ADVANCING HUMAN ACTIVITY RECOGNITION IN AMBIENT ASSISTED LIVING THROUGH MULTI-VIEW ROBOTICS: THE RHM DATASET AND DUAL-STREAM C3D MODEL

by

MOHAMMAD HOSSEIN BAMOROVAT ABADI

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

Robotics Research Group
School of Physics, Engineering & Computer Science
Department of Computer Science
University of Hertfordshire

Dec 2023

Abstract

This thesis investigates the intersection of Human Action Recognition (HAR) and Human-Robot Interaction (HRI), in Ambient Assistive Living (AAL) environments. The primary contribution of our research is the development of the Robot House Multi-View (RHM) dataset, featuring 26,804 RGB trimmed videos from four distinct views classified into 14 action classes: a dynamic robot view, static top view, and static front and back views.

Dataset: The RHM dataset addresses significant gaps in existing HAR datasets, particularly within the HRI domain. To validate the dataset, a comprehensive approach using Deep learning (DL) and Mutual Information (MI) was employed. The dynamic robot view presents unique challenges due to lower accuracy in comparison to static views, attributed to its inherent variability and motion. A novel MI metric was introduced to analyse temporal dependencies and information redundancy across video frames. State-of-the-art DL models, including C3D, R(2+1)D, R3D, and SlowFast, were tested on the RHM dataset.

Methodology: The thesis introduces a novel multi-stream model, the Dual-stream C3D, which integrates multiple views to enhance HAR accuracy. The combination of Front and Robot views in this model shows the highest accuracy, highlighting the potential of multi-view integration for improving action recognition performance. Specifically, the model demonstrated a 10% increase in Top-1 accuracy for the robot view when combined with other views, such as the front view. However, despite these improvements, consistent confusion patterns among certain action classes persist, suggesting the need for further refinement in feature extraction in recognition models.

Feature Extraction Techniques: Additionally, the research introduces and evaluates three

novel feature extraction techniques: Motion Aggregation (MAg), Differential Motion Trajectory (DMT), and Frame Variation Mapper (FVM). These techniques target different temporal aspects of video frames and are shown to significantly enhance the performance of HAR models. Experimental results indicate that the combination of **Normal frames** in the first stream and DMT in the second stream achieves the highest accuracy, particularly for the Front-Robot viewpoint pair. These findings underscore the adaptability and effectiveness of these feature extraction methods across various models and viewpoints.

Conclusion: In summary, this thesis presents the RHM dataset as a substantial contribution to HAR and HRI, offering innovative methodologies and insights that significantly improve action recognition accuracy in AAL scenarios. The integration of multi-view data, novel deep learning models, and advanced feature extraction techniques collectively advance the state-of-the-art in HAR within the context of assistive robotics.

In Loving Memory of My Mother

This thesis is solemnly dedicated to the memory of my dear mother. Her constant love, support, and encouragement have been the guiding force in my life and academic pursuits. Although she is no longer physically present, her spirit continues to inspire every success, learning experience, and moment of resilience I encounter. She is an everlasting source of inspiration, and this work is a testament to her enduring influence in my life.

May her soul rest in eternal peace.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Farshid Amirabdollahian, for his invaluable guidance, patience, and immense knowledge. His mentorship was pivotal in shaping not only this thesis but also my understanding of the subject.

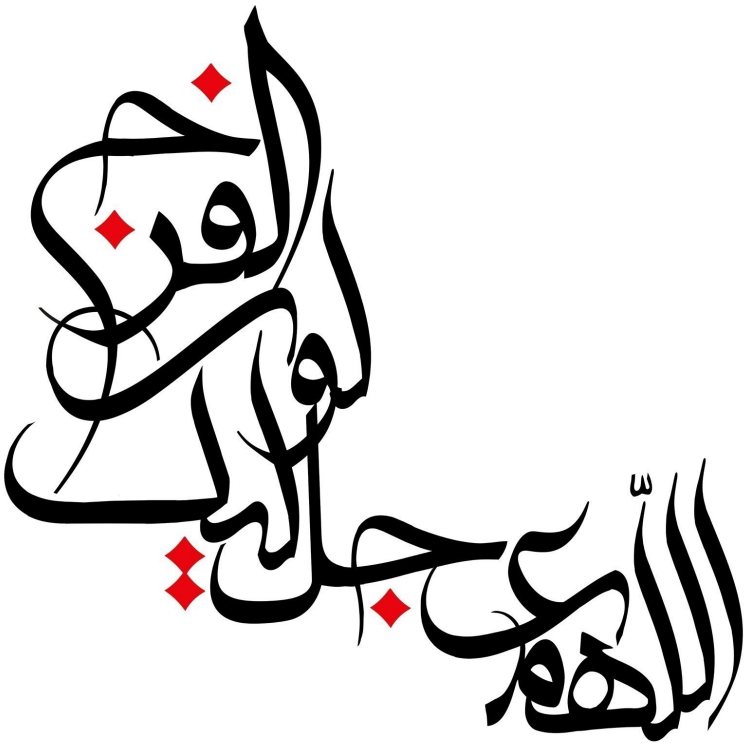
My heartfelt thanks go to my parents and family, whose love and encouragement have been my constant source of strength and motivation. Their sacrifices and unwavering belief in me have been the backbone of my success.

I am also profoundly thankful to my second supervisors, Dr. Catherine Menon and Dr. Patrick Holthaus, for their insightful feedback and unwavering support throughout this journey. Their perspectives and expertise have been crucial in refining my research.

I am also grateful to my friends, Mohammadreza Shahabian and Khashayar Ghamati, for their companionship, and moral support.

In conclusion, I address to the Last Savior, **Imam Mahdi**:

"At nights, with the hope of meeting you, my heart weeps,
Every moment, restless and intoxicated by your love.
O light of hope and love, my heart is bound to you,
O sun of the world, come, for my heart is restless."


اللّٰهمّ عجل لوليك الفرج

Contents

List of Figures

List of Tables

List of Acronyms

2D Two Dimensions. 50, 58, 96

2DCNN Two Dimension Convolutional Neural Networks. 5, 48, 50

3D Three Dimensions. 48, 50, 57, 58, 71, 75, 77, 95, 96

3DCNN Three Dimension Convolutional Neural Networks. 5, 48, 50

AAL Ambient Assistive Living. i, ii, 1–3, 7, 9, 10, 15, 27, 71, 72, 88, 91, 121, 125, 126, 128, 130

BoWs bag-of-words. 96, 97

C3D Convolutional Three Dimensions. i, xiii–xv, 9, 10, 12, 13, 48, 57, 58, 62, 64, 65, 68, 69, 71, 77–91, 94, 107, 110–113, 118–123, 125–130, 150–154

CNN Convolutional Neural Networks. 47–50, 57, 58, 76, 77, 82, 101, 107, 118, 121, 122, 128, 129

Conv-LSTM Convolutional Long Short-Term Memory Networks. 51

ConvNets Convolutional Networks. xiv, xv, 13, 51, 73–76, 84–88, 91, 94, 115–118

CRR Correct Recognition Rates. 96

DL Deep learning. i, 1, 3–7, 12, 13, 15, 19, 20, 31, 43, 48, 51, 52, 59, 69, 71, 83, 94, 107, 121, 123, 125–128

DMT Differential Motion Trajectory. ii, 9, 13, 102, 105–108, 110, 111, 113–119, 121–123, 125, 126, 129, 130

DNN Deep Neural Networks. 48

FP First Person. 21, 22

FPPW False Positive Per Window. 99

FVM Frame Variation Mapper. ii, 9, 13, 102, 104, 106–108, 110, 111, 113, 114, 116, 119, 121–123, 126, 129, 130

HAR Human Action Recognition. i, ii, xv, 1–10, 12–24, 27, 31, 32, 35, 37, 40, 41, 43, 48, 50, 51, 53, 57, 62, 64, 68, 69, 72, 74–77, 90, 94, 96, 97, 101–103, 108, 109, 115, 118, 121, 123, 125–128, 130–133

HMM Hidden Markov Model. 46

HOF Histograms of Optical Flow. 4, 98

HOG Histograms of Oriented Gradients. 4, 99

HOOF Histogram of Oriented Optical Flow. 4

HRI Human-Robot Interaction. i, ii, 1–3, 6–8, 10, 14, 15, 22, 27, 31, 32, 40, 41, 68, 71, 72, 88, 91, 121, 125, 126, 131–133

IAI Intensity-Accumulated Image. 100

iDT improved Dense Trajectories. 73–76

JSTAM Joint SpatialTemporal Attention Module. 51

KNN K-Nearest Neighbors. 5

LIREC Living with Robots and Interactive Companions. 2

LRCN Long-term recurrent convolutional networks. 49

LSTM Long Short-Term Memory Networks. 5, 49, 51

LTC-CNN Long-Term Temporal Convolution. 49

MAg Motion Aggregation. ii, 9, 13, 102–104, 106–108, 110, 111, 113, 114, 116, 119, 121–123, 125, 129, 130

mAP mean Average Precision. 74

MBH Motion Boundary Histograms. 98

MC Mixed Convolution. 50

MEI Motion Energy Images. 95, 98

MHI Motion History Images. 73, 95, 98, 99

MI Mutual Information. i, 8, 12, 43, 46, 48, 53–57, 61, 62, 69, 107, 127, 130

MIESW Mutual Information and Entropy-based adaptive Sliding Window. 46

ML Machine Learning. 1, 3–6, 59, 125

MSKVS mean shift-based keyframes for video summarization. 45

NN Neural Networks. 5, 19

OFT Optical Flow Tensor. 46, 49

PCA Principal Component Analysis. 51

R(2+1)D ResNets with (2+1) Dimension convolutions. i, 12, 57, 62, 63, 65, 69, 84, 127

R3D Three Dimensions ResNets. 12, 57, 62, 65, 69, 84, 127

RANSAC Random sample consensus. 97

RGB Red-Green-Blue Color Mode. i, 12, 16–18, 20, 21, 32, 40, 83, 93

RGB_D Red-Green-Blue-Depth Camera Mode. 16, 20

RH Robot House. 15, 31

RHM Robot House Multi-View. i, ii, xiii, xv, 8–10, 12, 13, 27, 31–41, 43, 48, 52, 57, 58, 61–69, 71, 77, 81, 82, 84, 85, 87, 88, 90, 93, 107, 113, 114, 116, 119, 121, 125–127, 129–132, 150–154

RNN Recurrent Neural Networks. 5, 49, 50

SbHI SURF-based History Image. 100

SGD Stochastic Gradient Descent. 82, 111

SIFT Scale-Invariant Feature Transform. 95, 96

SRS Socially Relevant Scenarios. 2

STCB Spatiotemporal Compact Bilinear. 76

STDAN Spatial-Temporal Dual-Attention Network. 51

STIPs Space-Time Interest Points. 95

StNet Spatial-Temporal Network. 50

SURF Speeded Up Robust Features. 46, 97, 100

SVM Support Vector Machines. 5, 99, 101

TAM Temporal Attention Module. 51

TP Third Person. 17, 21, 22

Chapter 1

Introduction

1.1 Introduction

Human-Robot Interaction (HRI) is an expanding field that combines sophisticated robotic technologies with the intricate dynamics of human actions. Its goal is to foster progress in a new era of intelligent assistance. HRI encompasses many interactions, extending beyond simple task execution to include social, emotional, and collaborative interactions between humans and robots. The changing needs of Ambient Assistive Living (AAL) environments have prompted the investigation of robotics as a key instrument for improving safety, autonomy, and the general quality of life, especially in ambient assisted living situations (Broadbent, Stafford, and MacDonald, 2009).

In the developing field of HRI, Human Action Recognition (HAR) is emerging as an important area. Powered by progress in Machine Learning (ML) and Deep learning (DL), HAR is essential for enabling robots to accurately comprehend, interpret, and respond to human actions. The combination of HRI and HAR is expected to lead to the creation of intelligent robotic systems skilled in navigating complex human-centred environments like AAL. This integration is likely to significantly contribute to the advancement of assistive robotics (Aggarwal and Xia, 2014).

This research is designed to assist robots in identifying human actions within an AAL

environment. The first step is introducing a new multiview dataset that includes a robot's perspective along with several static views. After establishing this dataset, the research then focuses on using these additional static viewpoints to enhance the accuracy of the robot's view. This improvement is achieved through the application of multi-stream networks, which are capable of processing information from multiple viewpoints simultaneously. Final step introduces and evaluates three innovative temporal feature extraction methods for static cameras to enhance the accuracy of the proposed deep learning models.

1.2 Background

1.2.1 HRI Projects in Ambient Assistive Living Settings

HRI is a key aspect of robotics, focusing on the interactions and responses between humans and robots. This is particularly important in social robotics, where effective human interaction is vital to the robots' functionality and purpose (Yan, Ang, and Poo, 2014). Understanding the role of robots in AAL environments involves looking at various significant projects. Each of these projects contributes uniquely to the development of HRI and HAR within AAL contexts. Below is an overview of some of these influential projects:

Living with Robots and Interactive Companions (LIREC) Project focused on developing interactive robots that can adapt to human behaviour and social contexts. A major emphasis of this project was on user acceptance and ethical considerations in the design and deployment of these robots (Van Oost and Reed, 2010). **Socially Relevant Scenarios (SRS)** Project dedicated to creating robots to engage in significant interactions with older adults. These robots are designed to assist in everyday tasks and provide cognitive and social support for the elderly (Qiu et al., 2012). **ACCOMPANY** Project created socially assistive robots to improve the independent living of older adults. The project aimed to provide support in daily activities, health monitoring, and social interaction for the elderly (Amirabdollahian et al., 2013). **GrowMeUp** Project tried to build a robotic platform designed to assist older adults in maintaining their independence and quality of life. It focused on providing personalised assistance and adapting to the unique

needs of each individual (Georgiadis et al., 2016). **EnrichMe** Project focused on developing an intelligent robotic platform to aid older adults in maintaining an active and independent lifestyle. The project aimed to provide personalised exercise recommendations, cognitive training, and opportunities for social engagement (Agrigoroaie, Ferland, and Tapus, 2016). **ACANTO** Project focused on creating a socially assistive robot designed to encourage physical activity, monitor health status, and foster social interaction among older adults (Pérez-Rodrıéguez et al., 2019). **RAMCIP** Project focused on studying human-robot interactions in ambient assisted living environments. RAMCIP aimed to help older adults with daily activities, thereby enhancing their quality of life (Kostavelis et al., 2019).

These projects collectively contribute to the evolving story of research and development in HRI and HAR within AAL environments. Each one offers unique insights and advancements, helping to develop robots that are proficient in effectively collaborating with humans.

1.2.2 Human Action Recognition

HAR is pivotal in distinguishing different human actions, significantly contributing to the progress in the HRI field (Abadi et al., 2021). In the context of recognition tasks, both ML and DL are crucial. The ML methodology involves two key stages: Feature Extraction and Classification. The Feature Extraction stage is crucial for identifying important attributes from the dataset that aid in the recognition process. Then, in the Classification stage, these extracted features are used to categorise the data into predefined classes, thus completing the task of recognition.

In contrast, DL employs an end-to-end approach, combining the Feature Extraction and Classification stages into a single, integrated framework. This unified method allows for the automatic detection and classification of data within the same architecture, potentially improving the efficiency and effectiveness of the recognition process. With this end-to-end model, DL reduces the need for handcraft feature selection and encompasses the entire recognition process within one cohesive model. This could lead to a more streamlined and automated recognition system.

Besides ML and DL, the dataset is another essential element in HAR. A strong dataset is vital for training and testing the models created using ML or DL approaches. The dataset includes numerous data points, each marked with the relevant activity label. This rich collection of data provides an ideal environment for the model to learn and generalise patterns related to human activities.

Handcraft Feature Extraction

Handcraft feature extraction involves creating and using algorithms to identify spatial and/or temporal features in video sequences that signify different actions. This traditional method needs a thorough understanding of the specific characteristics of the actions being analysed. This knowledge is crucial for carefully designing feature extractors. Some common handcraft feature extraction techniques used in HAR are listed below:

- **Spatial Features:** These methods focus on extracting static information from a scene. It derives from the spatial arrangement of pixels in an image and includes attributes like edges, corners, and textures (Dalal and Triggs, 2005).
- **Temporal Features:** Temporal features capture motion information over time and are typically extracted from sequences of images or video frames. A well-known method for obtaining temporal features is optical flow. This technique estimates the motion vector of each pixel or region between consecutive frames (Lucas and Kanade, 1981).
- **Spatiotemporal Features:** Spatiotemporal features are designed to simultaneously capture both spatial and temporal information. They are generally extracted from sequences of images and are used to encapsulate complex motion patterns along with the evolution of spatial features over time (Klaser, Marszałek, and Schmid, 2008).
- **Transform-based Features:** Features such as Histograms of Oriented Gradients (HOG), Histograms of Optical Flow (HOF), and Histogram of Oriented Optical Flow (HOOF) are extracted in a transformed domain, often utilised to capture both spatial and motion information (Laptev et al., 2008).

- **Trajectory-based Features:** Derived from the tracked motion of points or regions across frames, trajectory-based features capture the motion path of particular points over time, offering a detailed representation of motion patterns (H. Wang, Kläser, et al., 2013).

The design process for handcraft features often demands meticulous attention to ensure the effective capture of essential action characteristics. Despite the emergence of automated feature extraction techniques through deep learning, handcraft feature extraction continues to be a valuable approach in limited data scenarios, or where interpretability or lower computational resources are required. Skilful extraction of handcraft features can enable HAR models to achieve notable accuracy and efficiency, especially in constrained computational settings or specific application scenarios.

Machine Learning in HAR

Following the feature extraction phase, which entails the labelling or categorisation of data predicated on the extracted features, the classification stage is the second stage of the ML framework as elucidated by (Shi et al., 2011). Several effective algorithms have been deployed at this juncture to effectuate the classification task, notably, the Support Vector Machines (SVM) detailed in (Laptev et al., 2008; Marszalek, Laptev, and Schmid, 2009), and K-Nearest Neighbors (KNN) presented by (Tran and Sorokin, 2008).

Deep Learning in HAR

DL, a specialised branch of ML, leverages multi-layered artificial Neural Networks (NN) to unravel complex patterns from raw data. Various DL architectures have demonstrated significant potential in HAR by adeptly analysing spatial-temporal data. Noteworthy among these are Two Dimension Convolutional Neural Networks (2DCNN) (Karpathy et al., 2014), Three Dimension Convolutional Neural Networks (3DCNN) (Tran, Bourdev, et al., 2015), and Recurrent Neural Networks (RNN)s (Rumelhart, Hinton, and Williams, 1986). Particularly, Long Short-Term Memory Networks (LSTM)s (Hochreiter and Schmidhuber, 1997), a specialised variant of RNNs, and Dual-stream networks (Simonyan and Zisserman, 2014) have played a pivotal role

in augmenting action recognition accuracy by proficiently processing both spatial and temporal information.

1.2.3 Datasets for HAR

Advancements in ML and DL have catalysed the generation of an array of datasets tailored for HAR. Datasets such as UCF101 (Soomro, Amir Roshan Zamir, and Shah, 2012), YouTube-8M (Abu-El-Haija et al., 2016), and Kinetics-700 (Carreira, Noland, Hillier, et al., 2019) have rendered substantial contributions to the domain. Nonetheless, relatively few datasets from a robot's viewpoint, thus constraining the evolution of HAR models for robot interaction contexts (Abadi et al., 2021). Efforts are underway to address this gap by creating datasets from robot perspectives, which are essential for training and evaluating HAR models designed for HRI contexts. The advent of such datasets is foreseen to spur advancements in robot perception and action recognition proficiencies, thereby fostering the enrichment of the HRI domain.

1.3 Problem Statement

The growth in HRI and HAR highlights the need for improved methods to accurately recognise and interpret human actions. A key aspect of this is having robust and comprehensive datasets that cover a wide range of perspectives and scenarios found in real-world human-robot collaboration environments. However, there is a noticeable gap in terms of datasets from a robot's viewpoint, which slows down the development of models well-adapted to HRI scenarios. This leads to the research hypothesis:

> ***The accuracy of state-of-the-art deep models in robotic vision is generally lower compared to static views such as top-mounted or wall-mounted cameras, primarily due to the movement of the robot. Furthermore, incorporating additional viewpoints into multiview deep models can enhance the accuracy of observations from the robot's perspective. Additionally, the presence of static views allows for more rapid extraction of temporal information.***

1.4 Research Questions

The primary objective of this research is to enhance the understanding of robot recognising with human activities in AAL environments. This involves advancing the field of HAR within HRI contexts. To reach this goal, several key research questions have been formulated to direct the research. These questions are crucial for exploring ways to optimise HAR models from different camera perspectives, particularly in dynamic settings. The research questions are outlined as follows:

Question One:

How does the dynamics of a camera from a robot viewpoint impact the accuracy of DL models in HAR?

This question probes the effect of camera mobility on the performance metrics of DL models designated for HAR. It explores the accuracy disparities induced by a static versus dynamic camera setup in capturing and recognising human activities within a robot's movement. In Chapter 3 and Chapter 4 this question will be answered.

Question Two:

Does the inclusion of additional camera views in a multi-stream DL model for HAR enhance the accuracy of the robot view?

This inquiry evaluates the potential accuracy augmentation when multiple camera views are integrated with the robot view, with a specific emphasis on how additional perspectives bolster the robot view in a multiview DL framework for HAR. Chapter 5 will work on this question.

Question Three:

How does employing handcraft feature extraction temporal information on dual-stream DL model for a robot view and another view in parallel impact HAR?

This question ventures into the domain of feature engineering by evaluating the efficacy of handcraft feature extraction when applied concurrently to a robot view and another view. It examines how these handcraft features affect the accuracy and robustness of HAR models. Chapter 6 will discuss about this question.

1.5 Contribution

1- RHM Dataset Contribution:

The cornerstone contribution of this thesis lies in the creation of a novel dataset, named the RHM, designed for HAR. This dataset addresses three critical facets often missing in existing datasets: a dynamic perspective (Robot View), a top view (Fish Eye View), and redundancy across multiple views. The RHM dataset has been extensively documented in various research publications:

- "RHM: Robot House Multi-view Human Activity Recognition Dataset," (Abadi et al., 2023).
- "RHM-HAR-SK: A multi-view dataset with skeleton data for ambient assisted living research," (Alashti et al., 2023b).
- "Robot house human activity recognition dataset," (Abadi et al., 2021).

The dataset is accessible to researchers and practitioners through the following link: RHM Dataset. This dataset, employed by the research team (Alashti et al., 2023b; Alashti et al., 2023a), now supports collaborative international research between the University of Hertfordshire and the Multimedia University of Malaysia.

2- RHM Analysis Contribution:

A novel metric based on Mutual Information (MI) is introduced for analysing HAR datasets. This metric, which considers temporal dependencies between successive video frames, serves as a powerful tool for investigating information redundancy and the discriminative capacity across various actions and viewpoints. This analysis has been further validated and expanded upon in the publication titled *RHM: Robot House Multi-View Human Activity Recognition Dataset* (Abadi et al., 2023), presented at the ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions. This paper delves into the application of multi-view datasets for HAR in HRI, featuring comprehensive performance assessments and benchmarks for different views within the dataset.

3- Dual-Stream C3D Model Contribution:

In pursuit of enhancing accuracy, a novel multi-stream model termed the Dual-stream C3D is developed. This model combines multiple views with the robot view to improve accuracy within the RHM dataset. The methodology and findings related to this model are detailed in the paper titled *Robotic Vision and Multi-View Synergy: Action and Activity Recognition in Assisted Living Scenarios* (Abadi et al., 2024b), presented at the ACHI 2023. This study underlines the significance of robotic vision and multi-view synergy in AAL environments, providing a robust foundation for the methodologies employed in this thesis.

4- Multi-View Fusion and Feature Extraction for Enhancing HAR:

The methodologies and findings presented in Chapter 6 have been further validated and expanded upon in the recent publication titled *Multi-View Fusion and Feature Extraction: Enhancing HAR for Assistive Robotics* (Abadi et al., 2024a). This paper, presented at the 2024 IEEE RAS International Conference on Humanoid Robots, addresses the challenge of improving HAR in robotics by focusing on the integration of multi-view data and the extraction of temporal features from static cameras. Utilising the Robot House Multiview (RHM) dataset, this research introduces three innovative handcrafted feature extraction methods: Motion Aggregation (MAg), Differential Motion Trajectory (DMT), and Frame Variation Mapper (FVM). The results demonstrate that incorporating these methods into dual-stream models significantly boosts performance, with the DMT method exhibiting the most substantial improvement. A key finding is the superior efficacy of combining the Robot view with normal frames and the Front view with DMT frames, which consistently achieved the highest top-1 and top-5 results in the experiments. The detailed findings of this research provide a robust foundation for further investigations into more complex feature extraction methods and their applications in multiview HAR.

In summary, the contributions of this thesis span the creation of a novel dataset, the introduction of new analytical metrics and models, and the development of innovative feature extraction techniques, all aimed at advancing the field of Human Activity Recognition in the context of Human-Robot Interaction.

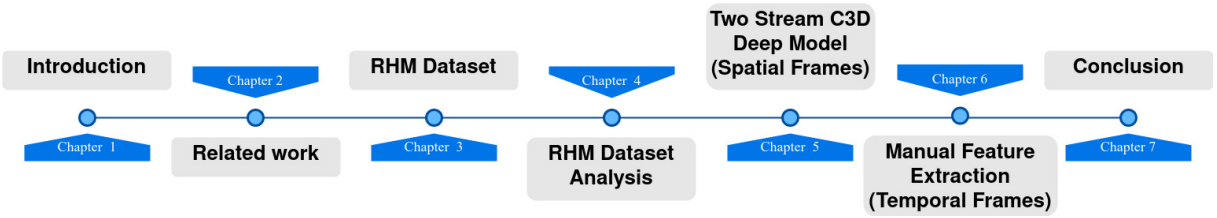


Figure 1.1: Thesis Chapters Map

1.6 Thesis Outline

This thesis unfolds over six meticulously structured chapters, each probing into vital facets and findings of this investigative journey. The chapter road map is depicted in Figure 1.1, providing a visual guide through the investigative terrain explored.

Figure 1.2 shows the process Diagram of the work. The green block represents the initial research inquiry concerning the RHM dataset, elucidated in Chapter 3 and 4. Following this, the red block navigates through the discourse on the Dual-stream C3D model, addressing the second research question in Chapter 5. The purple block unveils the realm of handcraft feature extraction in Chapter 6, engaging with the third research question. The thesis is sequenced as follows:

1.6.1 Chapter 1: Introduction

Chapter One introduces the key background, explains the motivation, states the problem, and poses the research questions for this research. It prepares for the upcoming investigation into HAR in HRI settings.

1.6.2 Chapter 2: Literature Review

Chapter Two conducts detailed reviews of significant HAR datasets, uncovering major gaps in the HRI domain. The review concludes that in the HRI domain, there is a notable shortage of HAR datasets from the robot's perspective. Additionally, for AAL environments, there is a lack of datasets offering a top view.

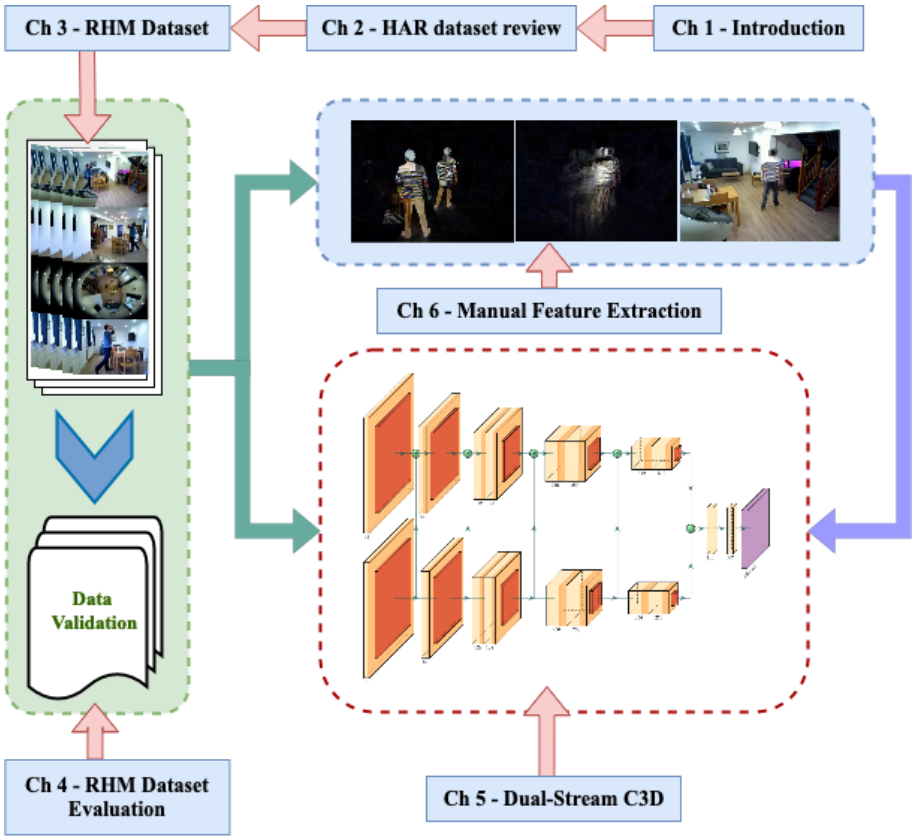


Figure 1.2: Chapter Structure Overview and Research Workflow
This figure presents the interconnected structure and workflow of the thesis, beginning with **Chapter 1** (Introduction) and progressing through the review of existing Human Activity Recognition (HAR) datasets in **Chapter 2**, which justifies the creation of the Robot House Multiview (RHM) dataset detailed in **Chapter 3**. The validated dataset from **Chapter 4** serves as the foundation for the experiments and model development in subsequent chapters. **Chapter 5** introduces the Dual-Stream 3D Convolutional Neural Network (C3D) model, while **Chapter 6** explores manual feature extraction techniques, both of which are applied to the RHM dataset. The diagram illustrates the logical progression and integration of each chapter, culminating in a comprehensive approach to enhancing HAR through both dataset development and advanced model evaluation.

