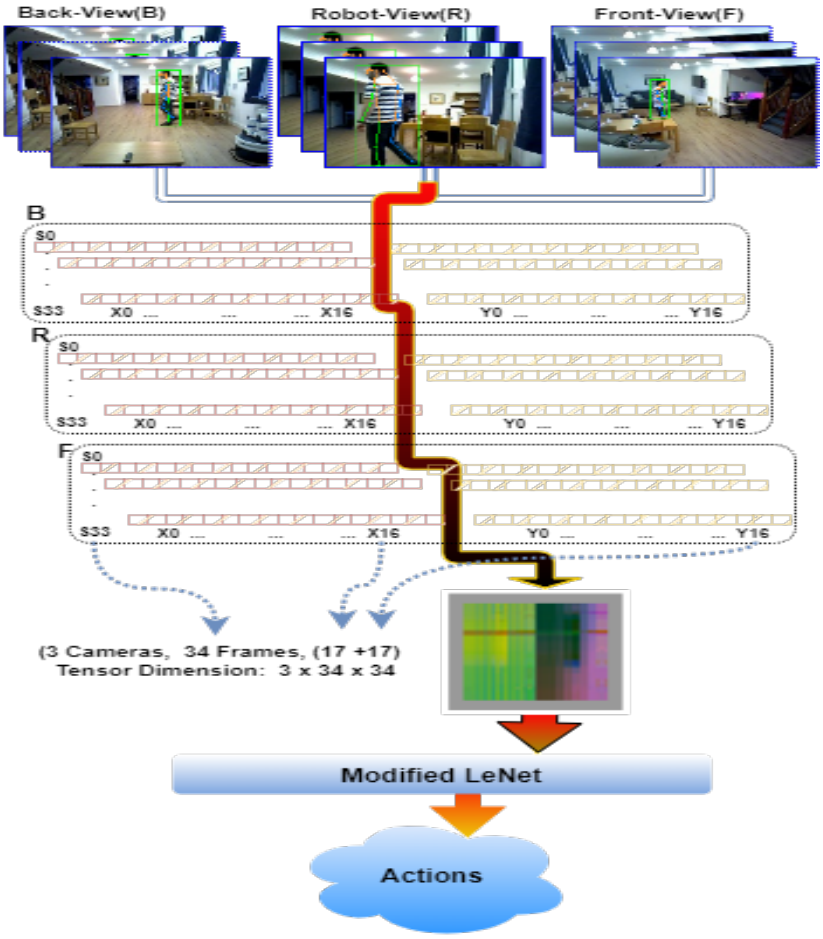


Figure 4.1: The MV-HAR pipeline,
as described in detail in Sec. 4.1 begins with capturing video from multiple views (Sec.4.1.1) and the extraction of skeletons from multiple viewpoints (Sec.4.1.2, followed by the conversion of each viewpoint into a spatial-temporal matrix (Sec.4.1.3). These matrices are subsequently combined into a single tensor file, which is finally classified by the modified LeNet model (Sec.4.1.4).

cameras (Front-view and Back-view), and an omnidirectional view (Omni-view) camera capture the activities synchronously. However, the analysis of that dataset in Sec. 3.2 reveals that the Omni-view data has low accuracy in the skeleton-based method. Consequently, it has been removed from the experiment in this Chapter and Chapter 5, and three other camera perspectives are being used.

4.1.2 Pose extraction

The utilization of RGB cameras is due to their simplicity, affordability, and accessibility. It is expected that by using high-performance pose extraction methods like HRNet applied to RGB data, better results can be achieved in human body skeleton extraction. In the RHM-HAR-SK dataset, a pre-trained HRNet model as described in [64] is utilized to extract poses from videos. This model has been trained on the COCO keypoint detection dataset [134] and the MPII Human Pose dataset [135]. In Section 2.3.3.1, the architecture of HRNet pose estimation is presented.

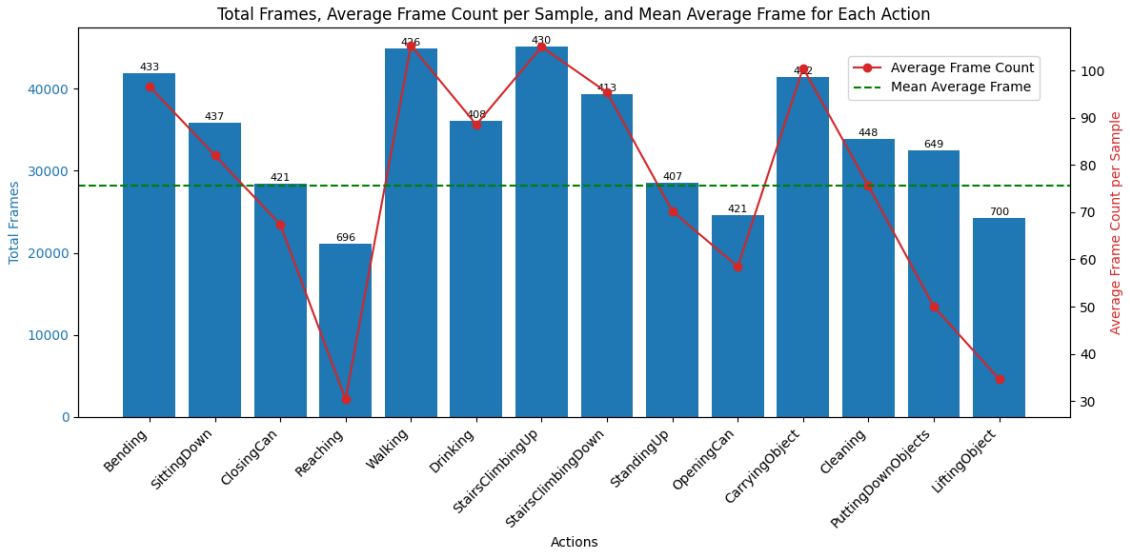


Figure 4.2: Analysis of the frame count in the RHM-HAR-SK dataset
Left vertical axes: count of all frames in action class. Right axes: Referring to the red point in the graph (total frames in a class divided by the number of samples in that action class). The value on top of each bar shows the number of samples in each action

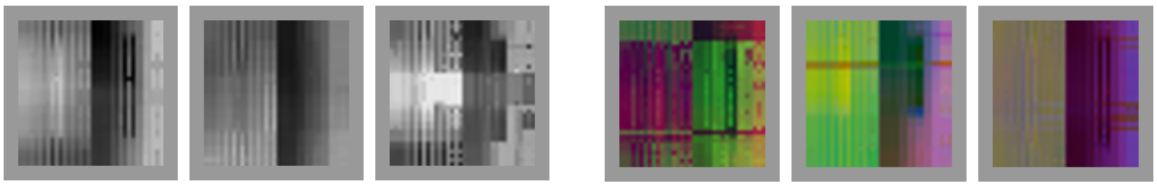
The HRNet model accepts an RGB image as input, with the image size adaptable based on the chosen model configuration and available computational resources. This work utilizes the Top-down HRNet model, generating outputs that include bounding boxes for detected humans, predicted locations of keypoints for each identified individual in the image, and associated confidence values for each keypoint. The number of key points is task and dataset-specific; for this work, the 17-keypoints model for the human body pose and the skeleton is employed.

In Chapter 5, YOLOv7 is used for pose extraction, with qualitative and quantitative analyses of both methods provided in Chapter 3 . The reason for using HRNet in this chapter is that the work initially started with HRNet, and YOLOv7 was subsequently integrated into the research.

4.1.3 Preprocessing

Following the extraction of the skeleton data from each frame in a video sample, the spatial and temporal input data was transformed into a $3 \times 34 \times 34$ tensor. Each row in this two-dimensional tensor refers to spatial data or a human skeleton and the sequence draws a temporal feature which is the full-size tensor. In Figure 4.1 the process of finding a single person from three cameras to make the tensor file is shown. The first digit (3) refers to three cameras, the 34×34 dimension refers to 34 frames of skeleton data, and two 17 columns related to 2-dimensional coordinates referring to 17 extracted poses human body. The first 17 columns belong to the X value and the second half to the Y value. The method involves selecting 34 frames randomly from videos of varying lengths while maintaining their original sequential order to preserve the temporal information.

Referring to human behaviour involving actions of varying durations and repetitions, certain actions occur more rapidly than others, while some repeat the same movement within a given period, as observed in Figure 4.2, activities that include stepping such as walking, stair climbing,



(a) Three samples of Single view of bending action.

(b) Three samples of combined three views as a RGB image of bending action

Figure 4.3: Synchronized skeleton output from different views of bending action.

and carrying object have more than 90 frame count in average while activities which include pick and place like reaching and lifting objects have about 30 frame count in average.

In Figure 4.2, the statistics of total frames in each action class (left axes), average frame count per sample (red points), and mean average frame for each action (green line) are illustrated. The lowest average frame count per action is approximately 30 frames for actions like reaching and lifting objects. This suggests that selecting frame samples fewer than 30 would result in data loss for these actions, and choosing frames exceeding 30 is also not feasible. The highest average frame count is less than 110, and the mean average frame count is about 76.

Selecting the appropriate frame count to capture action movements in all samples is crucial. The mutual information analysis (MICM) in Sec. 3.3.2.5 indicates that consecutive frames could convey similar information, and random sequential sampling is particularly effective in this context, as it simplifies the input without requiring extensive processing.

Furthermore, to maintain the input tensor in a square shape, a frame count of 34 was randomly selected. Subsequently, for samples below 30 frames, the count was increased by replacing previous frames until reaching 34 frames. Consequently, each input video stream is transformed into a $3 \times 34 \times 34$ tensor.

The three-channel matrix is illustrated as an RGB image in Figure 4.1, with each camera view being mapped to the red, green, and blue channels. Figure 4.3 illustrates three samples of two types of input data, the grayscale in 4.3a and RGB in 4.3b. The RGB refers to three channels, each indicating a camera view and the grayscale a single-view camera. Each 2D image depicts skeleton data frames in an action. After the extraction of skeletons from the video stream and preparing the 2D image, two general machine-learning models were applied as outlined below.

4.1.4 The Modified LeNet model

In this work, the spatial and temporal data is converted from a single skeleton into a 2D tensor(Sec. 4.1.3), enabling the use of a 2D CNN model for classification. In Table 4.1 the effect of choosing two different sizes of input image on the CNN model's complexity is shown. The lower the input size the lower the model complexity. It shows that by far some models such as LeNet and M-LeNet have lower complexity.

When considering the recognition model, whether it's image-based or skeleton-based, the preparation of the input data greatly impacts the model's performance. This includes factors such as using full-size images, smaller image sizes, or abstract representations of input data. In this work, the skeleton data is utilized, with the input video transformed into a 34 by 34 image. Comparing this with a single frame of image input, there is a significant reduction in input data.

	Image size: 640 × 480		**Image size: 34 × 34**	
	#Params(M)	FLOPs(G)	#Param(M)	FLOPs(G)
LeNet	35.2824	0.3478	0.0617	0.0004
M-LeNet	186.5713	0.404140	0.6211	0.0010
SqueezNet	1.2354	2.2753	0.7285	0.0055
MnasNet	2.2185	0.7080	2.2185	0.0070
MobileNet	3.5048	1.9984	2.2412	0.0150
DenseNet	7.9788	17.7315	6.9619	0.0639
ResNet18	11.6895	11.1649	11.1774	0.0787

Table 4.1: A comparison of computational characteristics.
($Params$ in million and $FLOPs$ in Giga) for different CNN models in two image sizes: 640 × 480 and 34 × 34

The base model that has been used in this experiment is LeNet [97]. This is a relatively simple CNN model with a limited number of training parameters [98]. Originally developed for handwritten digit recognition, it proves suitable for tasks with constrained computational resources [99]. Due to its status as a state-of-the-art (SOA) model, the LeNet is modified for use as the baseline classification model. The goal is to provide a basis for comparison by evaluating proposed modified LeNet model derivations against SOA models such as this.

Examining the images in Figure 4.3 reveals their dissimilarity to typical human life pictures, featuring objects and environments. Instead, these images illustrate movement patterns in a way that is not easily understood by humans. This work presumes that there may not be a significant necessity to expand the CNN model and maintaining a few layers for feature extraction could be sufficient. Then, the LeNet adaptation retained the fundamental structure of the model with notable alterations, including a reduction in the number of convolutional layers from 3 to 2 and a decrease in kernel size from 5 to 3. The kernel size was reduced so the model could capture more details in CNN layers.

To address the risk of overfitting, two dropout layers were introduced. Furthermore, to boost the learning capabilities of the model, an additional fully connected layer was added to two layers in the original LeNet, thereby increasing the number of trainable parameters.

Figure 4.4 illustrates the structure of the Modified CNN model. Two convolution layers are applied in this model, which is tested by two different configurations, 10 and 20 channels for the low parameter and 20 and 40 channels for the high parameter configuration.

4.1.5 Vision Transformers (ViT) Architecture

The paper "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale" proposes a novel approach to image recognition: Vision Transformer (ViT) [143]. Unlike dominant convolutional neural networks (CNNs) like LeNet and those used in ImageNet, ViT relies solely on Transformer architecture, commonly used for natural language processing.

In the ViT architecture, each input picture is divided into patches of sub-images. Then by applying the positional encoding, the model is trained. Each patch is considered as a word and projected to the feature space. Figure 4.5 illustrates the four random input data and their patches. In Figure 4.6 a random input data with its patches and the ViT classification architecture is shown. The process of preparing the input data for the ViT and M-LeNet is the same.

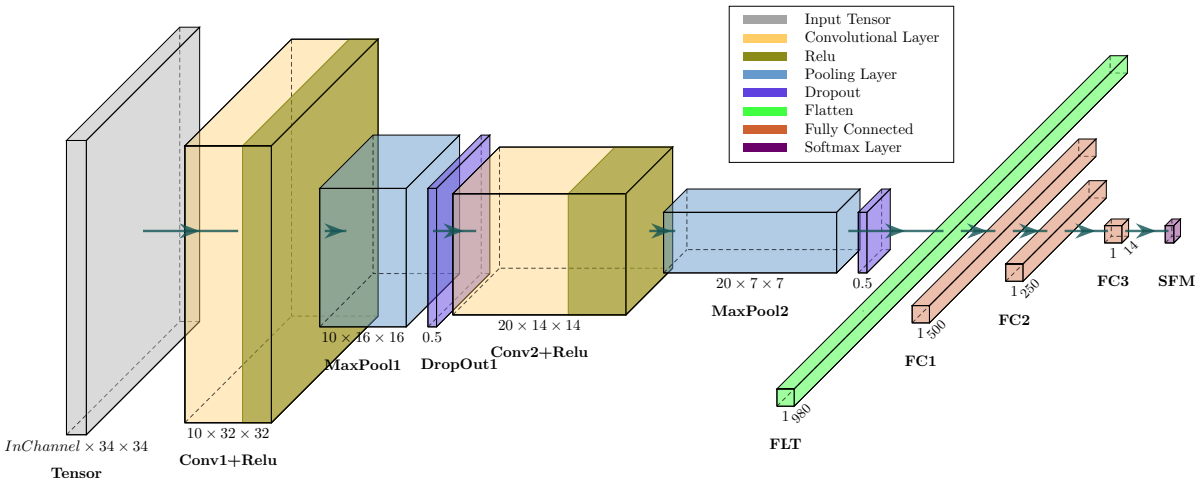


Figure 4.4: A Schematic Representation of a Modified LeNet (M-LeNet) Convolutional Neural Network (CNN) Architecture with Various Layers and their Corresponding Functions

Each patch is then flattened into a 1D vector. Subsequently, a linear projection is applied to these vectors, transforming them into a lower-dimensional embedding space. This step helps extract essential features from the patches.

Since spatial information is lost during splitting, positional encodings are added to each patch embedding. These encodings inject information about the relative position of each patch within the original image. There are various positional encoding techniques, but they generally involve sine and cosine functions based on patch indices. The final input sequence for the ViT model is formed by concatenating the processed patch embeddings with the positional encodings.

4.1.5.1 Transformers in ViT

The heart of ViT lies in the transformers architecture with the following structure.

Encoder: The transformer encoder is the core component. It processes the sequence of encoded patches through a series of self-attention layers followed by feed-forward networks.

Self-Attention: These layers allow each patch to attend to other patches in the sequence, capturing global dependencies and relationships between features across the entire image. Each attention layer performs three key operations:

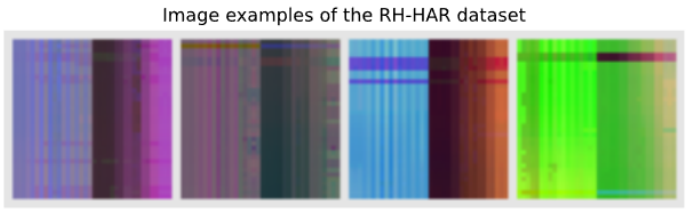
Query, Key, Value Projection: Each patch embedding is projected into a separate query (Q), key (K), and value (V) vectors.

Attention Scores: The model calculates attention scores between each query vector and all key vectors. These scores represent the relevance of each patch to the current query patch.

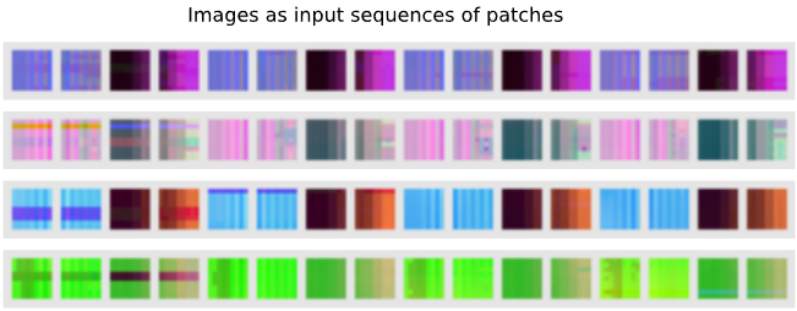
Weighted Sum: The attention scores are used to weight the corresponding value vectors, creating a context vector for each patch that incorporates information from relevant patches.

Feed-Forward Network (FFN): This network adds non-linearity to the model, allowing it to learn more complex relationships between features.

Decoder (Optional): In some ViT variants, a decoder might be used for tasks like image generation or segmentation. However, for standard image classification, the encoder output is sufficient.



(a) Four Input samples



(b) Four patches, each comprising 16 sub-images, extracted from the four samples in image 4.5a

Figure 4.5: The ViT sample input image and patches A- Four random samples of input B- The image patch of four random input

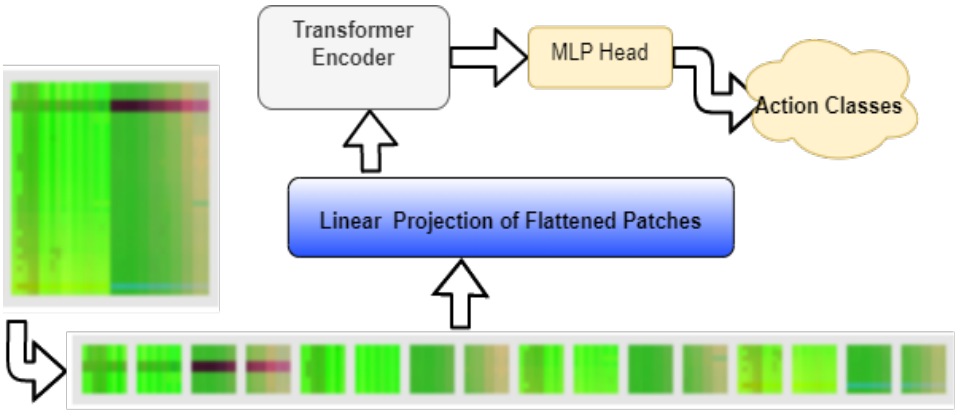


Figure 4.6: The ViT classification Architecture Applied on one of the RHM-HAR-SK dataset's sample [143]

4.1.5.2 N. Param. and FLOPs

There are many ViT variants with different architectures and sizes (e.g., ViT-Base, ViT-Large, and ViT-Huge). Each variant has its own FLOP count and parameter count. Similar to CNN models, the number of FLOPs and parameters in ViT can also vary depending on the specific dataset and training configuration. In Table 4.2, three primary ViT variants are presented, including the modified version introduced in this work for the HAR task, specifically designed for a lightweight structure, denoted as **ViT-HAR**.

Applying a similar logic employed in modifying the LeNet to create the M-LeNet (where the

Model	Layers	Hidden size (D)	Heads	Params (M)	FLOPs (G)
ViT-Base	12	768	12	43.2	0.37
ViT-Large	24	1024	16	155.1	1.29
ViT-Huge	32	1280	16	315.8	2.68
ViT-HAR	**4**	**512**	**8**	**6.82**	**0.06**

Table 4.2: Details of Vision Transformer model
, variants with batch size equal to 8 and 16 batches for 34×34 input image.

input image resembles a machine-like image), the ViT model is implemented in this study. Table 4.2 illustrates that the **ViT-HAR** model exhibits reduced parameters, with the input layer reduced to four layers, Hidden size to 512, and Heads to 8. These modifications lead to a significantly lighter model with 6.82 million parameters and 0.06 FLOPs, approximately six times less than the ViT base model. However, these parameters for M-LeNet are lower, 10 times smaller in number of parameters with 0.6 M, and 61 times smaller with 0.001 G FLOPs.

The size of input data significantly influences the parameters and computational operations (FLOPs) of the ViT model. For instance, consider the ViT-base model, renowned for its lightweight architecture. When utilizing input data from the RHM dataset which is the RGB-based version of the introduced skeleton dataset in this work (frames with dimensions 640×480 pixels), this model comprises a staggering 279 million parameters and requires 110.65 billion FLOPs.

However, through transformative measures applied to spatial and temporal image data, encapsulating all video frames into a compact 34×34 tensor, these figures are drastically reduced. In this scenario, the ViT-base model features a mere 43.2 million parameters and demands a modest 0.37 billion FLOPs. This represents a remarkable reduction of 6.5 times in parameter count and a staggering 300-fold decrease in FLOPs.

Comparing the ViT-base model with the ViT-HAR model further underscores this efficiency. The ViT-HAR model exhibits a parameter count and FLOPs approximately 40 and 1844 times smaller than those of the ViT-base model, respectively.

This revision clarifies the comparison between different ViT models and emphasizes the significant reduction in parameter count and FLOPs achieved through input data transformation by using human skeleton data.

4.2 Results

This section compares the results of the *M-LeNet* and *ViT-HAR* classification models in different conditions. Table 4.3 shows the comparison results of parameters like *accuracy*, number of total *trainable parameters*, different *camera views*, number of skeleton *positions*, and *classes*. The FLOPs parameters for all M-LeNet models are 0.001 G, and the ViT-HAR model is 60 times more complex by 0.06 G.