1.6.3 Chapter 3: RHM Dataset

Chapter Three introduces a new HAR dataset called RHM, created using the Red-Green-Blue Color Mode (RGB) approach for HAR. The RHM dataset addresses three key aspects missing in current datasets: a dynamic perspective (Robot View), a top view (Fish Eye View), and multiple view redundancy. It consists of four unique views: Front (static), Back (static), Top (fish-eye), and Robot (dynamic). Each view contains 6,701 videos, totalling 26,804 videos across all views, divided into 14 action classes. Every video, categorised by class and number, is carefully synchronised in time across the different views.

1.6.4 Chapter 4: RHM Dataset Analysis

Chapter Four validates the RHM dataset and introduces a new Mutual Information (MI) metric for evaluating HAR datasets. This metric accounts for the related information between video frames to analyse information redundancy. The analysis revealed that dynamic views, such as the Robot View, have lower MI values, indicating less redundancy between frames, unlike static views like the Front and Back Views, which show more redundancy. A thorough evaluation of various DL models, including Convolutional Three Dimensions (C3D), ResNets with (2+1) Dimension convolutions (R(2+1)D), Three Dimensions ResNets (R3D), and SlowFast, was conducted on the RHM dataset, providing key performance metrics for HAR. Notably, the Robot View consistently had lower accuracy in Top-1 and Top-5 measures across all models, due to the unique motion and frame changes from the robot's perspective. In contrast, the Top and Front views showed higher accuracy, highlighting the challenges and potential improvement strategies for action recognition accuracy. Moreover, a confusion matrix analysis of the C3D model showed consistent confusion patterns among certain action classes across different views and models.

1.6.5 Chapter 5: Two Stream C3D Deep Model

To address the low accuracy of the robot view in the RHM dataset, this chapter explores a multiview model based on the C3D model, specifically the Dual-stream C3D. The goal is to understand how adding multiple views affects the accuracy of the robot view in the RHM dataset. The research of the Dual-stream C3D model shows significant improvements in both top1 and top5 accuracy when different views are included, particularly the robot-front combination. This highlights the importance of extra views in improving the accuracy of the robot view. However, despite these improvements in accuracy, the models still display the same confusion patterns in their confusion matrices. This indicates that the fundamental challenges in recognising specific activities persist across different models, regardless of any differences in accuracy.

1.6.6 Chapter 6: Handcraft Feature Extraction on RHM

To improve the results in the proposed Dual-stream model and utilise temporal information for the static views, chapter Six explores how different feature extraction methods affect three key Deep Learning models for HAR: the SlowFast model, the Two stream Convolutional Networks (ConvNets), and the Two Stream C3D model. It examines Motion Aggregation (MAg), Differential Motion Trajectory (DMT), and Frame Variation Mapper (FVM) techniques, each designed to capture various types of temporal information in video frames. The research shows significant improvements in both Top1 and Top5 accuracy for many viewpoint pairs in both models, thanks to feature extraction. Specifically, the SlowFast model works best with normal frames in the first stream and the "DMT" method in the second stream. For the Two Stream C3D model, the combination of "Normal, DMT" frames is most effective, particularly for the Front-Robot viewpoint pair. These findings highlight how effective and adaptable feature extraction techniques are. They show that these techniques can be used across different DL models and viewpoints to improve performance. The results also suggest that certain combinations of feature extraction methods and viewpoints work better for accurate HAR, providing important insights for future research and practical use.

1.6.7 Chapter 7: Conclusion

The final chapter summarises the main discoveries, contributions, and impacts of this research. It outlines possible future research directions, connecting this thesis to the broader conversation about HAR in HRI. The chapter also includes reflections on the research process and offers ideas for further exploration in this field.

Chapter 2

Related Work

2.1 Introduction

Preparing a comprehensive and high-quality dataset is indeed a critical component of any DL research, including HAR. Recent research highlights the importance of having a robust dataset with essential parameters such as diversity and a sufficient number of data samples to effectively train DL models (Rosebrock, 2022).

In this chapter, I will review the existing HAR datasets available worldwide and perform a comparative analysis of their characteristics. This analysis will enable the identification of any limitations or gaps in the existing datasets and establish the rationale for developing a new HAR dataset.

By examining and comparing existing HAR datasets, their suitability for addressing the specific requirements and challenges of HAR in the context of HRI within an AAL setting, such as the RH, can be assessed. The evaluation will consider factors such as the range of activities captured, the diversity of subjects and environmental conditions, and the representation of real-world scenarios. This analysis will highlight the need for a new HAR dataset that specifically addresses the complexities and nuances of HRI in AAL environments.

2.2 Related Work

Upon investigation of existing HAR datasets, it becomes evident that these datasets are categorised based on various features, which encompass the activity's theme, camera properties, environment, subject, situation, or scenario (Abadi et al., 2021). These categorisations enable researchers to analyse and compare datasets based on specific criteria and requirements.

For instance, activity theme categorisation classifies the datasets according to the type of activities performed, such as daily activities, sports activities, industrial tasks, or surveillance scenarios. This classification helps researchers focus on specific domains and understand the challenges associated with different types of activities.

Camera properties categorisation involves considering the characteristics of the cameras used in data collection. This categorisation may include information about the camera types, such as RGB cameras or RGB_D cameras. Additionally, it considers the cameras' position, which can be static or dynamic, and whether the dataset includes single or synchronised multiple views. Understanding the camera properties is crucial for capturing different perspectives and ensuring comprehensive coverage of the activities (Abadi et al., 2023).

Another significant categorisation is based on the environment in which the activities take place. This classification differentiates between indoor and outdoor environments, controlled or uncontrolled settings, and even wild scenarios. Recognising the environment is essential for assessing the adaptability and generalisability of the HAR models in different real-world contexts (Abadi et al., 2023).

Subject, situation, or scenario categorisations consider the individuals or groups involved in the activities, specific situations in which the activities occur, or the broader context or scenario of the dataset. These categorisations provide additional context and enable researchers to analyse the impact of different subjects, situations, or scenarios on HAR performance (Abadi et al., 2023).

By considering these various categorisations, researchers can identify the datasets that align with their specific research goals and requirements. This understanding allows for more focused and targeted analysis of the datasets and facilitates the development of robust HAR models that

are tailored to specific activity themes, camera properties, environments, subjects, situations, or scenarios (Abadi et al., 2023).

The *KTH* dataset, introduced in 2004 by (Schuldt, Laptev, and Caputo, 2004), is a pivotal RGB-based HAR dataset consisting of 599 videos representing six activity classes. Collected in a controlled outdoor environment with a static background, the dataset has served as a benchmark for early HAR research. Similarly, the *Weizmann* dataset presented by (Gorelick et al., 2007), featuring ten activity classes and 90 videos, has made significant contributions to individual action recognition. Both datasets have provided researchers with valuable resources for developing and evaluating HAR algorithms. However, as the field progresses, there is a growing need for more diverse and realistic datasets that capture real-world complexities and environmental variations, enabling the development of robust models capable of handling a wide range of human activities in dynamic and uncontrolled scenarios.

The *INRIA XMAS* dataset, introduced by (Weinland, Ronfard, and Boyer, 2006), marks the first multiview RGB HAR dataset. It comprises 390 videos featuring 13 activities and encompasses five different views. The dataset was meticulously prepared in a controlled indoor environment, considering variations in actors, cameras, and viewpoints. The availability of multiple views in *INRIA XMAS* enables researchers to explore the challenges and benefits of utilising different camera perspectives for activity recognition. This dataset has contributed to advancing the field of HAR by providing researchers with a valuable resource for developing and evaluating multiview RGB-based HAR algorithms.

The *MuHAVi* dataset, presented by (S. Singh, Velastin, and Ragheb, 2010), is a significant contribution to the field of HAR. This dataset consists of 238 videos capturing 17 different activity classes performed by 14 actors. It features a multiview structure with eight Top-View Third Person (TP) perspectives, providing comprehensive coverage of the activities. The dataset was collected in a controlled indoor environment, ensuring consistent conditions for activity recognition. The availability of the *MuHAVi* dataset has facilitated research in HAR by offering a diverse set of activities and multiple camera views, enabling the development and evaluation of robust HAR algorithms capable of handling complex real-world scenarios.

(Kuehne et al., 2011) introduced the *HMDB51* dataset, a comprehensive collection of images by static camera used for human activity recognition. This dataset encompasses 51 activity classes and consists of a total of 6,849 videos. The videos were sourced from various platforms such as movies, YouTube, and Google Videos, providing a diverse range of activities and scenarios. The *HMDB51* dataset has been widely utilised in the field of HAR, serving as a valuable resource for training and evaluating activity recognition models. Its large-scale nature and diverse content have contributed to advancing the understanding and development of algorithms for human activity recognition tasks.

The *Hollywood* dataset, introduced by (Laptev et al., 2008), is a significant contribution to the field of HAR. It consists of 233 videos captured from Hollywood movies, making it a unique dataset for activity recognition. Each video in the dataset corresponds to one of the 10 activity classes. In 2009, an updated version of the dataset, known as *Hollywood2*, was released by (Marszalek, Laptev, and Schmid, 2009). *Hollywood2* expanded upon its predecessor by including 12 activity classes and containing a total of 3,669 videos, resulting in approximately 20 hours of footage. The dataset provides a diverse range of activities and scenes, with around 150 samples per action class and 130 samples per scene class. Both versions of the *Hollywood* dataset have played a crucial role in advancing activity recognition research, particularly in the domain of action recognition in movies.

The *UCF* series of HAR datasets encompass a wide range of variations in terms of class numbers, action types, modalities, and viewpoints. The early versions of the UCF datasets include *UCF11* (Jingen Liu, Luo, and Shah, 2009) with 11 classes and 1,160 videos, and *UCF50* (Reddy and Shah, 2013) with 50 classes and 6,676 videos. However, the most renowned dataset among the UCF series is *UCF101* (Soomro, Amir Roshan Zamir, and Shah, 2012), which comprises 101 activity classes and a staggering 13,000 videos. These datasets primarily consist of RGB videos sourced from YouTube clips, providing a diverse range of activities captured in various uncontrolled environments with both static and dynamic scenes.

In addition to the general UCF datasets, there are specific UCF variations tailored to certain domains. For example, *UCF Sport* focuses on sports actions, featuring ten classes and 150

videos (Soomro and Amir R Zamir, 2014). Another variation is *UCF-ARG*, a multiview dataset with ten actions and 480 videos captured from different viewpoints, including aerial, rooftop, and ground cameras (Nagendran, Harper, and Shah, 2010). The views in UCF-ARG are fixed, and the actions are recorded in a controlled outdoor environment. The UCF series of datasets has become prominent in the field of HAR due to their large-scale nature, diverse content, and inclusion of various real-world scenarios, facilitating the development and evaluation of activity recognition models.

The *ACT4* dataset, introduced by (Cheng et al., 2012), is a multiview dataset that consists of four different camera views, capturing 14 distinct human actions across 6,844 videos. This dataset was recorded in a controlled indoor environment, providing a diverse range of activities from various viewpoints. On the other hand, the *ASLAN* dataset, curated by (Kliper-Gross, Hassner, and L. Wolf, 2011), is a comprehensive HAR dataset containing 432 action classes and a total of 10,000 videos. The videos in ASLAN are collected from YouTube, representing a wide range of actions performed in uncontrolled and diverse real-world environments. These datasets contribute to the field of HAR by offering diverse perspectives, controlled settings, and a large number of action classes and videos for training and evaluation purposes.

In recent years, the demand for larger volumes of data has increased with the rise of NNs and DL models. As a result, new and sizable HAR datasets have emerged to meet these requirements. One notable dataset is **Sport-1M**, which is considered the first large-scale HAR dataset with over 1,000,000 videos and 487 action classes. This dataset focuses on sports activities and is exclusively annotated using YouTube clips. (Karpathy et al., 2014) introduced Sport-1M to facilitate research in HAR and support the training and evaluation of DL models on a vast and diverse collection of sports-related videos.

In addition to Sport-1M, another significant contribution in terms of dataset size is the **YouTube-8M** dataset, developed by (Abu-El-Haija et al., 2016). This dataset is incredibly large, containing over 8,000,000 annotated video clips spanning 4,800 different classes. YouTube-8M is known as one of the largest multi-label HAR datasets, providing a diverse collection of videos from various environments. The dataset offers approximately 500,000 hours of annotated video,

making it a valuable resource for training and evaluating HAR models, particularly those utilising DL approaches.

The **NTU** HAR datasets consist of two multiview RGB_D datasets that were developed specifically for HAR in a controlled indoor environment, focusing on daily activities. The first version, known as *NTU RGBD*, comprises 1,000,000 annotated samples spread across 60 activity classes (Shahroudy et al., 2016). This dataset includes RGB and depth data, providing richer information for activity recognition.

The second version, *NTU RGBD 120*, is an extension of the NTU RGB_D dataset, offering a larger and more diverse collection of annotated videos. It contains approximately 8,000,000 annotated RGB_D videos, covering 120 different activity classes (Jun Liu et al., 2019). This dataset allows for more comprehensive training and evaluation of HAR models, enabling researchers to explore a wider range of activities and achieve higher accuracy in their recognition tasks.

The **Kinetics** HAR dataset is a highly popular and widely used dataset for action recognition. It encompasses several versions with varying numbers of action classes and annotated videos. The first version, *Kinetics 400*, was introduced by (Kay et al., 2017) and contains 400 action classes with approximately 300,000 annotated videos collected from YouTube clips. This dataset serves as a benchmark for evaluating HAR DL models.

Subsequently, *Kinetics 600* was released in 2018 by (Carreira, Noland, Banki-Horvath, et al., 2018), expanding the number of action classes to 600 and including 496,000 annotated videos. The dataset was further extended to *Kinetics 700* in 2019, with 700 action classes and a significantly larger collection of 650,000 annotated videos (Carreira, Noland, Hillier, et al., 2019). These versions of the Kinetics dataset have contributed to advancing the field of HAR and facilitating the development of more accurate and robust action recognition models.

Additionally, *Ava_Kinetics* is a localised HAR dataset derived from Kinetics 700, annotated using the Ava Kinetics annotation protocol (A. Li et al., 2020). It consists of 230,000 annotated clips covering 80 classes. Furthermore, (Smaira et al., 2020) introduced a new edition called *Kinetics_700_2020*, which includes at least 700 videos for each action class, providing a more

comprehensive dataset for training and evaluation purposes. These versions of the Kinetics dataset have played a significant role in advancing the field of action recognition and have become popular choices for researchers in the HAR domain.

Some HAR datasets include an additional viewpoint known as the First Person (FP) or Ego view, which provides a perspective from the human's point of view and is particularly useful for capturing human-object interactions. The *20BN-Something-Something* dataset, introduced by (Goyal et al., 2017), focuses on the FP view of actions. It consists of 100,000 videos depicting 174 action classes. A subsequent version, *20BN-Something-Something-V2*, was published by (Mahdisoltani et al., 2018), featuring the same FP view and action classes but expanded to include 220,000 videos.

Another notable dataset, *Charades-Ego*, presented by (Sigurdsson et al., 2018), provides a multiview HAR dataset with both FP and TP views. It contains 8,000 videos and 68,500 annotated frames across 157 action classes. Including the FP view in these datasets allows for a more immersive and detailed representation of human activities, capturing the perspective of the person involved in the interaction.

These datasets with FP views contribute to advancing the understanding and recognition of human-object interactions, as they provide valuable insights into the actions performed from the human's point of view. They enable the development and evaluation of HAR models that can effectively analyse and interpret FP visual data, leading to improved performance in real-world scenarios.

Created by (Jia et al., 2020), *LEMMA* is a multiview HAR dataset that includes one FP view and two TP views. Baoxiong Jia and his team carefully curated this dataset, and it comprises 1,093 video clips along with 900,000 annotated frames across 641 action classes.

The *HOMAGE* dataset, introduced by (Rai et al., 2021), is a multiview HAR dataset that includes the FP view. It provides both the FP view and at least one TP view for each action. Nishant Rai and colleagues curated the dataset by incorporating 12 sensors, including RGB, infrared, microphone, acceleration, magnet, and more. The RGB modality alone consists of 5,700 annotated videos across 75 action classes, offering a rich resource for researchers to

explore and develop innovative approaches for activity recognition using diverse sensor data.

The *EPIC-KITCHENS-100* dataset, introduced by (Damen, Doughty, Farinella, Furnari, et al., 2022), is an Ego view HAR dataset specifically focused on kitchen actions. It is an extension of the original *EPIC-KITCHENS* dataset, curated by Dima Damen and colleagues, with 149 action classes (Damen, Doughty, Farinella, Fidler, et al., 2018). In EPIC-KITCHENS-100, the dataset consists of 700 videos capturing the FP view, encompassing a wide range of 4,053 action classes with approximately 90,000 instances. This dataset provides a valuable resource for studying activity recognition in kitchen settings, enabling researchers to explore the intricacies of human-object interactions and fine-grained action understanding.

One of the early examples of using a robot to capture data for Human Action Recognition is the *LIRIS* dataset, introduced by (C. Wolf et al., 2012). This dataset is a multiview HAR dataset that includes a Robot View, along with a depth TP view. It consists of 828 videos across ten different action classes. By incorporating the robot view, the LIRIS dataset offers a unique perspective on human activities, allowing researchers to investigate the challenges and benefits of utilising dynamic viewpoints for action recognition tasks. However, the dataset lacks coverage of motion views, and the robot remains static.

Another notable example of a HAR dataset that incorporates robots is the *InHARD* dataset, as presented by (Dallel et al., 2020). In the case of *InHARD*, a robot is utilised during data collection; however, the dataset does not cover the dynamic camera. Also, it is worth noting that all three views (Top, Left, and Right) are static camera. The focus of this dataset is on HRI, making it particularly suitable for research and exploration in the field of HRI.

Table 2.1 presents a comprehensive overview of 42 popular HAR datasets, highlighting their specific characteristics and attributes. The datasets are sorted by year to illustrate the evolution of HAR datasets over time. Each dataset is described by several parameters, including the number of videos, annotation type, number of action classes, fixed views, environment type, camera motion capability, point of view, and accessibility.

Explanation of Key Columns in Table 2.1:

- **An:** Number of annotations.

- **Act:** Number of activity 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).
- **PoV:** Point of view (FP: First Person, TP: Third Person).
- **Mode:** Mode of data collection (RGB, RGBD, etc.).
- **B:** Background (Dy: Dynamic, St: Static).
- **MV:** Multiview availability.
- **AT:** Atomic actions.
- **L:** Localisation of actions.
- **So:** Source of data (C: Created, W: Web, M: Movie, YT: YouTube).
- **U:** Usage of the dataset (T: Training, A: Annotation).
- **Acc:** Accessibility of the dataset.

This table allows for a detailed comparison of existing HAR datasets and provides insights into their suitability for various research needs.

Table 2.1: Overview of popular HAR datasets and their properties

Dataset Name	Year	Video	An	Act	FV	En	Si	Mot	PoV	Mode	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 S	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
multiviewPoint	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

Table 2.1 – continued from previous page

Dataset Name	Year	Video	An	Act	FV	En	Si	Mot	PoV	Mode	B	MV	AT	L	So	U	T	Acc
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

Table 2.1 – continued from previous page

Dataset Name	Year	Video	An	Act	FV	En	Si	Mot	PoV	Mode	B	MV	AT	L	So	U	T	Acc
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
multiview 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

An: Number of Annotations, 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), MV: Multiview, AT: Atomic, L: Localisation, So: Source (C: Created, W: Web, M: Movie, YT: YouTube), U: Usage, T: data preparation type (Tr: Trimmed, A: Annotation), Acc: Accessibility

2.3 Dataset Review Analysis

After assessing the existing HAR datasets as described in table 2.1, the following omissions have been identified:

- Dynamic Perspective (Robot View): Only the "LIRIS" (C. Wolf et al., 2012) and "In-HARD" (Dallel et al., 2020) datasets include a Robot View without motion. Recognising human actions from the perspective of a robot is crucial in the field of HRI, and the presence of motion frames is a prominent feature in such views. While some existing datasets in the motion category of Table 2.1 may include motions in certain videos, they do not specifically focus on providing a separate dataset for motion camera views.

- Top View (Fish Eye View): Fish eye or top views are commonly used in AAL scenarios. However, no HAR dataset with a dedicated top view was found in the assessment.

- Redundancy in Camera Type: To identify redundancy, examine multiview datasets is needed. Most multiview datasets include static cameras positioned at various angles from the sides, and some may include an ego view. No dataset has separate dynamic views, top views, and wall-mounted views.

To address these gaps and include diverse viewpoints, a new multiview dataset named RHM has been introduced. This dataset is designed to fill the existing deficiencies in HAR within the field of HRI in AAL environments. The RHM dataset is intended to contribute significantly to the advancement of HAR research, especially in dynamic perspectives, top views, and redundancy aspects, which will be detailed in the next chapter.

2.4 Related Work Summary

This section summarise the key related works reviewed in Chapters Four, Five, and Six, which encompass key frame extraction, deep learning models, multi-stream networks, and handcraft feature extraction methods for human action recognition (HAR).

2.4.1 Key Frame Extraction

Key frame extraction is a fundamental task in video analysis, aiming to identify the most representative frames that capture the essence of the video content. Several approaches have been proposed in the literature:

- **Shot Detection-Based Methods**: These methods, such as the one proposed by (Ejaz, Tariq, and Baik, 2012), rely on changes in colour histograms to detect key frames based on scene changes. While effective for simple videos, they struggle with more complex scenarios.
- **Clustering-Based Methods**: Techniques like those by (Amiri and Fathy, 2010) use clustering algorithms to group similar frames and select key frames from these clusters. These methods are computationally intensive and sensitive to noise.
- **Motion-Based Methods**: Methods like (Yanming Zhu, K. Li, and Jiang, 2014) combine dimensionality reduction with motion analysis to select key frames, but they may lose local details and are sensitive to content variations.
- **Feature Descriptor-Based Methods**: Approaches using descriptors such as SURF (Yu et al., 2018) and MIESW (W. Li et al., 2020) focus on capturing specific features within frames, offering high-quality summaries but potentially missing broader video context.

This research draw on the work of (W. Li et al., 2020), employing mutual information (MI) to analyse frames within the RHM dataset from both individual and group perspectives.

2.4.2 Deep Learning Models

Deep learning has significantly advanced the field of HAR, with several notable contributions:

- **3D Convolutional Neural Networks (CNNs)**: The introduction of C3D models by (Tran, Bourdev, et al., 2015) and LTC-CNN by (Varol, Laptev, and Schmid, 2017) demonstrates the superiority of 3D CNNs over 2D CNNs in capturing spatiotemporal features.

- **Recurrent Neural Networks (RNNs) and LSTMs**: Models like LRCN (Donahue et al., 2015) combine LSTMs with CNNs to handle sequential data, improving long-term action recognition.

- **Multi-Stream Networks**: The SlowFast networks by (Feichtenhofer, Fan, et al., 2019) leverage dual pathways to capture both slow and fast temporal dynamics, setting new performance benchmarks.

- **Spatiotemporal Fusion**: Techniques such as Spatiotemporal Pyramid Networks by (Y. Wang et al., 2017) and convolutional fusion by (Feichtenhofer, Pinz, and Zisserman, 2016) enhance the integration of spatial and temporal data, improving action recognition robustness.

2.4.3 Multi-Stream Networks

Multi-stream networks excel in recognising human actions by processing multiple data types simultaneously:

- **Depth Integration**: Multi-stream networks that incorporate depth data through specialised sub-networks, as discussed by (Kong and Fu, 2022) and (Gu et al., 2020), enhance the understanding of spatial relationships in videos.

- **Fusion Techniques**: Fusion methods such as early, mid-level, late, and lateral fusion improve accuracy by effectively combining spatial and temporal streams. Studies like (Feichtenhofer, Pinz, and Zisserman, 2016) and (L. Wang, Z. Wang, et al., 2015) demonstrate the effectiveness of these strategies.

- **Cross-Stream Interactions**: Architectures like the one proposed by (Feichtenhofer, Pinz, and Wildes, 2017) use cross-stream residual connections to enable nuanced spatiotemporal feature extraction.

2.4.4 Handcraft Feature Extraction

Despite the rise of deep learning, handcraft features remain relevant for HAR due to their ability to capture temporal details:

- **Local Features**: Techniques such as the 3D SIFT descriptor by (Scovanner, Ali, and Shah, 2007) and dense trajectories by (H. Wang and Schmid, 2013) focus on detailed local motion patterns crucial for complex action recognition.
- **Global Features**: Methods like Motion History Images (MHI) by (Bobick and Davis, 2001) and Histograms of Oriented Gradients (HOG) by (Dalal and Triggs, 2005) provide a comprehensive view of actions, though they can be sensitive to noise.
- **Computational Efficiency**: Research by (Peng et al., 2020) and (Z. Xu, Yang, and Hauptmann, 2015) highlights the need for efficient feature extraction methods to handle large datasets without compromising performance.

Drawing inspiration from prior works on background removal (M. Singh, Basu, and Mandal, 2008), MHI (Bobick and Davis, 2001), and motion capture (Peng et al., 2020), this research proposes novel, computationally efficient feature extraction methods for the dataset.

Chapter 3

Robot House Multiview Dataset

3.1 Introduction

Responding to the deficiencies in HAR datasets within the HRI domain identified in the previous chapter, this work has developed a new multiview dataset named RHM to address these shortcomings.

In the following discussion, this research will detail the key characteristics of the newly created HAR dataset, highlighting its unique strengths and contributions. This dataset is specifically designed to cover a broad range of human activities pertinent to the Robot House (RH) context, taking into account the particular requirements and challenges of HRI in an ambient assisted living setting. It features a variety of subjects, different environmental conditions, and real-life scenarios, ensuring the dataset's comprehensive nature and practical utility. The RHM dataset can be found at this link. Additionally, the skeleton-extracted RHM dataset is presented in (Shahabian Alashti et al., 2023) and can be accessed here.

The creation of this new HAR dataset is aimed at bridging the gap in existing resources and datasets. It is intended to serve as a foundational resource for future developments in HAR and aid in the creation of more precise and dependable DL models. These models are expected to enhance the interactions between robots and humans in ambient assisted living environments, thereby contributing significantly to the field.

3.2 Robot House Multiview (RHM) Dataset

Based on the analysis of existing datasets and by identifying a gap, the RHM is a novel multiview RGB benchmark dataset designed for HAR tasks, with a specific focus on HRI within the domain of ambient assisted living scenarios. Unlike existing datasets, the RHM dataset addresses the identified omissions mentioned in Section 2.3. It consists of four distinct viewpoints, providing a comprehensive perspective for analysing human actions. A frame from each class and viewpoint can be seen in Figure 3.3. Following subsections will provide a detailed description of the RHM dataset, including its characteristics and properties.

Figure 8.1 in the appendix provides a sequential representation of frames captured from the Front View of the RHM dataset, illustrating a diverse range of activities such as bending, carrying objects, and walking. Each row in the figure corresponds to a specific action, with five sample frames shown at regular intervals to depict the progression of each activity. This visual representation underscores the variety and complexity of human actions captured in the dataset, emphasising the challenges involved in recognising and classifying these activities. The Front View perspective offers a clear and consistent view of the subject's movements, making it a key component for training and evaluating action recognition models. The sequential frames provide essential insights into the temporal dynamics of the actions, which are crucial for understanding the effectiveness of the proposed feature extraction techniques discussed in the main text.

The RHM dataset is accessible via this link. Moreover, the skeleton-extracted RHM dataset, as discussed in (Shahabian Alashti et al., 2023), can be found here.

3.2.1 Robot House

The Robot House, a state-of-the-art facility at the University of Hertfordshire, serves as the environment for the RHM dataset collection. It is equipped with various home-based backgrounds under different lighting conditions, both day and night, to simulate real-life scenarios. The facility is designed to support research in ambient assisted living and human-robot interaction. More details about the Robot House can be found at Robot House.

3.2.2 Camera Types and Viewpoints

The RHM dataset incorporates various camera types and viewpoints to capture human actions from different perspectives. The Robot view camera utilises the Fetch robot[1], which is capable of movement and changes its position during the recording of actions. This dynamic camera is mounted on the robot and captures video at a resolution of 640×480 pixels with a frame rate of 30 frames per second (FPS).

Another unique viewpoint is the Top View, which employs a fish-eye camera mounted on the ceiling to provide a bird's-eye perspective of the scene. This static camera captures video at a resolution of 512×486 pixels with a frame rate of 30 FPS.

Additionally, two wall-mounted cameras are utilised to capture static side views of all actions, namely the Back view and Front view. Both of these static cameras are mounted on the walls, capturing video at a resolution of 640×480 pixels with a frame rate of 30 FPS.

The specific details of the cameras and viewpoints used in the RHM dataset are provided in Table 3.1. For a visual representation of the camera positions within the Robot House environment, refer to Figure 3.1.

Table 3.1: RHM viewpoints details

View Name	Motion	Position	Resolution	FR
FrontView	Static	Wall	640×480	30
BackView	Static	Wall	640×480	30
RobotView	Dynamic	Robot	640×480	30
TopView	Static	Ceiling	512×486	30

For the RHM dataset, the camera details are as follows: **Motion**: This indicates whether the camera has dynamic frame capability, capturing movement effectively. **Position**: This denotes the location of each camera, such as front, back, ceiling, or on the robot's head. **Resolution**: This specifies the size of each frame captured by the camera, providing an idea of image clarity and detail. **FR (Frame Rate)**: This indicates the number of frames captured per second by the camera, which is crucial for understanding the fluidity of the video capture.

[1] Fetch Robot

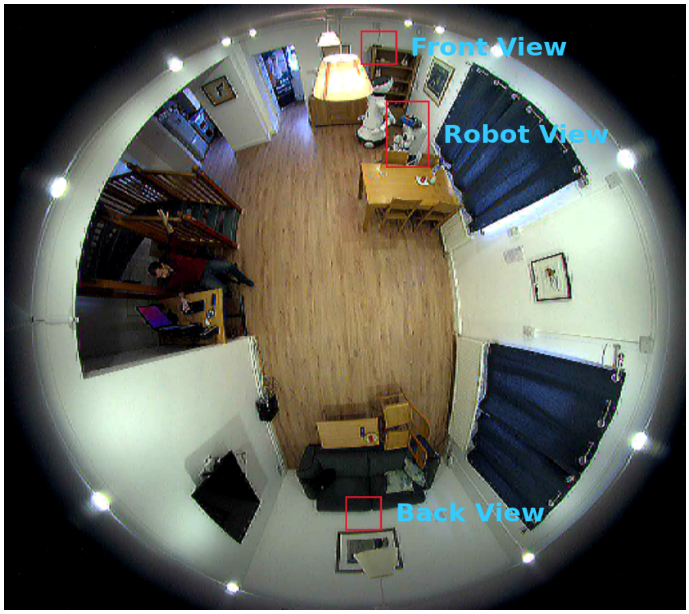


Figure 3.1: Camera positioning at Robot House from Top View
Four cameras were utilised, comprising Front and Back cameras that were mounted on the walls, a ceiling-mounted camera, and a camera mounted on the head of a robot

3.2.3 Participant

Due to the limitations imposed by the COVID-19 pandemic, the RHM dataset was populated with a single participant who performed the actions. Unfortunately, the inclusion of multiple participants was not feasible at the time. However, for future versions of the dataset, this work plans to address this limitation by involving external participants and incorporating multiple subjects into the dataset. This will enhance the diversity and generalisability of the dataset for a broader range of scenarios and applications.

3.2.4 Content

The activity classes in the RHM dataset were selected based on the work of (Bedaf et al., 2014), which focuses on identifying important daily activities for individuals living independently. The research emphasises the potential value of companion robots and ambient-assistive systems in detecting and supporting these activities. The dataset includes a comprehensive list of activities, which can be found in Figure 3.2. These activities represent key tasks and behaviours that are relevant to the home caring domain and serve as the basis for activity recognition and analysis

in the dataset.

The chosen activities were selected to cover a wide range of common daily tasks that are crucial for independent living. These activities include walking, drinking, carrying objects, climbing stairs, opening and closing cans, and more. Compared to other datasets, the RHM dataset offers a richer variety of activities, providing a more comprehensive resource for training and evaluating HAR models. The activities were chosen based on their relevance to everyday life and the ability to be clearly distinguished from one another, ensuring that the dataset is both practical and challenging for HAR tasks.

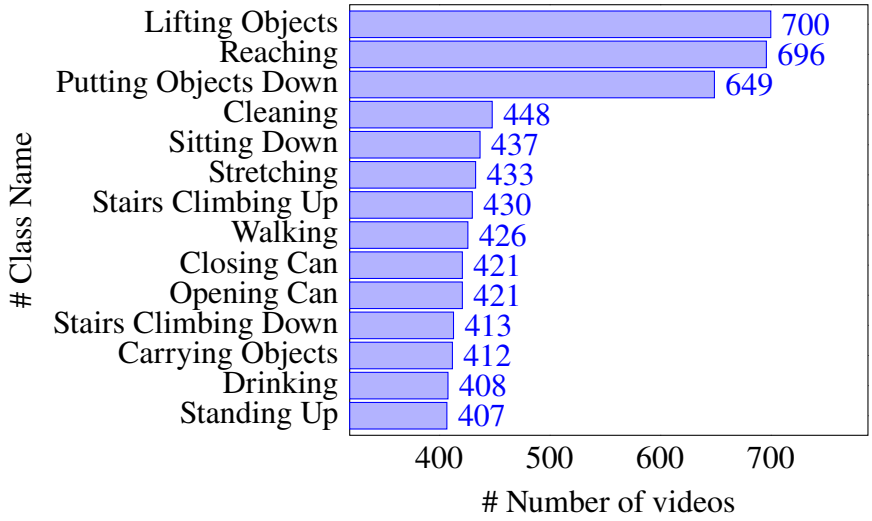


Figure 3.2: RHM Videos number in each class-view
The figure displays a comprehensive list of all the classes included in the dataset, along with the corresponding number of video samples available for each class in every view.

3.2.5 Statistics

The RHM dataset consists of 14 activity classes, as indicated in Figure 3.2. Each class is represented by a varying number of videos, ranging from 407 to 700, across different viewpoints. In total, there are 6701 videos for each individual viewpoint, resulting in a combined total of 26804 videos across all viewpoints. The duration of each video clip in the dataset varies between 1 to 5 seconds, capturing key moments of the activities performed. These statistics provide an overview of the dataset's size and distribution of videos among the different classes and viewpoints.

3.2.6 Training/Validation/Testing

To facilitate the evaluation of models and ensure unbiased performance assessment, the RHM dataset is divided into three subsets: training, testing, and validation. Each subset is partitioned separately for each view, maintaining consistency across the dataset.

The training set comprises 65% of the total videos in each view, providing a substantial amount of data for model training. The testing set consists of 20% of the videos, used for evaluating the performance of trained models on unseen data. Lastly, the validation set, which accounts for 15% of the videos, serves as an additional benchmark for fine-tuning and hyper parameter optimisation during the model development process.

Table 3.2 presents the specific number of videos allocated to each subset for both individual views and the combined dataset. This partitioning strategy ensures a balanced distribution of videos across the training, testing, and validation sets, enabling robust model evaluation and comparison.

Table 3.2: Number of videos in each View/Split

	Train	Validation	Test
Each View	4278	1076	1347
All Views	17112	4304	5388

The table provides detailed information on the distribution of data in the train, test, and validation splits for each view, as well as for the combination of all views.

3.2.7 Naming Protocol

To maintain consistency and facilitate easy identification of videos within the RHM dataset, a specific naming protocol is followed for each video clip. The naming convention is as follows:

ClassName_ViewName_clipNumber.avi

Each video clip is assigned a unique name based on the action class, view name, and clip number. For example, the clip named Drinking_RobotView_103.avi corresponds to clip number 103 of the action class 'drinking' from the Robot's viewpoint. This naming protocol enables

the efficient organisation and identification of videos within the dataset, making it easier for researchers to locate specific clips for analysis and model development purposes.

3.2.8 Time Synchronising

To ensure consistency and facilitate cross-view analysis, all clips within the RHM dataset that share the same action class and clip number are time-synchronised. For example, clips such as Reaching_FrontView_320.avi, Reaching_BackView_320.avi, Reaching_OmniView_320.avi, and Reaching_RobotView_320.avi are synchronised. This synchronisation ensures that corresponding clips from different viewpoints capture the same temporal sequence of actions, allowing for meaningful comparisons and analysis across views.

3.2.9 Data Pre-processing

The RHM dataset was provided as raw data, with minimal pre-processing to ensure the integrity and authenticity of the recorded actions. The videos were trimmed and classified into folders based on the specific action they depicted. No additional pre-processing steps, such as noise reduction or normalisation, were performed. This approach maintains the dataset's versatility, allowing researchers to apply their own pre-processing techniques as needed for their specific use cases.

3.2.10 Object-Manipulation Scenarios

Note that in the RHM dataset, most of the actions involve interactions with various objects, such as opening or closing a can, pouring a drink, or picking up an item. The primary focus of this work is on developing a HAR model that recognises and classifies these actions based on the sequence of movements performed by the individual, rather than on the specific details or state of the objects involved. This work aims to create a HAR model that is not intended to detect or verify the precise status of objects but rather to identify the overall pattern and motion associated with each action. For instance, when a person opens or closes a can, the model recognises the

Figure 3.3: A frame of all classes and all views of RHM dataset.

(e) Front View (f) Back View (g) Top View (h) Robot View

Figure 3.3: Continue a frame of all classes and all views of RHM dataset.

characteristic movements of the hand and arm that are typical of such actions. However, the model does not concern itself with whether the can is fully closed or opened. This approach allows the model to generalise and recognise actions across different scenarios, even when slight variations in object appearance or position occur. By prioritising the recognition of actions over the detailed examination of objects, the model effectively fulfils its role in applications like human-robot interaction and ambient assisted living, where understanding human behaviour and activities is more critical than monitoring specific object states. This approach ensures that the model remains versatile and capable of identifying a wide range of actions, regardless of the particular objects involved.

3.3 RHM Dataset Contribution

The RHM dataset has been extensively documented in various research publications. These include:

- "RHM: Robot House Multi-view Human Activity Recognition Dataset," (Abadi et al., 2023).
- "RHM-HAR-SK: A multi-view dataset with skeleton data for ambient assisted living research," (Alashti et al., 2023b).
- "Robot house human activity recognition dataset," (Abadi et al., 2021).

The dataset is accessible to researchers and practitioners through the following link: RHM Dataset.

3.4 Chapter Summary

This chapter conducted a comprehensive review of the most prominent HAR datasets and identified significant gaps in the context of HRI. To address these limitations, this research introduced a new RGB-based HAR dataset called RHM. The primary objective of the RHM

dataset was to encompass three crucial features that were lacking in existing datasets: the dynamic perspective (Robot View), the top view (Fish Eye View), and redundancy in multiple views.

The RHM dataset consists of four distinct viewpoints: Front (static), Back (static), Top (fish-eye), and Robot (dynamic). Each viewpoint contains a separate collection of 6,701 videos, resulting in a total of 26,804 videos across all views. The dataset comprises 14 distinct action classes, and clips with the same class and number are carefully time-synchronised across different viewpoints.

By introducing the RHM dataset, this work aimed to fill the existing gaps in HAR datasets within the HRI domain. The inclusion of dynamic perspectives, top views, and redundancy in multiple viewpoints enables researchers to tackle more complex and realistic human action recognition tasks. This research believe that the RHM dataset, with its diverse range of views and synchronised clips, will serve as a valuable resource for advancing research in the field of HAR in the context of HRI.

Chapter 4

Robot House Multiview Dataset Analysis

4.1 Introduction

DL model comparison is a commonly used strategy for exploring new datasets. However, when dealing with multiview datasets, an alternative method for comprehensive data analysis becomes imperative. Key frame selection emerges as a significant technique for feature extraction from multiview video content. Notably, the use of Mutual Information (MI) represents one of the most recent advancements in key frame selection methodologies. In this context, this work employs MI for the first time to analyse the multiview HAR dataset.

To gain a clear understanding of the impact of the robot viewpoint and its dynamic effects on the results of HAR models, this chapter will conduct an in-depth analysis of the RHM dataset. This research will specifically use the MI technique, based on information theory, and test various DL models with the RHM dataset. By employing these methods, the goal is to extract valuable insights and develop a thorough understanding of the inherent characteristics and patterns within the RHM dataset.

4.2 Related Work

4.2.1 Key Frame Extraction Review

A **keyframe** is a critical frame in a video sequence that represents significant content or changes within the video. Keyframes are typically selected to summarise the video effectively, capturing the most important scenes or actions while reducing the amount of data needed for storage or analysis. The extraction of keyframes is a fundamental process in video analysis, as it enables efficient video browsing, indexing, and summarising by selecting the minimal number of frames required to convey the essential visual information. Keyframes are crucial in various applications, such as video summarising, segmentation, and action recognition, as they help in reducing redundancy and computational load. This subsection reviews several keyframe extraction methods, categorised based on their underlying techniques, such as shot detection, clustering, motion features, and specialised feature descriptors, and discusses their relevance and applicability in different video analysis tasks.

The increasing amount of video content from various sources such as security cameras, data collections, and smartphones has made analysing these videos quite challenging. These difficulties are apparent in tasks like video search (Antani, Kasturi, and Jain, 2002), dividing videos into segments (W. Wang et al., 2015), and recognising actions in videos. Manually picking important parts from these videos is both time-consuming and hard work. To tackle these issues, specialised fields like video summarising (Mei et al., 2015), condensation (J. Zhu et al., 2014), and skimming (L. Zhang et al., 2016) have arisen. A critical part of these fields is *key frame extraction*, which offers ways to identify essential video content, automatically.

The purpose of video key frame extraction is to use as few video frames as possible to represent as much video content as possible, reduce redundant video frames, and reduce the amount of computation, so as to facilitate quick browsing, content summarising, indexing, and retrieval of videos (Yao, 2022).

One early method for key frame extraction uses shot detection as suggested by (Ejaz, Tariq, and Baik, 2012). This method analyses colour histograms to identify key frames based on

changes in scenes. It works particularly well for videos with simple content and few scene changes. Another research by (Hannane, Elboushaki, and Afdel, 2018) developed a system for segment identification and video condensation. This system works for various types of videos like films, documentaries, and sports. It uses a modified mean shift algorithm and specific orientation features to select key frames, termed as mean shift-based keyframes for video summarization (MSKVS). However, further research is needed to improve this system for real-time use. In action summarising, a method by (Meghdadi and Irani, 2013) creates a summary using static images that match the action frames. However, this technique may struggle if multiple subjects or obstacles are present in the video. In the area of visual positioning, key frames based on shot detection are used to build an offline database containing location information. Lastly, an algorithm by (L. Ma et al., 2018) sets initial key frames based on fixed-size clusters rather than content similarity. This could compromise the accuracy of key frame extraction.

The second main type of key frame extraction method relies on clustering techniques, where frames are grouped based on how similar they are. In one research by (Amiri and Fathy, 2010), sparse coding is combined with k-means clustering to select key frames. However, the strict rules and complex parameters needed make this approach challenging to use. Another research by (Zhou, Qiao, and Xiang, 2018) treats video summarising as a decision-making process, using k-medoids to pick cluster centres as key frames. Although this method is unsupervised, it is sensitive to noisy or inconsistent data and best suited for smaller datasets due to its computational demands. (Yin, Thapliya, and Zimmermann, 2016) proposes an algorithm that focuses on the relationships between elements, using a technique called Semantic Tree (SeTree). This method is quite comprehensive, as it considers visual attributes, text, and user preferences to pick key frames. However, its complexity makes it less suitable for real-time applications. In conclusion, clustering methods for key frame extraction mainly use unsupervised learning. They are sensitive to data quality and can sometimes miss the time-based context of the original video. These methods are also computationally demanding, making them better suited for smaller datasets.

The third main type of key frame extraction focuses on using motion features. In a research by (Yanming Zhu, K. Li, and Jiang, 2014), the method involves reducing the dimensions of the

original data and then clustering the motion information. This forms the basis for selecting key frames. However, this combination of dimensionality reduction and motion features might result in the loss of local details, potentially leading to inaccurate results. Another research by (Gao et al., 2009) presents a video summarising method that combines Optical Flow Tensor (OFT) with Hidden Markov Model (HMM). This combination effectively captures the video's dynamic motion. However, the method is very sensitive to the video content and works best for videos with subtle motion changes. It operates at the pixel level, measuring changes in grayscale values between frames. In conclusion, using optical flow for key frame extraction has its limitations. It requires the video's brightness and spatial features to remain almost constant, which restricts the types of videos it can effectively analyse.

The fourth main type of key frame extraction method centres on specialised feature descriptors designed for broad use. In one research by (Yu et al., 2018), the Speeded Up Robust Features (SURF) descriptor is used to identify local points in frames. These points are then analysed in a sequence using a sliding window technique to extract key frames. Another research by (Rao and Das, 2012) employs contour wave transformations to calculate energy and standard deviation for each sub-band. These metrics are used to form a feature vector, which then helps in extracting key frames for each shot. A different approach is taken in the research by (W. Li et al., 2020), which introduces the Mutual Information and Entropy-based adaptive Sliding Window (MIESW) algorithm. This method is tailored for summarising gesture videos. It starts by resizing video frames and then uses inter-frame MI to adaptively adjust a sliding window. Finally, SURF analysis is applied to remove any redundant frames. The method is shown to produce high-quality key frame summaries. However, it's worth noting that methods focusing solely on one feature descriptor may not capture all the nuances of complex video content. This is particularly true for videos with intricate or elaborate shots.

Overview of Key Frame Selection Methods

Key frame selection is a fundamental process in various video processing tasks such as video summarising, action recognition, and video compression. The objective is to identify a subset

of frames that effectively represent the entire video, capturing the most significant events while minimising redundancy. Several approaches have been developed for key frame selection, each with its own strengths and applications.

Shot Boundary Detection methods involve dividing a video into segments or "shots," which are sequences of frames captured continuously without interruption. Key frames are selected by identifying the boundaries between shots, typically using changes in visual features such as colour histograms, edge detection, or pixel differences (Ejaz, Tariq, and Baik, 2012; Hannane, Elboushaki, and Afdel, 2018).

Clustering-Based Methods employ algorithms like K-means or K-medoids to group similar frames together based on visual features. The centroid of each cluster, representing the most typical frame within that group, is chosen as the key frame (Amiri and Fathy, 2010; Zhou, Qiao, and Xiang, 2018).

Motion Analysis methods select key frames based on motion features, such as optical flow, which measures the motion between consecutive frames. Key frames are chosen where significant motion occurs or where motion patterns change (Yanming Zhu, K. Li, and Jiang, 2014; Gao et al., 2009).

Entropy-Based Methods utilise information theory to select frames that maximise the information content, measured as entropy, between frames. Higher entropy indicates more significant differences, leading to the selection of frames that capture critical changes in the scene (W. Li et al., 2020; Rao and Das, 2012).

Deep Learning Approaches involve using CNNs and other deep learning models to learn and identify key frames directly from data. These models can capture complex patterns and dependencies that traditional methods might miss (Peng et al., 2020; H. Wang and Schmid, 2013).

By reviewing these methods, this section provides a foundation that justifies the selection or development of key frame selection techniques in this thesis. This overview not only supports the methodologies discussed in subsequent chapters but also contextualises the contributions made in this work within the broader landscape of key frame selection research.

This research takes cues from the work of (W. Li et al., 2020) and use MI to analyse frames within the RHM dataset for both individual and group perspectives.

4.2.2 Deep Model Review

In recent years, deep learning has risen to prominence owing to its robust capabilities in feature engineering. Consequently, the domain of HAR has progressively transitioned towards the utilisation of DNN. Temporal modelling and convolutional operations serve as key elements in achieving efficacious action recognition. DL-based methods for HAR can generally be divided into supervised and unsupervised learning paradigms. Within the supervised learning framework, two notable sub-classes are Spatiotemporal Networks and Multiple Stream Networks, as elaborated upon by (Herath, Harandi, and Porikli, 2017).

The inclusion of Three Dimension Convolutional Neural Networks (3DCNN) plays a crucial role in capturing temporal information for action recognition. In the work conducted by (S. Ji et al., 2012), the researchers introduced a CNN-based approach that incorporated 3D convolutions between neighbouring frames, enabling the extraction of both spatial and temporal features.

Another notable contribution in this area is the introduction of a deep architecture called Convolutional Three Dimensions (C3D) by Tran (Tran, Bourdev, et al., 2015). The research presents an innovative yet uncomplicated technique for the extraction of spatiotemporal features through the use of 3DCNN, which are trained on a comprehensive supervised video dataset. The research is distinguished by three pivotal conclusions. Firstly, it confirms that 3DCNN are superior to Two Dimension Convolutional Neural Networks (2DCNN) for the purpose of spatiotemporal feature extraction, addressing an essential issue in this domain. Secondly, it reveals that a consistent architecture featuring small 3x3x3 convolutional kernels throughout all layers is among the most effective setups for 3DCNN, providing valuable guidance for future architectural decisions. Thirdly, the research introduces a novel set of attributes termed C3D, which, in conjunction with a basic linear classifier, surpass established benchmarks in four distinct evaluations and hold their own in two others. Moreover, C3D attributes are both concise and computationally economical, achieving a 52.8% success rate on the UCF101 dataset with

a mere 10 dimensions. Due to the rapid inference speed of ConvNets, these features are also highly practical for real-time applications.

In another work, (Varol, Laptev, and Schmid, 2017) tackles a fundamental shortcoming in the realm of action recognition, specifically the limitations of brief temporal frame evaluation, by incorporating Long-Term Temporal Convolution (LTC-CNN) into CNN. This novel methodology is engineered to fully grasp the extended time duration of human activities, which often unfold over multiple seconds and exhibit distinct spatiotemporal configurations. The findings reveal that the integration of LTC-CNN architectures markedly elevates the precision of action identification. Additionally, the investigation examines the effects of diverse low-level attributes, such as unprocessed pixel data and optical flow vectors, emphasising the vital importance of precise OFT calculations for reliable action representation. The approach achieves unparalleled results on two rigorous benchmarks: it registers a 92.7% accuracy level on UCF101 and a 67.2% accuracy level on HMDB51.

To incorporate temporal information, several studies have employed Recurrent Neural Networks (RNN) (Robinson and Fallside, 1988) and Long Short-Term Memory Networks (LSTM) networks (Hochreiter and Schmidhuber, 1997). (Donahue et al., 2015) introduced a novel model called Long-term recurrent convolutional networks (LRCN) Networks. The paper presents a pioneering investigation into the synergistic use of LSTM and CNN for sequence-oriented tasks. It introduces an innovative "temporally deep" architecture that is end-to-end trainable, effectively addressing both spatial and temporal data complexities. This marks a significant leap over existing models that either have fixed spatiotemporal receptive fields or rely on basic temporal averaging. The architecture's "doubly deep" nature allows for layered composition in both spatial and temporal dimensions, offering advantages for complex target concepts and scenarios with limited training data. Additionally, the integration of nonlinearities equips the model to learn long-term dependencies, enhancing its versatility for tasks that require variable-length inputs and outputs. Empirical evidence strongly supports the model's efficacy, demonstrating that the joint training of its temporal and convolutional components outperforms existing state-of-the-art models.

Tran et al. (Tran, H. Wang, et al., 2018) presented a comprehensive paper that discusses various spatiotemporal deep models for action recognition. The paper innovatively tackles spatiotemporal convolution by unveiling two new architectures: Mixed Convolution (MC) and "R(2+1)D" convolutional blocks. Through empirical validation on benchmark datasets like Kinetics and Sports-1M, the paper demonstrates the efficacy of these models in action recognition. The MC model, which uses early-layer 3DCNN followed by top-layer 2DCNN, achieves a 3-4% gain in clip-level accuracy over traditional 2D ResNets while matching the performance of more computationally intensive 3D ResNets. The "R(2+1)D" model, which factorises 3DCNN into separate spatial and temporal operations, shows even greater promise, outperforming MC and full 3D models by up to 4.7% in various settings. This model also surpasses traditional 2D ResNets by significant margins, up to 9.8% in some cases. In terms of computational efficiency, the "R(2+1)D" model strikes a balance between performance and computational cost, outperforming even state-of-the-art models in comparative tests.

(He et al., 2019) offers a significant contribution to the understudied area of spatial-temporal modelling in videos, despite existing advancements in deep learning for static images. The authors propose a novel Spatial-Temporal Network (StNet) that deviates from traditional CNN+RNN or 3DCNN approaches. Ingeniously, the StNet architecture stacks N consecutive video frames into a 'super-image' and uses 2D convolution to capture localised spatiotemporal relationships. For global modeling, a unique 'temporal Xception block' is introduced, employing separate channel-wise and temporal-wise convolutions. Empirical validation on the Kinetics dataset is compelling, indicating that StNet surpasses multiple state-of-the-art models in action recognition while maintaining an optimal trade-off between model complexity and recognition accuracy. The paper further validates StNet's robustness by showcasing its strong transfer learning performance on the UCF101 dataset, with mean class accuracies reaching up to 95.7%.

In different research conducted by (Feichtenhofer, Pinz, and Zisserman, 2016), a consecutive spatial fusion function was designed to create a channel at the corresponding pixel. The paper conducts an exhaustive analysis of ConvNet architectures for HAR in videos, with a particular

emphasis on the best practices for integrating spatial and temporal data. It presents a new Convolutional Networks (ConvNets) architecture informed by several key insights: the advantages of fusing information at the convolution layer level, the suitability of the last convolution layer for spatial fusion, and the performance gains achieved through pooling over spatiotemporal neighbourhoods. Empirical tests reveal that the proposed convolutional fusion strategy is highly effective, achieving an average accuracy of 85.94% on the first split of the UCF101 dataset. This not only surpasses other methods but also benefits from a shorter training period when the convolution kernel is initialised with identity matrices. The research further indicates that fusing at the ReLU5 layer marginally outperforms fusing at Fully Connected layers, likely due to the better retention of spatial correspondences. In comparison to existing state-of-the-art techniques, the proposed architecture demonstrates a performance improvement ranging from 3% to 6% on both the UCF101 and HMDB51 datasets.

(Z. Zhang et al., 2020) introduces the Spatial-Temporal Dual-Attention Network (STDAN), an innovative architecture for HAR in videos. This architecture uniquely combines Convolutional Long Short-Term Memory Networks (Conv-LSTM) and Fully-Connected LSTM with dual-attention mechanisms. Unlike prior models that mainly rely on high-level fully connected features, STDAN utilises both convolutional and fully connected layers to enhance video representation. The architecture incorporates a Temporal Attention Module (TAM) and a Joint SpatialTemporal Attention Module (JSTAM), both of which are further refined using Principal Component Analysis (PCA). Experimental evaluations reveal that STDAN surpasses existing state-of-the-art models on multiple benchmarks, achieving accuracies of 98.2% on UCF11, 56.5% on HMDB51, and 87.4% on UCF101. The paper also offers a detailed comparative analysis with other models, emphasising the efficacy of its dual-attention mechanisms and PCA-based optimisation.

The SlowFast DL model, introduced by (Feichtenhofer, Fan, et al., 2019), is a prominent contribution to the field of HAR. The paper unveils SlowFast networks, a two-pathway architecture designed for video recognition, which sets new performance benchmarks on multiple datasets including Kinetics-400, Kinetics-600, and Charades. The Slow pathway focuses on spatial

semantics at a reduced frame rate, whereas the Fast pathway, engineered for computational efficiency, captures motion at a high temporal resolution. Impressively, the SlowFast model exceeds the prior state-of-the-art in Top-1 accuracy on the Kinetics-400 dataset by 2.1%, even without the benefit of ImageNet pre-training. The architecture is also noted for its computational frugality, requiring fewer temporal clips during inference and achieving a low computational cost of 36.1 GFLOPs per space-time view. On the Kinetics-600 dataset, the model attains a Top-1 accuracy of 81.8%, outdoing the winner of the most recent ActivityNet Challenge 2018. In the case of the Charades dataset, the SlowFast model significantly elevates the mAP to 42.1, which further rises to 42.5 with the addition of Non-local layers. When pre-trained on Kinetics-600, the mAP jumps to 45.2, surpassing the previous best while being more computationally economical.

In another work by (Feichtenhofer, 2020), they present X3D, an innovative video network architecture designed to optimise the balance between accuracy and computational efficiency by expanding along four dimensions—space, time, width, and depth. Empirical tests reveal that X3D achieves exceptional efficiency without compromising on performance. For example, on the Kinetics-400 dataset, the X3D-XL model nearly equals the Top-1 accuracy of the leading SlowFast model but does so with 4.8× fewer FLOPs and 5.5× fewer parameters. Similar efficiency advantages are observed on the Kinetics-600 and Charades datasets, where X3D models either outperform or match current state-of-the-art models while demanding substantially fewer computational resources. Specifically, X3D-XL registers an average Top-1/5 accuracy of 85.3% on the Kinetics-400 test set. On the Charades dataset, it exceeds the previous best-performing system, SlowFast, by up to 1.9 mAP, while requiring up to 5.5× fewer parameters and 4.8× fewer FLOPs. Overall, the results affirm X3D's ability to deliver top-tier performance in video classification and detection tasks while maintaining high computational efficiency.

In DL, the performance of a model is often evaluated using various metrics such as loss functions, accuracy, and precision. These metrics are crucial in assessing how well the model performs on a given dataset. In this work, the primary focus is on validating the performance of the developed models using the RHM dataset, ensuring that the models are robust and capable of accurately recognising human activities within this dataset. Additionally, to enhance the model's

effectiveness, a multi-view fusion method is employed. This approach integrates information from different viewpoints to improve the accuracy and generalisability of the model. The fusion of multiple views allows the model to capture more comprehensive features from the data, which is elaborated upon in the subsequent sections. This strategy is particularly valuable in the context of HAR, where diverse perspectives can significantly contribute to the overall understanding and classification of actions.

4.3 RHM Dataset Analysis

When exploring fusion techniques for multiple views, it is crucial to take into account both MI and the performance of individual views using benchmark models. Before performing Dual-stream fusion, it is important to assess the MI between the views to determine their level of correspondence and relevance. Additionally, evaluating the performance of each view independently using benchmark models can provide valuable insights into their capabilities and strengths.

By considering MI and single-view performance, researchers can make informed decisions regarding the fusion of multiple views. This comprehensive analysis helps in understanding the interplay between views, identifying complementary information, and ensuring that the fusion process enhances the overall performance and effectiveness of the system.

While template matching (Brunelli, 2009), Least Square Error (Lucas and Kanade, 1981), and pair-wise comparison methods (Davidson, 1959) have their merits, they were tested in the early stages of this research. However, these methods did not yield significant results in capturing the complex temporal dynamics required for robust HAR in a robot-centric environment. Due to their limitations in handling the dynamic and diverse nature of the data, they were ultimately not included in this thesis. The methods chosen in this study are better suited to address the challenges posed by the specific goals of ensuring scalability and leveraging deep learning for effective HAR.

4.3.1 Mutual Information Analysis Methodology

MI is a fundamental concept in information theory that quantifies the amount of information one random variable contains about another. In the context of video analysis, MI is used to measure the degree of dependency or similarity between consecutive frames within a video sequence. The underlying principle of MI is to evaluate how much knowledge of one frame reduces the uncertainty about the next frame. This is particularly useful in assessing the temporal redundancy in videos. MI is computed by analysing the joint probability distribution of the pixel intensities across two frames and comparing it with the individual (marginal) distributions of each frame. The higher the MI value, the greater the redundancy between the frames, indicating that they share similar information. Conversely, lower MI values suggest less redundancy, indicating more variation between the frames. In this work, MI is leveraged to compare the dynamic (Robot View) and static views within the dataset, helping to quantify the differences in temporal coherence and validate the hypothesis that the dynamic view exhibits less redundancy between frames.