The results, presenting various evaluation metrics, including precision, recall, F1 score, and accuracy, were computed to assess the performance of the classification model. However, due to the negligible deviation among their values and for the sake of succinctness in the results table, only accuracy values were displayed. It is imperative to note that high precision signifies a low rate of false positives, high recall denotes a low rate of false negatives, and a high F1 score indicates a harmonized balance between precision and recall. Additionally, accuracy serves as a comprehensive measure of the model's predictive correctness. The convergence of these metrics

to similar values underscores the model's consistency and reliability across multiple dimensions of classification accuracy, reinforcing its robust performance.

In the analysis, accuracy is calculated as the proportion of correctly predicted instances out of the total instances. This metric evaluates the model's performance by measuring how many predictions align with the actual labels. To ensure reliable results, the reported accuracy is based on a single run. We use a standard training and testing split, where 80% of the data is allocated for training and validation and 20% for testing. This split ensures that the model is trained on a substantial portion of the data while being tested on a separate, unseen subset to evaluate its generalization capability.

These two models are applied to the RHM-HAR-SK dataset, including a synchronized three-view video stream. Table 4.3 shows the results of the models trained on full views, for all 14 classes, and all poses in the upper section and the lower section show the same comparison with excluded ankle poses (marked with 0-15 on poses column). The results show that the overall accuracy is between 69 and 77 percent for all views and 57 to 78 percent for single and double views respectively. Among them, the comparison of models with all poses and removed ankles shows that for the ViT model, the accuracy moderately increased by 3% in all views and remained the same in a single view. In contrast, the M-LeNet model decreased by about 2% in both high and low parameters models. The rationale behind removing these joints was to assess the impact of this action on model accuracy. The ankle joint was chosen because it exhibited lower confidence during the pose analysis conducted in Chapter 3. The difference between these M-LeNet classifiers is the number of parameters in linear layers, which one is double compared to the other.

In this experiment, the HRNet pose estimation method is used for the human skeleton extraction and all the results in Table 4.3 are based on that. In the last part of the lower and upper section of the table, the details of *single view* training models are shown. Interestingly, the ViT model results follow the missed poses statistics in the RHM-HAR-SK dataset in Figure 3.8, in which the front and robot views have fewer missed poses, and the highest accuracy among all views, and the Back-view is less accurate with more missed poses. Moreover, the Front-view accuracy is 78% in the ViT-HAR model, and robot-view accuracy is 70% in M-LeNet.

The double-view combination results are shown in the second part of the upper section. Combinations of Front (F), Back (B), and Robot (R) views are considered for assessing their impact on accuracy. The average accuracy of the double-view in the ViT models is higher than the lowest accuracy in the relevant single-view and less than the higher one, which means that the accuracy of the view with a lower value increased. For instance, the individual robot-view accuracy increased from 72% to 75% in combination with front-view and back-view increased from 61% to 69% when fused with front-view. For M-LeNet models, all single-view accuracy increased in the combination of double-views.

Although the results in all view combination settings don't show consistent improvement, there is evidence that multi-view can enhance accuracy. For example, the Back-view and its combinations with other views suggest that multi-view has potential for improvement.

The comparison of the upper section and lower one proves that removing low confidence joints like the ankle joints does not affect negatively, even in ViT all-views model accuracy increased by 3%. For M-LeNet and all single views, the accuracy fluctuated about 1%.

An examination of the number of parameters in Table 4.3 illustrates that the M-LeNet model exhibits a significantly lower number of parameters in comparison to the ViT model. Furthermore, the results of removing poses with lower accuracy further contribute to the reduction in the model's parameters.

Model	Accuracy	Params	Views	Poses	Classes
M-Lenet	70%	0.6M	ALL	ALL	14
ViT-HAR	71%	6.8M	ALL	ALL	14
M-Lenet	71%	0.6M	R+B	ALL	14
M-Lenet	70%	0.6M	R+F	ALL	14
M-Lenet	70%	0.6M	B+F	ALL	14
ViT-HAR	**75%**	6.8M	R+F	ALL	14
ViT-HAR	69%	6.8M	B+F	ALL	14
ViT-HAR	68%	6.8M	R+B	ALL	14
M-Lenet	70%	0.6M	Robot	ALL	14
M-Lenet	57%	0.6M	Back	ALL	14
M-Lenet	66%	0.6M	Front	ALL	14
ViT-HAR	72%	6.8M	Robot	ALL	14
ViT-HAR	61%	6.8M	Back	ALL	14
ViT-HAR	**78%**	6.8M	Front	ALL	14
M-Lenet	69%	**0.32M**	ALL	0-15	14
ViT-HAR	74%	6.8M	ALL	0-15	14
M-Lenet	69%	**0.32M**	Robot	0-15	14
M-Lenet	58%	**0.32M**	Back	0-15	14
M-Lenet	69%	**0.32M**	Front	0-15	14
ViT-HAR	73%	6.8M	Robot	0-15	14
ViT-HAR	61%	6.8M	Back	0-15	14
ViT-HAR	**77%**	6.8M	Front	0-15	14

Table 4.3: Results of ViT and M-LeNet classification methods on RHM-HAR Skeleton dataset in different conditions.

Here are the overall points, presented as bullet points for clarity:

- Combining different views can potentially increase overall accuracy (e.g., combining the back view with two other views).
- A simple CNN model with a low number of parameters can achieve comparable accuracy to a more complex model like ViT, which has about 10 times more parameters and 60 times more complexity.
- Eliminating the knee pose, which has lower confidence, does not significantly negatively impact overall accuracy.

4.2.1 Statistical Analysis of Model Performance

The purpose of this section is to compare the accuracy of two HAR models. To assess the normality of the data distributions, we conducted Kolmogorov-Smirnov and Shapiro-Wilk tests. The results shows that the M-LeNet dataset deviated from normality, whereas the ViT-HAR did not. Consequently, we applied a non-parametric method, the Mann-Whitney U test, to compare the accuracies of the two models.

4.2.1.1 Tests of Normality

To evaluate whether the data were normally distributed, we used the Kolmogorov-Smirnov and Shapiro-Wilk tests. The null hypothesis for these tests states that the data are normally distributed.

Table 4.4: Results of Normality Tests for MLeNet and ViT Datasets

HAR Model	Test	Statistic	df	Sig. (p-value)
M-LeNet	Kolmogorov-Smirnov	0.371	12	< 0.001
	Shapiro-Wilk	0.673	12	< 0.001
ViT-HAR	Kolmogorov-Smirnov	0.141	12	0.200
	Shapiro-Wilk	0.919	12	0.280

The results are summarized in Table 4.4. For the M-LeNet model, both tests indicated a significant deviation from normality (Kolmogorov-Smirnov $D(12) = 0.371, p < 0.001$; Shapiro-Wilk $W(12) = 0.673, p < 0.001$). Therefore, we reject the null hypothesis and conclude that the M-LeNet data are not normally distributed. For the ViT-HAR, the tests did not indicate a significant deviation from normality (Kolmogorov-Smirnov $D(12) = 0.141, p = 0.200$; Shapiro-Wilk $W(12) = 0.919, p = 0.280$). Thus, we retain the null hypothesis and conclude that the ViT-HAR data are normally distributed.

Figures 4.7 present histograms and Q-Q plots for the M-LeNet and ViT-HAR models, respectively. The histograms provide a visual assessment of the data distribution, while the Q-Q plots compare the observed quantiles with the expected quantiles under normal distribution. For the M-LeNet model, the histogram and Q-Q plot show clear deviations from normality, corroborating the results of the normality tests. In contrast, the ViT-HAR displays a more normal distribution pattern.

4.2.1.2 Non-Parametric Test

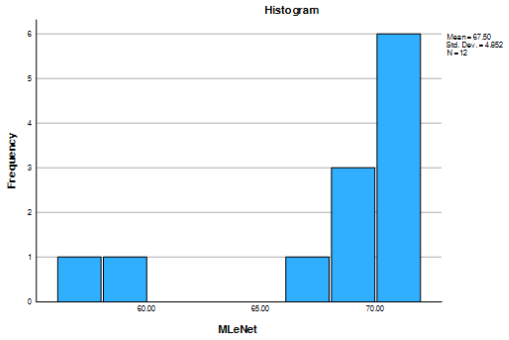
Given the non-normal distribution of the M-LeNet model, we employed the Mann-Whitney U test, a non-parametric method, to compare the accuracy of the two HAR models. The null hypothesis for this test states that the distributions of accuracy are the same across the two models.

The Mann-Whitney U test results are summarized in Table 4.5. The test statistic was 102.500, with an asymptotic significance (2-sided) of 0.076 and an exact significance (2-sided) of 0.078. The results of the Mann-Whitney U test indicate that we retain the null hypothesis that the distribution of accuracy is the same across the categories of the model (p = 0.078).

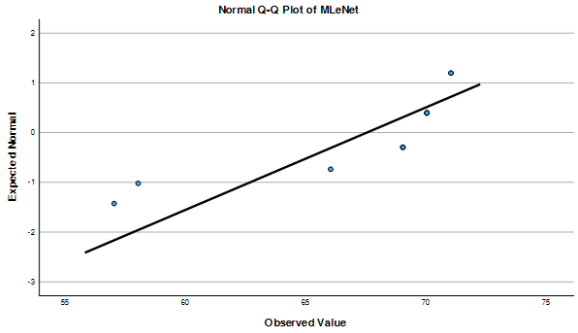
Figure 4.9 presents the box plots of the accuracy distributions for M-LeNet and ViT-HAR models. These box plots visually illustrate the spread and central tendency of the accuracy values for each model, reinforcing the statistical conclusion that there is no significant difference between the two distributions.

4.2.2 Computational Complexity Analysis

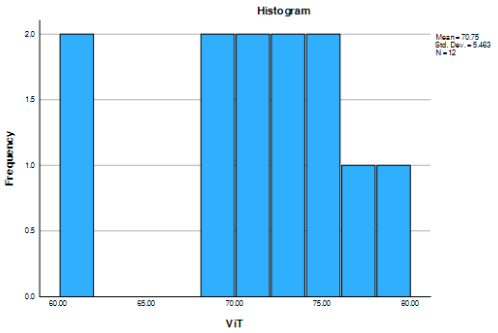
In addition to accuracy, it's essential to consider the computational complexity of the models. The M-LeNet models have a relatively low computational complexity, with FLOPs parameters of 0.001 G. In contrast, the ViT-HAR model is significantly more complex, with FLOPs parameters of 0.06 G. This indicates that the ViT-HAR model requires 60 times more computational resources compared to the M-LeNet models.



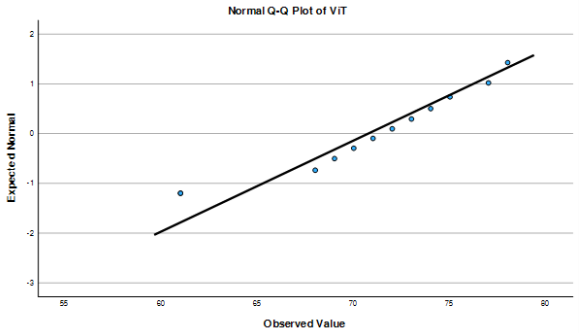
(a) Historgram of M-LeNet



(b) Q-Q Plot of M-LeNet



(c) Histogram of ViT-HAR



(d) Q-Q Plot of ViT-HAR

Figure 4.7: Histograms and Q-Q Plots for MLeNet and ViT

Statistic	**Value**
Total N	24
Mann-Whitney U	102.500
Wilcoxon W	180.500
Test Statistic	102.500
Standard Error	17.188
Standardized Test Statistic	1.774
Asymptotic Sig. (2-sided test)	0.076
Exact Sig. (2-sided test)	0.078

Table 4.5: Independent-Samples Mann-Whitney U Test Summary

4.2.3 Impact of Class Reduction on Model Accuracy and Action Similarity Analysis

Following the analytical methods for joint movement which is introduced in Sec. 3.3.1, the confusion matrix can also show valuable information about the action similarity which overlaps with the finding in Chapter 3. To examine it, some of the actions with higher similarity are removed from the original list which includes 14 actions. In the following, the number of classes was changed from 14 to 11, 10 and 9. The outcome reveals that removing these classes increases the accuracy of M-LeNet and the ViT models. Looking at Figure 4.10, the confusion matrix of these models, reveals the relation between classes. For example, in Figure 4.10 A and B, the

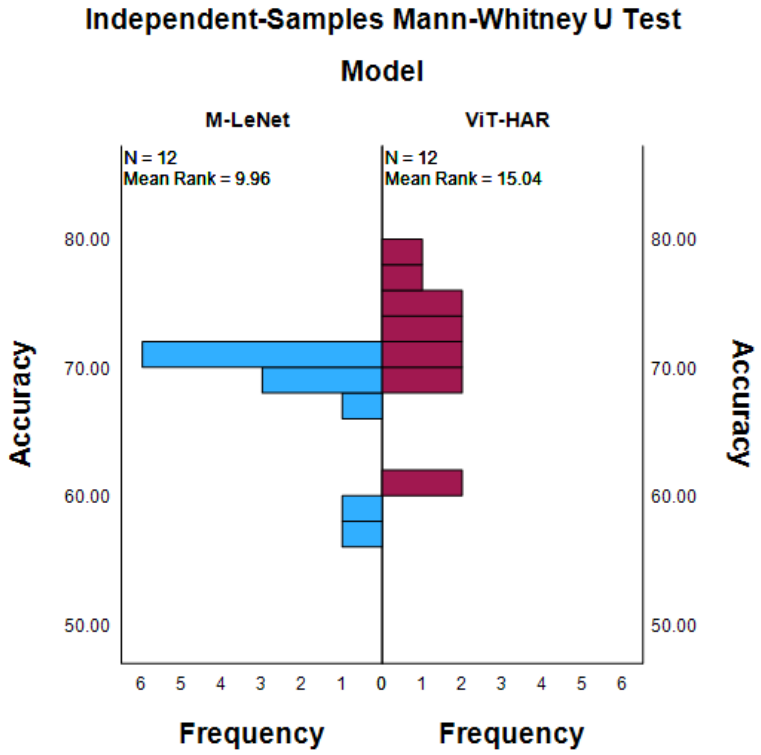


Figure 4.8: Comparison of Accuracy Distributions Between MLeNet and ViT Models Using the Independent-Samples Mann-Whitney U Test

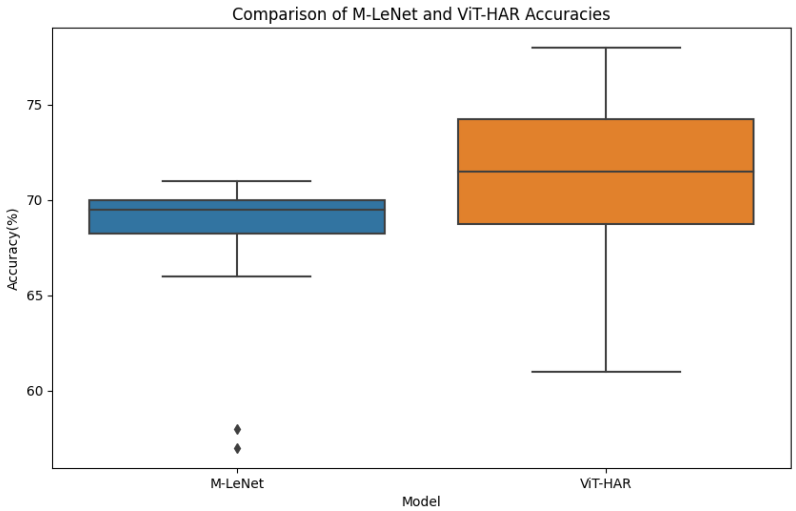
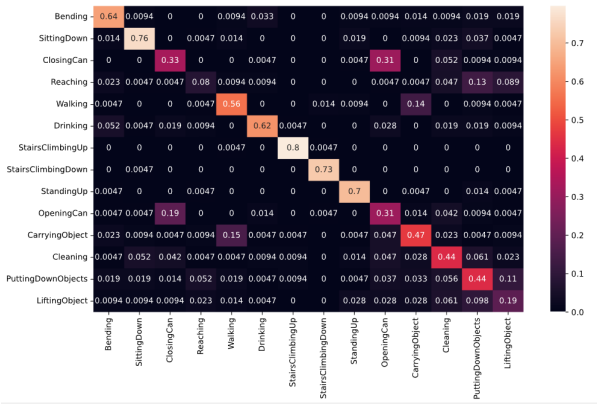
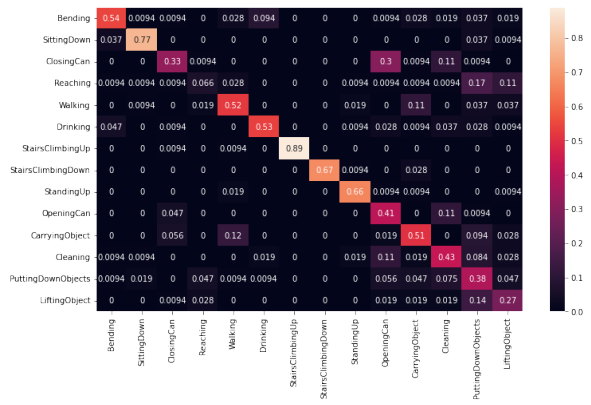


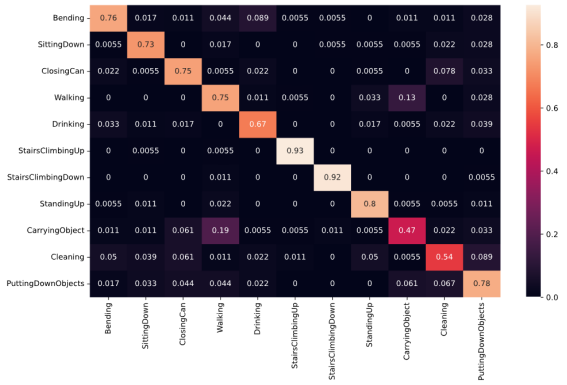
Figure 4.9: Comparison of Accuracy Distributions between M-LeNet and ViT-HAR Models



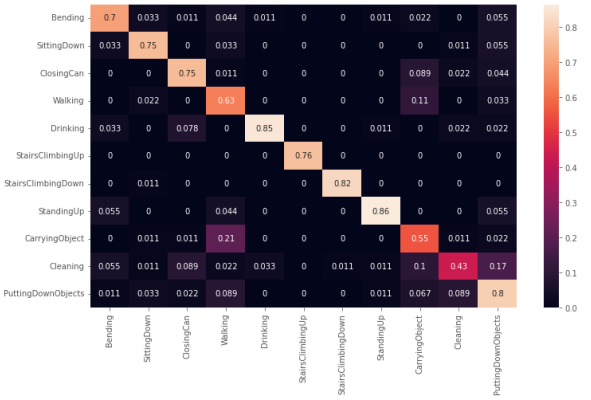
(a) The ViT-HAR model with 14 classes



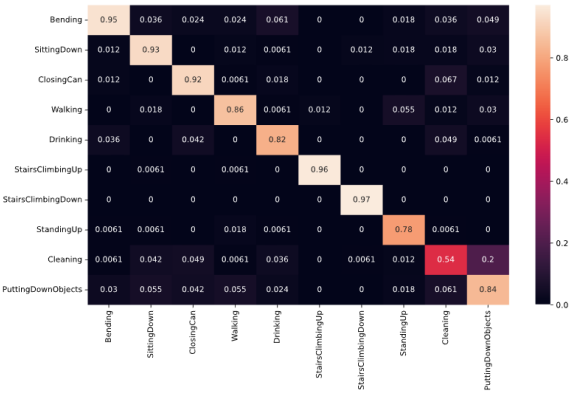
(b) The M-LeNet model with 14 classes



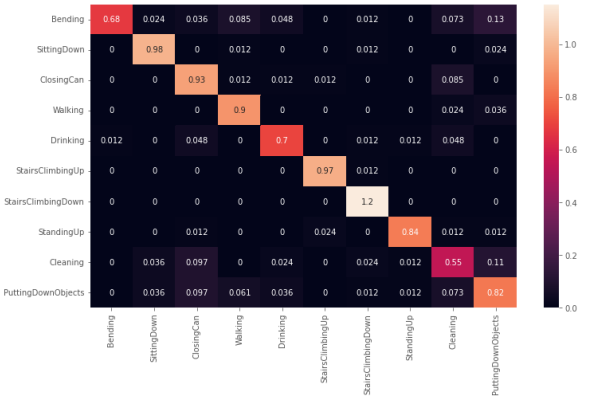
(c) The ViT-HAR model with 11 classes



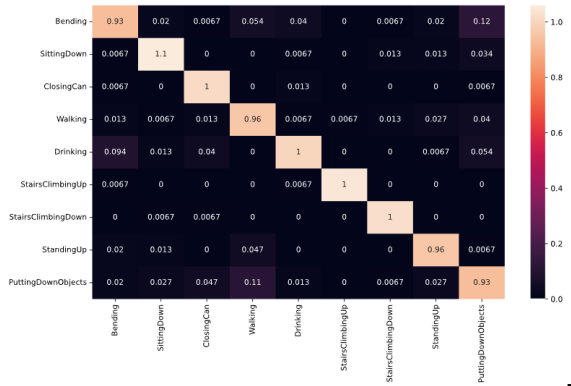
(d) The M-LeNet model with 11 classes



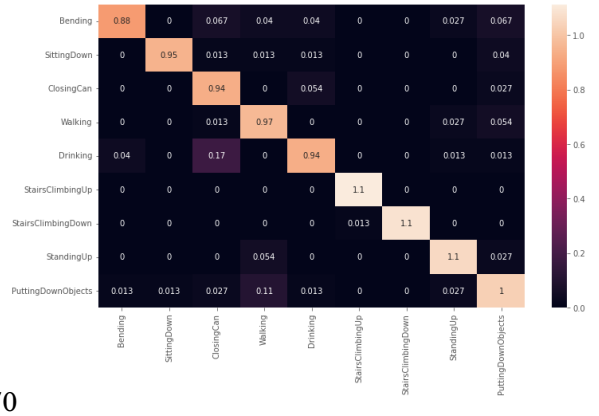
(e) The ViT-HAR model with 10 classes



(f) The M-LeNet model with 10 classes

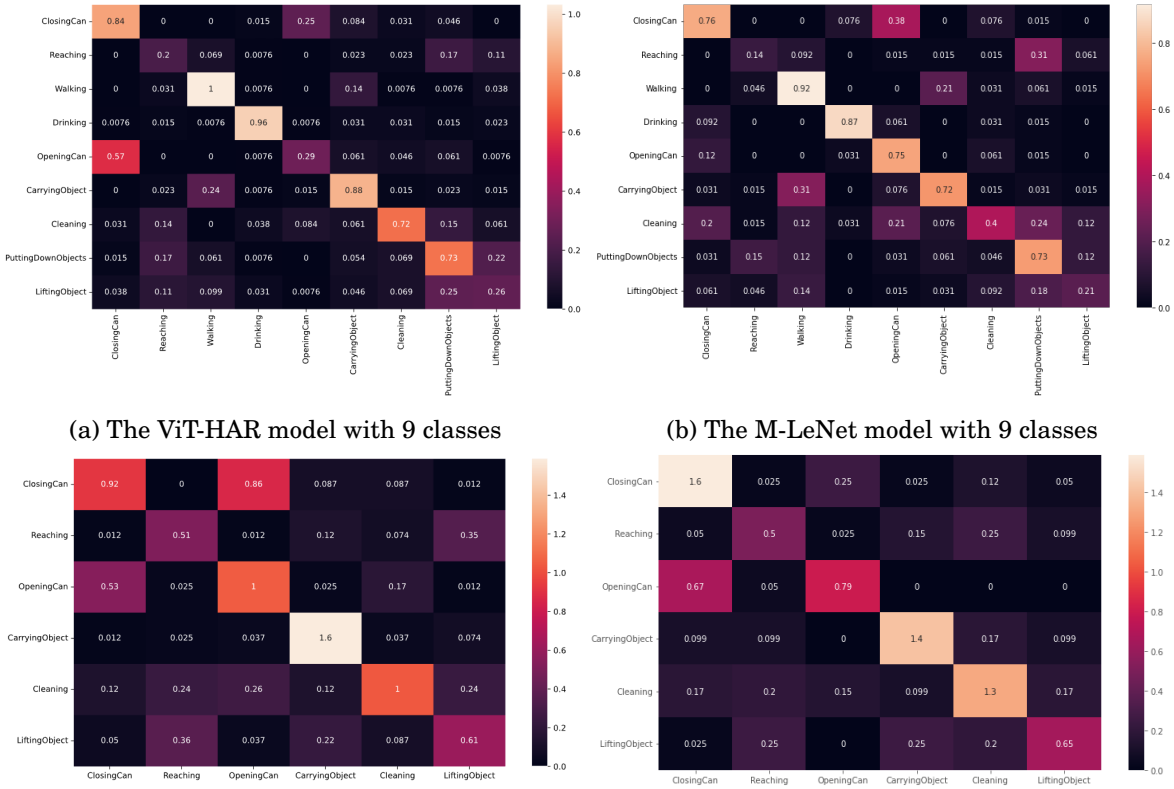


(g) The ViT-HAR model with 9 classes



(h) The M-LeNet model with 9 classes

Figure 4.10: The confusion matrix with different model configurations



(a) The ViT-HAR model with 9 classes

(b) The M-LeNet model with 9 classes

(c) The ViT-HAR model with 6 classes

(d) The M-LeNet model with 6 classes

Figure 4.11: The confusion matrix with random intervention

reaching action has similarities with many actions, specifically with actions that act with objects like pouting down objects and lifting objects. Likewise, the actions with more movements in hand and object have more similarities, like opening can and closing can and reaching.