The pipeline for applying the MI analysis in this research begins with the extraction of consecutive frames from each video sequence within the dataset. These frames are then processed to calculate their joint probability distribution, which is necessary for determining the MI between each pair of adjacent frames. Specifically, the pixel intensities of the frames are analysed using a 2D histogram method, where the joint distribution $P(x, y)$ is obtained. This distribution is then compared with the individual probability distributions $P(x)$ and $P(y)$ of the frames to compute the MI values. The MI calculation is repeated for all consecutive frame pairs within each video, and the results are aggregated to obtain the overall MI for the entire sequence. To ensure comparability across videos of different lengths, the total MI is normalised by the number of frame pairs, yielding the average MI value Ave_m. This pipeline allows for a systematic comparison of the temporal redundancy and similarity across the different views (static and dynamic) in the dataset, providing critical insights into the distinctive characteristics of each view.

This work hypothesises that the dynamic view (Robot View) of the dataset exhibits less

redundancy and similarity between consecutive frames compared to the three static views. To validate this hypothesis, a novel metric method inspired by the work of (Guo et al., 2016) is proposed to compare the different views in the dataset using mutual information.

Mutual information is a measure of the statistical dependency between two random variables, in this case, the frames within each view. By quantifying the MI (Cover, 1999) between consecutive frames, the level of redundancy and similarity present in each view can be assessed.

The proposed method involves calculating the MI between consecutive frames in both the static and dynamic views. By comparing these values across the different views, the degree of redundancy and similarity can be evaluated and compared. This analysis will provide insights into the uniqueness and distinctiveness of the dynamic view, supporting the hypothesis.

The MI $I(X;Y)$ quantifies the statistical dependence between two variables, X and Y, with a joint probability distribution $P(X,Y)$ (Cover, 1999). It is computed using the following formula:

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \tag{4.1}$$

In this research MI (Cover, 1999) to a video comprising m frames applies for the analysis. The MI calculation between consecutive frames f_i and $f_{(i+1)}$ is given by:

$$MI(f_i, f_{i+1}) = \sum_{i=1}^{m} P(f_i, f_{i+1}) \log \frac{P(f_i, f_{i+1})}{P(f_i)P(f_{i+1})} \tag{4.2}$$

This equation allows us to evaluate the MI between successive frames in the video and capture the degree of statistical dependency or similarity between them.

The $MI(f_i, f_m)$ represents the sum of all MI values between adjacent frames in a video. Here, f_1 denotes the first frame of the video, and f_m represents the last frame. By summing the MI values between each pair of adjacent frames from f_i to f_m, the overall MI for the video will obtain which is capturing the cumulative statistical dependency or similarity between consecutive frames.

To obtain the average MI between adjacent frames in a video, the calculated MI value $MI(f_i, f_m)$ will divide by $m - 1$, where m represents the number of frames in the video. This

yields the average MI value Ave_m defined as:

$$Ave_m = \frac{1}{m-1} MI(f_1, f_m) \tag{4.3}$$

Here, Ave_m represents the average MI between each pair of adjacent frames in the video. By dividing the total MI by $m - 1$, the MI values are normalised to account for the varying lengths of videos, providing a representative average measure of statistical dependence or similarity between consecutive frames.

In the context of Mutual information analysis, the probability of each frame refers to the likelihood of a specific frame occurring within a sequence, considering the distribution of pixel intensities or other features across the video. Mutual information is a measure that quantifies the amount of information obtained about one random variable (in this case, a frame or its features) through another random variable (such as the preceding or subsequent frame). The probability distribution of each frame is derived by analysing the frequency and distribution of pixel values or feature occurrences across the entire video or dataset. These probabilities are then used to calculate the joint probability distribution between pairs of frames, which is essential for determining the mutual information. The higher the mutual information between two frames, the more predictable one frame is given the other, indicating a strong dependency between them. This concept is crucial for understanding the redundancy or uniqueness of frames in a video, which in turn can inform decisions about key frame selection, data compression, or action recognition in video sequences.

Generally, Mutual information works by quantifying the amount of information obtained about one random variable through another random variable. In this analysis, the Python library Scikit-learn was used to perform MI calculations. The joint probability was calculated using the histogram method, where the joint probability distribution $P(x, y)$ was obtained by creating a 2D histogram of pixel intensities of consecutive frames.

4.3.2 Deep Model Analysis

An additional method to compare the viewpoints in the RHM dataset is to employ benchmark models. By applying these models to the videos captured from different viewpoints, the performance and effectiveness of each viewpoint can be assessed in various tasks such as HAR, object detection, or other relevant tasks.

Benchmark models, widely recognised in the field of computer vision, provide a standardised evaluation framework for assessing the performance of different approaches. These models are typically trained on large-scale datasets and have been validated on various challenging tasks. By evaluating the performance of these models on videos from different viewpoints, insights can be gained into the strengths and limitations of each viewpoint.

Through benchmark models, the accuracy, precision, recall, and other performance metrics achieved by each viewpoint in different tasks can be compared. This analysis provides valuable information about the suitability of each viewpoint for specific applications and highlights the effectiveness of each viewpoint in capturing relevant information.

By combining the results of benchmark models with the analysis of MI, a comprehensive understanding of the differences and characteristics of each viewpoint in the RHM dataset can be obtained. This multi-faceted approach enhances the ability to make informed decisions regarding viewpoint selection for specific tasks and applications.

In the comparative analysis of the viewpoints in the RHM dataset, several benchmark models have been utilised, including C3D (Tran, Bourdev, et al., 2015), R(2+1)D (Tran, H. Wang, et al., 2018), R3D (Tran, H. Wang, et al., 2018), and SlowFast (Feichtenhofer, Fan, et al., 2019) models. These models are widely recognised and used in the field of video understanding and HAR.

The C3D model, for instance, leverages 3D CNNs to extract spatiotemporal features from videos. The R(2+1)D and R3D models extend this concept further by incorporating residual connections and deeper network architectures. The SlowFast model introduces a Dual-stream architecture with separate pathways for spatial and temporal information, enabling it to capture both fine-grained details and long-term motion cues.

CNNs can be categorised based on the dimensions of the convolutions they perform, namely 1d, 2D, and 3D CNNs. A 1D CNN is typically used for processing sequential data, where the convolution operation is applied along a single spatial dimension, such as time-series data or text sequences. In contrast, a 2D CNN operates on two spatial dimensions—height and width—making it ideal for processing images, where the convolutional filters move across the image's spatial dimensions to capture spatial hierarchies and features. Finally, a 3D CNN extends this concept by performing convolutions across three dimensions—height, width, and depth (often the temporal dimension in video data). This allows 3D CNNs to capture spatiotemporal features, making them particularly effective for tasks like video analysis, where both spatial and temporal information are critical for understanding the content. Each type of CNN is suited to different kinds of data and tasks, depending on the dimensionality of the input data.

In this research, it is essential to distinguish between the terms "3D-CNN" and "C3D". A 3D-CNN refers to a general class of convolutional neural networks that perform 3-dimensional convolutions, meaning that the convolutional filters move through three dimensions (height, width, and depth) of the input data. This allows the network to capture spatial and temporal information simultaneously, making it particularly suitable for tasks involving video data. On the other hand, C3D is a specific architecture that employs 3D convolutions. It was introduced by Tran et al. and is designed specifically for learning spatiotemporal features from video clips. While 3D-CNN is a broad concept referring to any neural network using 3D convolutions, C3D is a particular implementation of such a network, optimised for video action recognition tasks.

4.4 Experiment

4.4.1 Experiment Conditions

The RHM dataset was prepared under varying conditions to ensure robustness. These variations include differences in clothing, lighting (day and night), and other environmental factors. This diversity helps in testing the models under different real-world scenarios, contributing to the high accuracy reported.

4.4.2 Hyperparameters and Experiment Setup

Hyperparameters were selected based on the related literature. The setup was as follows:

- **nEpochs**: 500
- **resume_epoch**: 0
- **useTest**: True
- **nTestInterval**: 20
- **snapshot**: 50
- **lr**: 1e-3
- **criterion**: CrossEntropyLoss
- **optimiser**: SGD with momentum=0.9 and weight_decay=5e-4
- **scheduler**: StepLR with step_size=10 and gamma=0.1

4.4.3 Comparison Dataset

The Kinetics_400 dataset (Kay et al., 2017) was used as a benchmark due to its widespread acceptance and use in related studies. It serves as a standard reference, allowing for a fair comparison of model performance across different datasets and studies.

4.4.4 Metrics

Top-1 & Top-5

In the field of ML and DL, accuracy is a commonly used metric to evaluate the performance of a classification model. The Top-1 and Top-5 accuracy are two specific variations of accuracy metrics widely utilised in image recognition and classification tasks (Hutchinson and Gadepally, 2021; Tran, Bourdev, et al., 2015; Feichtenhofer, Fan, et al., 2019; Feichtenhofer, 2020).

The Top-1 accuracy is a measure of how often the model correctly predicts the most probable class label for a given input sample. In other words, it measures the percentage of instances in which the model's highest confidence prediction matches the ground truth label. For example, if a model correctly predicts the class label for 80 out of 100 images, its Top-1 accuracy would be 80%.

On the other hand, the Top-5 accuracy provides a more relaxed evaluation metric by considering whether the correct class label is present within the Top-5 predictions of the model. This metric is particularly useful when dealing with large or fine-grained classification problems where there may be multiple plausible class labels for an input sample. The Top-5 accuracy measures the percentage of instances in which the correct label appears within the Top-5 predicted labels. For example, if a model correctly predicts the class label for 90 out of 100 images within the top 5 predictions, its Top-5 accuracy would be 90%.

Both Top-1 accuracy and Top-5 accuracy are valuable performance indicators for classification models. The Top-1 accuracy provides a strict measure of the model's ability to make precise predictions, while the Top-5 accuracy allows for some flexibility by considering a wider range of potential correct predictions. These metrics help researchers and practitioners assess the effectiveness and generalisation capability of their models in accurately classifying and recognising objects or patterns in the given data.

Confusion Matrices

The confusion matrices indicate that there is no significant difference between the views in terms of confusion between classes. The same classes exhibit confusion across all views, suggesting that the viewpoint does not affect the confusion patterns.

Mean Average Precision (mAP)

The Mean Average Precision (MAP) is the arithmetic mean of the average precision values for an information retrieval system over a set of query topics (Voorhees, n.d.). It provides a single measure of quality across recall levels and is widely used in evaluating models in information

retrieval and related fields.

4.5 Analysis Results

4.5.1 Mutual Information Analysis Results

For the RHM dataset, Equation 4.3 was applied to calculate the average MI between consecutive frames in each class. A video was randomly selected from each class, and its frames were extracted. The MI between two adjacent frames was then computed iteratively until the last two frames of the video. This process was repeated for the same video in each view. For example, video 100 from the walking class was considered for all four views. In general, video number 100 was selected for all classes and all views for the experiments.

The results of this method, which capture the differences between the same video in different views, are presented in Figure 4.1. A higher MI value indicates greater redundancy between frames, while a lower MI value suggests lower redundancy in the video. By analysing the MI values, insights can be gained into the degree of similarity or dissimilarity between frames within a video, highlighting the varying levels of information content and redundancy across different views.

Analysing the MI values reveals interesting patterns in the RHM dataset. The Robot Viewpoint exhibits the lowest MI among all actions, except for the reaching action. This finding can be attributed to the inherent motion of the camera in the Robot Viewpoint, leading to frames with diverse and distinct information. In actions involving significant movement, such as walking, the Robot Viewpoint's MI is particularly low due to the varying perspectives captured by the moving camera.

The Top View, on the other hand, demonstrates the second-lowest MI across actions. This can be attributed to the unique characteristics of the fish-eye lens, which captures a wide field of view but with some distortion. As a result, the frames in the Top View may contain less redundant information, contributing to lower MI values.

In contrast, the Front and Back Views exhibit higher MI values compared to the Robot and

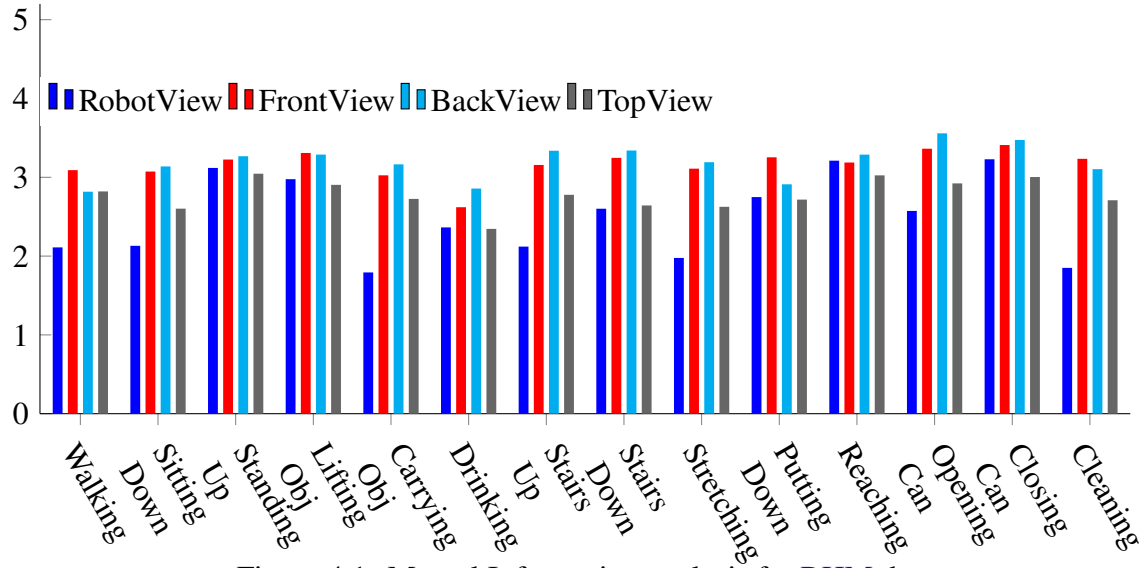


Figure 4.1: Mutual Information analysis for RHM dataset.
Video number 100 was selected from all activity classes and views for this experiment. The figure illustrates the redundancy of information between consecutive frames. Higher values indicate greater redundancy. The results clearly show that static views, such as Front and Back, contain the most redundant information between two consecutive frames, compared to the dynamic Robot view.

Top Views. These views are fixed on the wall, providing a more stable and consistent viewpoint. Consequently, the frames captured from these viewpoints contain more MI due to the relatively constant perspective and fewer variations in the scene.

4.5.2 Deep Model Performance Results

Table 4.1 presents the results of applying benchmark models, namely C3D (Tran, Bourdev, et al., 2015), R(2+1)D (Tran, H. Wang, et al., 2018), R3D (Tran, H. Wang, et al., 2018), and SlowFast (Feichtenhofer, Fan, et al., 2019), on the RHM dataset. To provide additional context, the results of the Kinetics_400 dataset are also included for comparison. Kinetic_400 is one the most famous benchmarks dataset in HAR which most of the models usually compare their results with this dataset.

The variants SF101 and SF50 refer to specific configurations of the SlowFast model, distinguished by their backbone architectures. SF101 utilises a ResNet-101 backbone, which is deeper and thus capable of capturing more complex patterns in the data, while SF50 uses a ResNet-50 backbone, which is shallower and computationally less intensive. These variants

offer a trade-off between model complexity and performance, allowing for adjustments based on the specific requirements of the task at hand.

The table displays various performance metrics, such as Top-1 accuracy and Top-5 accuracy, for each model on the RHM dataset. These metrics indicate the models' abilities to recognise actions within the RHM dataset. By comparing the results of the benchmark models to the performance on the Kinetics_400 dataset, it becomes possible to assess the relative performance and generalisation capability of each model on the RHM dataset.

From Table 4.1, it can be observed that the bold characters entries represent the highest accuracy for the Top-1 metric, indicating the models that achieve the best performance in correctly classifying the primary action label. Similarly, the bold character entries denote the highest accuracy for the Top-5 metric, which measures the models' ability to include the correct action label within the top five predictions.

Furthermore, the underlined values indicate the highest accuracy achieved among all the models and viewpoints, both for Top-1 and Top-5 metrics. These underlined entries represent the best overall performance in accurately recognising the actions in the RHM dataset.

The results obtained from the benchmark models on the RHM dataset reveal some interesting findings regarding the different viewpoints.

Firstly, the Robot View demonstrates the lowest Top-1 and Top-5 accuracy across all models. This can be attributed to the presence of motion in the robot's viewpoint, which introduces additional complexities and variability in the captured frames, making action recognition more challenging.

On the other hand, the Front view stands out with most of the highest Top-1 and Top-5 accuracy results. This can be attributed to the advantageous viewpoint provided by the overhead perspective, which offers a comprehensive view of the entire activity area. The Front View consistently achieves some of the best accuracy results, except in the case of the R(2+1)D model. The wall-fixed views, including the Front and Back Views, exhibit the highest accuracy results in terms of both Top-1 and Top-5. This can be attributed to the stationary nature of these views, eliminating motion-related challenges and providing a stable and well-positioned

viewpoint to capture action details effectively.

Specifically, the C3D model demonstrates the highest overall accuracy results when considering both Top-1 and Top-5 metrics, particularly when utilising the Front View. This suggests that the C3D model is well-suited for action recognition on the RHM dataset, leveraging the strengths of the fixed frontal viewpoint.

The 98.14% accuracy mentioned in this study refers specifically to the Top-5 accuracy metric, which has been previously clarified in the dataset and results sections. This high accuracy was achieved under controlled experimental conditions within the Robot House environment. The conditions include consistent lighting, minimal location variance, and controlled clothing variance among the participants, all of which are detailed in earlier sections of this thesis. These controlled variables were essential to focus on the evaluation of the model's ability to recognise actions rather than on external factors like environmental changes. It is important to note that while these conditions were kept consistent, the model's robustness to variations in lighting, location, and clothing is a crucial consideration for future work, especially for applications in more diverse and dynamic real-world environments. The experiments were conducted in a stable indoor setting with predefined lighting setups, and the clothing worn by the participants was kept consistent to minimise any potential biases or variability that could affect the model's performance. This controlled setup allowed for a focused assessment of the HAR model's capabilities, leading to the reported accuracy results.

Additionally, a confusion matrix was generated for the C3D model, depicting its performance across all views. The confusion matrix provides a comprehensive visual representation of the model's classification results, showcasing the relationship between predicted labels and ground truth labels for each class in the dataset. By analysing the confusion matrix, insights can be gained into the model's strengths and weaknesses in recognising different actions from various viewpoints.

Figure 4.2 presents the confusion matrices for each view obtained from the C3D model. These matrices illustrate the classification performance of the C3D model for the different views in the RHM dataset.

Table 4.1: Benchmark models on RHM and Kinetic_400

	RobotView		FrontView		BackView		TopView		Kinetic 400	
Model	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
C3D	44.61	89.38	**67.59**	**98.14**	66.63	97.99	62.21	96.95	71.4	NA
R3D	48.10	89.45	64.21	95.91	63.77	95.69	54.78	93.91	74.4	91
R(2+1)D	44.51	87.97	51.67	93.91	61.91	95.76	52.33	94.28	72	90
SF(50)	41.10	88.56	57.16	95.32	56.27	95.30	53.08	94.50	77	92.6
SF(101)	42.24	88.19	58.63	95.43	57.87	95.68	54.39	95.39	77.9	93.2

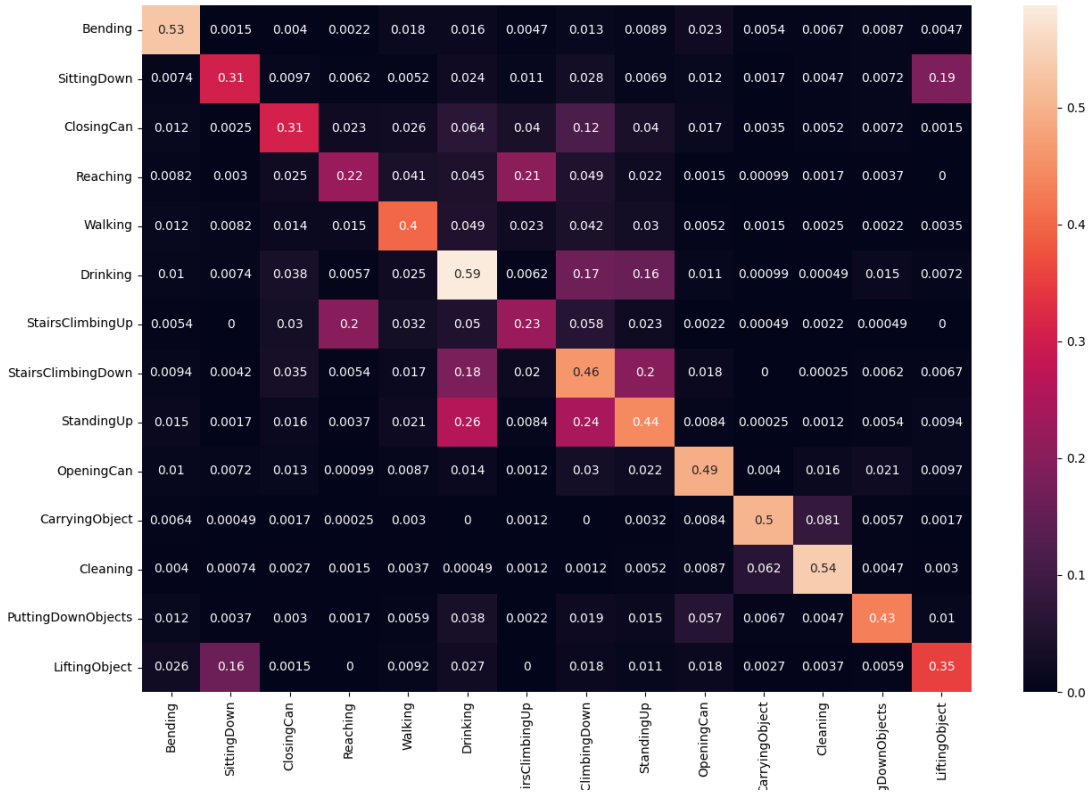
The results of using RHM dataset with C3D (Tran, Bourdev, et al., 2015), R3D and R(2+1)D (Tran, H. Wang, et al., 2018) and SlowFast (Feichtenhofer, Fan, et al., 2019) models. The robot view achieved the lowest results across all models. The best results were obtained with the C3D model using the front view. The abbreviation SF represents the SlowFast model, which was tested on both ResNet-50 and ResNet-101 architectures.

Upon examining the confusion matrices, it is observed that certain classes exhibit consistent confusion patterns across all views. This indicates that the confusion is not primarily influenced by the viewpoint. Notably, the following pairs of classes consistently exhibit confusion with each other across all views:

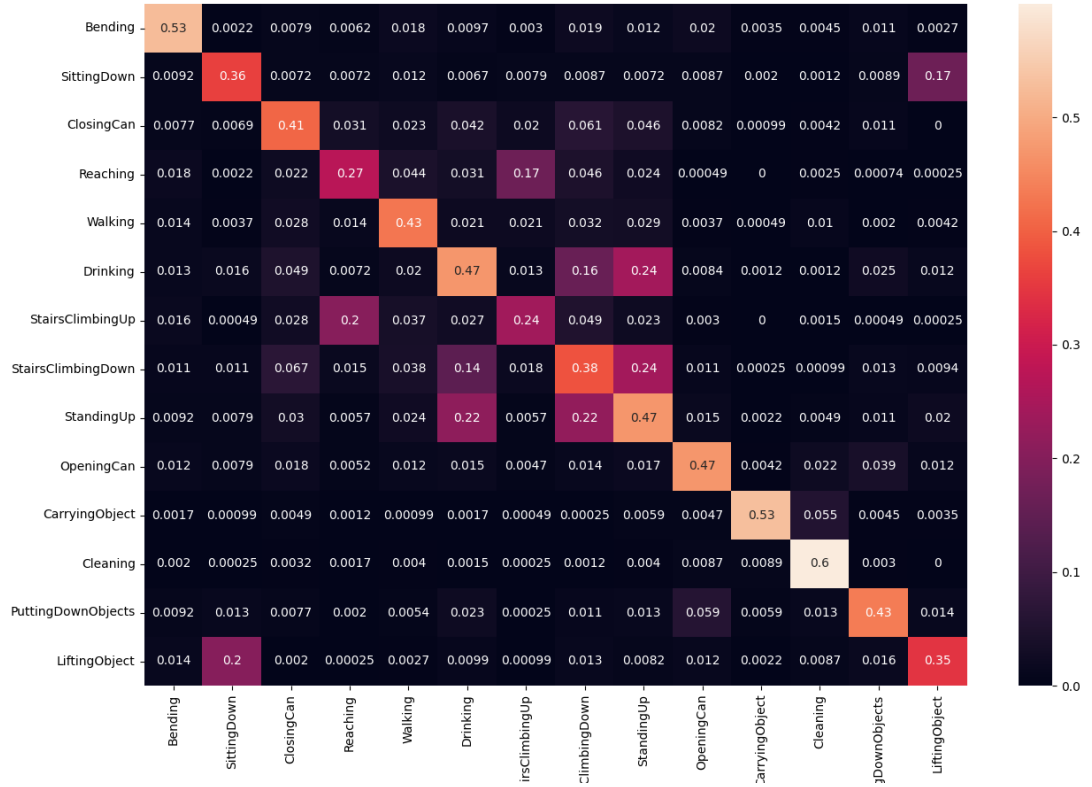
- Sitting down and Lifting objects
- Reaching and Stairs up
- Drinking and Standing Up
- Stairs Down and Opening Cans & Putting down objects

These findings suggest that the C3D model encounters challenges in accurately distinguishing between these pairs of actions, regardless of the viewpoint. The presence of consistent confusion patterns among these classes highlights potential areas for improvement in the model's discriminative capabilities for these specific action pairs.

The main conclusion from the confusion matrices is that there is no significant difference between the views regarding confusion between classes. This indicates that the viewpoint does not affect the confusion, and the same classes exhibit confusion across all views. Different camera views and types (static and dynamic) were used to determine whether they affected model results and confusion matrices. This chapter demonstrates that having a dynamic camera

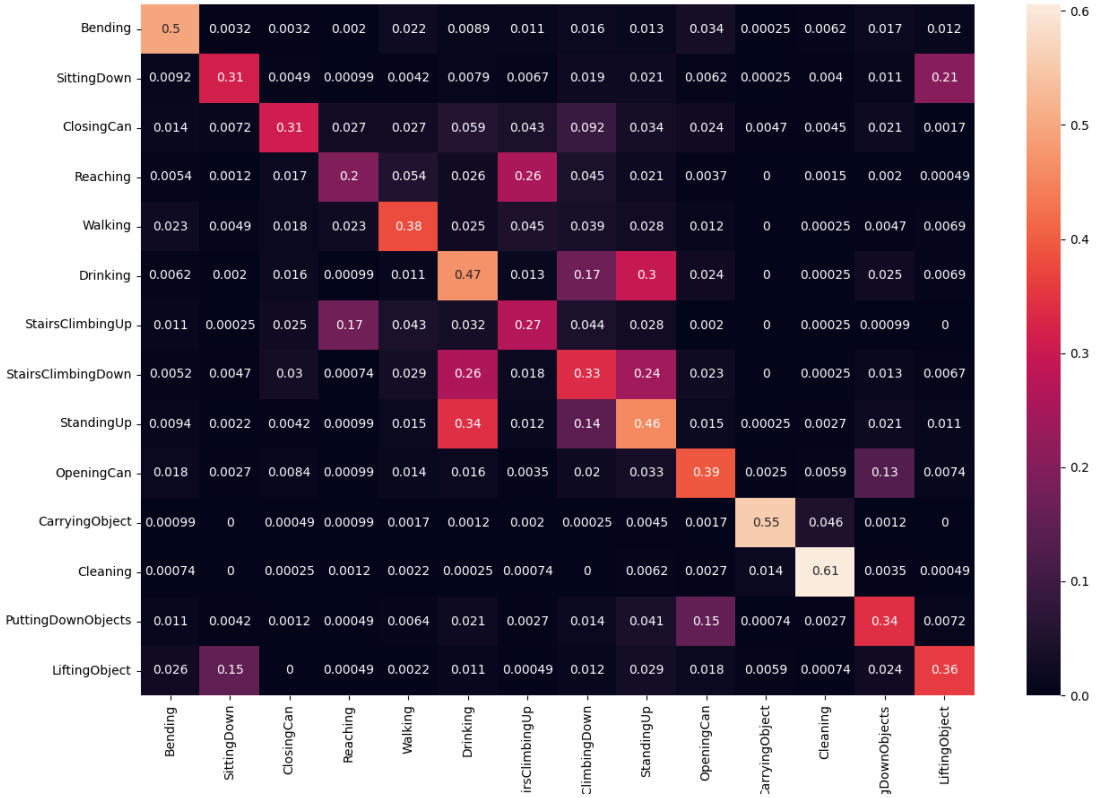


(a) Front View

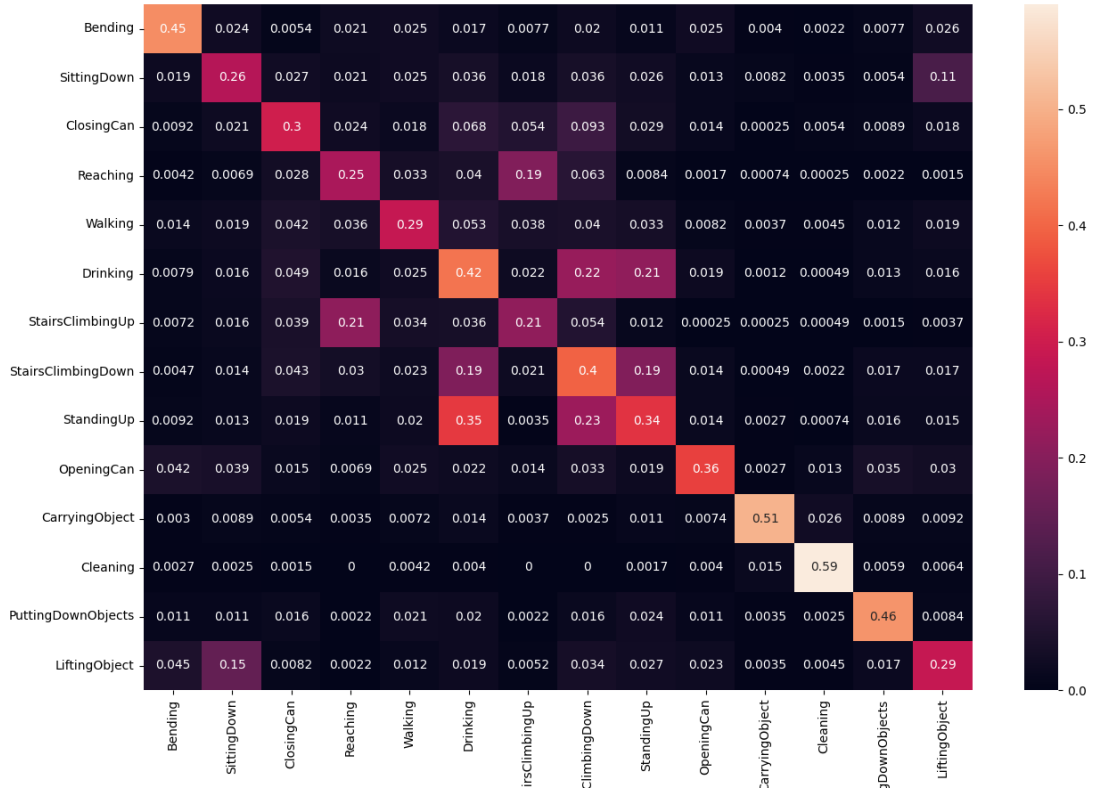


(b) Back View

Figure 4.2: RHM Confusion Matrix for all views with C3D Model



(c) Top View



(d) Robot View

Figure 4.2: RHM Confusion Matrix for all views with C3D Model

(robot view) results in the lowest performance. However, for confusion matrices, the results are consistent, with the same classes exhibiting confusion across all views.

For a comparison of model strengths and weaknesses, it can be stated that: **C3D**: Achieved the best results due to a higher number of trained parameters. **R2+1D and R3D**: Utilise ResNet layers, making them lighter and faster in training compared to C3D. **SlowFast Models**:

- **SF50**: Utilises 50 ResNet layers, balancing performance and computational efficiency.
- **SF101**: Utilises 101 ResNet layers, offering improved performance at the cost of increased computational resources.

These models demonstrate varied strengths, with C3D offering high accuracy but requiring more computational power, while SlowFast models provide a balance between accuracy and efficiency.

4.6 RHM Analysis Contribution

In addition to the comprehensive analysis conducted in this chapter, the findings and methodologies have been further validated and expanded upon in the recent publication titled *RHM: Robot House Multi-View Human Activity Recognition Dataset* (Abadi et al., 2023). This paper, presented at the ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions, delves deeper into the application of multi-view datasets for HAR in the context of HRI. The RHM encompasses four views: Front, Back, Ceiling, and Robot Views, featuring 14 classes with 6701 video clips for each view, totalling 26804 video clips. Each clip, lasting between 1 to 5 seconds, is synchronised across different views. The paper explores the performance of state-of-the-art models on these views, assesses the information content through mutual information concepts, and benchmarks the strengths and weaknesses of each view. The study concludes that multi-view and multi-stream activity recognition has the potential to significantly enhance HAR results.

4.7 Chapter Summary

In this chapter, the RHM dataset was assessed using two distinct approaches. The first approach involved a new method grounded in MI, while the second approach entailed the integration of state-of-the-art DL models into the RHM dataset.

In the first part, a novel MI metric for action recognition dataset analysis was proposed, based on findings on MI. This metric considers the temporal dependencies and contextual relationships between consecutive frames in a video sequence. By quantifying the MI between frames, a better understanding of the information redundancy and discriminative power within different actions and viewpoints was achieved. This metric serves as a valuable tool for dataset analysis, allowing researchers to assess the complexity and diversity of actions captured from different viewpoints, identify potential challenges in recognition, and guide the development of more effective models and algorithms for action recognition tasks.

The analysis of MI between consecutive video frames revealed insights into the redundancy and similarity of information within different viewpoints. It was found that the dynamic viewpoint, such as the Robot View, exhibited lower MI values, indicating less redundancy and greater diversity between consecutive frames. On the other hand, fixed viewpoints, like the Front and Back Views, had higher MI values, suggesting a higher degree of redundancy and similarity in consecutive frames.

Furthermore, the evaluation of various DL models, including C3D, R(2+1)D, R3D, and SlowFast, on the RHM dataset provided valuable performance metrics for HAR. The Robot View consistently yielded lower accuracy results in terms of Top-1 and Top-5 accuracy metrics across all models. This can be attributed to the inherent motion and variability in frames captured from the robot's viewpoint. In contrast, the Top view achieved higher accuracy, benefiting from its comprehensive top-down perspective.

Notably, the fixed Front view demonstrated superior accuracy in both Top-1 and Top-5 metrics, except for the R(2+1)D model. This outcome can be attributed to the absence of motion in these views and their optimal positioning for capturing action areas effectively.

Analysing the confusion matrices generated for the C3D model across different views,

consistent patterns of confusion between certain action classes were found, irrespective of the viewpoint. This indicates that the confusion was not solely attributed to the specific camera angle but rather to inherent similarities or complexities within those classes.

One intriguing aspect of the findings is the performance of the robot view, which consistently demonstrated lower accuracy and lower redundancy results compared to other viewpoints. The presence of motion in the robot view can be attributed to these outcomes. The dynamic nature of the camera introduces variations in the frames, making it more challenging for the models to accurately classify actions. Moreover, the lower redundancy suggests that consecutive frames in the robot view exhibit less similarity or repetitive patterns, potentially due to the camera's continuous movement. Understanding the challenges and limitations posed by the robot view can guide future research in developing specialised techniques to mitigate these factors and improve action recognition performance in such scenarios.

Chapter 5

Multi-Stream C3D Network

5.1 Introduction

Based on the analysis of DL models performance in RHM dataset in Chapter 4, the evaluation of different views in terms of Top-1 and Top-5 revealed notable variations in performance. Amongst the four views considered, the Robot View consistently exhibited the lowest accuracy, indicating a greater difficulty in accurately recognising actions from this perspective. Conversely, the Front View consistently demonstrated the highest accuracy, suggesting it provides a more informative and discriminative viewpoint for action recognition tasks.

On the other hand, in the context of DL models, the C3D model, proposed by (Tran, Bourdev, et al., 2015), emerged as the top-performing architecture across all views. Its effectiveness in capturing spatiotemporal information from video sequences was evident through its superior performance compared to other models. The C3D model leverages 3D convolutional layers to analyse both spatial and temporal features, making it well-suited for action recognition tasks.

To overcome the challenges associated with the Robot View and improve its usefulness in HRI and AAL contexts, additional view information will be integrated alongside the Robot View using the RHM dataset. This approach aims to enhance the overall performance and applicability of the Robot View in these specific scenarios.

To accomplish this objective, a new DL model will be developed as an extension of the C3D

model, featuring a *Dual-stream* architecture. This model is designed to leverage the advantages of different viewpoints to enhance the performance of the Robot View. By incorporating multiple streams and experimenting with different combinations of views, the model aims to capture a broader range of spatiotemporal information. This approach is intended to overcome the issues related to the lower accuracy typically seen with the Robot View, thereby providing a more accurate and reliable representation of human activities in HRI and AAL environments.

This chapter concentrates on the challenges related to recognising human activities using spatial information alone, particularly within the context of dual-stream models. The primary challenges addressed include the difficulty of accurately capturing and processing spatial features from different viewpoints and ensuring that these features are effectively utilised in dual-stream architectures. The methods discussed in this chapter focus on improving recognition accuracy by leveraging dual-stream models that integrate spatial information from multiple views, without incorporating temporal dynamics at this stage. This approach is essential for understanding the impact of spatial features on model performance, laying the groundwork for later integration with temporal data in subsequent chapters.

In this chapter, the exploration of multiview models for HAR is researched. The primary objective is to provide a review of the existing literature on multiview models in HAR, examining various approaches and methodologies employed in the field. The discussion begins by addressing related work and summarising the key findings and advancements made in multiview HAR in Section 5.2. Next, the proposed methodology is presented, entailing the development of *Dual-stream* in Section 5.3. Detailed explanations of the architecture and fusion mechanism employed in the model are provided. Subsequently, in Section 5.4, the implementation of the proposed model is discussed and its performance is evaluated using appropriate evaluation metrics. Finally, in Section 5.6, the chapter is summarised by synthesising the results obtained, discussing their implications, and outlining potential future research directions in the field of multiview HAR.

5.2 Related Work

One of the methods for enhancing the extraction of spatiotemporal data from videos involves the use of multi-stream networks. These specialised deep learning models excel at recognising human actions by simultaneously processing various types of data, such as skeletal formations, motion information, and object interactions. The capacity for handling multiple data streams makes multi-stream networks particularly effective in complex situations (Kong and Fu, 2022).

In addition to handling multiple data types, multi-stream networks can integrate depth information via specialised sub-networks to improve their video action recognition capabilities. For a more nuanced understanding of motion, some multi-stream network models utilise techniques like 3-channel Motion History Images (MHI) or Optical Flow. These capture both forward and backward movements and feature joint selection mechanisms to generate sparse skeleton graphs. This multi-faceted approach enables multi-stream networks to comprehend human actions by taking into account contextual, global, and local motion attributes (Gu et al., 2020).

The use of multi-stream networks is further enhanced by their ability to integrate depth information, courtesy of specialised sub-networks. This additional layer of depth data significantly bolsters the multi-stream networks' capabilities in video action recognition tasks. By capturing depth details, multi-stream networks can offer a more nuanced understanding of spatial relationships in the video content, thereby increasing the accuracy and robustness of their action recognition algorithms (Kong and Fu, 2022).

In a work presented by (L. Wang, Z. Wang, et al., 2015), a novel methodology for action recognition, tailored for the THUMOS15 challenge is introduced. This approach integrates very deep Dual-stream ConvNets with Fisher vector representations of improved Dense Trajectories (iDT) features (H. Wang and Schmid, 2013). Utilising advanced architectures like GoogLeNet and VGGNet, the authors find that while deeper networks significantly enhance the performance of spatial nets, they do not yield similar improvements for temporal nets, likely due to the constrained size of the UCF101 training dataset. The research also incorporates traditional iDT features, encoded using Fisher vectors, and introduces a new video segmentation technique based on colour and motion histograms. Experimental evaluations on the THUMOS15 validation

dataset reveal that deeper architectures excel for spatial nets but not for temporal nets. Notably, the Dual-stream ConvNets achieves a 10% performance improvement over traditional iDT features. Furthermore, combining both methods leads to an additional increase in the mean Average Precision (mAP) by approximately 5%, achieving a 68% mAP.

In another work by (L. Wang, Xiong, et al., 2015), the authors tackle the complexities of video-based HAR by unveiling a sophisticated model termed "very deep Dual-stream ConvNets," adapted from successful image recognition frameworks like GoogLeNet and VGGNet. In this dual-stream model, the spatial net is tasked with scrutinising individual video frames and largely mirrors the architecture employed for static image object recognition. Conversely, the temporal net is engineered to capture inter-frame motion, utilising a 10-frame stack of optical flow fields as its input. To mitigate the issue of overfitting, particularly pronounced due to the limited size of datasets like UCF101, the authors introduce specialised training strategies. These encompass pre-training both spatial and temporal nets on more extensive datasets, implementing smaller learning rates, and utilising a high dropout ratio to curb overfitting. The authors also employ various data augmentation methods to enrich the training dataset. On the technical front, they extend the Caffe toolbox to accommodate Multi-GPU setups, thereby improving computational efficiency and minimising memory usage. Empirically, the paper demonstrates that the very deep Dual-stream ConvNets attain a recognition accuracy of 91.4% on the UCF101 dataset, marking a 3.4% improvement over the original, less complex Dual-stream ConvNets and surpassing other leading methods by 2.8%.

Amongst architectures that employ Dual-stream or multi-stream networks, the amalgamation of results from different streams can generally be divided into four primary categories: early fusion, mid-level fusion, late fusion, and lateral fusion. Early fusion entails the merging of features from multiple streams before they are input into the network for final classification. Mid-level fusion integrates intermediate representations or features from each stream before making the ultimate prediction. Late fusion calculates a weighted mean of the individual stream predictions to generate the final result. Lateral fusion, in contrast, executes parallel processing of streams while intermittently sharing features or information between them. Each of these fusion

methods comes with its own set of pros and cons, and the selection of a particular technique often hinges on the specific demands of the task being addressed (Karpathy et al., 2014).

In an insightful discussion about fusion techniques presented in (Feichtenhofer, Pinz, and Zisserman, 2016), the authors investigate a range of methods for integrating ConvNets to enhance the recognition of HAR in videos. A novel ConvNets architecture is introduced that amalgamates both spatial and temporal data at varying layers. The authors demonstrate that fusion at the convolutional layer, as opposed to the softmax layer, retains performance while minimising parameter count. They also show that fusion at the terminal convolutional layer and the class prediction layer can elevate accuracy levels. Moreover, pooling across spatiotemporal neighbourhoods of high-level convolutional features further augments performance. Two pre-trained ImageNet models, VGG-M-2048 and VGG-16, are utilised. Various optimisation techniques, such as learning rate adjustments and dropout ratio modifications, are employed. For the fusion of the dual streams, a batch size of 96 is used, and the learning rate is adapted based on validation accuracy. Multiple fusion strategies like Max, Concatenation, Sum, and Conv fusion are examined, with Conv fusion emerging as the most effective. The paper also assesses the benefits of deeper models, revealing that while a deeper spatial model substantially improves performance, a deeper temporal model offers only marginal gains. The authors also explore different temporal fusion methods, finding that 3D pooling and 3D filtering further elevate recognition accuracy. The methodology is evaluated on two widely-used datasets: UCF101 and HMDB51, achieving state-of-the-art results that surpass both the original Dual-stream model and other existing techniques. Additionally, the authors find that their ConvNets-based approach can be further optimised by late fusion with hand-crafted iDT features, reaching an accuracy of 93.5% on UCF101 and 69.2% on HMDB51. The paper thus offers significant contributions to the understanding of effective ConvNets fusion strategies for video-based action recognition.

(Feichtenhofer, Pinz, and Wildes, 2017), present a groundbreaking ConvNets architecture specifically designed for video HAR, focusing on the multiplicative interactions of spacetime features. This approach diverges from the conventional Dual-stream architectures, which typically rely on late fusion of softmax predictions. Instead, the authors introduce cross-stream

residual connections that allow for early and more nuanced interactions between the appearance and motion pathways. This innovative approach enhances the model's ability to capture truly spatiotemporal features. The architecture is fully convolutional in both spatial and temporal dimensions, enabling a single-pass evaluation of entire videos, thereby increasing computational efficiency. Built upon 50 and 152-layer ResNets pre-trained on ImageNet, the architecture incorporates multiplicative gating functions into the residual networks, supported by both theoretical and empirical justifications. The model employs a dynamic learning rate and various data augmentation techniques during the training phase, while the motion network utilises 10-frame optical flow stacks and applies a dropout rate of 0.8 after the final classification layer. Rigorous evaluations on UCF101 and HMDB51 datasets reveal that the multiplicative gating functions outperform their additive counterparts, achieving error rates of 8.72% and 37.23% on the first splits of UCF101 and HMDB51, respectively. The paper also shows that the architecture significantly outperforms existing state-of-the-art methods. Even when fused with hand-crafted iDT features, the performance gains are minimal, suggesting that the model is nearing the performance limits of these datasets.

(Y. Wang et al., 2017) introduces the Spatiotemporal Pyramid Network, a groundbreaking architecture designed to overcome the limitations of traditional Dual-stream ConvNets in capturing complex spatial and temporal inter-dependencies for video HAR. The network is built upon well-established CNN architectures such as VGGnet, ResNets, and BN-Inception, and features a unique Spatiotemporal Compact Bilinear (STCB) operator for the efficient fusion of spatial and temporal features. This operator projects the high-dimensional outer product of these features into a lower-dimensional space using the Count Sketch function. Additionally, the architecture incorporates a visual attention mechanism that focuses on salient regions within the video, guided by the fused spatiotemporal features. Utilising a multi-stage training strategy and various data augmentation techniques, the model is rigorously evaluated on two standard datasets: UCF101 and HMDB51. The results demonstrate that the Spatiotemporal Pyramid Network achieves state-of-the-art performance, improving the average accuracy by 0.6% on UCF101 and 0.4% on HMDB51 compared to previous methods.

In another research, (Yi Zhu et al., 2019) presents a groundbreaking CNN architecture known as "hidden Dual-stream networks," specifically designed for real-time HAR. This innovative architecture captures motion information between adjacent video frames implicitly, thereby eliminating the need for pre-computed optical flow and achieving end-to-end trainability. The authors explore two methods for fusing motion features with action labels—stacking and branching—with stacking emerging as the more effective approach. The architecture undergoes rigorous testing across four challenging datasets: UCF101, HMDB51, THUMOS14, and ActivityNet v1.2. A series of ablation studies further validate the efficacy of various components of the proposed MotionNet, including specialised loss functions and operators. The results are highly promising, outperforming existing state-of-the-art real-time action recognition methods. Specifically, the architecture achieves a 6.1% improvement in accuracy on UCF101, a 14.2% increase on HMDB51, an 8.5% boost on THUMOS14, and a 7.8% enhancement on ActivityNet. The architecture also demonstrates its flexibility by seamlessly integrating with various backbones CNN architectures like VGG16, TSN, and I3D, all while maintaining real-time performance.

Leveraging the dual-stream CNN architecture proposed by (Simonyan and Zisserman, 2014), along with the SlowFast lateral connection mechanism introduced by (Feichtenhofer, Fan, et al., 2019), and incorporating the foundational 3D CNN model (C3D) presented by (Tran, Bourdev, et al., 2015), a novel Dual-stream model has been developed. The details of this innovative architecture are elaborated in the subsequent section.

5.3 Dual-Stream C3D Network Methodology

The C3D deep model is selected as the foundation for the multi-stream models because it not only achieved the highest results among the deep models evaluated on the RHM dataset in Chapter 4, but it also remains one of the leading models for spatiotemporal feature extraction (Kong and Fu, 2022). The C3D model has demonstrated its effectiveness in capturing both spatial and temporal information from video sequences, making it well-suited for action recognition tasks.

As the name suggests, the model consists of two distinct streams, each incorporating a

C3D design. The motivation behind this architecture is to leverage the complementary information captured by different streams, thereby enhancing the model's ability to understand and discriminate actions from multiple perspectives. Inspired by the successful SlowFast model (Feichtenhofer, Fan, et al., 2019), lateral connections from the first stream to the second stream are incorporated at each layer using the concatenation fusion method. This one-way lateral fusion strategy enables the transfer of information from the first stream to the second stream, promoting mutual enrichment and collaboration. By combining the outputs of these streams, the Dual-stream C3D model aims to capture both spatial and temporal features effectively, resulting in improved action recognition capabilities.

5.3.1 Training the Network in Two Streams

The Dual-stream C3D model is trained by feeding video frames into both streams simultaneously. Each stream processes the input frames through its C3D layers independently. The lateral fusion, implemented through concatenation at each layer, allows for the exchange of information from the first stream to the second stream. This method ensures that the temporal and spatial features extracted by each stream are shared, enhancing the overall representation of the video data.

5.3.2 Lateral Fusion Mechanism

Lateral fusion is a critical component of the Dual-stream C3D architecture. At each layer, the outputs from the first stream are concatenated with those of the second stream, allowing the model to integrate complementary information. This fusion strategy promotes mutual enrichment between streams, enabling the model to capture a more holistic representation of the actions. The lateral connections facilitate the transfer of spatial and temporal features from the first stream to the second, thereby improving the model's ability to discriminate between different actions.

5.3.3 Symmetry and Dominance of Streams

The two streams in the Dual-stream C3D model are not symmetric in terms of size because the second stream incorporates additional images, making it larger. The architecture ensures that both streams contribute significantly to the final representation, with no dominant stream. The upward arrows in the architectural diagram indicate the direction of information flow from the first stream to the second, emphasising the continuous integration of features through lateral connections.

The utilisation of the Dual-stream C3D model with lateral connections offers several advantages. Firstly, the incorporation of multiple streams allows for a more comprehensive representation of the input data. Each stream captures unique visual cues and temporal dynamics, enabling a richer understanding of the actions being performed. Additionally, the lateral connections facilitate the transfer of information between streams, promoting mutual enrichment. Furthermore, the fusion of the two streams provides a holistic representation that combines spatial and temporal information, enhancing the discriminative power of the model.

The architecture of the Dual-stream C3D model is visualised in Figure 5.1, highlighting the lateral connections between the streams. Table 5.1 provides a detailed overview of the model's design, including the number of parameters and layer configurations.

The model architecture includes the following layers for each stream:

- **Conv1**: 32 filters with a kernel size of (3,3,3) and padding of (1,1,1).
- **Pool1**: Kernel size of (1,2,2) and stride of (1,2,2).
- **Conv2**: 64 filters with a kernel size of (3,3,3) and padding of (1,1,1).
- **Pool2**: Kernel size of (2,2,2) and stride of (2,2,2).
- **Conv3a**: 128 filters with a kernel size of (3,3,3) and padding of (1,1,1).
- **Conv3b**: 128 filters with a kernel size of (3,3,3) and padding of (1,1,1).
- **Pool3**: Kernel size of (2,2,2) and stride of (2,2,2).
- **Conv4a**: 256 filters with a kernel size of (3,3,3) and padding of (1,1,1).
- **Conv4b**: 256 filters with a kernel size of (3,3,3) and padding of (1,1,1).
- **Pool4**: Kernel size of (2,2,2) and stride of (2,2,2).

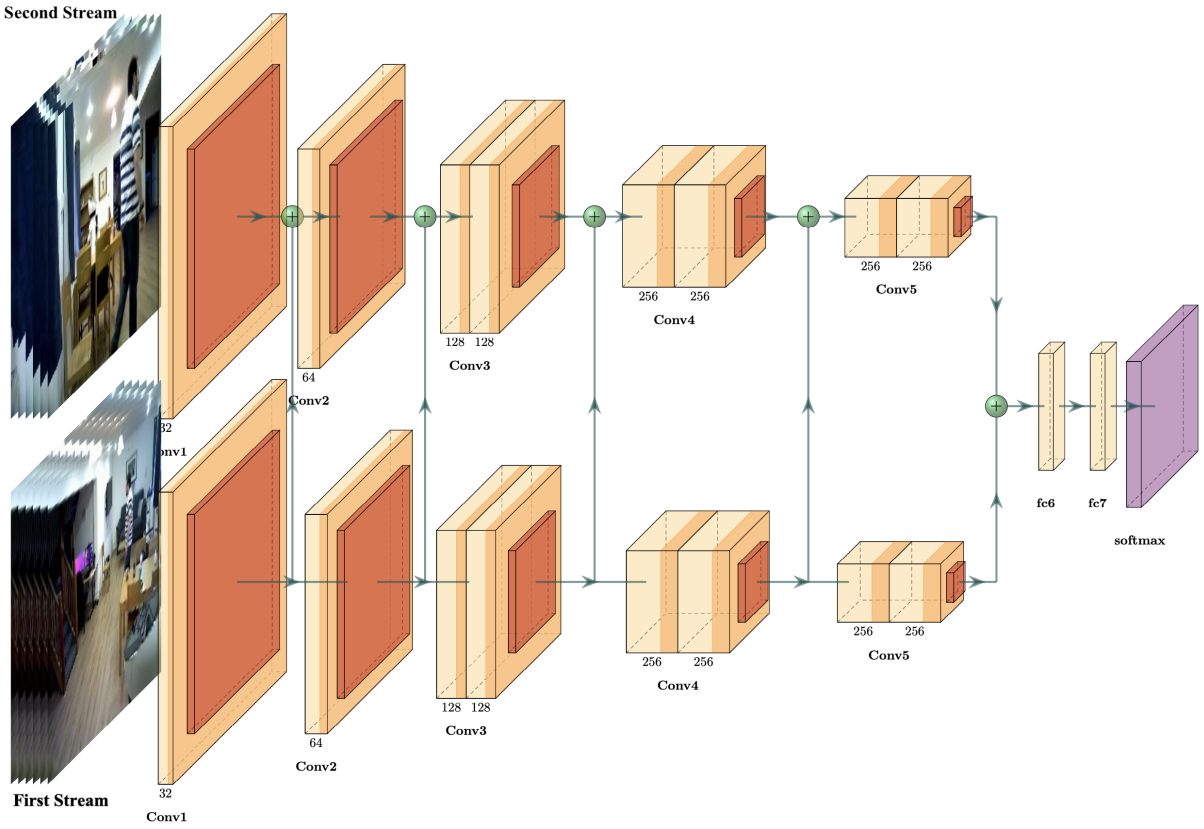


Figure 5.1: Dual-stream C3D Network Architecture.

The lower stream is designated as the first stream, and the upper stream serves as the second stream. The model features yellow boxes representing the convolutional layers, with orange boxes indicating the padding layers. There are two fully connected layers followed by a softmax layer for classification purposes. A key component of this model is the fusion mechanism, represented as a lateral connection. This fusion uses the concatenation method to combine layer information from both streams, effectively integrating the features extracted from each to enhance the overall model's performance in recognising and interpreting human activities.

- **Conv5a**: 512 filters with a kernel size of (3,3,3) and padding of (1,1,1) in the second stream and 256 filters in the first stream.

- **Conv5b**: 512 filters with a kernel size of (3,3,3) and padding of (1,1,1) in the second stream and 256 filters in the first stream.

- **Pool5**: Kernel size of (2,2,2) and stride of (2,2,2).

The lateral fusion of the two streams, coupled with the C3D architecture, contributes to the overall expressive power and effectiveness of the model.

Table 5.1: Dual-stream C3D Details.

Stage	Second Stream	First Stream
Clip Input	16*112*112	16*112*112
$Conv_1$	I=3, O=32 K=(3,3,3), P=(1,1,1)	I=3, O=32 K=(3,3,3), P=(1,1,1)
$pool_1$	K=(1,2,2), S=(1,2,2)	K=(1,2,2), S=(1,2,2)
$Conv_2$	I=32, O=64 K=(3,3,3), P=(1,1,1)	I=64, O=64 K=(3,3,3), P=(1,1,1)
$pool_2$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)
$Conv_{3a}$	I=64, O=128 K=(3,3,3), P=(1,1,1)	I=128, O=128 K=(3,3,3), P=(1,1,1)
$Conv_{3b}$	I=128, O=128 K=(3,3,3), P=(1,1,1)	I=128, O=128 K=(3,3,3), P=(1,1,1)
$pool_3$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)
$Conv_{4a}$	I=128, O=256 K=(3,3,3), P=(1,1,1)	I=256, O=256 K=(3,3,3), P=(1,1,1)
$Conv_{4b}$	I=256, O=256 K=(3,3,3), P=(1,1,1)	I=256, O=256 K=(3,3,3), P=(1,1,1)
$pool_4$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)
$Conv_{5a}$	I=256, O=256 K=(3,3,3), P=(1,1,1)	I=512, O=512 K=(3,3,3), P=(1,1,1)
$Conv_{5b}$	I=256, O=256 K=(3,3,3), P=(1,1,1)	I=512, O=512 K=(3,3,3), P=(1,1,1)
$pool_5$	K=(2,2,2), S=(2,2,2)	K=(2,2,2), S=(2,2,2)
Concatenate & FC6 & FC7		Classes
Parameter		92.81M

I: Stands for Input, representing the input received by each layer. **O**: Denotes Output, which refers to the number of filters in the layer. **K**: Indicates the Kernel size, which is the dimension of the convolutional filters. **P**: Illustrates the Padding size, determining the amount of padding applied to the input. **S**: Demonstrates the Stride size, specifying the step size the convolutional filters take across the input. **FC**: Stands for Fully Connected Layer.

5.4 Experiments & Results

5.4.1 Experiments

Different experiments are performed on the Dual-stream C3D models using the RHM dataset. The objective is to evaluate the performance of the model in terms of Top-1 and Top-5 accuracy metrics. By analysing these metrics, the effectiveness of the multi-view impact on the Dual-stream C3D model in accurately classifying actions in the RHM dataset is determined.

Focusing on improving the accuracy of the robot view in the RHM dataset, the experiments

are specifically tailored to the robot view pairs.

Additionally, given that the dual-stream C3D network outperforms the standard single stream C3D due to its increased complexity and number of parameters, the model is initially tested using the same views in both streams. This initial test, called the same view, provides the baseline results of the model on the RHM dataset.

Following this, in a subsequent experiment known as the different view, static view frames are used alongside the robot view. This helps to assess the impact of incorporating multi-views in dual-stream networks on performance enhancement.

For these experiments, the training parameters are configured as follows: a batch size of 30, a frame count of 16, a learning rate set at 0.0001, and the use of the SGD optimiser. Also, the Top-1 and Top-5 accuracy are the evaluated metrics in these experiments.

Hyperparameters were selected based on the related literature. The setup was as follows:

- **nEpochs**: 500
- **resume_epoch**: 0
- **useTest**: True
- **nTestInterval**: 20
- **snapshot**: 50
- **lr**: 1e-3
- **criterion**: CrossEntropyLoss
- **optimiser**: SGD with momentum=0.9 and weight_decay=5e-4
- **scheduler**: StepLR with step_size=10 and gamma=0.1

To ensure a thorough comparison and analysis of the results from the proposed model, identical tests are conducted on both the SlowFast (Feichtenhofer, Fan, et al., 2019) and Dual-stream CNN (Simonyan and Zisserman, 2014) models, which are among the top-tier and well-known models in multi-stream networks. It is important to note that in the experimental setup,

normal RGB frames are used in both streams of all three models under consideration. This approach allows for a comprehensive evaluation of the performance and effectiveness of these DL models in the given context.

The experiments were conducted using the University of Hertfordshire GPU Cluster, specifically utilising gpu2 and gpu3. These machines are equipped with three Tesla V100 units, with gpu3 having 16 GB VRAM per unit and gpu2 having 32 GB VRAM per unit. This powerful computational setup ensured efficient handling of the large video datasets and the complex dual-stream model architectures, facilitating thorough and accurate training and evaluation processes.