In the next step, three classes were removed, *reaching*, *the opening can*, and *lifting the object*. Overall accuracy increased by almost 10% in both models, and each action had better results. Figure 4.10 *C* and *D* shows the differences clearly. The *carrying object* and *walking* actions also have high similarities. Then the carrying object action was removed in the next step, and the results show about a 5% increase in accuracy overall.

The last step is removing the *cleaning* action, and then *nine* classes remain. Individual class accuracy hovers around 97% in the confusion matrix, and the overall accuracy is almost 90%.

Figure 4.11 *A* and *B* shows the confusion matrix of two models classification in nine classes. These classes include the five removed classes from the first experiment. The results compared to the 4.10 *G* and *H* which also have nine classes, reveal that the reason for increasing the accuracy of G and H is not about the decreasing number of classes and emphasizes the importance of selecting classes wisely. Images C and D in Figure 4.11 show the six similar classes confusion matrix, which is inhomogeneous. The low accuracy results for ViT and M-LeNet, prove that similar actions in a model decrease prediction accuracy. Overall, this emphasizes that a lower number of classes does not guarantee higher accuracy, whereas having less similar classes can.

4.3 Conclusion and Addressing the Research Questions

In this chapter, two lightweight HAR models are introduced, namely M-Lenet and ViT-HAR, derived from modified versions of the LeNet and ViT Image classifiers. The M-Lenet is CNN-based, while ViT-HAR is based on transformers. The modifications aimed to enhance learning parameters for M-Lenet and reduce complexity for ViT-HAR compared to their base models. Transforming spatial and temporal frames into 2D images streamlined the pipeline, enabling effective handling of multi-view combinations.

RQ1:How effective is the use of skeleton-based human activity recognition? Enhancing skeleton-based models allows us to transform input data into a versatile 2D tensor, offering a pathway to leverage streamlined classifiers. This modification enhances scalability, enabling the application of various CNN models traditionally used in image classification to the transformed skeleton data. This adaptability amplifies the potential for effective human activity recognition within AAL environments.

The exploration of various ViT variants has illuminated their diverse architectural configurations and sizes, each characterized by distinct FLOP and parameter counts. This study introduces a modified ViT variant, ViT-HAR, specifically designed for Human Activity Recognition (HAR) tasks. Through strategic adjustments, ViT-HAR emerges as a significantly lighter model compared to other ViT variations, boasting reduced parameters and FLOPs. Specifically, ViT-HAR features a parameter count that is 6.5 times smaller and demands a computational workload 300 times lighter than the ViT-Base model. The input data size plays a pivotal role in shaping ViT model parameters and computational operations, with the transformation to a compact tensor resulting in substantial reductions. These findings underscore the efficacy of input data transformation in minimizing parameter count and FLOPs, leveraging human skeleton data for HAR tasks.

RQ2: What is the impact of different camera perspectives on the quality and richness of data captured for human activity recognition in multi-view scenarios? The findings underscore the impact of additional views in bolstering overall accuracy, notably elevating the lower single-view accuracy. This compelling outcome serves as a catalyst, inspiring further exploration and development of advanced multi-view architectures, a focal point detailed in Chapter 5.

RQ3: What are the optimal models for multi-view human activity recognition, and how do different levels of data combination approaches influence their suitability, advantages, and drawbacks in multi-view HAR?

Despite the ViT-HAR model's higher computational complexity, the lack of significant differences in accuracy between the M-LeNet and ViT-HAR models suggests that, from an accuracy standpoint, both models perform comparably across various conditions. Therefore, when choosing between these models for practical deployment in Human Activity Recognition tasks, considerations should extend beyond accuracy to include factors such as computational complexity, ease of implementation, and specific application requirements.

The CNN model exhibits approximately 10 times fewer parameters and about 60 times less complexity compared to their transformer counterparts. In Chapter 5, the focus is on the CNN-based model within a multi-view architecture, leveraging its inherent strengths for robust and effective human activity recognition across multiple perspectives.

CHAPTER 5

MULTI-VIEW STRUCTURE

This chapter introduces an efficient multi-view structure employing CNN models such as LeNet, M-Lenet, ResNet, MobileNet, SqueezeNet, DenseNet, and MnasNet. A comparative analysis is conducted to benchmark their performance, utilizing the AAL dataset detailed in Chapter 3. The evaluation considers key factors such as the number of training parameters, FLOPs, performance time, and accuracy across the selected CNN models. By harnessing the advantages of skeleton-based models and the integration of multiple camera views, it is shown that the proposed system exhibits improved accuracy of the Robot-view, reduced training parameters, and faster performance.

5.1 Multi-view CNN-based HAR structure

In this work, the multi-view learning (MVL) structure is systematically constructed, employed and developed. Including the dataset specifications in Section 5.1.1, the interpretation and preprocessing of spatial and temporal data in Section 5.1.2, the establishment of the multi-view CNN models in Section 5.1.3, including feature level fusion, and the multi-view co-learning methods. Collectively, these components compose a comprehensive structure crucial for deploying the benchmark and evaluating models across various variables.

5.1.1 AAL multi-view dataset

To effectively address the recognition of human activities based on skeletal data with multiple perspectives in ambient assistive living scenarios, where a robot is also involved, it is crucial to select a dataset that encompasses all relevant variables. After careful consideration, the RHM-HAR-SK dataset [144] was opted for this work (Chapter 3, an extension of the RHM RGB data [1]). This dataset offers several advantages, primarily focusing on the classification of multi-view human activities, comprising trimmed videos from four distinct cameras: two wall-mounted (Front-view and Back-view), a mobile robot (Robot-view), and a ceiling fish-eye camera (Omni-view). Similar to the Chapter 4 the Omni-view is excluded. These cameras' strategic placement ensures comprehensive coverage of a typical living room, creating overlapping views.

Furthermore, the RHM-HAR-SK dataset encompasses a diverse range of activity classes, a key aspect influenced by research conducted by Bedaf et al. [145]. Their study focused on identifying crucial daily activities that were essential for independent living. Using the list of activities in the study by [145], the dataset prepared and used in this submission captures a total of fourteen daily activities captured indoors, underscoring the potential advantages that companion robots and ambient-assistive systems can offer if they can successfully detect and interpret these activities.

5.1.2 Human skeleton stream to tensor

Capturing the spatial and temporal changes of a human skeleton within a video stream is of paramount importance, as it allows us to preserve intricate details of human body movements. To effectively convey this information to a learning model, we must ensure the data is structured optimally. The input data structure is similar to the method that introduced in Chapter3. Referring to Table 4.1 helps us understand how the input data size directly influences a model's complexity. As input dimensions increase, so does the model's complexity. This insight is vital for optimizing performance in a multi-view setting using CNN models. Achieving the right input size is crucial, striking a balance between computational efficiency and accurate human activity recognition. This balance becomes even more critical when handling data from various camera perspectives, demanding additional processing. This work focuses on designing a multi-view structure that excels in both efficiency and effectiveness by addressing this intricate balance.

5.1.3 Multi-view CNN configurations

At the heart of the multi-view structure lies the integration of information derived from diverse perspectives. To achieve this objective and gain a deeper understanding of crafting an effective and efficient structure, a systematic approach is adopted. This method involves combining data at various stages of the HAR pipeline: at the feature level (Section5.1.3.1), during the batch-level training process (Section 5.1.3.2), at the high-level probability stage (Section 5.1.3.3), and through a combination of the last two levels (Section 5.1.3.4). In Section 2.5.1, two primary structures for Multi-View CNN were introduced, namely the "one-view-one-net" mechanism and the "multi-view-one-net" mechanism. The specific emphasis on the "multi-view-one-net" mechanism is placed, where all input data is learned by a single model. The lightweight and less complex nature of this approach is a crucial feature, contributing to the overall streamlined design of the structure by having only one model.

5.1.3.1 Multi-view low-level fusion (LW)

To implement feature-level or low-level fusion, the tensor data extracted from the initial stage of the HAR process is input into the classification model. This is the first fusion method that is developed in Chapter 4, where each input camera view, represented as a tensor file as previously described in Section 5.1.1, is treated as an individual input channel. In practical terms, if we consider, for instance, three camera views, the model's input configuration is adjusted to accommodate three channels, and subsequent training, validation, and testing procedures are conducted accordingly.

The advantage of this fusion method lies in its simplicity. For instance, we can apply standard image classification techniques without any modification for the three input channels, effectively processing them as RGB data. However, it's important to acknowledge the potential drawback:

The fusion of input data does not encompass feature extraction beyond the initial transformation of human joint data into the tensor. As the experiments proceed, they present comparative results alongside other methods, shedding light on the trade-offs between simplicity and the potential loss of certain features and details.

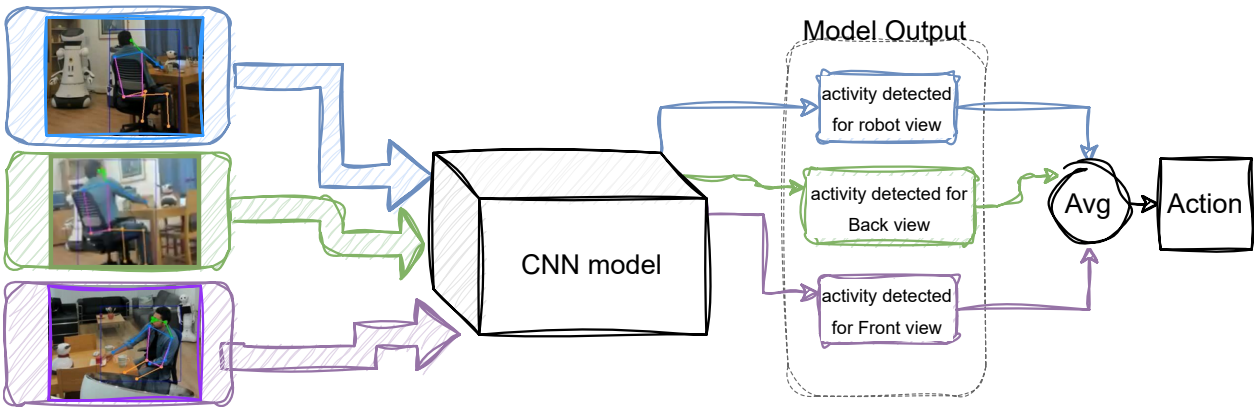


Figure 5.1: The structure of high-level multi-view co-learning.
Input data (34 × 34 tensor) from the same subject is sequentially fed into a single model. After obtaining the average output from all inputs (represented by a specific colour), the resultant output is used in the subsequent training or performance processes.

5.1.3.2 Multi-view mid-level co-learning(MD)

An approach that is developed to combine all views involves utilizing input data from all perspectives during the training process simultaneously, which is called *multi-view (MV) mid-level co-learning*. In other words, in each batch multiple synchronised input data in coming and one by one going through the prediction, loss and backpropagation. Since there is only one model, through each view input, the model weights will update for the next view. Figure 5.2, showing the training process of the Mid-level. This process repeats for all views j ($j = 1,2,\ldots,v$) within each batch i ($i = 1,2,\ldots,m$) for each epoch.

5.1.3.3 Multi-view high-level co-learning(HG)

In this MV high-level co-learning method, multiple views are combined at the highest level of the pipeline, using only the average output in the learning process, unlike mid-level co-learning, which involves individual view outputs in the learning process. However, similar to mid-level co-learning, this method follows the multi-view-one-net structure, as depicted in Figure 5.1 for MV-HG co-learning.

Let $V_1, V_2, \ldots, V_v$ represent the views (e.g., Robot, Back, Front) for a given dataset. Each view has a set of outputs, $O_1, O_2, \ldots, O_v$, where O_j is the set of outputs for view V_j. For the MV High-level Co-learning approach, the learning process involves averaging the predictions from all views. This can be mathematically expressed as:

$$O_{\text{avg}} = \frac{1}{v} \sum_{j=1}^{v} O_j \tag{5.1}$$

Where:

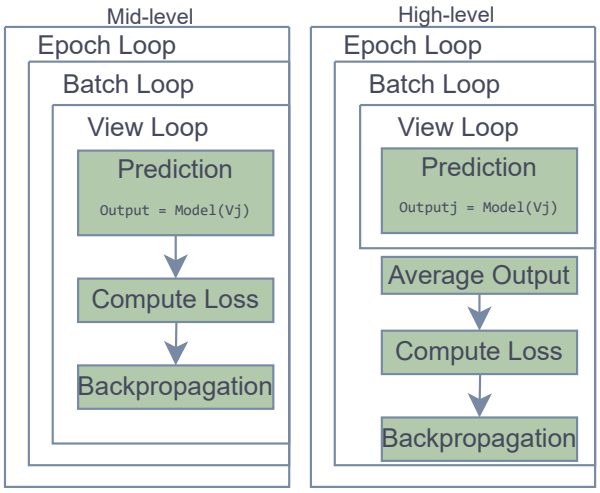


Figure 5.2: Process diagram of Mid-level and High-level methods during training (HG)

- O_{avg} represents the average predictions for all views.
- v is the total number of views.
- O_j is the set of predictions for view V_j.

This process involves obtaining the average predictions, O_{avg}, which is then used for further learning and model updates. Figure 5.2 illustrates the training process of the High-level co-learning. Within the view loop, the prediction output of each view is calculated. After the loop, the average value undergoes the loss computation and backpropagation.

5.1.3.4 Combining multi-view mid-level and high-level co-learning(MH)

The MD method faces a limitation in its inability to utilize outputs from all perspectives within the one-net structure simultaneously, rendering it unsuitable for testing. Thus, the High-level co-learning model strategy becomes necessary during the testing phase. To address this and capitalize on the strengths of both approaches, we propose the MH model, which combines elements of both Mid-level and High-level co-learning methodologies. Mid-level training enhances the model's training and validation process by incorporating diverse input data, while High-level combination is utilized for test set classification. Separating the test data from the training and validation sets in both Mid and High-level co-learning methods mitigates the risk of overfitting while maximizing the model's performance.

5.2 Experiments

This section presents some experiments and their results obtained as well as analysis focusing on comparative model performance. As established state-of-the-art and each representing a distinct architecture, the following CNN models are compared in the experiment as introduced in Section 2.4: LeNet, M-LeNet, ResNet18, MobileNet, SqueezNet, DenseNet, and MnasNet. To measure model performance as a dependent variable, the recognition *accuracy*, model *complexity* (number of parameters), *computational complexity* (FLOPs), and *performance time* as individual

metrics were compared. To assess potential influences on MV structures' performance, the R was considered as the base single view for improvement, along with LW, MD, HG, and the MH methods as additional independent variables in comparing MV methods.

5.2.1 Experimental Settings

Individual tests for CNN models have been performed under identical conditions, randomly sampling 34 frames from the dataset RHM-HAR-SK [144], which integrates YOLOv7 pose estimation to extract human skeleton data. In Chapter 4, HRNet was used for pose extraction. However, based on the comparison in Chapter 2, which demonstrated the superiority of YOLO, this chapter follows the YOLO-based pose extraction method. This comparison was conducted after the completion of Chapter 4.

Hyperparameters encompass a dropout rate of 0.0135, the learning rate has been set to 7.20e-05, with a weight decay of 5.21e-05, a batch size of 128, and a training duration spanning 128 epochs. The optimization method employed is Adam [146]. The experimental design incorporates stratified K-Fold cross-validation with five folds to ensure a uniform distribution of classes in each fold. Throughout the training, model performance is assessed on a validation set after each epoch, utilizing the cross-entropy loss as the chosen loss function. The dataset is split into training and validation sets during cross-validation, with 20% of the data reserved for subsequent testing. The trained model is then evaluated on the test set, and key metrics, including test loss, accuracy, precision, recall, F1-score, and the confusion matrix, are logged for analysis. Tests have been performed on a high-performance computer with the following specifications:

- Architecture: $X86\_64$
- CPU: AMD Ryzen Threadripper PRO 5975WX (32 Cores, 64 Threads, 7006.64 MHz)
- Graphics Card: NVIDIA GA102GL [$RTXA6000$] (10,752 CUDA cores, memory size of 48 GB GDDR6)
- Storage: PC801 NVMe SK Hynix 2TB SSD
- Memory: 125 GiB RAM

5.2.2 Model complexity & HAR

In this segment of the experiment, the relationship between model complexity and human activity recognition (HAR) is explored. Our hypothesis challenges the conventional belief that more complex CNN models inherently yield superior accuracy. To test this hypothesis, various CNN models are rigorously trained and tested, focusing initially on single-view HAR with the robot view as the base perspective. The results, as depicted in Table 5.1, paint a compelling picture. Models with fewer parameters and complexity, such as MnasNet (77.72%) and M-LeNet (77.14%), not only hold their own but also outperform single-view models even with higher parameters and complexity, like ResNet (76.91%) and DenseNet (74.81%).

5.2.3 Views & HAR accuracy

When analyzing the impact of additional views on HAR accuracy, the primary goal is to see the effect of enhancing a robot's perception by introducing multiple external cameras. As referenced

in Sec. 3.3, diverse perspectives capture distinct spatial and temporal features, impacting the overall model's performance. The results in Table 5.1 reveal that, except for some models in the LW method, all other view combination structures have improved the accuracy, showcasing the potential of multi-view perspectives.

The MH method notably demonstrates the highest improvement, ranging from 10% to 25% across various models. The top four highest accuracies are consistently achieved using the MH method, with notable performances from ResNet (90.90%), MnasNet (90.08%), DenseNet (89.93%), and M-LeNet (88.59%).

Conversely, the HG method showcases improvement of 5 to 14 percent, with ResNet (85.25%), M-LeNet (83.95%), and DenseNet (83.94%) securing the top spots in terms of accuracy. Notably, SqueezeNet exhibits a remarkable jump in accuracy within the HG structure.

For the MD method, improvements are slightly lower, ranging from 1 to 10.5 percent, with MnasNet (82.28%), DenseNet (81.05%), and M-LeNet (80.14%) claiming the top accuracy spots in the MD method. However, the LW structure's results are less consistent, with some models experiencing improvement and others exhibiting declines. Despite challenges, certain models like MobileNet (76.91%) and ResNet (76.91%) demonstrate reasonable improvement within the LW structure.

5.2.3.1 Enhancing Performance with MH Structure

This section presents the results obtained by applying the MH structure, which combines the Mid-Level (MD) and High-Level (HG) architectures, described in Sec. 5.1.3.4. The MH structure demonstrates the highest accuracy, as indicated in Table 5.1, highlighting its effectiveness in leveraging diverse perspectives.

Figure 5.5 and 5.6 showcase the Confusion Matrix, portraying the accuracy achieved in three individual views alongside the combination of all views by M-LeNet and ResNet models. Notably, the classification models for individual views are trained using the Mid-Level Co-Learning (MD) method. In contrast, the combined view (Figure 5.5 and 5.5) reflects the results of both Mid-Level and High-Level Co-Learning, denoted as MH. This MH structure computes the average output across all individual perspectives. The outcomes highlight the increase in accuracy across almost all actions and highlight the efficacy of the multi-view technique in boosting overall performance.

5.2.4 CNNs in multi-view trade-offs

In this phase of the experiment, the variation in the trade-off between model accuracy and complexity across different CNN architectures in the context of multi-view HAR has been investigated. Specific CNN architectures may demonstrate superior trade-offs between accuracy and model complexity in the multi-view HAR scenario. To identify the optimal structure, the accuracy density [147, 148] was assessed, and calculated as the ratio of accuracy to the number of parameters. A higher accuracy density signifies greater efficiency. Figure 5.3 plots the top accuracy of each tested model (MH method) against their accuracy density. The accuracy density serves as a metric to gauge the efficiency of parameter utilization for each model. In this figure, the run time of each model during the training processes is depicted using colour mapping, varying between 6 minutes (blue) for LeNet and 43 minutes for DenseNet (yellow).