It is important to clarify that the experiments conducted in this Chapter focused exclusively on the spatial domain, where only spatial frames were used for testing the deep learning models. No temporal information or temporal dynamics were incorporated into these tests. The analysis was designed to evaluate the models' performance based purely on spatial features extracted from individual frames, providing a baseline for understanding the impact of spatial information alone on action recognition accuracy.

5.4.2 Results

Before evaluating the results, it is essential to present the baseline outcomes derived from the single-stream model, as detailed in the previous chapter. These foundational results serve as a benchmark, providing a point of comparison for assessing the improvements achieved through the dual-stream architecture.

In the prior chapter, the performance metrics of the single-stream C3D model were thoroughly analysed across various viewpoints. This analysis revealed critical insights into the model's capabilities and limitations when processing video data from a single perspective. By establishing these baseline results, the impact of integrating additional streams and advanced feature extraction techniques in the current experiments can be more effectively gauged.

The single-stream model's performance data, 5.2 is summarised to highlight its accuracy and efficiency in action recognition tasks. These results form the reference point against which the enhancements introduced by the dual-stream C3D model, including the implementation of

Table 5.2: Benchmark models on RHM

	RobotView		FrontView		BackView		TopView	
Model	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
C3D	44.61	89.38	**67.59**	**98.14**	66.63	97.99	62.21	96.95
R3D	48.10	89.45	64.21	95.91	63.77	95.69	54.78	93.91
R2+1D	44.51	87.97	51.67	93.91	61.91	95.76	52.33	94.28
SF(50)	41.10	88.56	57.16	95.32	56.27	95.30	53.08	94.50
SF(101)	42.24	88.19	58.63	95.43	57.87	95.68	54.39	95.39

The results of using RHM dataset with C3D (Tran, Bourdev, et al., 2015), R3D and R(2+1)D (Tran, H. Wang, et al., 2018) and SlowFast (Feichtenhofer, Fan, et al., 2019) models. The robot view achieved the lowest results across all models. The best results were obtained with the C3D model using the front view. The abbreviation SF represents the SlowFast model, which was tested on both ResNet-50 and ResNet-101 architectures.

lateral fusion and multi-view integration, will be measured.

The training times for the three dual-stream models reveal significant differences in computational requirements. The SlowFast (101) model demonstrates the shortest training duration, taking 6 hours and 16 minutes. This efficiency can be attributed to its optimised architecture, designed for balancing computational load while maintaining high performance. In contrast, the Dual-stream ConvNets model requires 8 hours and 47 minutes, reflecting its increased complexity and the additional computational overhead associated with handling dual streams of input data. The Dual-stream C3D model exhibits the longest training time, extending to 10 hours and 5 minutes. This extended duration is indicative of the intensive processing needed for capturing detailed spatiotemporal features across both streams. The significant increase in training time for the Dual-stream C3D model underscores its advanced capability in feature extraction, which is crucial for enhancing action recognition accuracy, albeit at the cost of higher computational resources.

Same views

The results of the same viewpoints experiments are presented in Table 5.3. Additionally, the corresponding confusion matrix for the robot-robot frame in the Dual-stream C3D model can be observed in Figure 5.2. The other pairs of confusion matrices are presented in appendices in Figure 8.2.

Table 5.3: Dual-stream Model Performance Results for same views in RHM dataset

Inputs		SlowFast (101)		DS ConvNets		DS C3D	
Second stream	First stream	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
Robot View	Robot View	42.24	88.19	48.26	89.02	**54.91**	**89.16**
Front View	Front View	58.63	95.43	63.82	96.38	**68.05**	**98.26**
Back View	Back View	57.87	95.68	62.59	96.27	**67.13**	**98.17**
Top View	Top View	54.39	95.39	60.44	95.98	**64.80**	**97.17**
Training Time		6 h 16 min		8 h 47 min		10 h 5 min	

The table presents the outcomes of experiments conducted with the same view in the SlowFast, dual-stream C3D, and dual-stream ConvNets models. **Bold numbers** indicate the top results within each model. **Underlined numbers** represent the highest results across all model pairs and experiments. **DS**: indicates Dual-Stream Network.

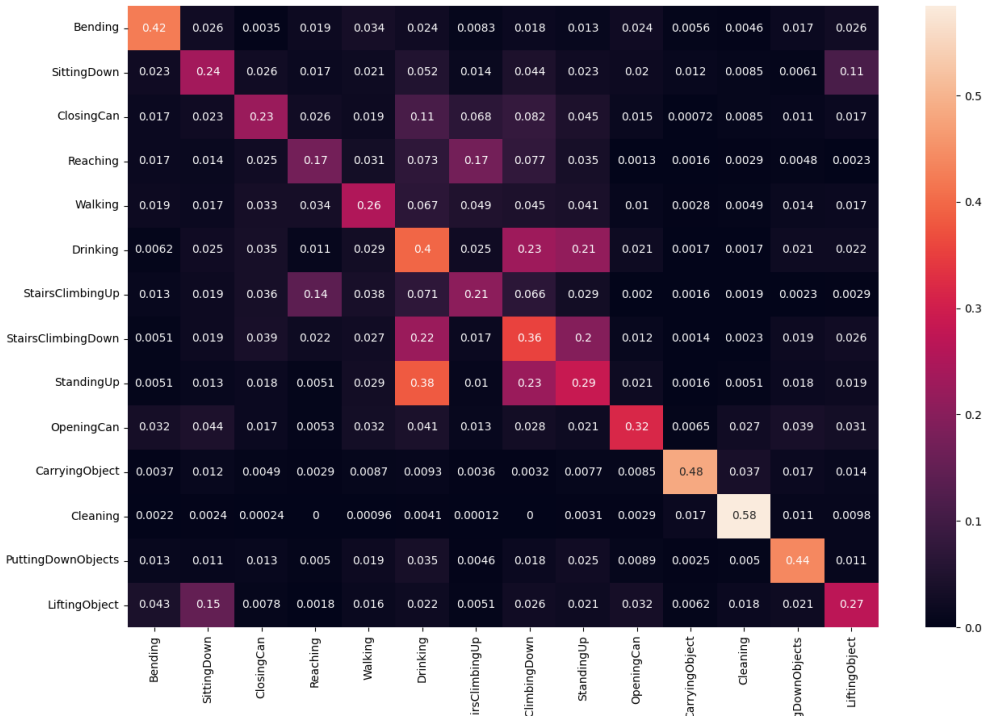


Figure 5.2: Confusion Matrix for robot-robot views with Dual-stream C3D Model

A comparative analysis of the results, as presented in Table 5.3, demonstrates that the Dual-stream C3D model outperforms the single-stream original C3D model in terms of both Top-1 and Top-5 accuracy metrics across various views. Specifically, the model shows Top-1 accuracy improvement of 10% in the Robot view, 1% in the Front view, 1% in the Back view, and 2% in the Top view. This empirical evidence substantiates that the Dual-stream C3D model is more effective than its single-stream counterpart. In contrast, the Top-5 accuracy metrics between the Dual-stream C3D and normal C3D models are nearly identical. This empirical evidence underscores the effectiveness of the Dual-stream C3D model in improving Top-1 accuracy while maintaining comparable performance in Top-5 accuracy.

Additionally, as evidenced in Table 5.3, the proposed model demonstrated superior performance across all views when compared to other models. It achieved an enhancement exceeding 15% relative to the SlowFast model. Furthermore, there was an improvement of over 5% in comparison with the Dual-stream ConvNets model across all views. Notably, in all three models, the front view yielded the most favourable outcomes, whereas the robot view was associated with the least results.

Figure 5.2 displays the confusion matrices for the experiments conducted with the same views using the Dual-stream C3D model. These matrices provide insights into the patterns of confusion between different action classes. Upon examining the matrices, it becomes apparent that there is a consistent structure of confusion across the same views.

Different views

In this section, two sets of experiments are conducted. In the first set, the robot view is used as the first stream, while in the second set, this view is swapped into the second stream. This approach allows for the exploration of the impact of the stream positioning of the robot view on the overall performance and results of the experiments. Overall, these experiments with robot view pairs enable the exploration of the contribution of other views in improving the accuracy of the robot view within a dual-stream framework.

A comprehensive summary of the experimental results can be found in Table 5.4. The

Table 5.4: Dual-stream Models Results with Different Viewpoints using RHM Dataset

Inputs		SlowFast (101)		DS ConvNets		DS C3D	
Second stream	First stream	Top 1	Top 5	Top 1	Top 5	Top 1	Top 5
Front View	Robot View	45.28	91.31	62.77	94.51	**71.06**	**98.14**
Back View	Robot View	44.69	90.64	61.02	93.89	66.25	97.17
Top View	Robot View	44.91	87.75	59.76	92.21	67.91	97.20
Robot View	Front View	41.86	89.95	58.77	91.98	65.09	95.95
Robot View	Back View	40.87	88.59	57.51	91.70	62.70	94.42
Robot View	Top View	40.27	88.02	56.68	90.79	64.60	95.70
Training Time		6 h 16 min		8 h 47 min		10 h 5 min	

The table presents the outcomes of experiments conducted with the different views in the SlowFast, dual-stream C3D, and dual-stream ConvNets models. The experiments are divided into two groups regarding robot view positioning in the model, the first stream or the second one. **Bold numbers** indicate the top results within each model. **Underlined numbers** represent the highest results across all model pairs and experiments. **DS**: indicates Dual-Stream models.

table provides detailed information on the performance of each view pair and highlights the improvements achieved in the Robot View accuracy for the SlowFast, Dual-stream ConvNets, and Dual-stream C3D model.

The results presented in Table 5.4, along with the baseline results from Table 5.3, provide a basis for meaningful comparison. Using another view frame alongside the robot view in the dual-stream setup leads to higher Top-1 and Top-5 accuracies compared to the scenario where both streams utilise only one view. This demonstrates the effectiveness of integrating multiple viewpoints in enhancing the model's performance.

The positive impact of incorporating an additional view into multi-stream networks is also observed in both the SlowFast and Dual-Stream ConvNets models. This indicates that the integration of multiple viewpoints is a beneficial strategy across different types of multi-stream network models.

The results of the experiments demonstrate that incorporating additional viewpoint streams significantly improves the accuracy of the Robot View in action recognition. In all six viewpoint pairs, a notable enhancement in the performance of the Robot View was observed compared to using it in the same view experiments for all three deep models. Specifically, when the Robot View was used as the first stream and combined with other views, a substantial increase in

accuracy was noted. This finding suggests that the complementary information provided by the other views enhances the discriminative power of the Robot View, resulting in improved action recognition performance.

In the experiments where the Robot View was employed as the second stream, there was a noticeable decline in accuracy compared to scenarios where the Robot View served as the first stream. This observation implies that incorporating information from the Robot View at each layer of the model exerts a diminished positive impact on the overall performance of the model.

Furthermore, in all pairwise tests, the proposed Dual-stream C3D model consistently outperformed the other two models. The proposed model achieved a Top-1 accuracy that was over 10% higher compared to the SlowFast model and exceeded the Dual-stream ConvNets model by more than 5% across all pairs.

In the comparative analysis of view pairs across all models, the Robot-Front combination consistently delivered the highest performance in both Top-1 and Top-5 metrics. Notably, the most exemplary results among all model and view pair configurations were achieved by the Dual-stream C3D model with the Robot-Front view pairing, which attained a Top-1 accuracy of 71.06% and a Top-5 accuracy of 98.14%.

Overall, the experiments highlight the importance of considering other static views to incorporate the robot view in action recognition models. By incorporating additional viewpoint streams and exploring their interactions within the Dual-stream model, the complementary information from different views can be leveraged to enhance the performance of action recognition systems in HRI and AAL scenarios.

The training times for the models, which are the same as those for the same view experiments, demonstrate the consistency in computational requirements due to the identical nature of the models and the input data. Each model, including SlowFast (101) with a training time of 6 hours and 16 minutes, Dual-stream ConvNets with 8 hours and 47 minutes, and Dual-stream C3D) with 10 hours and 5 minutes, was trained under the same conditions. This includes using the same size of input data and the same RHM dataset. Additionally, all hyperparameters, such as learning rates, batch sizes, and optimisation algorithms, were kept consistent across

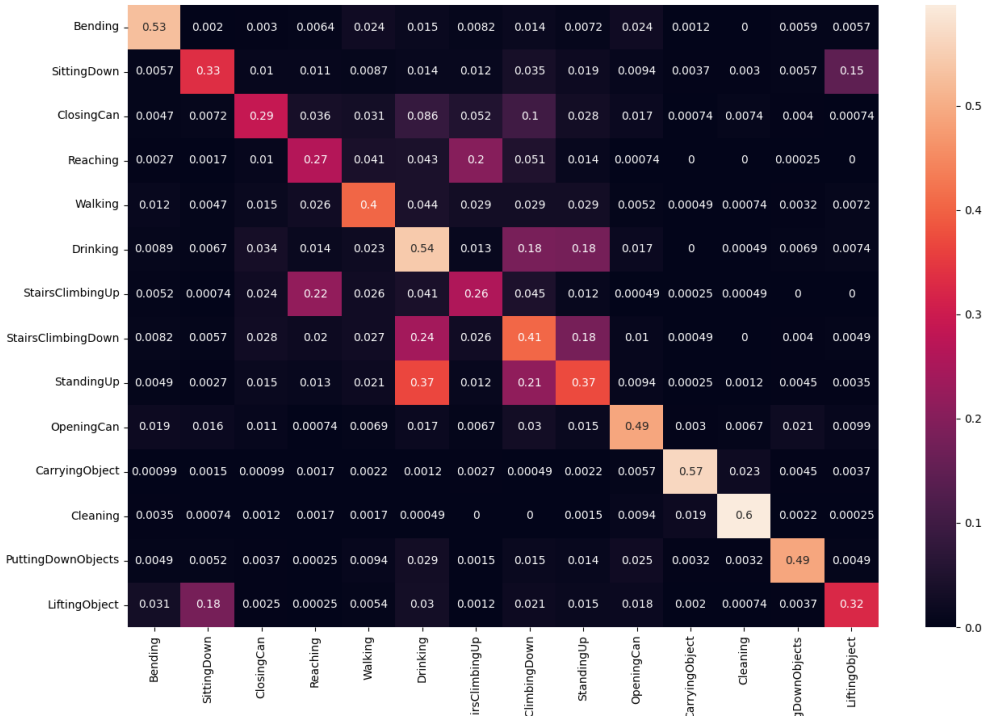


Figure 5.3: Confusion Matrix for Robot-Front views with Dual-stream C3D Model

experiments to ensure a fair comparison and accurate assessment of each model's performance. This uniformity allows for a direct evaluation of how the inclusion of different views and the dual-stream architecture impact model accuracy and efficiency.

Furthermore, the Dual-stream C3D model confusion matrices for the Robot-Front view pair can be observed in Figure 5.3. Despite the observed enhancements in the Top-1 and Top-5 accuracy results for the Robot View, the confusion patterns remain relatively unchanged. The confusion matrices demonstrate that certain classes continue to exhibit similar confusion, indicating that the improvements in accuracy are not solely attributed to a reduction in confusion between different action categories. This suggests that while the proposed approach successfully enhances the performance of the Robot View, additional factors beyond view selection may contribute to the existing confusion within the dataset. Other view pair confusion matrix results are presented in the appendices in Figure 8.3.

The relatively unchanged confusion patterns, despite improvements in accuracy, highlight a key insight into the performance of the Dual-stream C3D model. While the introduction

of dual streams has effectively increased the Top-1 and Top-5 accuracy, particularly for the Robot View, it appears that the confusion between certain action categories remains persistent. This persistence suggests that the underlying issues identified earlier in the thesis, such as the inherent similarities between specific actions or the limitations of the dataset's complexity, are not entirely mitigated by simply combining different views. The value of the Dual-stream approach, therefore, lies more in its ability to enhance the discriminative power of the model through multi-view integration rather than directly resolving the class-specific confusion. This insight underscores the importance of addressing dataset-specific challenges and considering additional strategies, such as refining action class definitions or improving dataset diversity, to further reduce confusion and enhance overall model performance.

5.5 Dual-Stream C3D Model Contribution

In addition to the comprehensive analysis conducted in this chapter, the findings and methodologies have been further validated and expanded upon in the recent publication titled *Robotic Vision and Multi-View Synergy: Action and Activity Recognition in Assisted Living Scenarios* (Abadi et al., 2024b). This paper, presented at the ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions, delves deeper into the application of robotic vision and multi-view synergy for action and activity recognition in ambient assisted living environments. The insights gained from this study provide a robust foundation for the methodologies employed in this chapter, reinforcing the significance and applicability of the research in real-world scenarios.

5.6 Chapter Summary

In conclusion, this research aimed to enhance the performance of HAR in robot-centric perspectives using the RHM dataset. The experiments evaluated various tests on the proposed multiview model, namely the Dual-stream C3D model.

The primary finding from the research is that integrating different viewpoints in Dual-stream

models significantly enhances the performance of the robot view. This suggests that using complementary static views can be advantageous for action recognition tasks in HRI within AAL environments.

Interestingly, the robot view performed better when used as the first stream, indicating its importance in each layer of the model. These improvements were observed across all the models (SlowFast, Dual-stream ConvNets, and Dual-stream C3D) tested in the work. However, despite these improvements, the confusion patterns between certain action classes remained consistent, indicating that viewpoint alone is not enough to overcome inherent classification challenges.

Finally, this research demonstrates that the proposed dual-stream C3D model outperformed the SlowFast and dual-stream ConvNets models in all tests, both in experiments with the same views and with different views. This indicates the superior effectiveness of the dual-stream C3D approach in the context of the experiments conducted.

Chapter 6

Handcraft Feature Extraction on RHM

6.1 Introduction

Reflecting on the insights from Chapter 5, it's clear that incorporating other static views in a multi-stream network has led to significant improvements in both Top-1 and Top-5 accuracy for the robot view in the RHM dataset. However, in these previous experiments, temporal information wasn't utilised. Simon et al. in their research (Simonyan and Zisserman, 2014) explored the beneficial effects of including temporal information in multi-stream networks.

In this chapter, the addition of temporal information to the model introduced in Chapter 5 will be explored. There are well-known methods for extracting temporal features from RGB frames, as discussed in the next chapter, such as (Bobick and Davis, 2001; Sun, Roth, and Black, 2010). However, these methods were developed for general purposes and are particularly suited for scenarios where the cameras are in motion, which can increase the computational demands of these models.

Given that there are three static views supporting the motion camera (robot view) in the RHM dataset, the aim is to access temporal information in a more computationally efficient way. This approach intends to leverage the static views to enhance the temporal analysis while minimising the computational load.

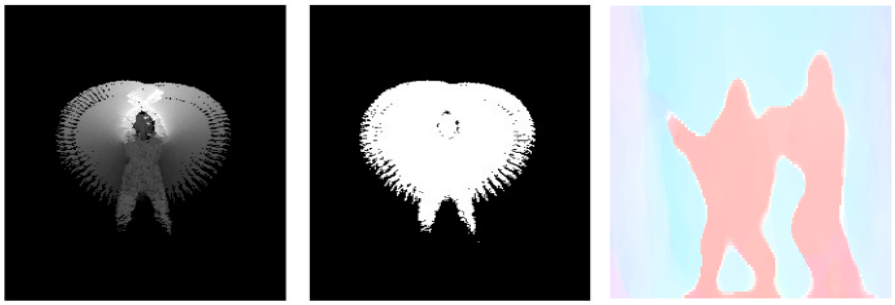
In this chapter, three new handcrafted feature extraction methods are presented. These

techniques are designed to capture more temporal information by sequentially processing frames. The aim is to develop a richer and more diverse representation of actions, addressing the challenges identified in Chapter 5.

This chapter is structured to first review existing literature on handcraft feature extraction for HAR in Section 6.2. Then, Section 6.3 will explore the detailed processes of the handcraft feature extraction techniques used, focusing on how they enhance the feature set of video frames. Following this, Sections 6.4 and 6.5 will detail a series of experiments and their outcomes for four DL models: the single stream C3D, the Dual-stream C3D, the Dual-stream ConvNets, and the SlowFast models. These experiments aim to assess the efficacy of the introduced handcraft feature extraction methods in both one-stream and Dual-stream scenarios. The chapter will conclude with a discussion in Section 6.7, summarising the main insights.

6.2 Related Work

In the domain of HAR, the complexity arises from the wide spectrum of human actions, necessitating a robust and versatile solution (Hutchinson and Gadepally, 2021). This challenge is typically tackled through a two-step process: feature representation and action classification. In the first phase, the goal is to extract pertinent attributes from action videos and transform them into feature vectors. Subsequently, in the action classification stage, these features are used to categorise the actions into predefined classes (Hutchinson and Gadepally, 2021). Despite the surge in deep learning, handcraft features remain pertinent in this context, as they excel at capturing temporal nuances, a task that persists as a challenge in the deep learning landscape. This section delves into handcraft feature extraction, categorising these features into two main types: global feature representation methods and local feature representation methods. Both categories will be examined in detail, emphasising their enduring significance and unique challenges within the overarching framework of human action recognition (Kong and Fu, 2022).



(a) MHI (b) MEI (c) Optical Flow

Figure 6.1: Examples of Local and Global Feature Representations.
The figure illustrates two types of feature representations used in human action recognition. (a) MHI and (b) MEI represent global features (Bobick and Davis, 2001), capturing motion patterns over time for the entire sequence. These methods emphasise the accumulation and recency of motion, making them effective for understanding overall motion trends in a video. (c) Optical Flow (Sun, Roth, and Black, 2010) represents a local feature, which captures motion at a pixel level by calculating the apparent velocity of movement between consecutive frames. Optical flow is highly sensitive to local changes, making it useful for capturing fine-grained motion details.

6.2.1 Local Representation

Local Representation tries to find a sensitive motion in adjacent frames as shown in Figure 6.1c. This kind of method started with Space-Time Interest Points (STIPs) algorithm (Laptev, 2005). The paper presents an innovative approach that extends spatial interest point operators into the spatiotemporal domain to detect significant events within video sequences. It adapts well-established interest point methods, such as Harris and Förstner (Harris, Stephens, et al., 1988), to identify local structures exhibiting notable variations in both space and time. These detected events are further characterised using a normalised spatiotemporal Laplacian operator across multiple scales, defining their spatial and temporal extents. These events are represented using scale-invariant spatiotemporal N-jets and classified based on their jet descriptors. The approach is showcased in human motion analysis, particularly in challenging scenarios like detecting and estimating the pose of walking individuals in videos with occlusions and dynamic backgrounds, demonstrating stability and effectiveness even without manual initialisation.

Another work presented by (Scovanner, Ali, and Shah, 2007) introduces a 3D Scale-Invariant Feature Transform (SIFT) (Lowe, 1999) descriptor specifically tailored for video or 3D imagery, such as MRI data, and highlights its superior performance in action recognition tasks. By

extending the 2D SIFT descriptor to 3D, the authors encapsulate spatiotemporal information more effectively. Using a bag-of-words (BoWs) approach to video representation, the paper also proposes a method to identify relationships between spatiotemporal 'words' to enhance the description of video data. Experimental results, performed on a dataset containing 92 videos of 10 different actions, demonstrate that the 3D SIFT descriptors significantly outperform traditional 2D SIFT and other state-of-the-art descriptors in classifying actions. Specifically, the 3D SIFT achieved an average precision of 82.6%, considerably better than other methods tested. Additionally, the authors note the descriptor's computational efficiency and suggest that further optimisations could make it even faster.

In the paper by (M. Singh, Basu, and Mandal, 2008), a novel, nonintrusive algorithm for HAR is presented, employing computer vision techniques. The method utilises adaptive background-foreground separation to isolate human silhouettes from video frames, and extracts directionality-based feature vectors from these silhouettes. These feature vectors are then clustered and recognised in a vector space. Designed to be resilient to changes in view angles, zoom levels, backgrounds, and frame rates, the algorithm also employs temporal smoothing to enhance decision-making accuracy over time. Experimental evaluation yielded high results, with an overall accuracy rate of 95.5% and Correct Recognition Rates (CRR) ranging from 85% to 100% across multiple scenarios, including outdoor tests. Specifically, in the UoA-DS3 dataset, the algorithm achieved a CRR of 96.9%. Although requiring retraining for significantly different body shapes, the algorithm shows significant promise for real-world applications, such as monitoring activities in special care homes.

In a work by (Klaser, Marszałek, and Schmid, 2008), an innovative video descriptor is introduced that utilises histograms of oriented 3D spatiotemporal gradients. The paper offers four primary contributions. Firstly, it develops a memory-efficient algorithm based on integral videos for computing 3D gradients at varying scales. Secondly, it suggests a generic 3D orientation quantisation grounded in regular polyhedrons. Thirdly, it provides an exhaustive evaluation of all descriptor parameters, fine-tuning them specifically for action recognition. Fourthly, it applies the optimised descriptor to three distinct action datasets—KTH, Weizmann,

and Hollywood. In terms of experimental results, the descriptor significantly outperforms existing methods on the KTH and Weizmann datasets, even matching the best-known KTH accuracy of 91.8%. On the Hollywood dataset, although it doesn't surpass all existing methods, it still performs better in three out of eight classes. The experiments use a BoWs approach for video representation and employ non-linear support vector machines with an x2-kernel for classification. Despite its robust performance, the descriptor does require parameter adjustments for optimal results on the Hollywood dataset.

In (Sun, Roth, and Black, 2010), the authors conduct an exhaustive examination of the factors contributing to the effectiveness of contemporary optical flow estimation algorithms, with a particular emphasis on the Middlebury optical flow benchmark. They observe that while the foundational algorithms, like those of Horn and Schunck, have remained mostly static over the years, the application of modern optimisation and implementation techniques has significantly enhanced their performance. A pivotal component in this improved accuracy is the use of median filtering after each warping step, which, although beneficial for accuracy, results in a higher energy solution. This insight led to the formulation of a new objective function that formally incorporates median filtering. This new objective includes a non-local term that allows for more robust flow estimation across a broader spatial area. The newly developed algorithm, which also takes into account considerations for image and flow boundaries, currently ranks at the top of the Middlebury benchmark for both angular and end-point errors. The paper concludes by speculating that while classical 2-frame methods may experience incremental improvements in the years to come, significant advancements will likely necessitate algorithms that take into consideration the complex spatial and temporal relationships of moving surfaces and boundaries.

(H. Wang and Schmid, 2013) introduces a novel approach to enhance HAR in videos by explicitly accounting for camera motion in dense trajectories. The technique leverages Speeded Up Robust Features (SURF) descriptors (Bay, Tuytelaars, and Van Gool, 2006) and dense optical flow (Horn and Schunck, 1981) for feature matching between frames and utilises Random sample consensus (RANSAC) (Fischler and Bolles, 1981) for robust homography estimation. The introduction of a human detector further refines the camera motion estimates by

filtering out inconsistent matches stemming from human activity. The updated method comprises two main components: "WarpFlow," which adjusts optical flow based on the camera motion, and "RmTrack," which removes irrelevant background trajectories. Extensive experimental evaluations on four challenging datasets—Hollywood2, HMDB51, Olympic Sports, and UCF50—demonstrate considerable performance gains. For instance, on the Hollywood2 dataset, the combined approach increased the accuracy of trajectory descriptors from a baseline of 42.2% to 48.5%. Similarly, improvements were noted in Histograms of Optical Flow (HOF) (H. Wang, Kläser, et al., 2013) (from 51.4% to 58.8%) and Motion Boundary Histograms (MBH)(Dalal, Triggs, and Schmid, 2006) (from 57.4% to 60.5%). On the Olympic Sports dataset, the trajectory accuracy remarkably improved from a baseline of 62.4% to 77.2%. These gains were consistently observed across different feature encoding techniques like the bag of features and Fisher vector methods.

6.2.2 Global Representation

As shown in Figures 6.1a and 6.1b, the concept of global representation is founded on the comprehensive portrayal of the entire human anatomy, encompassing both bodily shape and motion in whole action duration (Herath, Harandi, and Porikli, 2017). This approach has been observed to exhibit high sensitivity to noise, thereby affecting its robustness in various applications (Kong and Fu, 2018).

(Bobick and Davis, 2001) presents a groundbreaking, view-based method for capturing and identifying human movements through temporal templates, which are static vector images where each vector at a specific spatial point is determined by motion properties in a sequence of images. Using aerobics exercises as a testing ground, the research highlights the capability of a simplified, two-component template for real-time human action identification. The first component signifies the existence of motion, and the second quantifies its recency. Additionally, the paper delves into Motion Energy Images (MEI) and Motion History Images (MHI), both of which significantly enhance the discriminatory power in recognising motion. While MEIs encapsulates areas of cumulative motion, MHIs detail the temporal history of motion at individual pixels; their

integration proves to be especially effective. Notably, the approach excels in computational efficiency due to its recursive architecture, which eliminates the requirement to store historical data. Although there is some loss of information regarding the history of motion, the research concludes that the introduced methodology offers substantial potential for robust human motion recognition, with opportunities for future enhancements.

The paper by (Dalal and Triggs, 2005) delves into examining how effective different feature sets are for recognising visual objects, particularly honing in on detecting humans using a linear SVM framework. It thoroughly reviews existing edge and gradient-based descriptors and concludes that Histograms of Oriented Gradients (HOG) do a significantly better job compared to other methods. The HOG descriptors function by splitting the image window into small spatial regions, termed "cells," and then build a local 1-D histogram of gradient orientations within each cell. These cells are grouped into larger "blocks" to balance out local histogram "energy," making the method more robust to changes in lighting and shadows. The success of this approach hinges on the use of fine-scale gradients, precise orientation binning, rather coarse spatial binning, and high-quality local contrast normalisation in overlapping descriptor blocks. Through empirical testing, the paper demonstrates that this feature set nails near-perfect separation on the MIT pedestrian database, even introducing a tougher dataset to further back up the results. The paper points out that traditional smoothing techniques are a stumbling block to performance, and that gradients should be calculated at the finest scale available, followed by spatial blurring. It also finds that having strong local contrast normalisation is key and boosts performance from 84% to 89% at 10^{-4} False Positive Per Window (FPPW). Looking ahead, the authors suggest that although their current linear SVM is fairly efficient, there's still space for fine-tuning, maybe through a coarse-to-fine or rejection-chain style detector. They also mention the need to bring in motion information and a parts-based model to better capture the flexible nature of the human body for more generalised scenarios.