The findings indicate that simpler models like LeNet exhibit greater effectiveness and efficiency compared to more complex CNN architectures such as MobileNet in the HAR task. Despite LeNet achieving an accuracy density of nearly 1400, making it a more effective model,

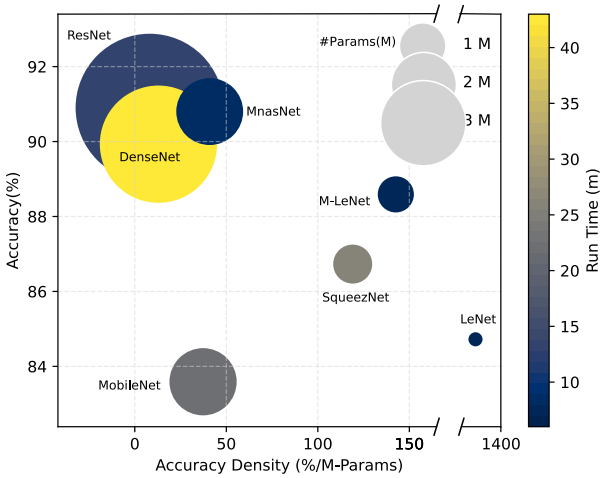


Figure 5.3: Accuracy density analysis for CNN models in the MH method
The centre of the circles represents the relevant value on the axes, denoting accuracy and accuracy density. Circle diameters correlate with the number of parameters, while their colours reflect the time spent during the training process.

its accuracy falls approximately 5% lower than the highest-performing model (ResNet at 90.9%). However, the top three accuracy models (ResNet, MnasNet, and DenseNet) display lower efficiency, with an accuracy density lower than 50. Notably, among these, DenseNet emerges as the slowest model in terms of performance, requiring approximately 40 minutes for training. Additionally, while M-LeNet and SqueezeNet demonstrate similar levels of efficiency, the runtime of SqueezeNet is nearly three times longer than that of M-LeNet. This observation highlights the direct impact of FLOPs on model performance time, considering that the FLOPs value in M-LeNet is approximately five times smaller than that in SqueezeNet.

Models	**Acc. (%)**	**#Params (M)**	**FLOPs(G)**
LeNet-LW	75.91	0.062006	0.00075
LeNet-HG	79.57	0.061706	0.00048
LeNet-MD	75.87		
LeNet-MH	84.71	-	-
LeNet-R	74.74		
M-LeNet-LW	76.67	0.621364	0.00125
M-LeNet-HG	**83.95**	**0.621184**	**0.00106**
m-LeNet-MD	80.14		
M-LeNet-MH	**88.59**	-	-
M-LeNet-R	**77.14**		
ResNet18-LW	82.04	11.689512	0.08102
ResNet18-HG	**85.25**	11.177422	0.07870
ResNet18-MD	77.52		
ResNet18-MH	**90.90**	-	-
ResNet18-R	76.91		
MobileNet-LW	76.91	3.504872	0.01644
MobileNet-HG	70.53	2.24123	0.01501
MobileNet-MD	65.18		
MobileNet-MH	**83.59**	-	-
MobileNet-R	58.76		
SqueezNet-LW	55.47	1.235496	0.00791
SqueezNet-HG	76.38	0.728526	0.00559
SqueezNet-MD	73.25		
SqueezNet-MH	**86.73**	-	-
SqueezNet-R	62.68		
DenseNet-LW	70.72	7.978856	0.06677
DenseNet-HG	**83.94**	6.961934	0.06395
DenseNet-MD	81.05		
DenseNet-MH	**89.93**	-	-
DenseNet-R	74.81		
MnasNet-LW	35.32	2.218512	0.00702
MnasNet-HG	81.51	**2.138512**	**0.00622**
MnasNet-MD	82.28		
MnasNet-MH	**90.08**	-	-
MnasNet-R	**77.72**		

Table 5.1: Summary of performance metrics for analyzed CNN models in human activity recognition.
listing accuracy, and model complexity expressed in the number of parameters and floating-point operations. R: Robot-View, MD: Mid-Level, HG: High-Level, MH: Mid and High-Level. "-"indicates the same value as above. Green: Highest accuracy, Orange: Highest growth, Grey: Highest Robot-view, Pink: Highest LW accuracy. Yellow: Most parameters and FLOPs, Light blue: Fewest.

5.2.5 Skeleton-based Versus RGB-based HAR

In this section, a comparative analysis between two distinct methodologies utilized for HAR is presented: the RGB-based approach and the Skeleton-based approach, both evaluated on the RHM dataset—a dataset specifically curated to simulate the AAL environment.

The outcomes from the RGB-based method were visualized through a confusion matrix (Figure5.4), showcasing the performance of the model across various actions. Similarly, the developed Skeleton-based model also yielded a confusion matrix (Figure 5.5d and Figure 5.6d), highlighting its performance in recognizing different activities within the AAL context.

The comparison highlights the performance disparity between the RGB-based (Dual-stream C3D) model, achieving approximately 70% accuracy with 92 million parameters, and the skeleton-based method utilized here, which achieves around 90% accuracy with only 0.6 million parameters, both using the same dataset. This indicates a nearly 20% higher performance for the skeleton-based approach, alongside significantly fewer model parameters and reduced complexity.

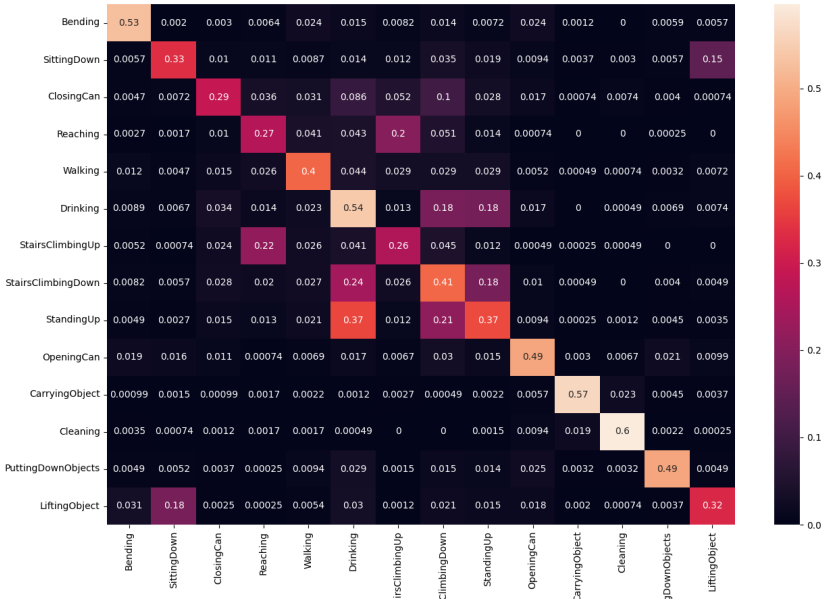
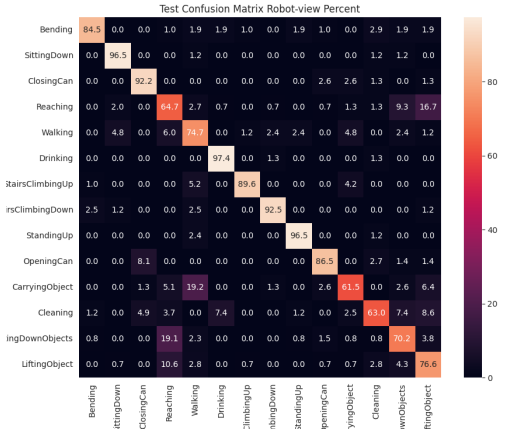


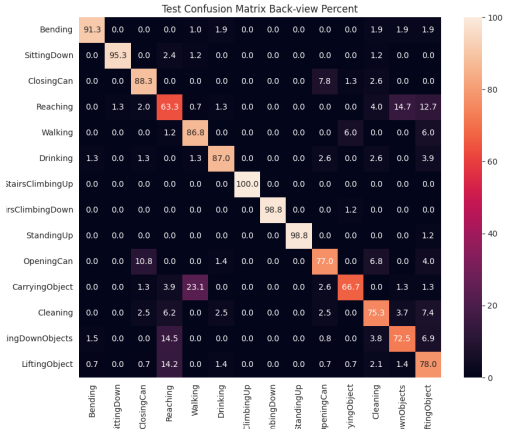
Figure 5.4: RHM Confusion Matrix for all Pair views with Dual-stream C3D Model in two-stream RGB-based with combining the Robot-view and Front-view

Upon thorough comparison, it became evident that the Skeleton-based model significantly outperformed the RGB-based counterpart in HAR tasks. The detailed analysis of the confusion matrices depicted superior performance metrics across multiple actions for the Skeleton-based model. Notably, the Skeleton-based model demonstrated heightened accuracy rates in recognizing 14 human activities within AAL environments chosen for this experiment.

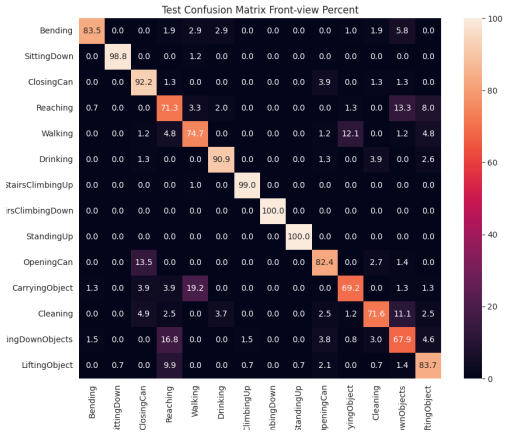
The observed supremacy of the Skeleton-based approach in HAR tasks within AAL environment underscores its efficacy and effectiveness. The skeletal representation of human movements appears to offer richer and more informative data for accurate activity recognition compared to the RGB-based approach. This finding highlights the potential of skeleton-based methodologies in delivering more precise and nuanced HAR solutions tailored specifically for AAL scenarios.



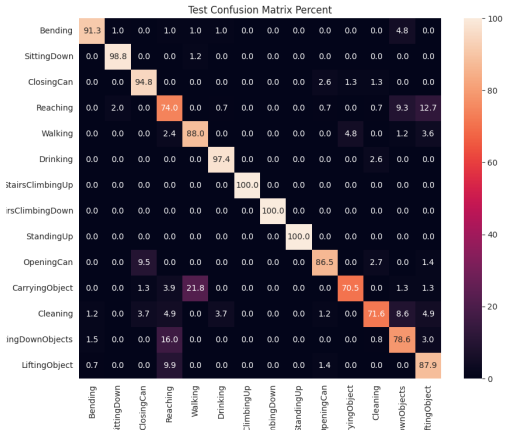
(a) Robot-view



(b) Back-view

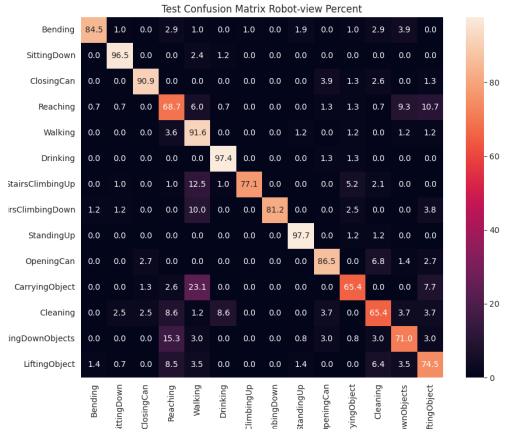


(c) Front-view

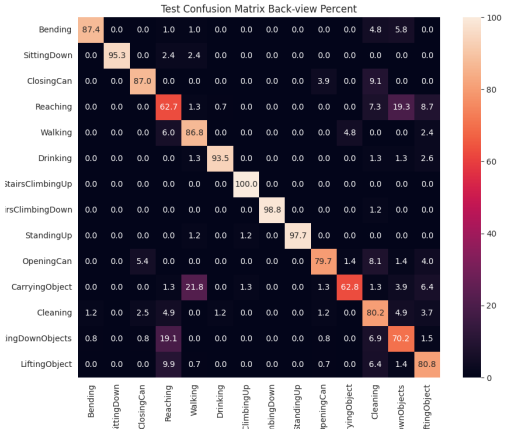


(d) All views

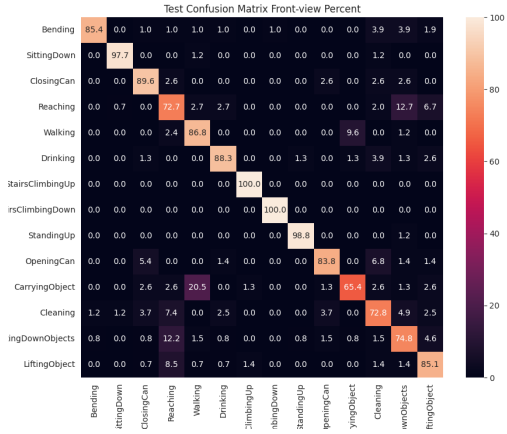
Figure 5.5: Confusion Matrix of MD and MH methods Using M-LeNet CNN model. Results of MD Method on Individual Views shown on Figures 5.5a, 5.5b, and 5.5c are referring to Robot, back and Front view respectively, and Figure5.5d referring to MH Method Across All Views



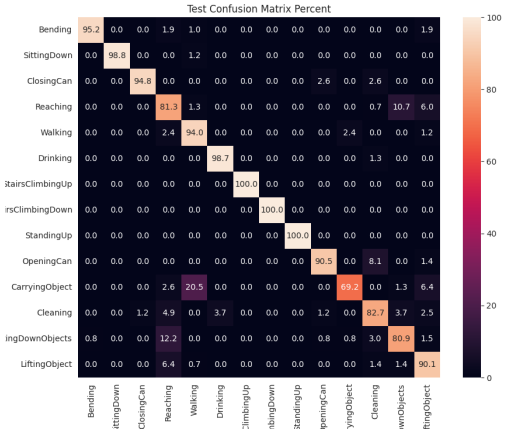
(a) Robot-view



(b) Back-view



(c) Front-view



(d) All views

Figure 5.6: Confusion Matrix of MD and MH methods Using M-ResNet CNN model. Results of MD Method on Individual Views shown on Figures 5.6a, 5.6b, and 5.6c are referring to Robot, back and Front view respectively, and Figure 5.6d referring to MH Method Across All Views

5.3 Conclusion and Addressing the Research Questions through multi-view structure

This chapter presents a multi-view CNN-based human activity recognition structure for enhancing robot perception in ambient assistive living scenarios. The overall content emphasizes the promising performance of the Multi-View CNN-based HAR structure, highlighting the significance of multi-view information, nuanced model complexities, and the necessity for a balanced approach in selecting architectures for optimal HAR performance in real-world applications. In the following, the research questions are addressed with the findings in this chapter.

RQ1:How effective is the use of skeleton-based human activity recognition?

The transformation of human skeleton stream data into a tensor format within the structure emphasized the potential effectiveness of the "multi-view-one-net" architecture which is used in this work. This structured approach enables the development of a multi-view system with high performance and reduced complexity, as shown in table 5.1, a simple CNN model like M-LeNet in a "multi-view-one-net" structure can perform effectively.

This streamlined structure lays the groundwork for future investigations aimed at refining skeleton-based models for multi-person HAR deployment, ensuring scalability and efficient management of multiple tasks. Ensuring that parallel HAR recognition can perform with low computational process in multi-view.

The comparative analysis between Skeleton-based and RGB-based (Figure 5.4) HAR methodologies on the RHM dataset reaffirms the superiority of the Skeleton-based approach in recognizing human activities within Ambient Assisted Living contexts. The remarkable performance exhibited by the Skeleton-based model signifies its effectiveness and potential for fostering advancements in automated monitoring and support systems aimed at enhancing the quality of life within AAL environments.

RQ2: What is the impact of different camera perspectives on the quality and richness of data captured for human activity recognition in multi-view scenarios?

Comparing the results of a single view and multi-view in Table 5.1 indicates a significant enhancement in HAR accuracy through additional camera views particularly highlighting the superiority of the MH method in consistently improving multi-view recognition. This observation underscores the crucial role of incorporating various perspectives in multi-view HAR tasks.

In addition, the accuracy and density results shown in Figure 5.3 demonstrate the effectiveness of the introduced multi-view HAR structure, referred to as "multi-view-one-net," which is being utilized for the first time in multi-view human activity recognition in this study. The efficacy of a simple model such as M-LeNet, as evidenced by its accuracy, performance time, complexity (FLOPs), and number of parameters, underscores the practicality of the MH pipeline in a multi-view setting. Furthermore, owing to its lightweight architecture, this pipeline has the potential to accommodate more than three input views, depending on the available computational resources.

RQ3: What are the optimal models for multi-view human activity recognition, and how do different methods for data combination influence their performance?

The comparison and evaluation of various CNN models like MnasNet and M-LeNet in Table 5.1 and Figure 5.3, showcasing that models with fewer parameters can outperform higher-complexity models, provide insights into suitable models for multi-view HAR tasks. Additionally, emphasizing the importance of resource constraints and specific applications for optimal model selection hints at the suitability of models concerning computational efficiency and accuracy.

In exploring the effectiveness of various data combination approaches within the Multi-View CNN structure, distinct strategies were employed to integrate multi-view information. The Low-Level Fusion approach (LW), focusing on feature-level integration, simplifies the process but may limit advanced feature extraction. To improve that, Mid-level co-learning (MD) was introduced which simultaneously trains models using multi-view data, enhancing data diversity during training without increasing the model parameters count and complexity.

On the other hand, High-Level Co-Learning (HG) simplifies training by averaging predictions from different views, possibly overlooking individual view nuances. Combining Mid-Level and High-Level Co-Learning strategies aims to benefit from both approaches, offering enriched training data while minimizing complexity during the model performance, as shown in Table 5.1. These diverse methods under the "multi-view-one-net" and CNN structure demonstrate trade-offs between simplicity and accuracy, impacting the effectiveness of human activity recognition across multiple perspectives.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This chapter provides concluding remarks and revisits the research questions and results obtained in support of their answers. Based on the findings, directions for future work are also considered.

6.1 Conclusion

This thesis concludes that a streamlined multi-view skeleton-based human activity recognition system, optimized for efficiency, is pivotal in enhancing accuracy without significantly escalating training parameters or performance time. This conclusion draws from the synthesis of findings across Chapters 3, 4, and 5, underscoring the potential of leveraging skeleton stream data transformed into tensor formats within the "multi-view-one-net" architecture and multiple data combination methods.

This research start with the creation of a novel open dataset comprising 26,801 videos depicting 14 daily living activity classes. The dataset features synchronized data from four camera perspectives, including one from a mobile robot and three fixed views.

In addressing **Research Question 1**, the significance of human skeleton stream data, extensively explored in Chapter 3 through the analysis of skeleton extraction methods and biomechanics, emerges as a pivotal factor in the evolution of Human Activity Recognition (HAR) systems.

In Chapter 3, two metrics were introduced for the qualitative and quantitative analysis of the dataset. The examination of extracted human skeleton data distinctly illustrates the capability of each camera view to capture the human body's skeletal structure.

In Chapters 4 and 5, the seamless integration of skeleton-based models demonstrates that skeleton data not only streamlines multi-view pipelines but also underscores their scalability and efficiency, significantly enhancing human activity recognition (HAR) in ambient assisted living (AAL) environments. Specifically, leveraging the skeleton model facilitates the transformation of spatial action data into a compact 2D tensor. This approach allows for processing a significantly reduced amount of data per action—only a small (34×34) tensor instead of processing 30 or 10 frames per second. Furthermore, while the presence of multiple views could potentially exacerbate

processing demands, this study illustrates that employing an efficient multi-view model based on the skeleton-to-2D formulation can effectively deploy HAR.

The utilization of a single image (tensor), coupled with its small size (34×34), presents opportunities for adapting existing CNN-based and transformer-based models such as LeNet and ViT. This adaptation capitalizes on the reduced computational complexity afforded by smaller input sizes. Specifically, in the case of CNN models, the Floating Point Operations (FLOPs) decrease by a significant factor, between 100 to 400 times—when employing a small image size compared to the original image size of 640×480. Similarly, for ViT models, the reduction in FLOPs reaches a remarkable 1800-fold decrease, accompanied by a substantial decrease in the number of parameters, approximately 297 times in the base ViT model. These findings underscore the effectiveness of the skeleton model and its transformation into a 2D tensor, enabling more efficient and scalable model architectures.

Comparison of the results with the RGB-based counterpart HAR in this dataset further highlights the efficacy of the skeleton-based model, exhibiting at least a 20% increase in accuracy in human action recognition. Moreover, the efficiency comparison reveals notable gains with the skeleton-based approach. For instance, a dual-stream C3D model comprises 93 million parameters. In contrast, the M-LeNet model incorporating skeleton data can accommodate more than three views with fewer than 1 million parameters, showcasing a significant reduction in model complexity. Additionally, the flexibility inherent in the skeleton-based model, particularly its capability to detect multiple individuals in video frames compared to the RGB-based model, inspires us to pursue the development of multi-person Human Activity Recognition (HAR) in future works.

Research Question 2 considers the role of perspective, e.g. robot-view, front and back views, and top-view, in multi-view HAR, properly illuminated across Chapters 3, 4, and 5. The consistent demonstration of superior multi-view recognition with additional camera views reinforces the imperative nature of diverse perspectives in fortifying robust HAR models, pivotal for effective human-robot interaction in AAL settings.

The spatial and temporal analysis conducted in Chapter 3 underscores the significance of each perspective in conveying activity-related information. Notably, the mutual information analysis reveals that synchronized videos do not necessarily convey identical information in each frame. Additionally, the utilization of analytical methods for representing joint movements, such as pairwise and min-max distance methods, highlights the varied perspectives through which human body movements are observed. This underscores the importance of combining data from multiple perspectives to capture richer information, resulting in enhanced recognition accuracy.

Meanwhile, the findings presented in Chapters 4 and 5 substantiate the notion that combining views enhances the accuracy of human action recognition. As detailed in Chapter 2.5.1, integrating the robot-view with two other perspectives led to a notable accuracy increase of up to 24%. Furthermore, in scenarios where a human is not visible in one or more camera perspectives, the presence of others compensates, contributing to a more robust HAR module.

The insights obtained from **Research Question 3**, detailed in Chapters 4 and 5, underscore the critical importance of model selection criteria. For instance, the examination of the Transformers model (ViT) in Chapter 4, comparison with a CNN model, illuminates the CNN model's capacity to effectively capture spatial and temporal features through a simplified data structure, transformed data into a 2D tensor. This finding highlights the potential of enhancing a simple CNN model (e.g., modifying LeNet to M-LeNet) to fulfil the HAR task with fewer parameters and reduced complexity compared to advanced models like ViT.

Subsequently, in Chapter 5, the comprehensive evaluation of various CNN models (including

LeNet, M-LeNet, ResNet, MobileNet, SqueezeNet, DenseNet, and MnasNet) reveals that models with fewer parameters, such as M-LeNet, exhibit superior performance in multi-view scenarios. This reaffirms the significance of computational efficiency in both system design and operation.

Lastly, the exploration of diverse data combination approaches, a central focus of Research Question 3 as detailed in Chapter 5, underscores the substantial impact of methodological choices on HAR effectiveness across multiple perspectives. The investigation into four data combination methods, low-level (LW), mid-level (MD), high-level (HG), and a combination of MD and HG (MH), demonstrates the effectiveness of different approaches and underscores the importance of multi-view topology in the HAR context. Specifically, the MH method emerges as particularly effective, yielding higher accuracy (between 5 to 20 per cent) across various CNN models.

In synthesis, the comprehensive insights from this multi-view exploration underscore the pivotal role of skeleton-based models, efficient model selection, diverse perspectives, and data combination methodologies in advancing robust and efficient HAR systems within AAL environments. These findings significantly contribute to enhancing human-robot interaction and assistive scenarios, aligning with the envisioned goals of optimizing HAR systems for practical deployment in real-world settings.

6.2 Contribution of the work to the body of knowledge

This research makes significant contributions to the field by focusing on the development and analysis of an efficient, comprehensive pipeline for skeleton-based multi-view human activity recognition.

6.2.1 Creation of a Multi-view Skeleton-Based Dataset

The primary contribution involves the establishment of a novel multi-view skeleton-based dataset within the domain of Ambient Assisted Living (AAL). This dataset serves as a foundational resource for further exploration and experimentation in this area.

6.2.2 Dataset Analysis and Metric Introduction

This work pioneers the introduction of several essential metrics and structures for dataset analysis. Notably, the introduction of metrics such as **miss frames** and **miss poses** enables both quantitative and qualitative analyses. Additionally, novel spatial and temporal illustration techniques, employing **pairwise distance** and **min-max distance** analyses, offer deeper insights into the dataset. The introduction of the MICM approach, grounded in mutual information theory, enriches the methodological toolbox for dataset examination at the frame level.

6.2.3 Development of Lightweight HAR pipeline

A significant contribution lies in the development of an innovative lightweight Human Activity Recognition (HAR) pipeline. This structure optimizes the input data by transforming it into a 2D tensor format, thereby enhancing the efficiency and effectiveness of the recognition process. The M-LeNet and ViT-HAR classifiers have been developed with high efficiency in this pipeline.

6.2.4 Creation of Various Multi-view structures

The research advances the field by introducing and evaluating diverse multi-view structures capable of comparing different levels of data combination employing various CNN models. This contribution not only provides novel insights into multi-view systems but also establishes a structure for comparative analysis across different data combination strategies.

These contributions collectively advance the understanding and implementation of skeleton-based multi-view human activity recognition systems, offering novel datasets, metrics, methodologies, and pipelines for future research endeavours in this domain.

6.3 Future Work

Advancements in technology have paved the way for innovative solutions aimed at enhancing the quality of life, particularly within Ambient Assisted Living (AAL) environments. Among these, Human Activity Recognition (HAR) systems stand as fundamental components, enabling the automated monitoring and understanding of human activities to support independent living, health monitoring, and personalized assistance.

The development of a streamlined and scalable HAR pipeline that operates across multiple views signifies a significant jump forward in the field. This pipeline, underpinned by its adaptability and efficiency, holds immense promise for seamless integration into various aspects of AAL scenarios, presenting multifaceted opportunities and addressing critical challenges.

The forthcoming sections delve into the transformative potential of integrating this advanced HAR pipeline into the fabric of AAL environments. This integration goes beyond conventional activity recognition, expanding into personalized dataset creation for diverse individuals, enabling multi-person recognition, deploying continual HAR with real-time anomaly detection, and fostering adaptive learning for enhanced human-robot interaction.

In the following, the potential future work that can be applied on top of the proposed pipeline is introduced.

6.3.1 Enhancing Dataset Usability

The concept of personalizing the RHM-HAR-SK Dataset holds tremendous potential for broadening the inclusivity and adaptability of Human Activity Recognition (HAR) systems within Ambient Assisted Living (AAL) environments.

By deploying generative models atop the existing dataset, a pioneering avenue emerges to accommodate individuals, particularly those in the third-age demographic or with a disability, who exhibit distinct movement behaviours not reflected within the current dataset's confines.

The envisioned approach of personalizing datasets stands as a pivotal strategy to enhance the pipeline's usability and applicability across a wider spectrum of individuals. This advancement can foster more inclusive AAL solutions that cater to the diverse movement characteristics and behavioural patterns observed within varying demographics.

The pivotal advantage of this dataset personalization lies in its potential to significantly bolster recognition accuracy and prediction performance. For instance, the current RHM-HAR-SK dataset, captured by a singular subject, inherently encapsulates only one individual's body dimensions and movement behaviours. However, through the integration of a small yet representative sample from another individual, the dataset's scope can be expanded.

By infusing this new individual's characteristic movements and behaviours into the dataset through generative models, a personalized dataset can be achieved. Consequently, this enriched dataset paves the way for developing a more versatile and adaptable activity recognition model. The model, trained on a personalized dataset, exhibits an increased capacity to discern and categorize activities across a broader range of individuals, transcending the limitations posed by a dataset captured solely by a single subject.

In essence, the process of dataset personalization through the integration of additional individual samples serves as a pioneering strategy to elevate the versatility and accuracy of HAR systems within AAL environments. Through this approach, the pipeline not only becomes more inclusive but also fosters more robust and refined activity recognition models, marking a pivotal step toward personalized and adaptive AAL solutions. On the other hand, combining multiple personalized datasets could offer a more comprehensive and generalized representation of diverse movement behaviours and body dimensions in a dataset.

6.3.2 Enabling Multi-Person Activity Recognition

The extension of the scalable HAR pipeline to encompass multi-person recognition represents a significant stride toward a more comprehensive understanding of activities within AAL environments. By empowering the pipeline to support multi-person HAR, the system transcends the conventional bounds of individual activity tracking and ventures into a domain where the activities and interactions of multiple individuals within a group can be discerned and analyzed.

This evolution not only facilitates the identification and tracking of individual activities within a group setting but also opens new vistas for analyzing the intricate dynamics and interactions among multiple persons. By enabling the pipeline to decipher and categorize activities performed by various individuals simultaneously, a deeper understanding of group behaviours and interactions within AAL settings can be garnered.

The ability to discern individual activities within a group environment is invaluable in the context of AAL scenarios. This advancement empowers the HAR system to identify distinct activities performed by each individual and explore the nuances of their interactions, collaborations, or dependencies within a shared space. For instance, it can discern activities such as social interactions, caregiving moments, or collaborative tasks undertaken by multiple individuals, shedding light on the dynamics and behavioural patterns prevalent in such settings.

Furthermore, this extension facilitates a more holistic approach to activity recognition within AAL environments. The system's capability to track and interpret activities across multiple persons offers a comprehensive perspective, enabling caregivers, healthcare professionals, or researchers to gain deeper insights into the collective behaviours and interactions that unfold within these settings.

6.3.3 Deploying Continual HAR

The integration of a continual model onto the streamlined pipeline represents a strategic move towards enhancing the pipeline's adaptability in real-time scenarios. Efficiency remains paramount, especially when the HAR model interfaces with robotic systems, particularly in real-time tasks. The streamlined pipeline, characterized by its efficiency and scalability, plays a pivotal role in ensuring the seamless integration and operation of the Continual HAR model.

Efficiency becomes a cornerstone attribute, especially in scenarios where robots are dynamically engaged in real-time tasks alongside individuals within AAL settings. The integration of

the Continual HAR model onto the streamlined pipeline ensures that the system operates with optimal efficiency, enabling swift recognition, analysis, and response to human activities without compromising the system's real-time performance.

On the other hand, deploying Continual HAR within the complex landscape of AAL scenarios involves navigating various challenges, especially in synchronizing temporal information sourced from multiple cameras. This synchronization intricately intertwines various streams of data originating from disparate viewpoints, necessitating robust mechanisms to harmonize and align temporal sequences across these diverse sources.

Adapting the HAR model for real-time anomaly detection is another pivotal facet of deploying Continual HAR within AAL settings. Anomalies, deviations from established behavioural patterns, or sudden changes in routine activities could signify potential risks or health concerns, highlighting the importance of real-time detection and response. Integrating anomaly detection mechanisms into the HAR pipeline enables the system to swiftly identify and flag unusual or unexpected activities, prompting timely intervention or alerts.

Furthermore, enabling the system for online learning from observed human behaviour forms a cornerstone of Continuous HAR deployment. Reinforcement learning models stand out as instrumental tools in facilitating adaptive learning from human activity routines. These models are adept at learning from interaction and feedback, allowing the HAR system to dynamically adapt and refine its recognition capabilities based on observed human behaviour over time.

The integration of reinforcement learning models empowers the system to autonomously adapt to variations or changes in human behaviour patterns within AAL environments. By continuously learning and updating its recognition algorithms based on real-time observations, the system can fine-tune its activity recognition capabilities, enhancing adaptability and anomaly detection while ensuring responsiveness to dynamic changes in the environment.

6.3.4 HRI Study: Exploring the Efficacy of HAR Methods

As a last suggestion for future work, an exploration into the effectiveness of Human Activity Recognition (HAR) methods within the context of Human-Robot Interaction (HRI) is suggested. This study aims to create scenarios that simulate real-life situations, allowing for a nuanced analysis of human satisfaction with the activities recognized by the robot.

In this investigation, it is proposed to not only assess the accuracy of HAR methods but also to understand the subjective experience of users. Human feedback on the robot's responsiveness to their behaviour will be crucial in gauging the overall success of the interaction. Examining the qualitative aspects of human-robot engagement aims to provide insights into the nuanced dynamics of HRI beyond mere technical performance metrics.

Simultaneously, the effectiveness of adaptive behaviour methods, particularly those grounded in reinforcement learning (RL) or recommender systems, will be explored. In a parallel scenario, an evaluation will be conducted to assess how these adaptive techniques enhance the overall user experience, taking into account factors such as responsiveness, proactiveness, and adaptability.

This comprehensive HRI study aims to advance the technical aspects of HAR and adaptive behaviour and bridge the gap between technological functionality and the human experience, paving the way for more intuitive and user-friendly robotic interactions.

6.3.5 Ethical and Privacy Considerations

Incorporating multi-view skeleton-based human activity recognition into research and practical applications inevitably raises ethical concerns and privacy issues. One primary concern revolves around data privacy and consent. While skeleton data may not capture identifiable facial features, it still raises privacy concerns regarding individuals' bodily movements and actions. Multi-view HAR often involves the collection and analysis of video data, which may include sensitive information about individuals, such as their movements, behaviors, and interactions.

Ensuring that individuals are adequately informed about the data collection process, the purposes for which their data will be used, and their rights regarding data privacy is important. Additionally, researchers must implement data anonymization and protection measures to minimize the risk of unauthorized access or misuse of personal information. Furthermore, there are ethical implications regarding the potential impact of HAR technologies on individual freedom. Deploying HAR systems in public spaces or private environments raises questions about surveillance, consent, and the balance between security and privacy. It is essential to engage in transparent discussions and ethical assessments to address these concerns and develop guidelines for responsible research and deployment of multi-view skeleton-based HAR systems that uphold ethical principles, respect individual rights, and promote societal well-being.

CHAPTER 7

APPENDIX A

This chapter presents the results of the missed poses analysis on the RHM-HAR-SK dataset and code snippets of different models. Each section is presented in a section with details in each subsection.

7.1 Missed Posses analysis

The analysis evaluated two pose estimation methods: HRNet and YOLOv7. A total of 28 bar charts, 14 for each method, display the performance of these methods across different activity classes. Further details on the analysis methodology are outlined in Chapter 3. The bar graphs provide a deep insight into the effects of the camera view, action types and the pose extraction methods on the quality of the extracted poses.

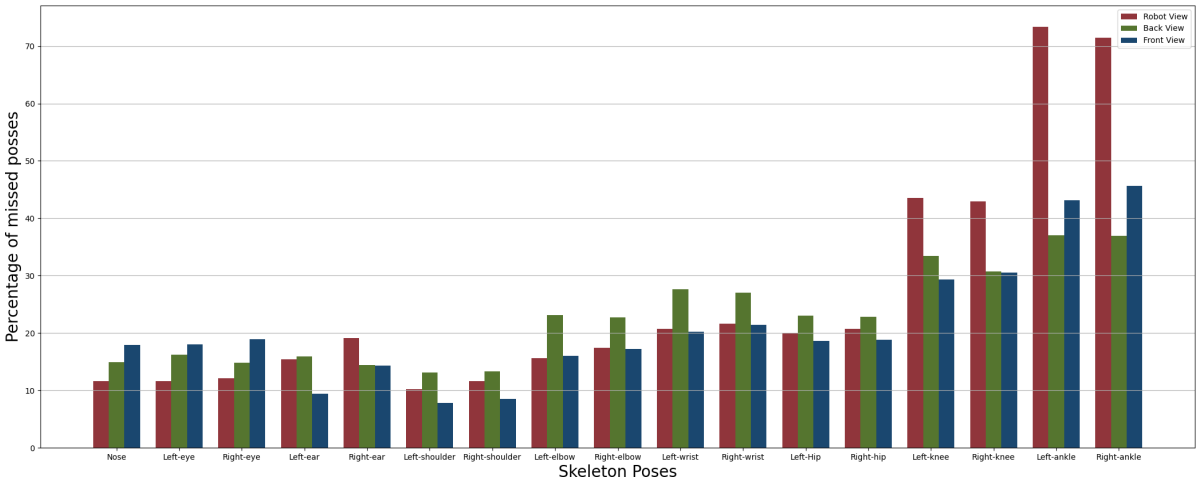


Figure 7.1: Percentage of frames with missed skeleton poses of "Bending" action across various views extracted via HrNet.

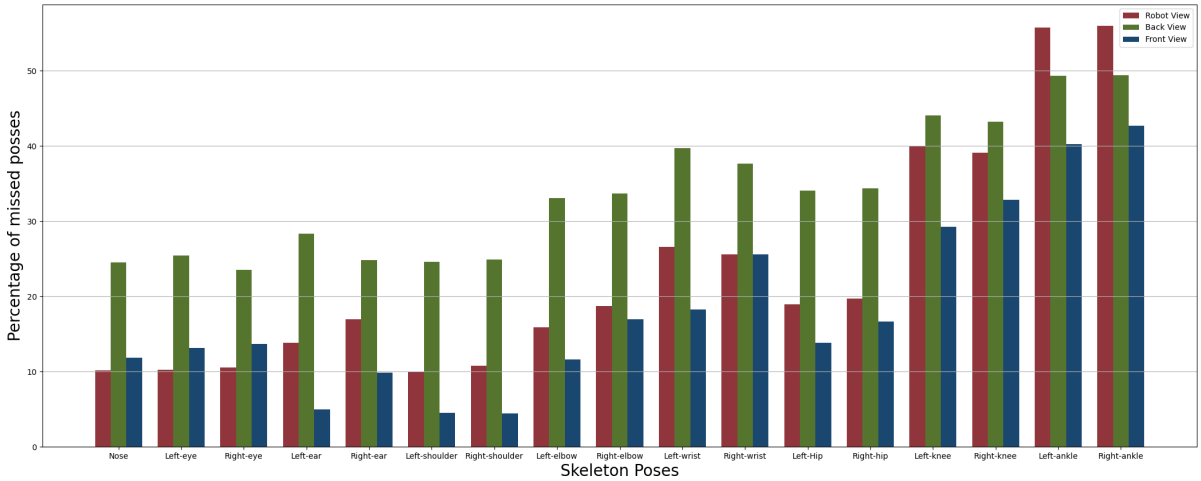


Figure 7.2: Percentage of frames with missed skeleton poses of "Sitting Down" action across various views extracted via HrNet.

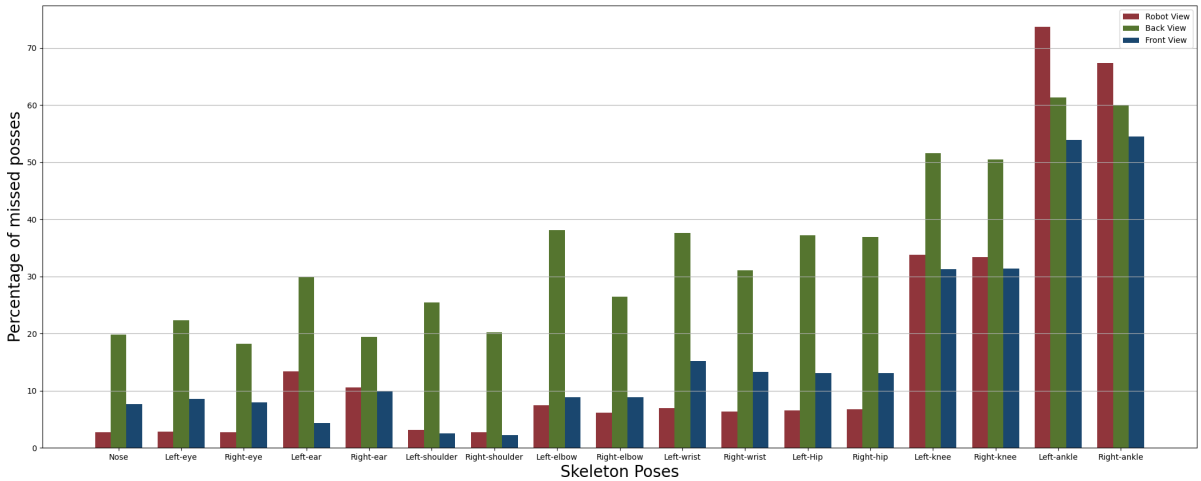


Figure 7.3: Percentage of frames with missed skeleton poses of "Closing Can" action across various views extracted via HrNet.

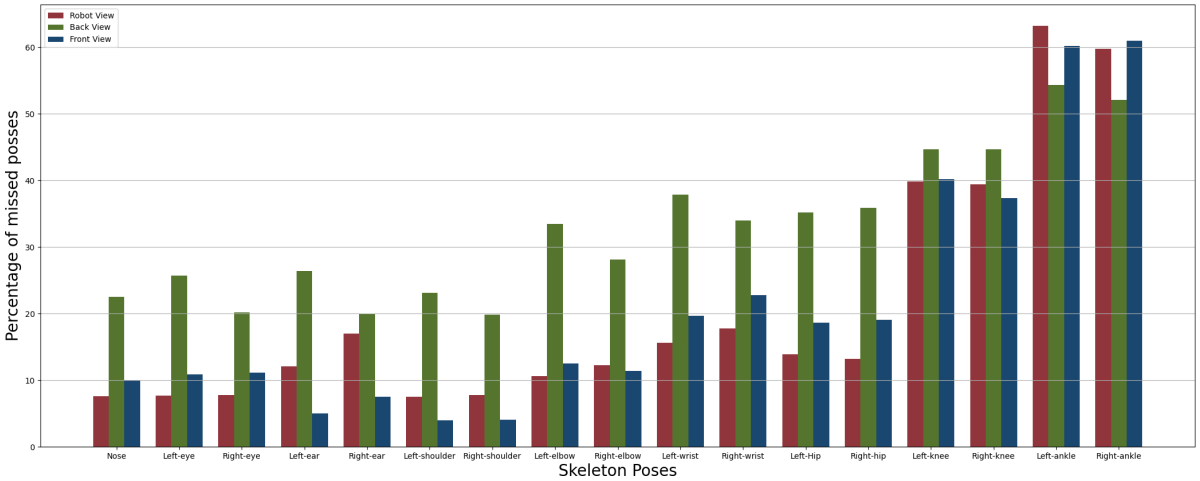


Figure 7.4: Percentage of frames with missed skeleton poses of "Reaching" action across various views extracted via HrNet.

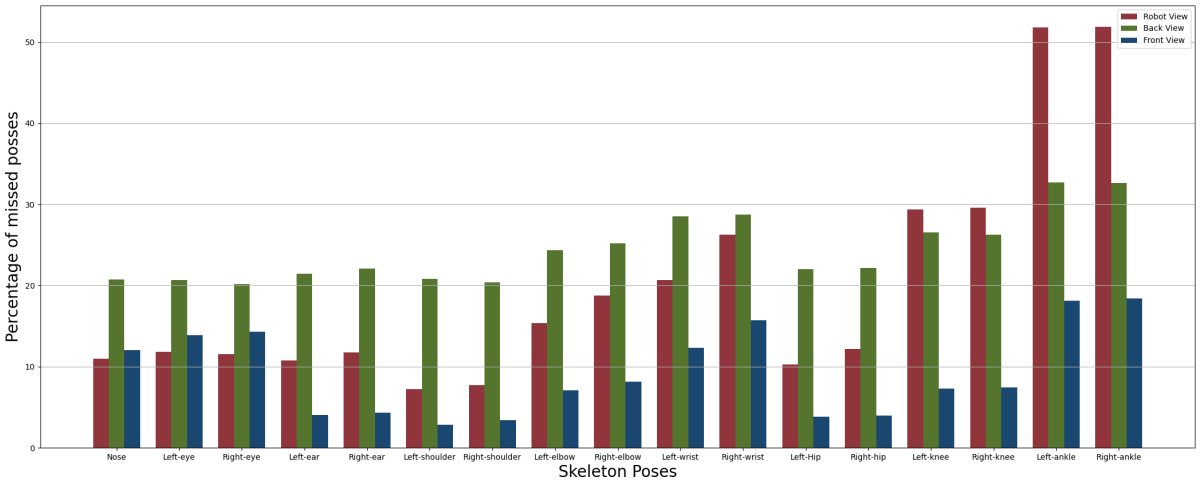


Figure 7.5: Percentage of frames with missed skeleton poses of "Walking" action across various views extracted via HrNet.