The manuscript delineated by (Ahad et al., 2011) unveils an advanced formulation of spatiotemporal (XYT) feature descriptors aimed at global-based action recognition, drawing foundational insights from the MHI technique. Precisely, it unfolds two innovative methodologies:

the SURF-based History Image (SbHI) technique and the Intensity-Accumulated Image (IAI) technique. The SbHI technique leverages the SURF detector for the selection of candidate points, and harnesses optical flow computations for motion vector derivation, with a targeted resolve towards ameliorating the motion-overwriting quandary engendered by self-occlusion phenomena. Conversely, the IAI technique is honed towards the management of occlusion or the absence of pertinent information in motion history. The manuscript posits that these methodologies exhibit a significant aptitude for real-time deployment in domains such as gaming and gesture recognition, albeit the global-centric paradigm of the approach engenders constraints in scenarios encompassing multiple individuals within the frame. Through an array of empirical evaluations, the authors elucidate the efficacy of the proposed methodologies, manifesting their operability within cluttered and diversely illuminated environments. Nonetheless, the discourse acknowledges the exigency for enhancements in computational efficacy and edge discernment capabilities. In summation, the manuscript underscores that the SbHI and IAI methodologies furnish propitious conduits for a myriad of applications within games, gesture, and action comprehension spheres, thereby advocating for continued investigative and developmental endeavours.

The paper by (Asumang et al., 2017) tackles the tough task of figuring out human poses, focusing on accurately identifying parts of the human body, which often gets tricky due to busy backgrounds and unclear detectors. The research suggests a three-step approach: a new part-learning technique, an evidence-supporting method, and a sub-graph pruning technique. The part-learning technique uses a special framework to efficiently spot human part candidates even when there's shape twisting and image misalignment. The evidence-supporting method boosts the certainty of detected human parts by using shared information between connected parts, which helps keep weaker parts from being wrongly pruned. The sub-graph pruning strategy works in a step-by-step way to handle the parts, reducing the computational work by narrowing down the state space early on and then using a step-by-step strategy for quicker detection. Tests on three public datasets show that this approach not only betters the detection rate of human body parts but also makes the process faster. However, the paper admits there's more work to

be done, especially in accurately locating lower arms and legs in situations with blockages and low contrast.

(Peng et al., 2020) unfolds a three-stream architecture tailored to enhance HAR tasks within video data. One stream zeroes in on spatial feature extraction from individual video frames, utilising a deep CNN, while the second and third streams delve into motion pattern analysis by processing optical flow fields generated through two distinct techniques: MBEpicflow and Flownet 2. To rigorously evaluate the model's performance, the research leverages four diverse datasets that span varying complexity and application domains, thereby illuminating the pivotal role of precise optical flow field generation on action recognition efficacy. The findings reveal that a surge in the accuracy of optical flow fields bolsters recognition rates by up to 2%, marking a substantial stride in machine learning model performance. Besides, the paper posits that traditional machine learning classifiers like SVMs trump deep learning classifiers in scenarios with smaller training datasets, a common occurrence in HAR tasks. This stance veers from the customary focus on deep learning algorithms within this research realm. Empirical results are furnished to substantiate this claim, demonstrating SVMs outshining deep learning classifiers in their setups. Additionally, the paper benchmarks the model against other state-of-the-art approaches, with results either comparable or superior in most instances, hence showcasing the model's robustness. An in-depth analysis of misclassifications is also embarked upon, shedding light on the challenges tied to identifying certain actions, thereby propelling suggestions for prospective research avenues. Conclusively, the research accentuates the integration of global temporal behaviour into the model as a viable strategy to further refine generalisation performance, underlining the criticality of long-term temporal structures in video data—a facet often sidelined in analogous models.

Recognising that traditional feature extraction methods like optical flow can be computationally intensive, particularly when dealing with large video datasets (Sun, Roth, and Black, 2010), this research seeks to explore new, more efficient techniques. The goal is to develop methods that reduce computational complexity while maintaining or improving the effectiveness of action recognition tasks. Drawing inspiration from recent advancements in background removal

(M. Singh, Basu, and Mandal, 2008), motion history techniques (Bobick and Davis, 2001), and the integration of motion information over time (Peng et al., 2020), three innovative feature extraction methods are proposed. These methods aim to strike a balance between computational efficiency and the ability to capture critical data features effectively. The proposed techniques are detailed in the subsequent sections and are designed to address the limitations of existing methods by offering a more scalable approach to feature extraction in complex video datasets, leveraging insights from both classical and contemporary research in the field (Kong and Fu, 2022; Hutchinson and Gadepally, 2021).

6.3 Methodology

This section provides a comprehensive overview of the handcraft feature extraction techniques employed in this research. These techniques are designed to capture more temporal information, thereby enriching the feature space of video frames for improved HAR.

Three distinct handcrafted feature extraction methods are introduced: Motion Aggregation (MAg), Differential Motion Trajectory (DMT), and Frame Variation Mapper (FVM). MAg represents a local feature, while DMT and FVM represent global features. Each method offers a unique approach to capturing temporal dynamics and will be explored in detail in the subsequent sections.

The three proposed methods have parallels in prior image processing and computer vision research. Similar techniques have been utilised in various applications where capturing and analysing motion dynamics are critical. For example, Motion Accumulation-like techniques are often employed in video summarising and activity recognition tasks to condense sequences by emphasising significant motion events (Meng et al., 2016). Differential motion analysis, akin to DMT, is widely used in object tracking and motion segmentation, where detecting changes over time between consecutive frames is crucial for identifying moving objects or segments (H. Wang and Schmid, 2013). Frame variance-based approaches, similar to FVM, have been applied in background subtraction and dynamic scene analysis, where distinguishing foreground

motion from static backgrounds is essential (Stauffer and Grimson, 1999). These techniques are foundational in many computer vision applications, ranging from surveillance to automated video analysis, demonstrating their broad utility beyond the specific context of HAR in robotics.

6.3.1 Motion Aggregation

Influenced by the SlowFast methodology, one of the main techniques in the MAg approach is the iterative adding of consecutive frames. This strategy captures the temporal dynamics between the frames, offering a richer representation of actions over time. As demonstrated by (Szeliski, 2022), to control the influence of each frame on the final concatenated frame, a weighted average scheme is employed.

Let $F(X)$ represent an individual frame in the video sequence, where X denotes the spatial coordinates of the frame. The process starts by generating an initial aggregated frame $G_1(X)$ from the first two frames $F_0(X)$ and $F_1(X)$ using the following equation:

$$G_1(X) = (1 - \alpha_1)F_0(X) + \alpha_1 F_1(X) \tag{6.1}$$

This equation is a linear interpolation between the two frames, with the weight α_1 allowing fine-tuning of each frame's contribution to $G_1(X)$.

After the first adding, the resulting frame $G_1(X)$ is then concatenated with the subsequent frame $F_2(X)$ using a similar equation:

$$G_2(X) = (1 - \alpha_2)G_1(X) + \alpha_2 F_2(X) \tag{6.2}$$

This process is generalised and iteratively applied, allowing for the incorporation of N frames into a single enriched frame $G_N(X)$, according to:

$$G_i(X) = (1 - \alpha_i)G_{i-1}(X) + \alpha_i F_i(X) \tag{6.3}$$

Here, i ranges from 2 to $N - 1$, and α_i controls the contribution of the i^{th} frame $F_i(X)$ in

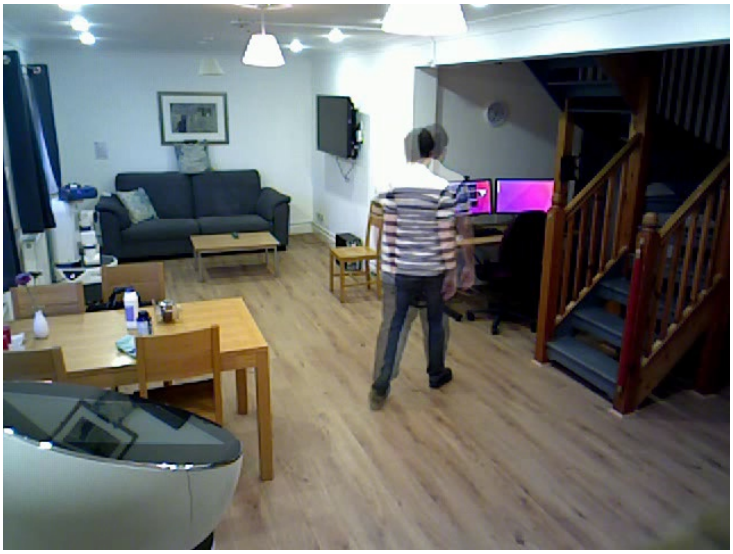


Figure 6.2: Trajectory Aggregation Frame Example with $\alpha = 0.5$

$G_i(X)$.

Figure 6.2 provides a visual representation of the result from the Front View, demonstrating the effectiveness of this iterative adding approach in capturing complex motions over time.

As depicted in Figure 6.2, the MAg technique offers a method for aggregating human motion trajectories across continuous frames, complete with background context. This approach eschews the random selection of individual frames in favor of a concatenated frame generated through MAg.

6.3.2 Frame Variation Mapper

The second technique, known as FVM, subtracts each new frame from the first frame of the action sequence.

Let $F(X)$ represent an individual frame in the video sequence, where X denotes the spatial coordinates of the frame. The equation for this subtraction method is:

$$H_i(X) = |F_1(X) - F_i(X)| \tag{6.4}$$

Here, $F_1(X)$ is the first frame of the action sequence, and $F_i(X)$ is any subsequent frame. The resulting frame $H(X)$ captures the absolute differences between the first frame and each

Figure 6.3: Frame Variation Mapper Frame Example for frame $H_i(X)$

subsequent frame in the action sequence.

This method allows for a more comprehensive understanding of how an action evolves over time, emphasising the changes relative to the initial frame. Figure 6.3 displays the method results' frame.

In Figure 6.3, the method is demonstrated to effectively remove background noise while providing a quantifiable variation between the initial and current positions of a dynamically moving object.

6.3.3 Differential Motion Trajectory

Another technique employed in this research for extracting motion trajectories involves background elimination through the absolute subtraction of consecutive frames. This method, referred to as DMT, focuses on minimising irrelevant information within the frame while retaining the essential temporal information crucial for action recognition.

Let $F(X)$ represent an individual frame in the video sequence, where X denotes the spatial coordinates of the frame. The initial step in this technique is to calculate the absolute difference between two consecutive frames $F_i(X)$ and $F_{i+1}(X)$. This is mathematically represented as:

$$D_i(X) = |F_i(X) - F_{i+1}(X)| \tag{6.5}$$

This operation yields a difference frame $D_i(X)$, capturing the absolute differences between $F_i(X)$ and $F_{i+1}(X)$.

The process is then iteratively applied to the next pair of consecutive frames $F_{i+1}(X)$ and $F_{i+2}(X)$, yielding another difference frame $D_{i+1}(X)$.

Subsequently, the two resulting difference frames $D_i(X)$ and $D_{i+1}(X)$ are combined using a weighted average parameter α, as given by:

$$Y_i(X) = (1 - \beta)D_i(X) + \beta D_{i+1}(X) \tag{6.6}$$

This composite frame $Y_i(X)$ encapsulates the dynamic changes occurring between multiple consecutive frames. The iterative process of subtraction and weighted averaging continues throughout the action sequence, thereby creating a set of enriched frames $Y_i(X)$ that effectively represent the entire action sequence.

One significant advantage of the DMT method is that it produces outputs similar to those obtained through optical flow extraction methods. However, DMT is computationally more efficient, making it a more practical choice for capturing temporal information in video frames. A snapshot of the results utilising this methodology is shown in Figure 6.4.

As shown in Figure 6.4, the methodology efficiently isolates motion trajectories by prioritising the temporal information from adjacent frames in a computationally efficient manner.

It is important to note that the three methods introduced—MAg, FVM, and DMT—are distinct from optical flow techniques. Unlike optical flow, which computes the motion of objects between consecutive frames based on pixel displacements, these methods focus on different strategies to capture temporal dynamics and motion information. MAg utilises iterative adding to create enriched frames that reflect temporal changes, FVM emphasises the absolute differences between the first and subsequent frames to highlight variations over time, and DMT isolates motion by subtracting consecutive frames and applying weighted averaging. These approaches

Figure 6.4: Differential Motion Trajectory Frame Example with $\beta = 0.65$

do not rely on the traditional principles of optical flow, offering alternative, computationally efficient means to analyse temporal features in video sequences.

6.4 Experiments

In this section, the arrangement of tests for the MAg, DMT, and FVM techniques using four different DL models: C3D (Tran, Bourdev, et al., 2015), SlowFast (Feichtenhofer, Fan, et al., 2019), Dual-stream CNN (Simonyan and Zisserman, 2014), and the proposed Dual-stream C3D in Chapter 5.3 is explained. The aim is to assess the impact of these techniques on the model's performance.

It is important to clarify that while mutual information (MI) was employed in the earlier stages of this research for dataset analysis—particularly to assess redundancy and similarity between video frames—it was not used in the deep learning model experiments presented in this chapter. The MI analysis was crucial for understanding the characteristics of the RHM dataset, but it did not play a role in the feature extraction techniques or in the evaluation of the deep models.

6.4.1 Preprocessing Time Analysis

In the context of HAR, the computational efficiency of feature extraction methods is a critical consideration, particularly when dealing with large-scale video datasets. Traditional methods like optical flow, while effective, can be computationally intensive, especially when applied to high-resolution video frames. This subsection presents an analysis of the preprocessing time required for the three proposed feature extraction methods—MAg, FVM, and DMT—compared to the standard optical flow method.

To provide a clear comparison, the preprocessing times for each method were measured on the same GPU system, specifically configured with Tesla V100 units. The preprocessing time for each method was calculated based on two consecutive frames with a resolution of 640x480 pixels. The results are summarised in Table 6.1.

Table 6.1: Preprocessing Time for Feature Extraction Methods

Feature Extraction Method	**Time (ms)**
MAg	0.3 ms
FVM	0.2 ms
DMT	0.7 ms
Optical Flow	1.6 ms

The data presented in Table 6.1 highlights the significant computational efficiency of the proposed methods compared to the optical flow method. The MAg method processes two frames in just 0.3 ms, which is more than five times faster than the 1.6 ms required by the optical flow method. Similarly, the FVM method is the most efficient, with a processing time of only 0.2 ms. The DMT method, although slightly more complex, still processes frames in 0.7 ms, which is significantly faster than optical flow.

These results demonstrate that the proposed feature extraction methods not only provide competitive accuracy in HAR tasks but also significantly reduce the computational overhead associated with preprocessing. This efficiency is crucial for real-time applications and large-scale datasets, where processing time can become a bottleneck. By reducing the time required for feature extraction, the proposed methods offer a practical alternative to traditional approaches like optical flow, making them well-suited for deployment in scenarios where computational

resources are limited or where real-time processing is required.

6.4.2 Robot Speed and Movement Considerations

The experiments conducted in this research were performed in an indoor environment, specifically designed to simulate elder care scenarios. Given the nature of these scenarios, the robot's movement was intentionally kept slow to mirror realistic conditions. The robot's speed and angular velocity were calibrated to align with typical human movement in such environments, ensuring that the interaction between the robot and the human subjects was safe and effective.

The robot's movement was designed with two key components in mind: pan-tilt (head) movement and whole-body movement. The pan-tilt mechanism, responsible for adjusting the robot's view, operated at a slow speed, with the pan (horizontal rotation) ranging from 5° to 10° per second, and the tilt (vertical rotation) between 3° to 5° per second. This slow and controlled movement ensured that the robot could effectively monitor its surroundings without causing abrupt changes in the field of view.

The whole-body movement of the robot was equally cautious, with a linear speed between 0.1 to 0.2 meters per second, and an angular speed for turning at 10° to 15° per second. These speeds were selected to allow smooth navigation in tight spaces, maintain stability, and ensure that the robot could stop or change direction quickly if necessary.

It is important to note that the robot's speed is not constant but rather adaptive, depending on the pace and actions of the human participants. This adaptive movement allows the robot to maintain a consistent and appropriate distance from the humans, which is crucial for accurate HAR in these controlled settings.

However, it is acknowledged that if the robot were to move faster or operate in outdoor environments, the current setup might encounter challenges. Faster movements could lead to increased motion blur and reduced frame-to-frame correspondence, potentially impacting the accuracy of the proposed feature extraction methods and overall HAR performance. Therefore, while the current setup is effective for the specific indoor, slow-paced scenarios studied, further adjustments and optimisations would be necessary for faster or more dynamic environments.

Figure 6.5: Sample of Extracted Temporal Feature Frames, Feeding to the models
The frames are prepared for feeding to the models as temporal frame data. The columns show the frame number. Each row displays the corresponding method in the first column. Also, the first row shows the normal frame.

6.4.3 One-Stream C3D model

In the initial phase of the experimental investigation, the efficacy of the newly introduced handcrafted feature extraction techniques was evaluated by deploying them on a C3D network model.

A sample of frame numbers 0, 4, 8, and 13 from the walking class of the front view is illustrated for all methods in Figure 6.5.

6.4.4 Dual-Stream C3D models

SlowFast Model

The SlowFast model was tested using three different pairs of views: Front-Robot, Back-Robot, and Top-Robot. The model comprises two streams, and various combinations of Normal/MAg and MAg/DMT/FVM frames were used for testing.

Dual-stream C3D

A similar experimental setup was followed for the Dual-stream C3D model with the SlowFast model. The same pairs of viewpoints—Front-Robot, Back-Robot, and Top-Robot—were used for testing. Additionally, the frame feeding strategy was kept consistent with the SlowFast experiments.

Dual-stream ConvNets

The Dual-stream ConvNet model, based on the architecture proposed by (Simonyan and Zisserman, 2014), was evaluated using the same experimental conditions as the SlowFast and Dual-stream C3D models. This model was tested with three pairs of views: Front-Robot, Back-Robot, and Top-Robot. Various combinations of frames, including Normal/MAg, MAg/DMT, and MAg/FVM, were utilised to explore the impact on model performance. The results were compared against those obtained from the SlowFast and Dual-stream C3D models to assess the efficacy of the dual-stream approach in enhancing action recognition accuracy.

6.4.5 Parameter details

For these experiments, the training parameters are configured as follows: a batch size of 30, a frame count of 16, a learning rate set at 0.0001, and the use of the SGD optimiser. Also, the Top-1 and Top-5 accuracy are the evaluated metrics in these experiments.

Hyper parameters were selected based on the related literature. The setup was as follows:

- **nEpochs**: 500

- **resume_epoch**: 0
- **useTest**: True
- **nTestInterval**: 20
- **snapshot**: 50
- **lr**: 1e-3
- **criterion**: CrossEntropyLoss
- **optimiser**: SGD with momentum=0.9 and weight_decay=5e-4
- **scheduler**: StepLR with step_size=10 and gamma=0.1

The experiments were conducted using the University of Hertfordshire GPU Cluster, specifically utilising gpu2 and gpu3. These machines are equipped with three Tesla V100 units, with gpu3 having 16 GB VRAM per unit and gpu2 having 32 GB VRAM per unit. This powerful computational setup ensured efficient handling of the large video datasets and the complex dual-stream model architectures, facilitating thorough and accurate training and evaluation processes.

In contrast to Chapter 5, the experiments in this Chapter introduced the use of temporal information alongside spatial data. This chapter explored the effectiveness of incorporating both raw temporal frames and temporally enriched frames generated through handcrafted feature extraction methods. The inclusion of temporal dynamics allowed for a more comprehensive evaluation of the deep learning models, highlighting the added value of temporal information in enhancing action recognition accuracy.

6.5 Results

6.5.1 One-Stream C3D

The outcomes of the C3D model are presented in Table 6.2. This table elucidates the model's performance in terms of both Top-1 and Top-5 accuracy metrics.

Table 6.2: Results of applying new feature frames on One Stream C3D

View	Frame Status	Top-1	Top-5
Front	Normal	67.59	97.92
Front	MAg	65.62	**98.14**
Front	DMT	65.1	96.43
Front	FVM	57.23	95.39
Back	Normal	66.63	97.77
Back	MAg	64.66	97.99
Back	DMT	65.18	96.36
Back	FVM	**70.37**	97.55
Top	Normal	62.21	96.95
Top	MAg	63.17	97.4
Top	DMT	59.53	95.76
Top	FVM	67.4	97.1
Robot	Normal	44.61	89.38
Robot	MAg	47.43	89.6
Robot	DMT	45.95	87.08
Robot	FVM	42.76	86.56

Feeding Normal, MAg, FVM and DMT frames into single stream C3D model with RHM dataset. The first line of each group demonstrates the basic results of feeding the normal frame to have a correct comparison. In all views, MAg feature shows the best Top-5 results.

As illustrated in Table 6.2, the incorporation of new handcraft feature extraction techniques yielded favourable outcomes on the performance of the C3D model. In terms of the Top-5 accuracy metric, across all viewpoints, the superior results were attributed to the application of the MAg technique. However, the Top-1 accuracy presented more varied results. For the Back and Top viewpoints, the FVM method emerged as the most effective. In contrast, the Robot viewpoint was best optimised using the MAg method, while the Front viewpoint yielded the highest accuracy when utilising the Normal frame.

Specifically focusing on the Robot viewpoint, both MAg and DMT methods outperformed the Normal frame in terms of Top-1 accuracy. Similarly, for the Top viewpoint, methods FVM and MAg demonstrated superior performance over the Normal frame. In the case of the Back viewpoint, only the FVM method surpassed the Normal frame. Interestingly, none of the applied methods were able to exceed the performance of the Normal frame for the Front viewpoint.

In a broader context, when evaluating the amalgamation of various viewpoints and frame techniques, the Back viewpoint utilising the FVM method secured the highest Top-1 accuracy.

Table 6.3: Results of applying new feature frames on SlowFast Model

View2	View1	Status1	Status2	Top1	Top5
Front	Robot	Normal	Normal	45.28	91.31
Front	Robot	Normal	MAg	49.81	90.64
Front	Robot	Normal	DMT	**53.97**	**93.31**
Front	Robot	Normal	FVM	47.58	91.83
Front	Robot	MAg	Normal	45.28	88.86
Front	Robot	MAg	MAg	50.11	92.5
Front	Robot	MAg	DMT	51.29	90.12
Front	Robot	MAg	FVM	50.18	90.94
Back	Robot	Normal	Normal	44.69	90.64
Back	Robot	Normal	MAg	46.84	90.42
Back	Robot	Normal	DMT	52.7	92.87
Back	Robot	Normal	FVM	49.51	91.83
Back	Robot	MAg	Normal	45.28	89.38
Back	Robot	MAg	MAg	47.06	90.27
Back	Robot	MAg	DMT	50.92	91.61
Back	Robot	MAg	FVM	48.4	91.23
Top	Robot	Normal	Normal	44.91	87.75
Top	Robot	Normal	MAg	45.73	88.49
Top	Robot	Normal	DMT	51.3	91.83
Top	Robot	Normal	FVM	47.8	90.49
Top	Robot	MAg	Normal	42.16	85.89
Top	Robot	MAg	MAg	45.87	89.45
Top	Robot	MAg	DMT	51.07	90.5
Top	Robot	MAg	FVM	49.59	91.61

Feeding Normal, MAg, FVM and DMT frames into SlowFast model with RHM dataset. The first line of each group demonstrates the basic results of feeding the normal frames to have a correct comparison. In all pairs, Normal frame form robot view and DMT temporal feature for static view shows the best Top-1 and Top-5 results.

Conversely, the Front viewpoint achieved the best Top-5 accuracy using the MAg method.

6.5.2 SlowFast Model

The results for the SlowFast model are presented in Table 6.3. The table shows the Top-1 and Top-5 accuracy metrics for different views and frame status combinations.

The results indicate several key findings. First, the use of new handcraft feature extraction methods led to significant improvements in both Top-1 and Top-5 accuracy metrics compared to the baseline scenario where normal frames were used in both streams. This suggests that new

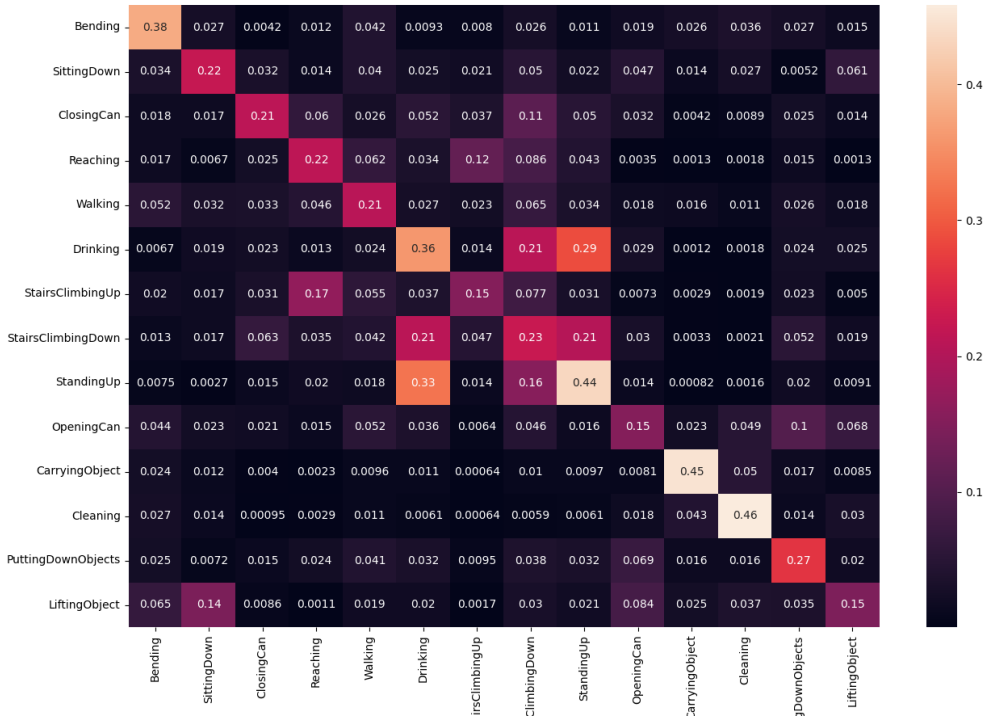


Figure 6.6: Confusion Matrix for Robot(Normal)-Front(DMT) views with SlowFast Model

techniques effectively enhance the model's performance.

Second, the best results for both Top-1 and Top-5 accuracy were achieved when "Normal" frames were used in the first stream, and the "DMT" method was used in the second stream. This configuration consistently outperformed other combinations across different pairs of viewpoints, indicating its effectiveness in capturing relevant features for human activity recognition.

Lastly, the Front-Robot configuration yielded the best overall results among all the pairs of viewpoints tested. This suggests that this particular viewpoint combination is most conducive to accurate HAR using the SlowFast model.

6.5.3 Dual-stream ConvNets

The outcomes of the Dual-stream ConvNets model are delineated in Table 6.4. This table delineates the performance of the model across various views and frame status combinations, quantified through the Top-1 and Top-5 accuracy metrics.

Table 6.4: Results of applying new feature frames on Dual-stream ConvNets Model

View2	View1	Status1	Status2	Top1	Top5
Front	Robot	Normal	Optical Flow	68.02	96.42
Front	Robot	Normal	MAg	62.91	94.88
Front	Robot	Normal	DMT	**68.35**	**97.50**
Front	Robot	Normal	FVM	66.07	96.1
Front	Robot	MAg	Optical Flow	67.33	96.21
Front	Robot	MAg	MAg	62.98	95.08
Front	Robot	MAg	DMT	67.42	96.36
Front	Robot	MAg	FVM	64.61	95.87
Back	Robot	Normal	Optical Flow	64.13	96.67
Back	Robot	Normal	MAg	60.48	92.89
Back	Robot	Normal	DMT	65.71	97.32
Back	Robot	Normal	FVM	63.47	95.1
Back	Robot	MAg	Optical Flow	63.24	96.22
Back	Robot	MAg	MAg	60.12	91.57
Back	Robot	MAg	DMT	63.89	96.46
Back	Robot	MAg	FVM	62.45	94.73
Top	Robot	Normal	Optical Flow	62.74	93.45
Top	Robot	Normal	MAg	58.11	89.92
Top	Robot	Normal	DMT	63.44	94.14
Top	Robot	Normal	FVM	61.06	91.62
Top	Robot	MAg	Optical Flow	61.64	92.2
Top	Robot	MAg	MAg	58.79	90.63
Top	Robot	MAg	DMT	62.46	93.19
Top	Robot	MAg	FVM	61.84	92.17

Feeding Normal, MAg, FVM and DMT frames into Dual-stream ConvNets model with RHM dataset. The first line of each group demonstrates the basic results of feeding the normal-optical flow frames to have a correct comparison. In all views, MAg feature shows the best Top-5 results.