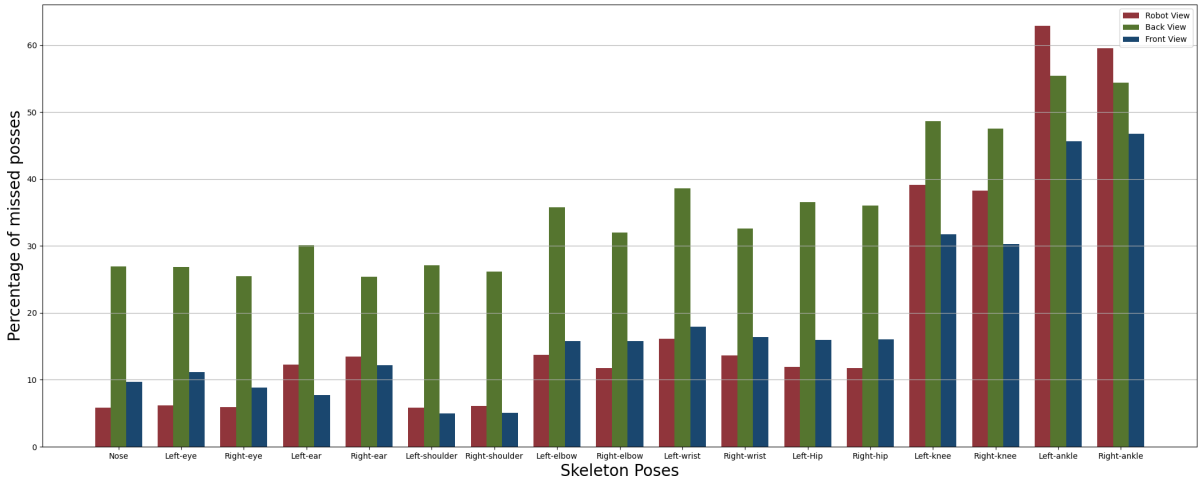


Figure 7.6: Percentage of frames with missed skeleton poses of "Drinking" action across various views extracted via HrNet.

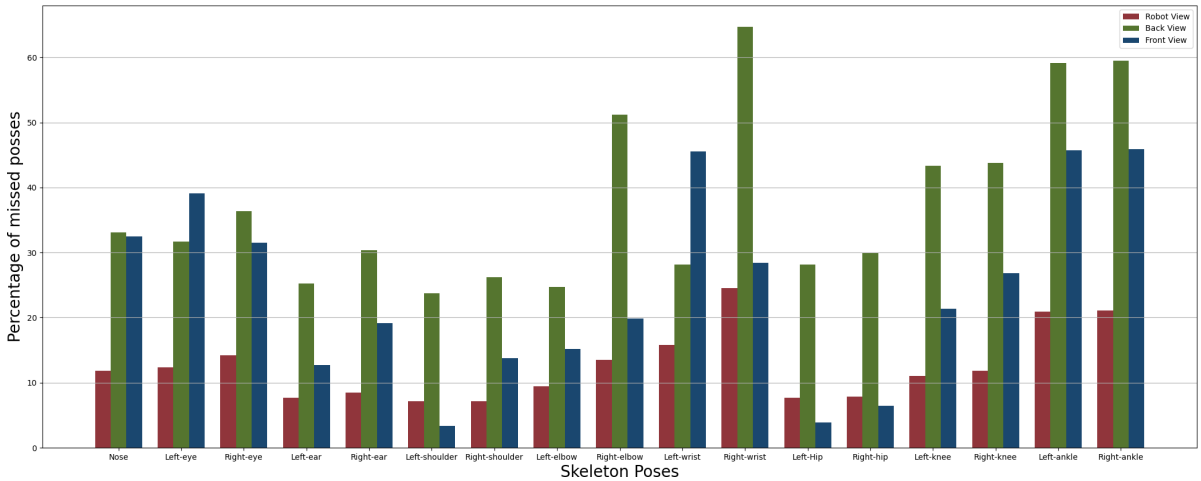


Figure 7.7: Percentage of frames with missed skeleton poses of "Stairs Climbing Up" action across various views extracted via HrNet.

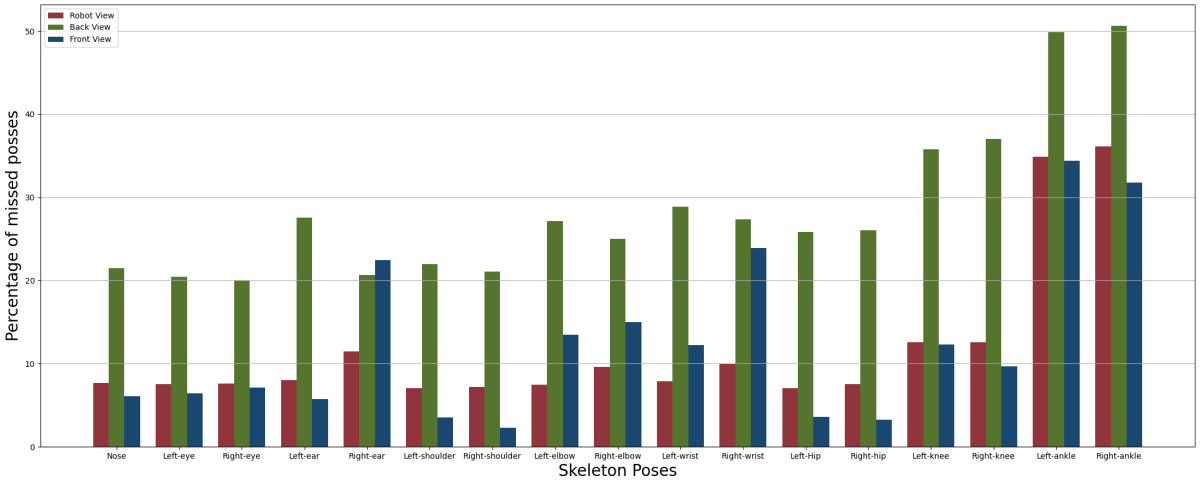


Figure 7.8: Percentage of frames with missed skeleton poses of "Stairs Climbing Down" action across various views extracted via HrNet.

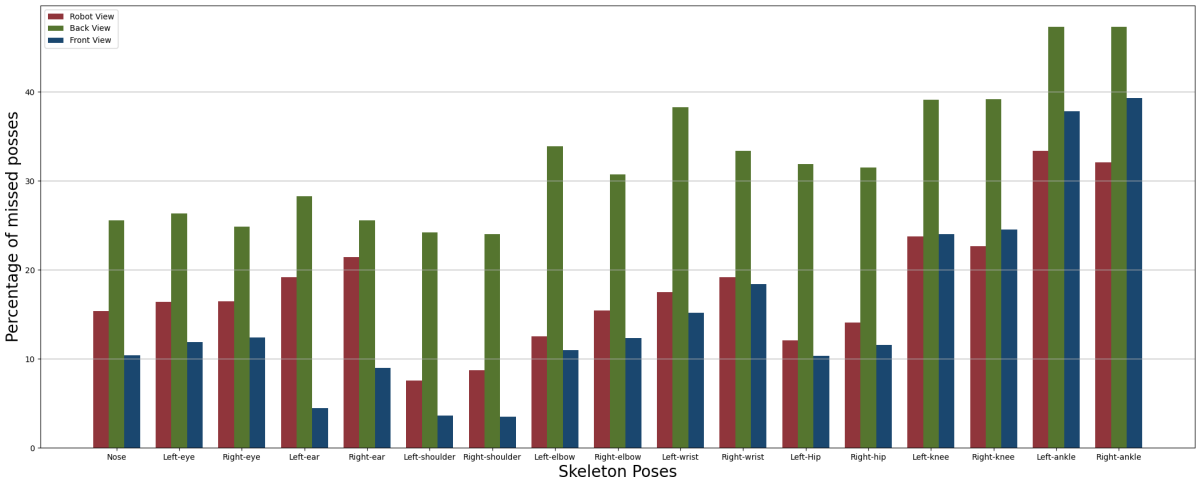


Figure 7.9: Percentage of frames with missed skeleton poses of "Standing Up" action across various views extracted via HrNet.

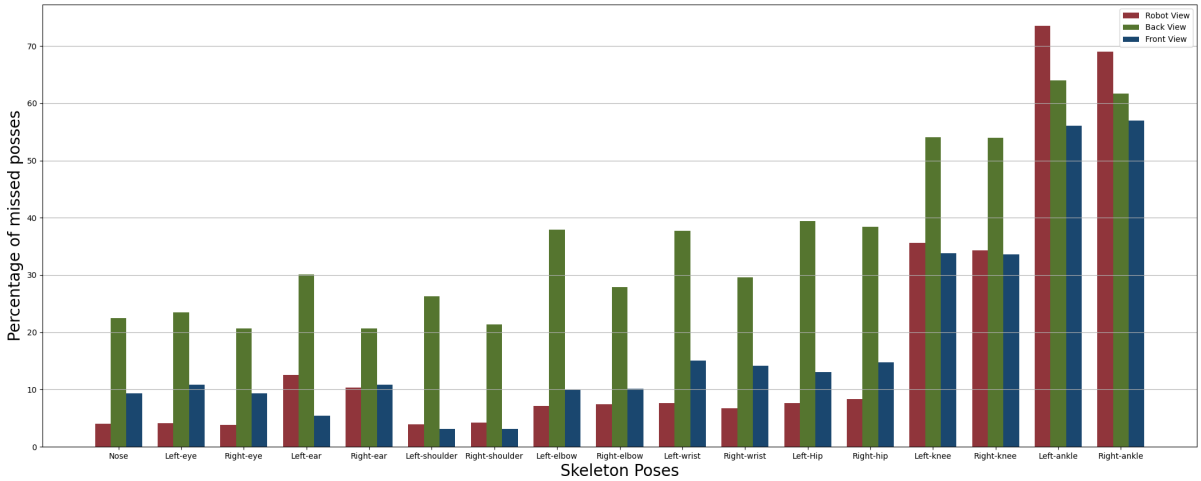


Figure 7.10: Percentage of frames with missed skeleton poses of "Opening Can" action across various views extracted via HrNet.

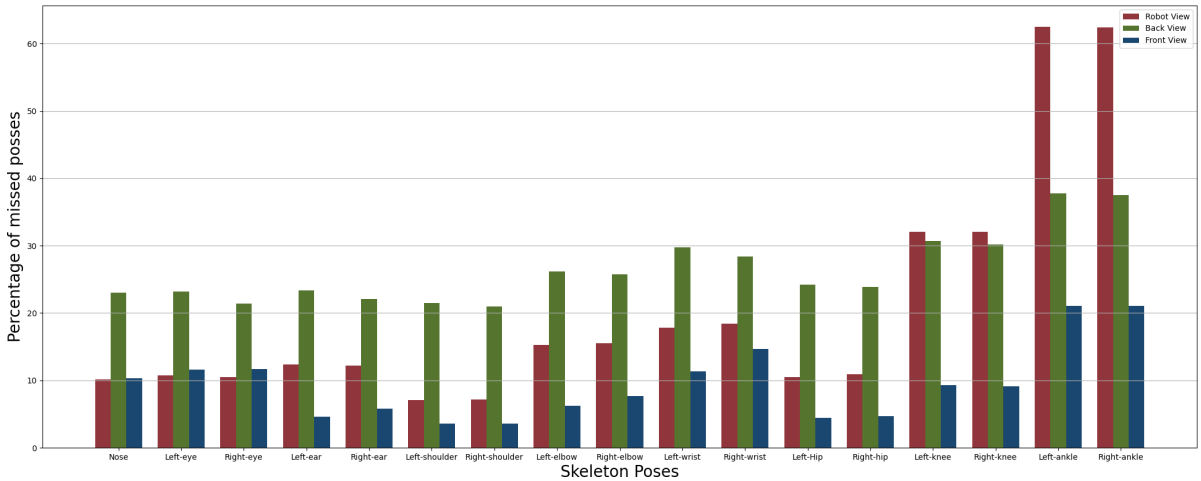


Figure 7.11: Percentage of frames with missed skeleton poses of "Carrying Object" action across various views extracted via HrNet.

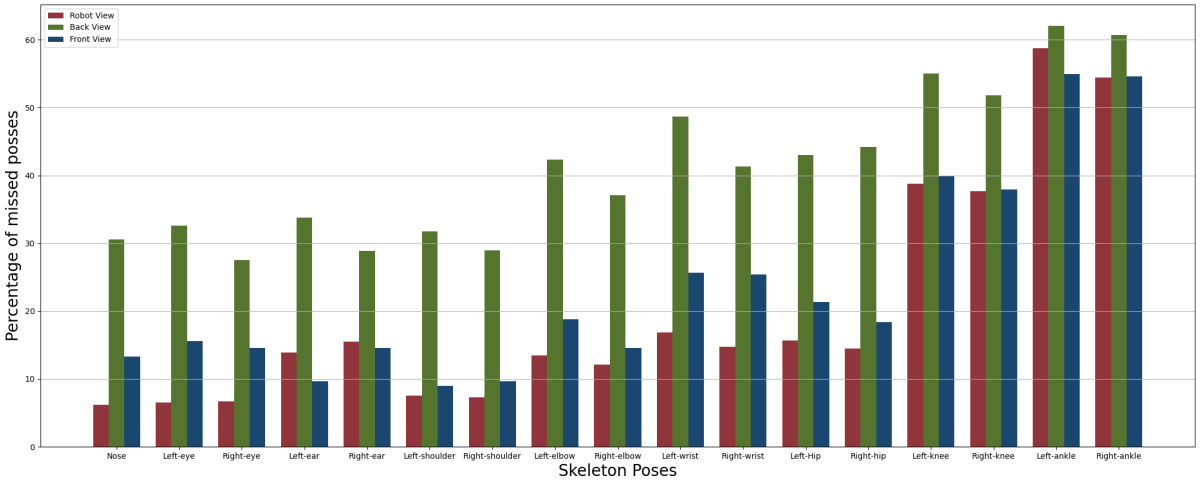


Figure 7.12: Percentage of frames with missed skeleton poses of "Cleaning" action across various views extracted via HrNet.

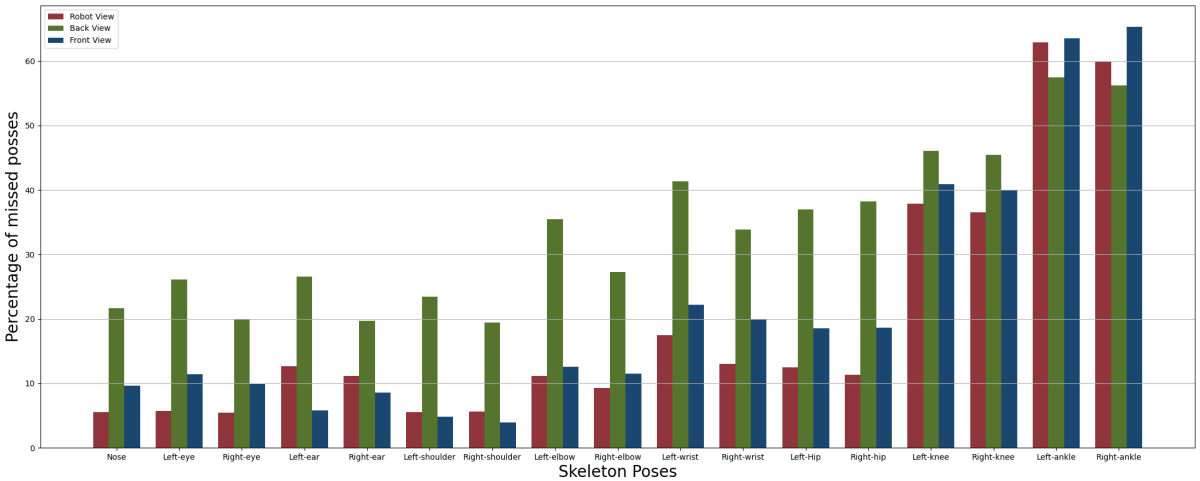


Figure 7.13: Percentage of frames with missed skeleton poses of "Putting Down Object" action across various views extracted via HrNet.

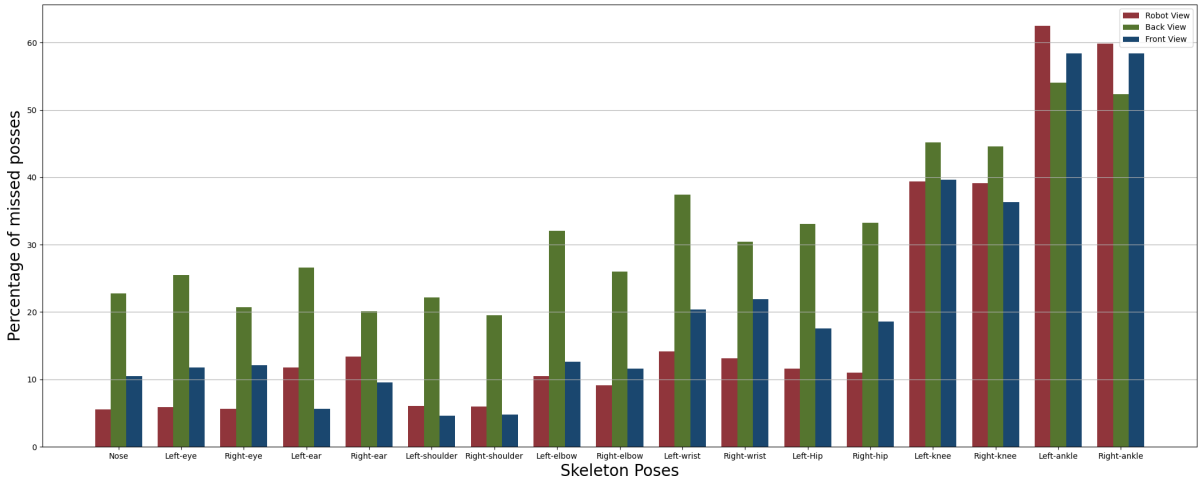


Figure 7.14: Percentage of frames with missed skeleton poses of "Lifting Object" action across various views extracted via HrNet.

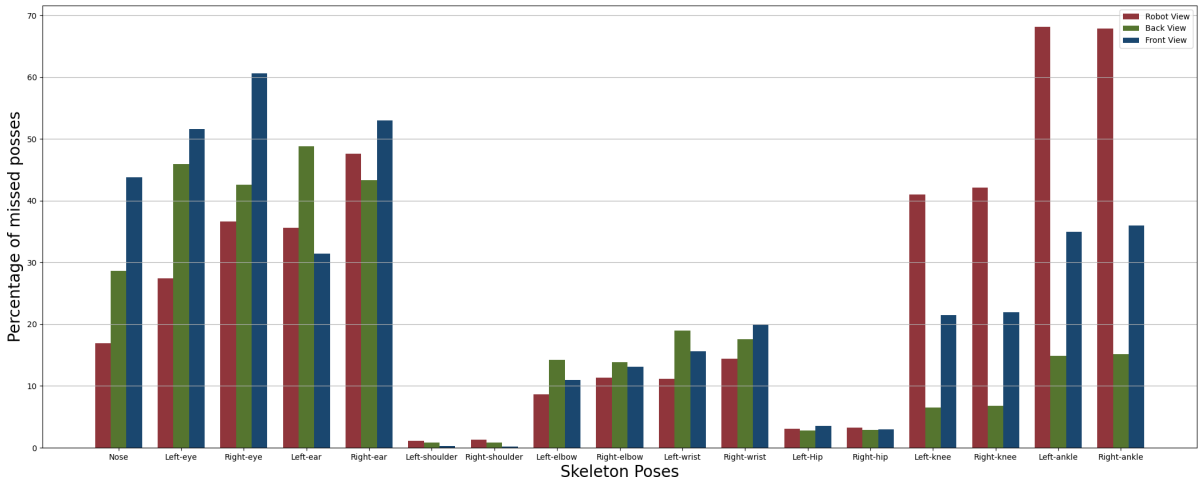


Figure 7.15: Percentage of frames with missed skeleton poses of "Bending" action across various views extracted via YOLOv7.

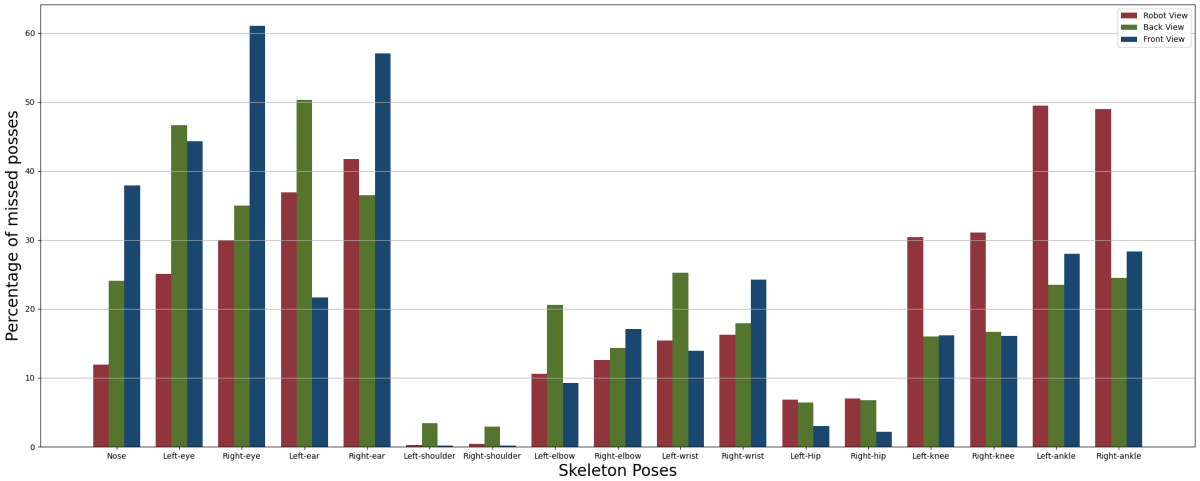


Figure 7.16: Percentage of frames with missed skeleton poses of "Sitting Down" action across various views extracted via YOLOv7.

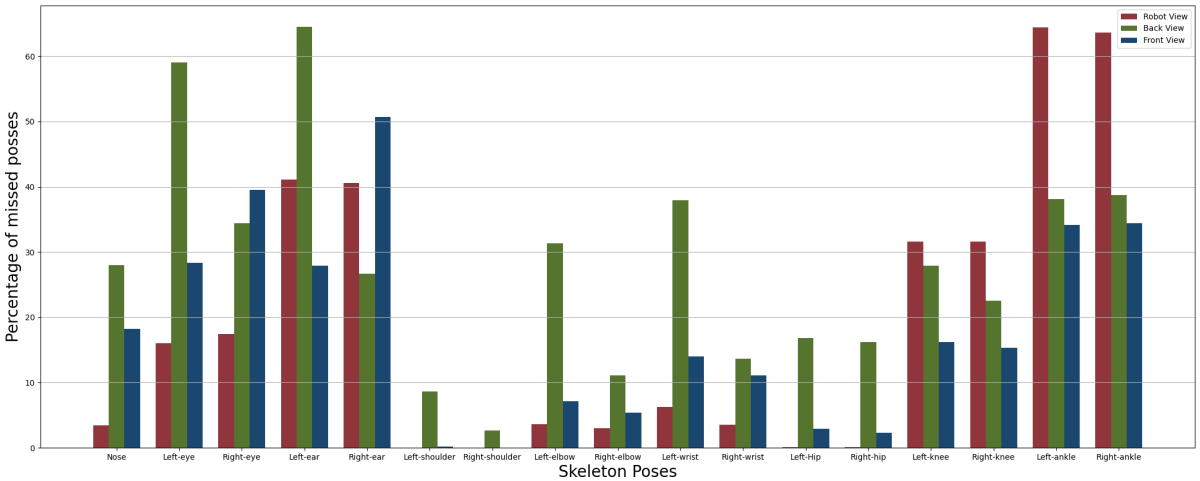


Figure 7.17: Percentage of frames with missed skeleton poses of "Closing Can" action across various views extracted via YOLOv7.

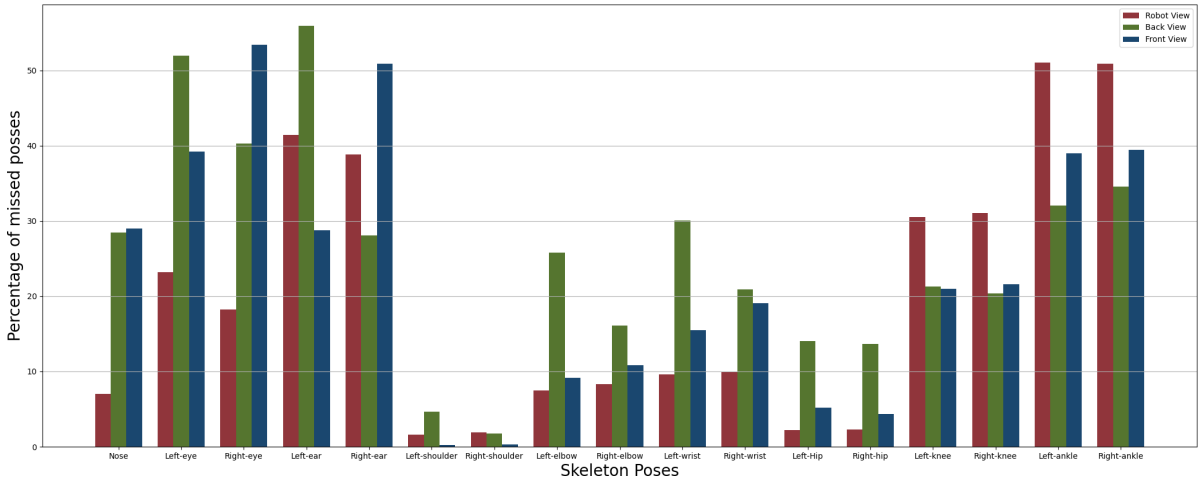


Figure 7.18: Percentage of frames with missed skeleton poses of "Reaching" action across various views extracted via YOLOv7.

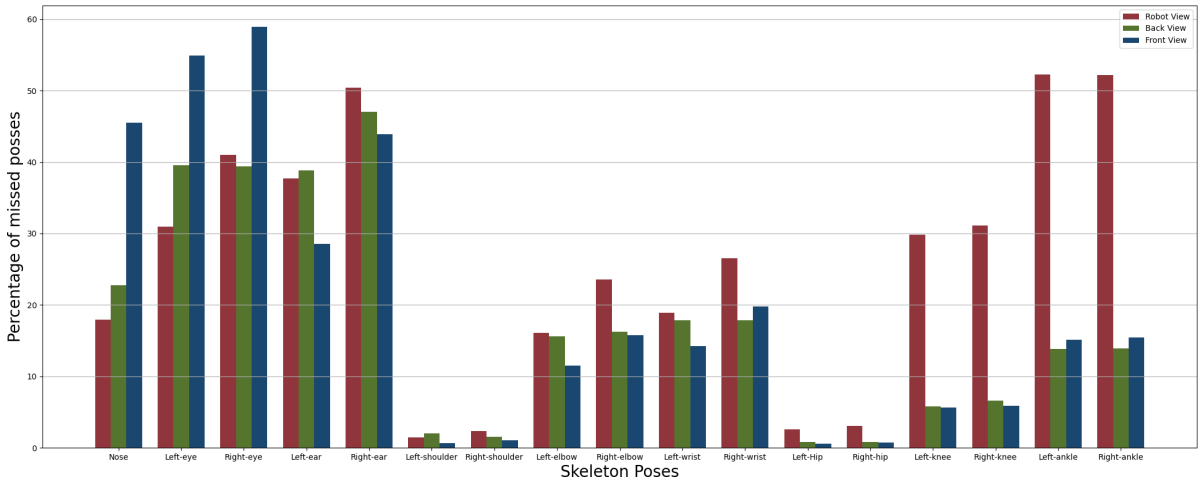


Figure 7.19: Percentage of frames with missed skeleton poses of "Walking" action across various views extracted via YOLOv7.

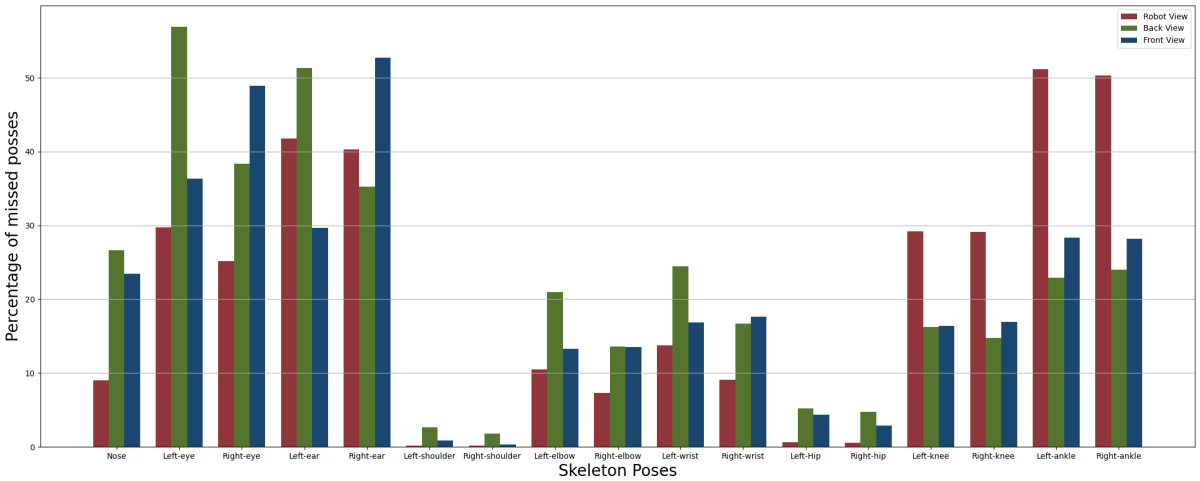


Figure 7.20: Percentage of frames with missed skeleton poses of "Drinking" action across various views extracted via YOLOv7.

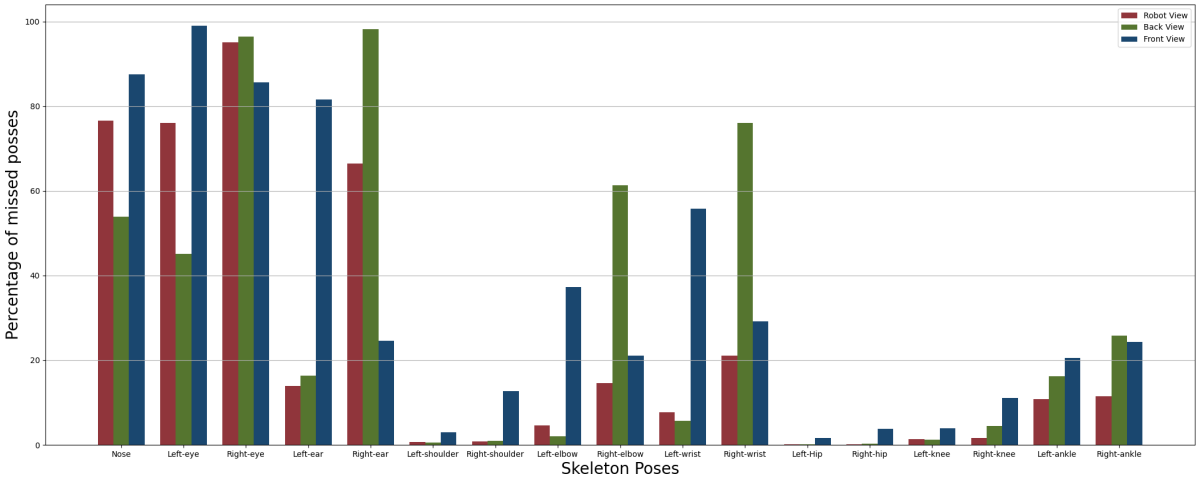


Figure 7.21: Percentage of frames with missed skeleton poses of "Stairs Climbing Up" action across various views extracted via YOLOv7.

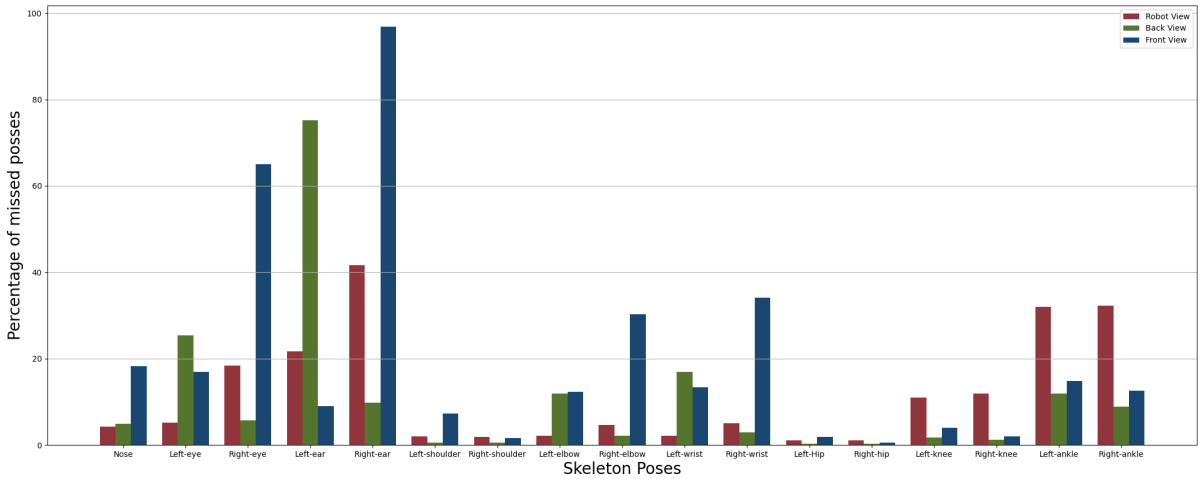


Figure 7.22: Percentage of frames with missed skeleton poses of "Stairs Climbing Down" action across various views extracted via YOLOv7.

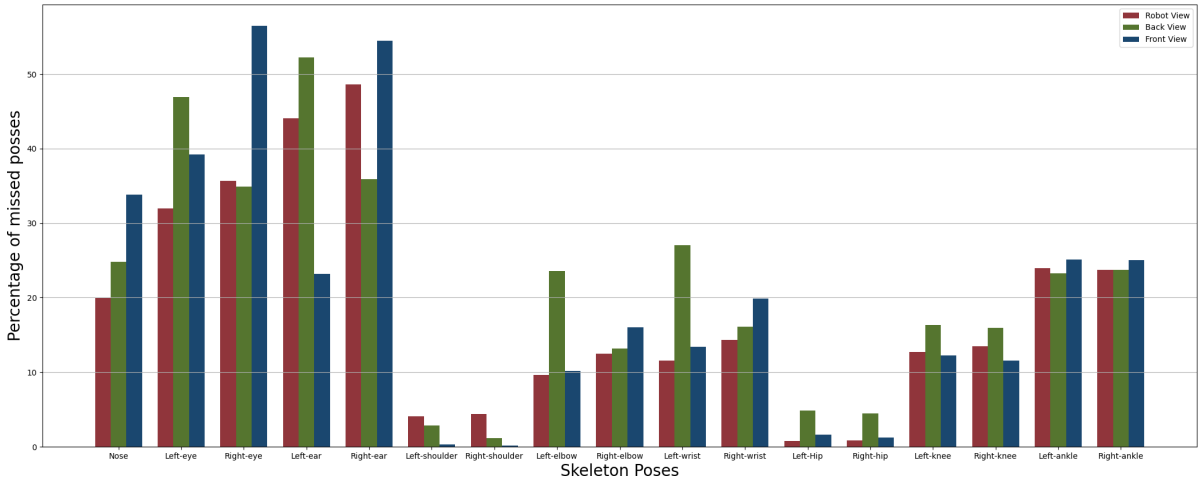


Figure 7.23: Percentage of frames with missed skeleton poses of "Standing Up" action across various views extracted via YOLOv7.

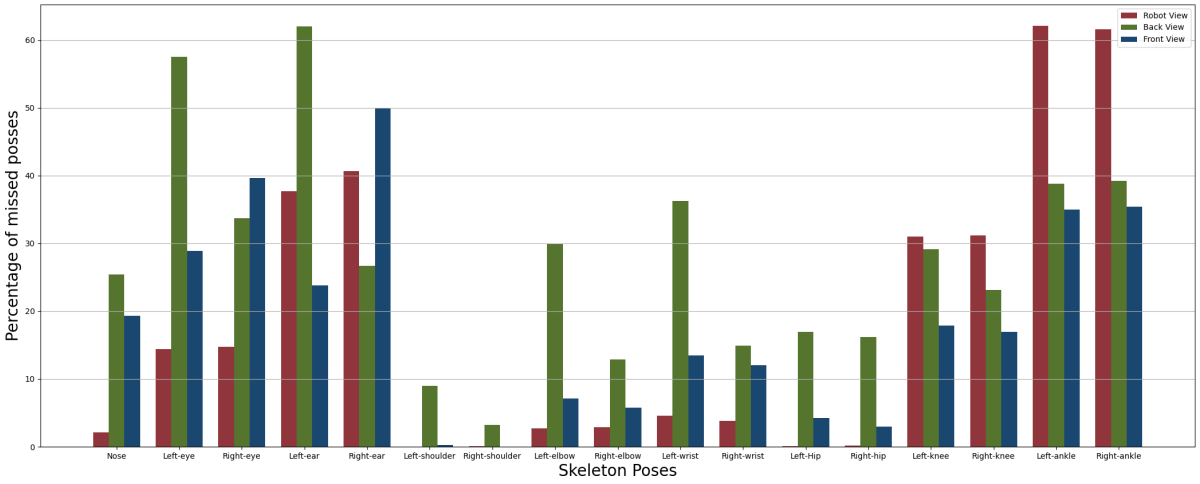


Figure 7.24: Percentage of frames with missed skeleton poses of "Openning Can" action across various views extracted via YOLOv7.

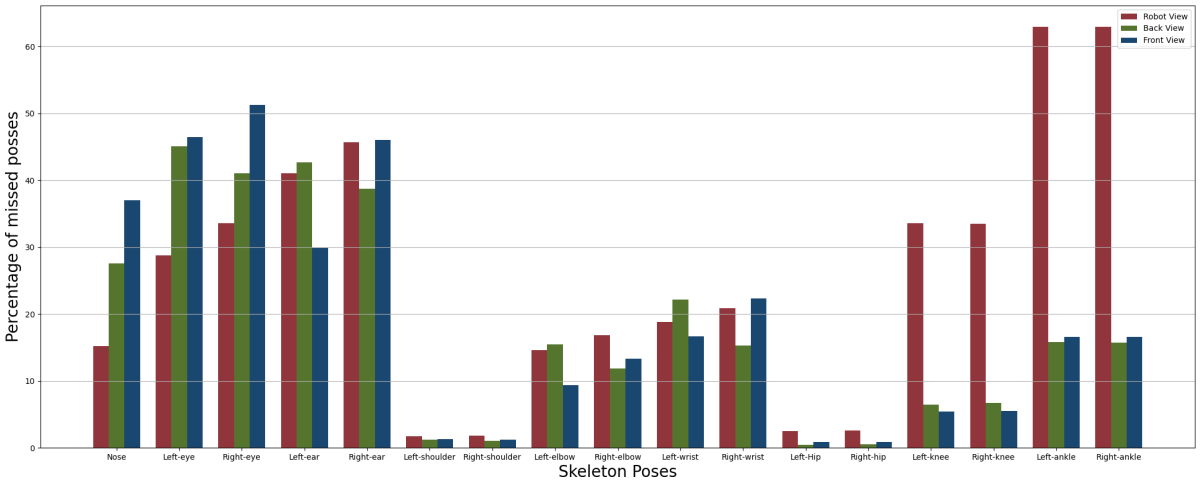


Figure 7.25: Percentage of frames with missed skeleton poses of "Carrying Object" action across various views extracted via YOLOv7.

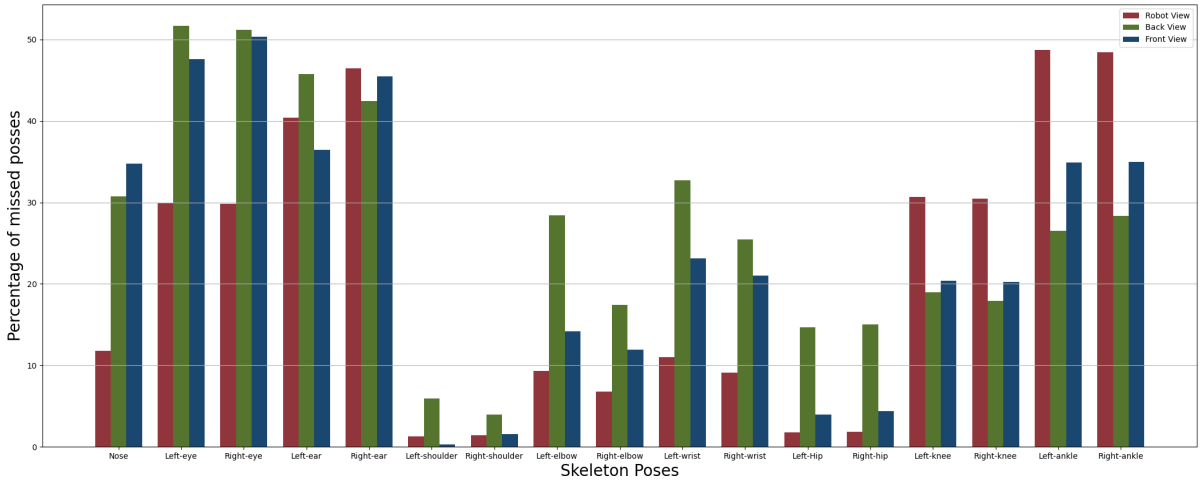


Figure 7.26: Percentage of frames with missed skeleton poses of "Cleaning" action across various views extracted via YOLOv7.

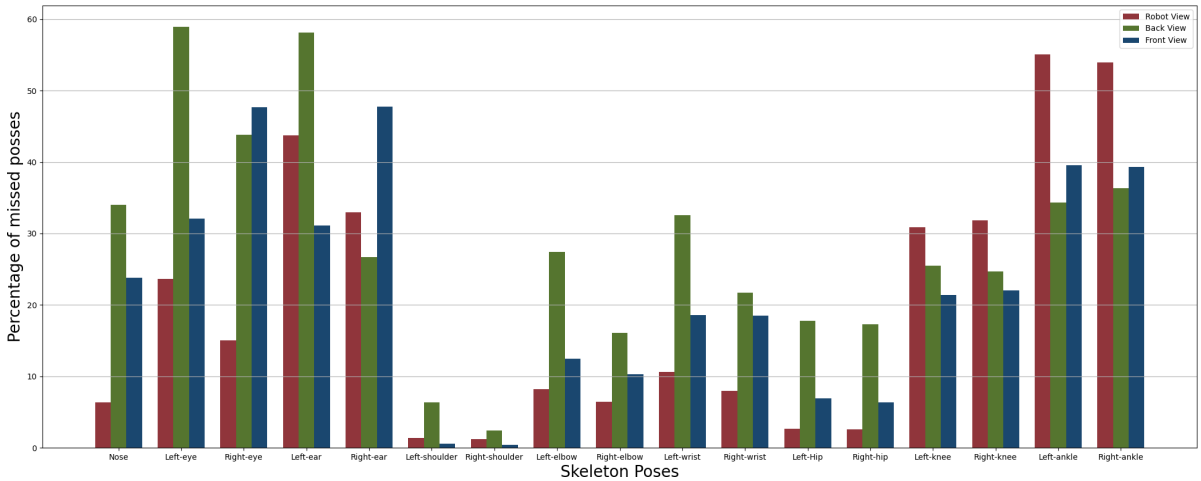


Figure 7.27: Percentage of frames with missed skeleton poses of "Putting Down Object" action across various views extracted via YOLOv7.

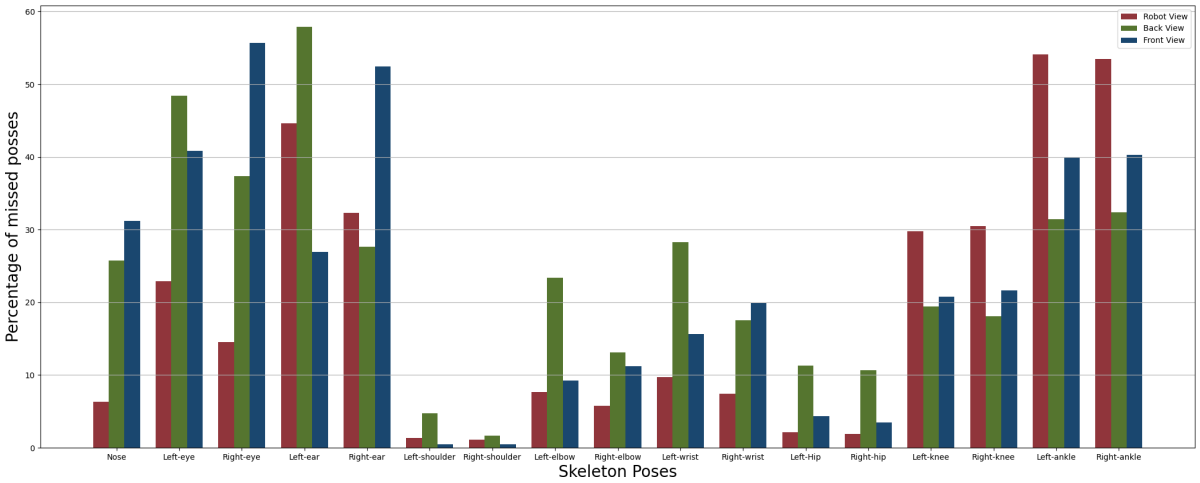


Figure 7.28: Percentage of frames with missed skeleton poses of "Lifting Object" action across various views extracted via YOLOv7.

7.2 Code Snippet

In the following, the main code structure for model training, validation and testing is presented. The below Python code snippet (Code 7.1) demonstrates the main structure of a machine learning model training process (it also shows the low-level fusion method (LW in Sec. 5.1.3.1)). It includes data preprocessing steps such as stratified k-fold cross-validation and data splitting into training, validation, and test sets. The training loop iterates over epochs, optimizing the model parameters using gradient descent. Additionally, the code evaluates model performance on the validation and test sets, calculating metrics such as loss, accuracy, precision, recall, and F1-score. Data loaders are utilized to efficiently handle batches of input data during training and evaluation.

Listing 7.1: The main structure/LW multi-vew Model

```
skf = StratifiedKFold(n_splits=5, shuffle=True,  random_state=42)
X_train_val, X_test, y_train_val, y_test = train_test_split(data, labels,
                                    test_size=0.20, random_state=42)
for train_index, val_index in skf.split(X_train_val, y_train_val):
    X_train, X_val = X_train_val[train_index], X_train_val[val_index]
    y_train, y_val = y_train_val[train_index], y_train_val[val_index]

    train_loader = DataLoader(X_train, batch_size=batch_size, shuffle=True)
    val_loader = DataLoader(y_train, batch_size=batch_size, num_workers=4)

    for epoch in range(num_epochs):
        model.train()
        train_loss = 0.0
        for inputs, targets in train_loader:
            inputs = inputs.to(device)  # Move inputs to GPU
            targets = targets.to(device)  # Move targets to GPU
            optimizer.zero_grad()
            outputs = model(inputs)
            loss = criterion(outputs, targets)
            loss.backward()
            optimizer.step()
            train_loss += loss.item()

            train_loss /= len(train_loader)
            train_losses.append(train_loss)

        model.eval()
        val_loss = 0.0
        with torch.no_grad():
            val_predictions = []
            val_targets = []
            for inputs, val_tags in val_loader:
                inputs = inputs.to(device)  # Move inputs to GPU
                outputs = model(inputs)
                outputs = outputs.to('cpu')
```

```
36                 v_loss = criterion(outputs, val_tags)
37                 predictions = torch.argmax(outputs, dim=1)
38                 val_predictions.extend(predictions.tolist())
39                 val_targets.extend(val_tags.tolist())
40                 val_accuracy = accuracy_score(val_tags, predictions)
41
42         val_loss = v_loss / len(val_loader)
43         val_losses.append(val_loss)
44
45     val_accuracy = accuracy_score(val_targets, val_predictions)
46     val_precision = precision_score(val_targets, val_predictions)
47     val_recall = recall_score(val_targets, val_predictions)
48     val_f1_score = f1_score(val_targets, val_predictions)
49
50     accuracies.append(val_accuracy)
51     precisions.append(val_precision)
52     recalls.append(val_recall)
53     f1_scores.append(val_f1_score)
54
55     # Evaluate the model on the test set
56     test_loader = DataLoader(X_test, batch_size=batch_size, num_workers=4)
57     model.eval()
58     test_loss = 0.0
59     test_predictions = []
60     test_targets = []
61     with torch.no_grad():
62         for inputs, targets in test_loader:
63             inputs = inputs.to(device)  # Move inputs to GPU
64             targets = targets.to(device)  # Move targets to GPU
65             outputs = model(inputs)
66             loss = criterion(outputs, targets)
67             test_loss += loss.item()
68             predictions = torch.argmax(outputs, dim=1)
69             test_predictions.extend(predictions.tolist())
70             test_targets.extend(targets.tolist())
71
72     cm_test = confusion_matrix(test_targets, test_predictions)
73
74     test_loss /= len(test_loader)*3
75     test_accuracy = accuracy_score(test_targets, test_predictions)
76     test_precision = precision_score(test_targets, test_predictions)
77     test_recall = recall_score(test_targets, test_predictions)
78     test_f1_score = f1_score(test_targets, test_predictions)
```

To show how other multi-view architectures have been deployed, lines 15 to 18 in the training, lines 33 to 25 for the validation, and lines 63 to 65 for different architecture which is shown in the following sections.

7.2.1 The MD architecture

To show the code snippets of the Mi-level data combination method (MD) which have been described in Sec.5.1.3.2 the below Code 7.2 is provided. This code which includes a for loop should replace line 17 in the main code. The for loop refers to each input going through the training phase separately and the model weight updates with each view. The same should apply to the validation and test section by replacing the code in lines 32 and 63 of the main code, respectively.

Listing 7.2: MD model

```
...
...
for i in range(3):
    input = inputs[:, i, :, :].unsqueeze(1)
    outputs = model(input)
    ...
    ...
```

7.2.2 The HG architecture

To show the code snippets of the high-level data combination method (hg) which have been described in Sec.5.1.3.3 the below Code 7.3 is provided. The code should replace the same lines that the MD replaced, lines 17, 32 and 64 of the main code for train, validation and testing respectively. It shows that each camera view feeds the single model and the output is the average of those individuals' view predictions probability.

Listing 7.3: HG model

```
...
...
input1 = inputs[:, 0, :, :].unsqueeze(1)
input2 = inputs[:, 1, :, :].unsqueeze(1)
input3 = inputs[:, 2, :, :].unsqueeze(1)

ouput1 = model(input1)
ouput2 = model(input2)
ouput3 = model(input3)

outputs = (ouput1 + ouput2 + ouput3) / 3
...
...
```

BIBLIOGRAPHY

[1] M. H. Bamorovat Abadi, M. R. Shahabian Alashti, P. Holthaus, C. Menon, and F. Amirabdollahian, "RHM: Robot House Multi-view Human Activity Recognition Dataset," in *The Sixteenth International Conference on Advances in Computer-Human Interactions (ACHI 2023)*. Venice, Italy: IARIA, 2023, pp. 159–166. [Online]. Available: https://www.thinkmind.org/index.php?view=article&articleid=achi_2023_4_160_20077

[2] F. Amirabdollahian, R. o. d. Akker, S. Bedaf, R. Bormann, H. Draper, V. Evers, J. G. Pérez, G. J. Gelderblom, C. G. Ruiz, D. Hewson, N. Hu, B. Kröse, H. Lehmann, P. Marti, H. Michel, V. Prevot, U. Reiser, J. Saunders, T. Sorell, J. Stienstra, D. S. Syrdal, M. L. Walters, and K. Dautenhahn, "Assistive technology design and development for acceptable robotics companions for ageing years," *Paladyn, Journal of Behavioral Robotics*, vol. 4, 2013.

[3] S. Blackman, C. Matlo, C. Bobrovitskiy, A. Waldoch, M. Fang, P. Jackson, A. Mihailidis, L. Nygård, A. Astell, and A. Sixsmith, "Ambient assisted living technologies for aging well: a scoping review," *Journal of Intelligent Systems*, vol. 25, pp. 55–69, 2016.

[4] T. Kleinberger, M. Becker, E. Ras, A. Holzinger, and P. Müller, "Ambient intelligence in assisted living: enable elderly people to handle future interfaces," *Universal Access in Human-Computer Interaction. Ambient Interaction*, pp. 103–112, 2007.

[5] K. Denecke, "What characterizes safety of ambient assisted living technologies?" *Studies in Health Technology and Informatics*, 2021.

[6] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, pp. 579–590, 2013.

[7] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, *et al.*, "The grand challenges of science robotics," *Science robotics*, vol. 3, no. 14, p. eaar7650, 2018.

[8] G. Castellano, R. Aylett, K. Dautenhahn, A. Paiva, P. W. McOwan, and S. Ho, "Long-term affect sensitive and socially interactive companions," in *Proceedings of the 4th International Workshop on Human-Computer Conversation*, 2008, pp. 1–5.

[9] C. Jaschinski, S. B. Allouch, O. Peters, R. Cachucho, and J. v. Dijk, "Acceptance of technologies for aging in place: a conceptual model," *Journal of Medical Internet Research*, vol. 23, p. e22613, 2021.

[10] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, pp. 30 509–30 555, 2020.

[11] M. A. Huber, M. Rickert, A. Knoll, T. Brandt, and S. Glasauer, "Human-robot interaction in handing-over tasks," *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 2008.

[12] K. E. Adolph, W. G. Cole, M. Komati, J. S. Garciaguirre, D. Badaly, J. M. Lingeman, G. L. Chan, and R. B. Sotsky, "How do you learn to walk? thousands of steps and dozens of falls per day," *Psychological science*, vol. 23, no. 11, pp. 1387–1394, 2012.

[13] C. Schneider, B. Trukeschitz, and H. Rieser, "Measuring the use of the active and assisted living prototype carimo for home care service users: evaluation framework and results," *Applied Sciences*, vol. 10, p. 38, 2019.

[14] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis," *ACM Computing Surveys*, vol. 43, pp. 1–43, 2011.

[15] M. J. Santofimia, J. M. d. Rincon, and J. Nebel, "Episodic reasoning for vision-based human action recognition," *The Scientific World Journal*, vol. 2014, pp. 1–18, 2014.

[16] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, pp. 976–990, 2010.

[17] S. Mekruksavanich and A. Jitpattanakul, "Lstm networks using smartphone data for sensor-based human activity recognition in smart homes," *Sensors*, vol. 21, p. 1636, 2021.

[18] K. Lai and S. Yanushkevich, "Cnn+rnn depth and skeleton based dynamic hand gesture recognition," *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018.

[19] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14, pp. 11 735–11 759, 2014.

[20] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: a survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[21] M. Ehatisham-ul Haq, A. Javed, M. A. Azam, H. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60 736–60 751, 2019.

[22] T. Ahmad, L. Jin, L. Lin, and G. Tang, "Skeleton-based action recognition using sparse spatio-temporal gcn with edge effective resistance," *Neurocomputing*, vol. 423, pp. 389–398, 2021.

[23] F. A. Van-Horenbeke and A. Peer, "Activity, plan, and goal recognition: a review," *Frontiers in Robotics and AI*, vol. 8, 2021.

[24] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, p. 115, 2016.

[25] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: a survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[26] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks*, vol. 12, no. 8, p. 1550147716665520, 2016.

[27] S. Aleksic, M. Atanasov, J. C. Agius, K. Camilleri, A. Cartolovni, P. Climent-Peerez, S. Colantonio, S. Cristina, V. Despotovic, H. K. Ekenel, *et al.*, "State of the art of audio-and video-based solutions for aal," *arXiv preprint arXiv:2207.01487*, 2022.

[28] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects," *Computers in Biology and Medicine*, p. 106060, 2022.

[29] M. H. Arshad, M. Bilal, and A. Gani, "Human activity recognition: Review, taxonomy and open challenges," *Sensors*, vol. 22, no. 17, p. 6463, 2022.

[30] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, pp. 1–43, 2011.

[31] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *arXiv preprint arXiv:1806.11230*, 2018.

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

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

[34] G. M.V., P. E.A., K. Yu.A., and T. O.Yu., "The category of sense and meaning in psychological discourse," *"Educational bulletin "Consciousness"*, 2023.

[35] P. Bharti, D. De, S. Chellappan, and S. K. Das, "Human: Complex activity recognition with multi-modal multi-positional body sensing," *IEEE Transactions on Mobile Computing*, vol. 18, no. 4, pp. 857–870, 2018.

[36] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE communications surveys & tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[37] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2016.

[38] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern recognition letters*, vol. 119, pp. 3–11, 2019.

[39] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6212–6221.

[40] R. Hou, Y. Li, N. Zhang, Y. Zhou, X. Yang, and Z. Wang, "Shifting perspective to see difference: A novel multi-view method for skeleton based action recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4987–4995.

[41] J. Cui, S. Li, Q. Xia, A. Hao, and H. Qin, "Learning multi-view manifold for single image based modeling," *Computers & Graphics*, vol. 82, pp. 275–285, 2019.

[42] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *European conference on computer vision*. Springer, 2016, pp. 286–301.

[43] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *International conference on artificial neural networks*. Springer, 2011, pp. 44–51.

[44] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[45] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.

[46] Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li, "A comprehensive study of deep video action recognition," *arXiv preprint arXiv:2012.06567*, 2020.

[47] Y. Bengio, Y. LeCun, *et al.*, "Scaling learning algorithms towards ai," *Large-scale kernel machines*, vol. 34, no. 5, pp. 1–41, 2007.

[48] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[49] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-t@inproceedingszolfaghari2018eco, title=Eco: Efficient convolutional network for online video understanding, author=Zolfaghari, Mohammadreza and Singh, Kamaljeet and Brox, Thomas, booktitle=Proceedings of the European conference on computer vision (ECCV), pages=695–712, year=2018 erm recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[50] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[51] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European conference on computer vision*. Springer, 2016, pp. 20–36.

[52] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[53] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[54] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 305–321.

[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[56] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[57] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Asian conference on computer vision*. Springer, 2018, pp. 363–378.

[58] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[59] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 203–213.

[60] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Tiny video networks," *arXiv preprint arXiv:1910.06961*, 2019.

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

[62] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5386–5395.

[63] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[64] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[65] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7103–7112.

[66] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4903–4911.

[67] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[68] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR 2011*. Ieee, 2011, pp. 1297–1304.

[69] W. Zheng, L. Li, Z. Zhang, Y. Huang, and L. Wang, "Relational network for skeleton-based action recognition," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 826–831.

[70] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5344–5352.

[71] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *IEEE Transactions on Image Processing*, vol. 29, pp. 15–28, 2019.

[72] G. Lev, G. Sadeh, B. Klein, and L. Wolf, "Rnn fisher vectors for action recognition and image annotation," in *European Conference on Computer Vision*. Springer, 2016, pp. 833–850.

[73] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[74] M. H. B. Abadi, M. R. S. Alashti, P. Holthaus, C. Menon, and F. Amirabdollahian, "Robot house human activity recognition dataset," *UKRAS21*, 2021.

[75] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[76] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 143–152.

[77] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Computer Vision and Image Understanding*, vol. 208, p. 103219, 2021.

[78] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond, "Unik: A unified framework for real-world skeleton-based action recognition," *arXiv preprint arXiv:2107.08580*, 2021.

[79] F. Shi, C. Lee, L. Qiu, Y. Zhao, T. Shen, S. Muralidhar, T. Han, S.-C. Zhu, and V. Narayanan, "Star: Sparse transformer-based action recognition," *arXiv preprint arXiv:2107.07089*, 2021.

[80] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. T. Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Information Sciences*, vol. 480, pp. 287–304, 2019.

[81] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[82] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[83] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2969–2978.

[84] Y. Obinata and T. Yamamoto, "Temporal extension module for skeleton-based action recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 534–540.

[85] T. Chen, D. Zhou, J. Wang, S. Wang, Y. Guan, X. He, and E. Ding, "Learning multi-granular spatio-temporal graph network for skeleton-based action recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4334–4342.

[86] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[87] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 359–13 368.

[88] A. Zeng, X. Sun, L. Yang, N. Zhao, M. Liu, and Q. Xu, "Learning skeletal graph neural networks for hard 3d pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 436–11 445.

[89] Z. Qin, Y. Liu, P. Ji, D. Kim, L. Wang, B. McKay, S. Anwar, and T. Gedeon, "Fusing higher-order features in graph neural networks for skeleton-based action recognition," *arXiv preprint arXiv:2105.01563*, 2021.

[90] D. Yang, M. M. Li, H. Fu, J. Fan, and H. Leung, "Centrality graph convolutional networks for skeleton-based action recognition," *arXiv preprint arXiv:2003.03007*, 2020.

[91] M. R. S. Alashti, M. H. B. Abadi, P. Holthaus, C. Menon, and F. Amirabdollahian, "Human activity recognition in robocup@ home: Inspiration from online benchmarks," *UKRAS21*, 2021.

[92] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[93] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th international conference on mobile computing, applications and services*. IEEE, 2014, pp. 197–205.

[94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[95] H. Zhou, L. Jiao, S. Zheng, S. Chen, L. Yang, W. Shen, and X. Yang, "Weight-variable scattering convolution networks and its application in electromagnetic signal classification," *IEEE Access*, vol. 7, pp. 175 889–175 896, 2019.

[96] O. Bazgir, R. Zhang, S. R. Dhruba, R. Rahman, S. Ghosh, and R. Pal, "Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks," *Nature communications*, vol. 11, no. 1, p. 4391, 2020.

[97] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[98] A. Hryniowski and A. Wong, "Seeing convolution through the eyes of finite transformation semigroup theory: An abstract algebraic interpretation of convolutional neural networks," *arXiv preprint arXiv:1905.10901*, 2019.

[99] A. Albanese, M. Nardello, and D. Brunelli, "Low-power deep learning edge computing platform for resource constrained lightweight compact uavs," *Sustainable Computing: Informatics and Systems*, vol. 34, p. 100725, 2022.

[100] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[101] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.

[102] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le, "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2820–2828.

[103] S. Targ, D. Almeida, and K. Lyman, "Resnet in resnet: Generalizing residual architectures," *arXiv preprint arXiv:1603.08029*, 2016.

[104] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[105] L. Hedegaard, N. Heidari, and A. Iosifidis, "Continual spatio-temporal graph convolutional networks," *Pattern Recognition*, vol. 140, p. 109528, 2023.

[106] X. Yan, S. Hu, Y. Mao, Y. Ye, and H. Yu, "Deep multi-view learning methods: A review," *Neurocomputing*, vol. 448, pp. 106–129, 2021.

[107] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE transactions on knowledge and data engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[108] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *Ieee Access*, vol. 7, pp. 63 373–63 394, 2019.

[109] C. Ahuja, L. P. Morency, *et al.*, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, pp. 1–20, 2017.

[110] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[111] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[112] S. Sun, "A survey of multi-view machine learning," *Neural computing and applications*, vol. 23, pp. 2031–2038, 2013.

[113] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: a survey," *IEEE Access*, vol. 7, pp. 63 373–63 394, 2019.

[114] 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*, vol. 45, no. 3, pp. 3200–3225, 2023.

[115] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021.

[116] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Information Fusion*, vol. 50, pp. 158–167, 2019.

[117] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," *arXiv preprint arXiv:1702.02519*, 2017.

[118] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," *arXiv preprint arXiv:2002.07017*, 2020.

[119] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view cnns for object classification on 3d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5648–5656.

[120] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 945–953.

[121] W. Wang, X. Wang, G. Chen, and H. Zhou, "Multi-view softpool attention convolutional networks for 3d model classification," *Frontiers in Neurorobotics*, vol. 16, p. 1029968, 2022.

[122] W. Zou, D. Zhang, and D.-J. Lee, "A new multi-feature fusion based convolutional neural network for facial expression recognition," *Applied Intelligence*, vol. 52, no. 3, pp. 2918–2929, 2022.

[123] K. Liu and G. Kang, "Multiview convolutional neural networks for lung nodule classification," *International Journal of Imaging Systems and Technology*, vol. 27, no. 1, pp. 12–22, 2017.

[124] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 1963–1978, 2019.

[125] L. Bai, L. Yao, X. Wang, S. S. Kanhere, B. Guo, and Z. Yu, "Adversarial multi-view networks for activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, pp. 1–22, 2020.

[126] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[127] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. T. Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Information Sciences*, vol. 480, pp. 287–304, 2019.

[128] A. A. Chaaraoui, J. R. Padilla-López, F. Ferrández-Pastor, M. Nieto-Hidalgo, and F. Flórez-Revuelta, "A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context," *Sensors*, vol. 14, pp. 8895–8925, 2014.

[129] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Transactions on Image Processing*, vol. 29, pp. 3835–3844, 2020.

[130] A. Subasi, D. H. Dammas, R. D. Alghamdi, R. A. Makawi, E. A. Albiety, T. Brahimi, and A. Sarirete, "Sensor based human activity recognition using adaboost ensemble classifier," *Procedia computer science*, vol. 140, pp. 104–111, 2018.

[131] W. Huang, L. Zhang, S. Wang, H. Wu, and A. Song, "Deep ensemble learning for human activity recognition using wearable sensors via filter activation," *ACM Transactions on Embedded Computing Systems*, vol. 22, no. 1, pp. 1–23, 2022.

[132] K. K. Sahoo, R. Ghosh, S. Mallik, A. Roy, P. K. Singh, and Z. Zhao, "Wrapper-based deep feature optimization for activity recognition in the wearable sensor networks of healthcare systems," *Scientific Reports*, vol. 13, no. 1, p. 965, 2023.

[133] A. Shakerian, V. Douet, A. Shoaraye Nejati, and R. Landry Jr, "Real-time sensor-embedded neural network for human activity recognition," *Sensors*, vol. 23, no. 19, p. 8127, 2023.

[134] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

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

[136] F. M. Reza, *An introduction to information theory*. Courier Corporation, 1994.

[137] D. Huang and T. W. Chow, "Effective feature selection scheme using mutual information," *Neurocomputing*, vol. 63, pp. 325–343, 2005.

[138] H. Liu, J. Sun, L. Liu, and H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, no. 7, pp. 1330–1339, 2009.

[139] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger, "Hierarchical clustering using mutual information," *Europhysics Letters*, vol. 70, no. 2, p. 278, 2005.

[140] A. Kraskov and P. Grassberger, "Mic: Mutual information based hierarchical clustering," *Information theory and statistical learning*, pp. 101–123, 2009.

[141] F. Maes, D. Vandermeulen, and P. Suetens, "Medical image registration using mutual information," *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1699–1722, 2003.

[142] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[143] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[144] M. R. S. Alashti, M. B. Abadi, P. Holthaus, C. Menon, and F. Amirabdollahian, "Rhm-har-sk: A multi-view dataset with skeleton data for ambient assisted living research," *IARIA, March*, 2023.

[145] S. Bedaf, G. J. Gelderblom, D. S. Syrdal, H. Lehmann, H. Michel, D. Hewson, F. Amirabdollahian, K. Dautenhahn, and L. De Witte, "Which activities threaten independent living of elderly when becoming problematic: inspiration for meaningful service robot functionality," *Disability and Rehabilitation: Assistive Technology*, vol. 9, no. 6, pp. 445–452, 2014.

[146] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[147] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," *arXiv preprint arXiv:1605.07678*, 2016.

[148] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE access*, vol. 6, pp. 64 270–64 277, 2018.