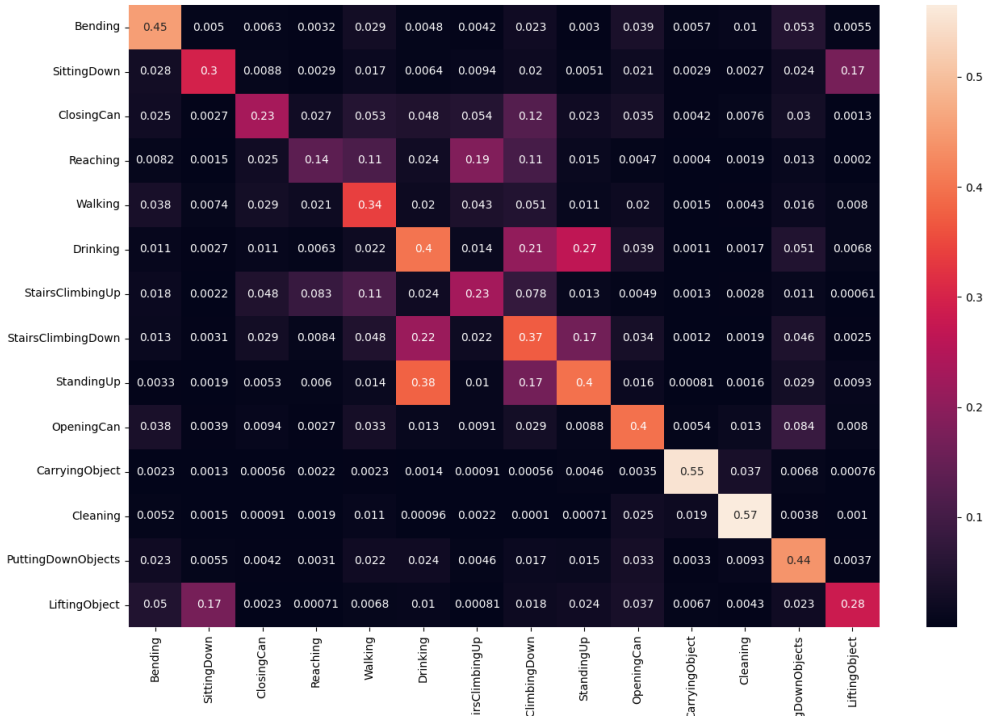


Figure 6.7: Confusion Matrix for Robot(Normal)-Front(DMT) views with Dual-stream ConvNets Model

The conducted tests incorporated spatiotemporal information within the model. For the Dual-stream ConvNets model, as detailed in (Simonyan and Zisserman, 2014), the temporal information is represented through Optical Flow.

The findings reveal that the only methodology surpassing the performance of Optical Flow is DMT. Across all view pairs, the combination of Normal frames with DMT (Normal-DMT) consistently yielded higher Top-1 and Top-5 results compared to the Normal-Optical Flow configuration.

Furthermore, the most outstanding results amongst all view pairs and status were observed with the Robot view using Normal frames and the Front view utilising DMT frames.

Overall, the experimental outcomes for the Dual-stream ConvNets model indicate that incorporating DMT as a means of temporal information within the model is more effective than using Optical Flow.

6.5.4 Dual-stream C3D

The results for the Dual-stream C3D model are presented in Table 6.5. The table shows the Top-1 and Top-5 accuracy metrics for different views and frame status combinations. Additionally, another test was performed using optical flow frames to provide a better comparison within the presented Dual-stream model.

The results from the Dual-stream C3D model present several key insights. Primarily, the incorporation of new techniques has resulted in notable enhancements in both Top-1 and Top-5 accuracy metrics across various viewpoint pairs. Specifically, the integration of the DMT method led to a Top-1 accuracy of 72.85%, which is an improvement of approximately 1.8% over the standard dual-stream configuration without DMT. Additionally, the model achieved a 71.06% Top-1 accuracy when combining the Front and Robot views, which is a significant 10% increase compared to the baseline C3D model's performance (refer to table 4.1). The Top-5 accuracy also saw a slight improvement, with a 0.81% increase when integrating temporal information. These figures highlight the substantial impact of incorporating temporal information and advanced feature extraction techniques in enhancing the model's efficacy in human action recognition, particularly in robot-centric environments.

Mirroring the performance trends seen in the SlowFast and Dual-stream ConvNets models, the findings indicate that the "Normal, DMT" frame status achieves the highest Top-1 and Top-5 accuracy across all viewpoint pairs.

Furthermore, when considering all view pairs and status, the paramount results in both Top-1 and Top-5 accuracies were observed in the Front-Robot pair using the "Normal, DMT" frame status. This outcome emphasises its effectiveness in capturing new temporal feature extraction for HAR using the Dual-stream C3D model.

The confusion matrices provided in Figures 6.6, 6.7, and 6.8 illustrate the performance of the SlowFast, Dual-stream CNN, and Dual-stream C3D models, respectively, when processing the Robot(Normal)-Front(DMT) views. A comparative analysis of these matrices reveals that their structural patterns remain consistent with the confusion matrices observed in earlier experiments, specifically those in Chapter 4 (Fig 4.2) and Chapter 5 (Fig 5.3). This consistency indicates that

Table 6.5: Results of applying new feature frames on Dual Stream C3D Model

View2	View1	Status1	Status2	Top1	Top5
Front	Robot	Normal	Normal	71.06	98.14
Front	Robot	Normal	Optical Flow	72.56	98.41
Front	Robot	Normal	MAg	72.18	98.72
Front	Robot	Normal	DMT	**72.85**	**98.95**
Front	Robot	Normal	FVM	71.59	97.88
Front	Robot	MAg	Optical Flow	71.14	97.68
Front	Robot	MAg	Normal	69.07	95.46
Front	Robot	MAg	DMT	71.22	97.73
Front	Robot	MAg	FVM	72.38	98.1
Front	Robot	MAg	MAg	70.03	98.7
Back	Robot	Normal	Normal	66.25	97.17
Back	Robot	Normal	Optical Flow	71.67	97.88
Back	Robot	Normal	MAg	69.81	97.65
Back	Robot	Normal	DMT	72.61	98.4
Back	Robot	Normal	FVM	70.33	97.51
Back	Robot	MAg	Optical Flow	71.05	96.91
Back	Robot	MAg	Normal	65.21	96.79
Back	Robot	MAg	DMT	72.19	97.73
Back	Robot	MAg	FVM	70.45	97.58
Back	Robot	MAg	MAg	71.63	97.95
Top	Robot	Normal	Normal	65.09	95.95
Top	Robot	Normal	Optical Flow	67.97	96.74
Top	Robot	Normal	MAg	67.5	97.65
Top	Robot	Normal	DMT	68.2	97.92
Top	Robot	Normal	FVM	68.17	97.58
Top	Robot	MAg	Optical Flow	67.36	97.23
Top	Robot	MAg	Normal	65.1	95.6
Top	Robot	MAg	DMT	68.1	97.8
Top	Robot	MAg	FVM	67.8	97.15
Top	Robot	MAg	MAg	68.17	96.84

Feeding Normal, MAg, FVM and DMT frames into proposed Dual-stream C3D model with RHM dataset. The first line of each group demonstrates the basic results of feeding the normal-normal flow frames to have a correct comparison. In all views, MAg feature shows the best Top-5 results. For better analysis and comparison of the proposed methods, optical flow frames were tested for the temporal information stream.

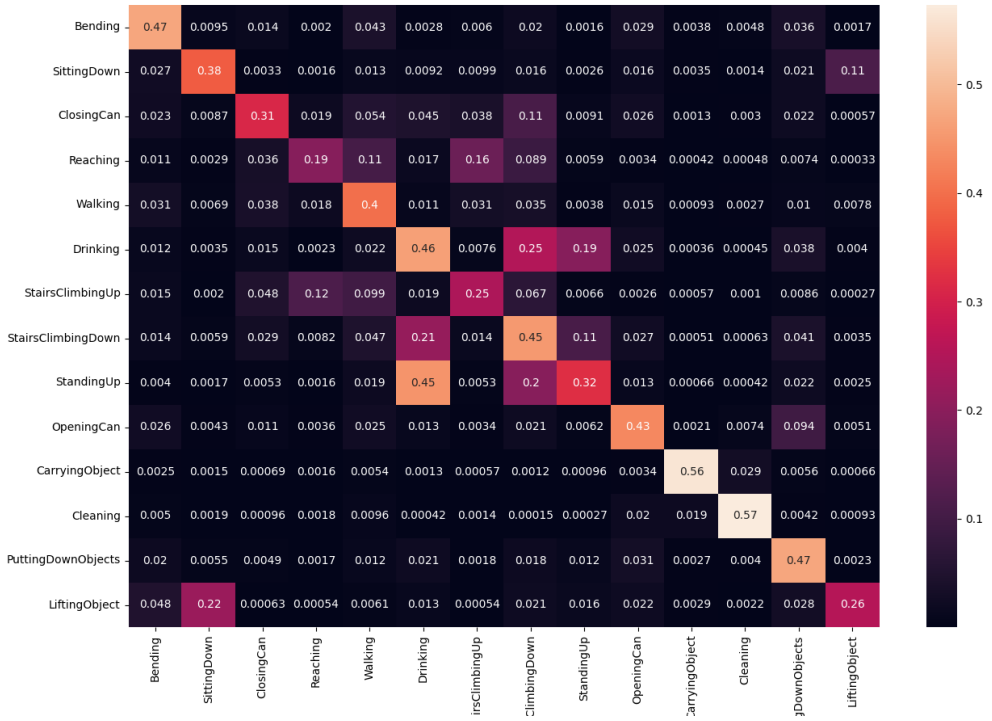


Figure 6.8: Confusion Matrix for Robot(Normal)-Front(DMT) views with Dual-stream C3D Model

despite the introduction of different dual-stream architectures in the current chapter, the overall confusion patterns have not undergone significant changes. The results across these models show a stable performance trend, suggesting that the dual-stream approach does not drastically alter the classification outcomes when compared to the base C3D model and other dual-stream configurations explored in previous chapters.

6.6 Chapter Contribution

The findings and methodologies presented in this chapter have been further validated and expanded upon in the recent publication titled *Multi-View Fusion and Feature Extraction: Enhancing HAR for Assistive Robotics* (Abadi et al., 2024a). This paper, presented at the 2024 IEEE RAS International Conference on Humanoid Robots (Humanoid 2024), delves into the intricacies of HAR in the context of HRI, especially within AAL scenarios. The accurate recognition of human activities is a pivotal challenge for enhancing interaction and cooperation between humans and autonomous systems in these environments.

The paper addresses the challenge of improving HAR in robotics by focusing on the integration of multi-view data and the extraction of temporal features from static cameras. Utilising the RHM dataset, which includes a robotic perspective alongside three other static views (Front, Back, Top), three innovative handcrafted feature extraction methods are introduced: MAg, DMT, and FVM. These methods are designed to enhance the representation of temporal information in static frames.

6.7 Chapter Summary

This chapter explored the impact of various handcrafted feature extraction techniques on the performance of four DL models for HAR: the C3D model (Tran, Bourdev, et al., 2015), the SlowFast model (Feichtenhofer, Fan, et al., 2019), the Dual-stream CNN model (Simonyan and Zisserman, 2014), and the proposed Dual-stream C3D model. The methods investigated include MAg, DMT, and FVM, each designed to capture different aspects of temporal information in

video frames.

For the first part, the experimental analysis underscores the efficacy of the newly introduced handcrafted feature extraction techniques on the single C3D model to establish baseline results for comparison. Notably, the MAg method consistently yielded superior results in terms of Top-5 accuracy across all viewpoints. However, Top-1 accuracy presented a more nuanced picture, with FVM emerging as the most effective method for Back and Top viewpoints, and MAg for the Robot viewpoint. Intriguingly, the Normal frame outperformed all methods in the Front viewpoint for Top-1 accuracy, suggesting a potential avenue for further investigation. Overall, the Back viewpoint employing the FVM method achieved the highest Top-1 accuracy, while the Front viewpoint utilising the MAg method excelled in Top-5 accuracy, highlighting the nuanced interplay between feature extraction methods and viewpoint-specific performance.

Table 6.6: Summary of Best Model Results from Each Chapter.

This table presents the top-performing models and methods, along with their Top-1 and Top-5 accuracy percentages, highlighting the advancements made in enhancing human action recognition in robot-centric environments.

Chapter	Best Model/View(Pair)	Top-1 Accuracy	Top-5 Accuracy
Chapter 4	C3D Model with Front View	67.59%	98.14%
Chapter 4	R3D Model with Robot View	48.10%	89.45%
Chapter 5	Dual-stream C3D Model (Front - Robot)	71.06%	98.14%
Chapter 6	DMT - Normal (Front - Robot)	72.85%	98.95%

For the next part of the experiments, three dual-stream models were tested: SlowFast, Dual-stream CNN, and the proposed Dual-stream C3D models. For all three models, the experiments revealed that the new handcrafted feature extraction techniques significantly improved both Top-1 and Top-5 accuracy metrics across almost all pairs of viewpoints. Specifically, the best performance for all models was generally achieved when "Normal" frames were used in the first stream and the "DMT" method was used in the second stream. The primary reason for this improvement is attributed to the incorporation of temporal information within the dual-stream networks.

Additionally, when comparing the presented feature extraction method and optical flow for temporal frames on the proposed dual-stream C3D model, the results indicate that the DMT

method achieved superior outcomes compared to both the other methods and the optical flow method. However, it is important to note that the FVM and MAg methods did not outperform the optical flow method.

Significant improvement in the accuracy of the Robot view was observed with the presented methods, particularly with the Dual-stream C3D model and the DMT method. The Dual-stream C3D model, which integrates the Robot view with other static views, achieved a remarkable 71.06% Top-1 accuracy, showcasing a substantial enhancement over traditional single-stream methods. The DMT method further boosted the performance, achieving the highest Top-1 accuracy of 72.85%. These results underline the effectiveness of the proposed dual-stream architecture and feature extraction techniques in capturing the dynamic nature of the Robot view, thereby improving action recognition accuracy in robot-centric environments.

While the removal of the background has provided substantial benefits, future work could further enhance these results by statistically optimising background alignment before removal. This could ensure that any residual motion blur or misalignment is corrected, thereby refining the quality of the extracted temporal features and potentially leading to even higher accuracy in more dynamic or uncontrolled environments.

In summary, this chapter contributes to the growing body of knowledge on optimising DL models for HAR by introducing and evaluating novel handcrafted feature extraction techniques for temporal information. The promising results pave the way for further investigations into more complex handcrafted feature extraction methods and their applications in multiview HAR.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The field of HRI in the context of assistive robotics merges advanced robotics with human activities to create intelligent assistance. This aim of HRI extends beyond simple tasks to include social and emotional interactions between humans and robots. This is especially important in AAL environments, where robotics has the potential to improve safety and contribute to a better quality of life. In this growing field, powered by ML and DL, HAR is crucial for robots to identify and respond to human actions. Combining HRI and HAR is key for developing smart robots that can navigate environments focused on human needs, like AAL, significantly advancing assistive robotics.

This research aimed to aid robots in recognising human actions in an AAL environment. An initial evaluation of existing HAR datasets revealed a lack of data from a robot's perspective. Prompted by previous studies emphasising important activities in daily living, the RHM dataset was developed specifically for HAR applications. Following its creation, the RHM dataset underwent thorough testing to determine its effectiveness and performance. To achieve better results from the robot's viewpoint, this study introduced an innovative method using the Dual-stream Convolutional Three Dimensions (C3D) network. Additionally, the research developed three new handcrafted feature extraction methods—Motion Aggregation (MAg), Differential

Motion Trajectory (DMT), and Frame Variation Mapper (FVM)—designed to capture temporal aspects. These methods were integrated into the Dual-stream C3D model, leading to enhanced performance in processing and recognising human activities from the perspective of a robot. The following research questions were answered using this formulation:

> **Q.1:** *How does the dynamics of a camera from a robot viewpoint impact the accuracy of Deep learning models in HAR?*

In the investigation, current datasets in HAR were assessed, revealing significant shortcomings in the area of HRI (Abadi et al., 2023). The evaluation highlighted three main areas of concern:

Firstly, concerning the dynamic perspective or Robot View, only the LIRIS (C. Wolf et al., 2012) and InHARD (Dallel et al., 2020) datasets which does not cover motion in the dataset. Recognising human actions from a moving robot's perspective is crucial in HRI. While some of the datasets mentioned in Table 2.1 might include motion, none provide a specialised collection specifically for moving camera perspectives. Secondly, there is an absence of the Top View or Fish Eye View in the examined datasets. This perspective is particularly important for AAL scenarios, and its lack in existing HAR datasets highlights a significant gap in current research resources. Thirdly, in terms of redundancy, a detailed examination of multiview datasets revealed critical insights. While many datasets offer multiple static camera angles and sometimes an egocentric viewpoint, it is primarily the LIRIS (C. Wolf et al., 2012) and InHARD (Dallel et al., 2020) datasets that provide Robot Views without motion.

In response to these gaps, a new RGB-based HAR dataset called the Robot House Multi-View (RHM) dataset was developed. The RHM dataset is carefully designed to incorporate the missing dynamic Robot View, an overhead Fish Eye View, and redundancy across multiple views, thereby filling the identified gaps in existing datasets.

The RHM dataset includes four distinct viewpoints: a static Front view, a static Back view, an overhead Fish Eye view, and a dynamic Robot view. From each of these perspectives, a substantial collection of 6,701 video recordings was compiled, resulting in a total of 26,804 videos across all the views. The dataset covers 14 different action categories, and importantly,

the videos within each category are time-synchronised across various viewpoints, providing a comprehensive and unified dataset.

Additionally, state-of-the-art DL models such as C3D (Tran, Bourdev, et al., 2015), ResNets with (2+1) Dimension convolutions (R(2+1)D) (Tran, H. Wang, et al., 2018), Three Dimensions ResNets (R3D) (Tran, H. Wang, et al., 2018), and SlowFast (Feichtenhofer, Fan, et al., 2019) were used to test their performance in HAR tasks using the RHM dataset. Results in Table 4.1 showed that the dynamic Robot View presented challenges for these models. The Robot View led to lower accuracy, both in Top-1 and Top-5 metrics, due to its inherent variability. In contrast, the Top View, which provides a broader perspective, consistently yielded higher accuracy scores.

In assessing the RHM dataset, evaluations were conducted using two distinct approaches. First, a new measurement technique based on Mutual Information (MI) was implemented. This method focuses on analysing the temporal and contextual features among video frames. The MI-based metric provides detailed insights into the flow of information, emphasising both the unique and shared content between frames. This approach is particularly useful for understanding the complexity of actions captured from different viewpoints. As Figure 4.1 showed, the Robot View, a dynamic angle, registered lower MI values, suggesting a richer diversity in its frame sequence compared to the higher MI values of static Front and Back Views, which indicated greater redundancy.

In analysing the confusion matrices for the C3D model, consistent classification patterns among certain actions were observed, regardless of the camera angle used. This recurring pattern suggests that the confusion in classification is less about the camera's viewpoint and more about the inherent complexities within the action classes themselves.

> **Q.2:** *Is there an enhancement in the accuracy of the robot view in a multi-stream DL model for HAR when other camera views are incorporated?*

The research focused on enhancing HAR in robot-centric scenarios. The effectiveness of a dual-stream deep learning architecture utilising the RHM dataset was explored. A Dual-stream C3D network model was introduced, integrating multiple views into a cohesive framework.

The Dual-stream C3D model used two C3D networks, each focusing on a different view. The architecture of this network is detailed in Table 5.1 and Figure 5.1. This dual-stream setup captured features specific to each view and combined insights through cross-connections, enhancing the representation of the robot view. For a comprehensive analysis of multiview impacts on multi-stream models, tests were solely performed with spatial frames. The results, presented in Table 5.3 and Table 5.4, showed a significant improvement in Top-1 accuracy, with the robot view experiencing a 10% increase using this dual approach. However, Top-5 accuracy did not show much change, suggesting that the dual-stream model's strength lies in its precision for the most likely classification.

For a more comprehensive comparison and evaluation, tests were conducted using the SlowFast (Feichtenhofer, Fan, et al., 2019) and Dual-stream CNN (Simonyan and Zisserman, 2014) models under the same experimental conditions. In these tests, the model demonstrated superior performance compared to both the SlowFast and the Dual-stream CNN models across all view pairs.

In the context of the experiments conducted, the combination of the top-view with the robot-view did not yield as significant an improvement in performance as some of the other view pairings. This outcome can be attributed to the inherent characteristics of the top-view perspective, which, while offering a comprehensive bird's-eye view of the environment, may lack the detailed, close-range information that is more effectively captured by views closer to the action, such as the front or side views.

In conclusion, the findings strongly support the use of multiview data in multi-stream models to enhance performance in robot-centric environments such as AAL. Additionally, the proposed Dual-stream C3D model achieved the highest results when compared to the SlowFast (Feichtenhofer, Fan, et al., 2019) and Dual-stream CNN (Simonyan and Zisserman, 2014) models.

> **Q.3:** *How does employing handcrafted feature extraction as temporal information on dual-stream DL model for a robot view and another view in parallel impact HAR?*

Knowing that the other views in the RHM dataset are static, three methods for temporal feature extraction that are low in computational cost were proposed. These methods are Motion Aggregation (MAg), Differential Motion Trajectory (DMT), and Frame Variation Mapper (FVM). Each method provides a unique approach to capturing the temporal dynamics present in video data.

In the first part of the experimental analysis, the effect of the new handcrafted feature extraction methods on the C3D (Tran, Bourdev, et al., 2015) model was focused on. Among these methods, MAg consistently achieved the best Top-5 accuracy for all camera views. For Top-1 accuracy, the results varied: the FVM method performed best for the Back and Top views, while MAg was most effective for the Robot view. For the Front view, normal frames showed the highest Top-1 accuracy, suggesting a potential area for further study. The Back view using FVM recorded the highest Top-1 accuracy, and the Front view with the MAg method excelled in Top-5 accuracy. This highlights the complex relationship between different feature extraction methods and their performance depending on the viewpoint.

In the research involving three models—the SlowFast (Feichtenhofer, Fan, et al., 2019), Dual-stream CNN (Simonyan and Zisserman, 2014), and the Dual-stream C3D—significant improvements in both Top-1 and Top-5 accuracy metrics for almost all viewpoint pairs were observed. This was largely due to the implementation of the new handcrafted feature extraction techniques. Notably, the best performance was achieved when 'Normal' frames were used in the first (robot) stream, combined with the DMT method in the second (static view) stream. This enhancement can primarily be attributed to the effective integration of temporal information within the Dual-stream networks.

Additionally, when the newly introduced feature extraction methods were compared with optical flow for temporal frames, tests were conducted using both the Dual-stream CNN and Dual-stream C3D models. In these tests, the DMT method outperformed both the alternative methods and optical flow. However, it is important to note that the FVM and MAg methods did not achieve better results than the optical flow method.

7.2 Contribution to the Body of Knowledge

The key contributions of this thesis are summarised below:

- **RHM Dataset Contribution:** Creation of the RHM dataset, which includes dynamic Robot View, top view (Fish Eye View), and redundancy across multiple views. This dataset addresses critical facets often missing in existing datasets.
  - "RHM: Robot House Multi-view Human Activity Recognition Dataset" (Abadi et al., 2023): Describes the development and validation of the RHM dataset for HAR tasks.
  - "RHM-HAR-SK: A multi-view dataset with skeleton data for ambient assisted living research" (Alashti et al., 2023b): Introduces the RHM dataset with added skeleton data, enhancing its application in AAL scenarios.
  - "Robot house human activity recognition dataset" (Abadi et al., 2021): Provides an overview of the dataset's structure and potential applications.
- **RHM Analysis Contribution:** Introduction of a novel metric based on Mutual Information (MI) for analysing HAR datasets, focusing on temporal dependencies and information redundancy. Detailed in the publication "RHM: Robot House Multi-View Human Activity Recognition Dataset" (Abadi et al., 2023).
- **Dual-Stream C3D Model Contribution:** Development of a multi-stream model, the Dual-stream C3D, combining multiple views to improve accuracy. Detailed in "Robotic Vision and Multi-View Synergy: Action and Activity Recognition in Assisted Living Scenarios" (Abadi et al., 2024b).
- **Multi-View Fusion and Feature Extraction for Enhancing HAR:** Introduction of three innovative handcrafted feature extraction methods: Motion Aggregation (MAg), Differential Motion Trajectory (DMT), and Frame Variation Mapper (FVM). These methods significantly boost performance in dual-stream models. Detailed in "Multi-View Fusion and Feature Extraction: Enhancing HAR for Assistive Robotics" (Abadi et al., 2024a).

In summary, the contributions of this thesis span the creation of a novel dataset, the introduction of new analytical metrics and models, and the development of innovative feature extraction techniques, all aimed at advancing the field of Human Activity Recognition in the context of Human-Robot Interaction.

In response to concerns about the generalisation of the models, it is important to note that all trained models have been saved and are available for further analysis and testing. These models, along with the RHM dataset, have been shared openly to facilitate reproducibility and external validation. The GitHub repository containing the code used for training and evaluation is also publicly accessible, allowing other researchers to apply these models to different datasets and assess their performance in diverse scenarios. While this study primarily focused on the RHM dataset, the availability of the models and code provides a valuable resource for testing generalisability on other datasets. This openness to external validation underscores the robustness of the proposed methods and their potential for broader application in the field of HAR.

7.3 Limitations

While this research has made significant strides in advancing HAR within HRI, particularly in elder care scenarios, certain limitations must be acknowledged. One of the primary limitations is the controlled nature of the RHM dataset, which was specifically designed to simulate indoor environments with slow, deliberate robot movement. The robot's speed was calibrated to approximately 0.1 m/s, with minimal rotation, to ensure safe and effective interaction with human subjects. While this setup is well-suited for the intended elder care scenarios, it may not fully represent the challenges posed by faster robot speeds or more dynamic environments, such as outdoor settings or situations requiring rapid rotational movements.

This controlled environment and limited robot motion may restrict the generalisability of the developed models to more diverse or complex real-world scenarios. For instance, faster robot movements could introduce motion blur, reduce the frame-to-frame correspondence, and

potentially impact the accuracy of HAR tasks. Additionally, the dataset's focus on a narrow range of activities and interactions may limit the robustness of the models when applied to broader contexts.

Another limitation is the relatively small scale of the dataset, which could lead to potential overfitting of the models. While the models have demonstrated high accuracy within the confines of the RHM dataset, their performance on larger and more varied datasets remains untested. This raises concerns about the models' ability to generalise to different environments and scenarios beyond those simulated in this research.

These limitations highlight the need for future work to explore and validate the models in more diverse settings and to consider the implications of more dynamic robot movements in HAR tasks.

7.4 Future Work

This dissertation establishes a strong base for future studies in the combined fields of Human-Robot Interaction (HRI) and Human Action Recognition (HAR). The intersection of these areas provides ample scope for ongoing innovation and the development of advanced assistive robotic systems. Here are some recommended paths for future research in these domains.

7.4.1 Extension of the Range and Number of Activities for RHM Dataset

Incorporating additional action categories will enrich the dataset, offering a broader spectrum of human activities. This expansion will facilitate a more comprehensive analysis of HAR algorithms and their capability to generalise across various human behaviors.

Also, increasing the number of videos per action class will contribute to the dataset's diversity and complexity, enabling the development of more robust and fault-tolerant HAR systems.

7.4.2 Multiple People Interaction Expansion for RHM Dataset

Currently, the dataset predominantly focuses on singular human actions. Future work should include interactions between multiple humans, reflecting the complexities and dynamics of real-world interactions.

7.4.3 Human-Robot Interaction Expansion for RHM Dataset

To advance the utility of the dataset in HRI scenarios, the introduction of Human-Robot Interaction scenarios is crucial. These interactions would provide invaluable data for training and evaluating HAR systems in collaborative tasks.

7.4.4 Activities Monitoring for Multiple People in Space

A key challenge will be differentiating between individuals and understanding their interactions within the same environment. This will involve leveraging sophisticated HAR techniques and potentially incorporating elements of machine learning like object recognition and HAR.

7.4.5 Mutual Information Use in DL Models

The use of mutual information will focus on optimising cluster selection within the model. This strategy is intended to minimise confusion and increase the model's efficiency, resulting in a more streamlined and lightweight design.

References

Abadi, Mohammad Bamorovat et al. (2021). "Robot house human activity recognition dataset". In: *The 4th UK-RAS Conference for PhD Students & Early-Career Researchers on'Robotics at Home'*. EPSRC UK-RAS Network, pp. 19–20.

— (2023). "Rhm: Robot house multi-view human activity recognition dataset". In: *IARIA, March*.

— (2024a). "Multi-View Fusion and Feature Extraction: Enhancing HAR for Assistive Robotics". In: *2024 IEEE RAS International Conference on Humanoid Robots*. Institute of Electrical and Electronics Engineers (IEEE).

— (2024b). "Robotic Vision and Multi-View Synergy: Action and activity recognition in assisted living scenarios". In: *2024 IEEE RAS EMBS 10th International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. Institute of Electrical and Electronics Engineers (IEEE).

Abu-El-Haija, Sami et al. (2016). "Youtube-8m: A large-scale video classification benchmark". In: *arXiv preprint arXiv:1609.08675*.

Aggarwal, Jake K and Lu Xia (2014). "Human activity recognition from 3d data: A review". In: *Pattern Recognition Letters* 48, pp. 70–80.

Agrigoroaie, Roxana, François Ferland, and Adriana Tapus (2016). "The enrichme project: Lessons learnt from a first interaction with the elderly". In: *International Conference on Social Robotics*. Springer, pp. 735–745.

Ahad, Md Atiqur Rahman et al. (2011). "Approaches for global-based action representations for games and action understanding". In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, pp. 753–758.

Alashti, Mohamad Reza Shahabian et al. (2023a). "Lightweight human activity recognition for ambient assisted living,"" in: *IARIA, March*.

— (2023b). "Rhm-har-sk: A multi-view dataset with skeleton data for ambient assisted living research". In: *IARIA, March*.

Amirabdollahian, Farshid et al. (2013). "Accompany: Acceptable robotiCs COMPanions for AgeiNG Years—Multidimensional aspects of human-system interactions". In: *2013 6th International Conference on Human System Interactions (HSI)*. IEEE, pp. 570–577.

Amiri, Ali and Mahmood Fathy (2010). "Hierarchical keyframe-based video summarization using QR-decomposition and modified-means clustering". In: *EURASIP Journal on Advances in Signal Processing* 2010, pp. 1–16.

Antani, Sameer, Rangachar Kasturi, and Ramesh Jain (2002). "A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video". In: *Pattern recognition* 35.4, pp. 945–965.

Asumang, Emmanuel Kofi Nii et al. (2017). "Human pose estimation based on evidence supporting and sub-graph pruning". In: *2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, pp. 20–27.

Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool (2006). "Surf: Speeded up robust features". In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer, pp. 404–417.

Bedaf, Sandra et al. (2014). "Which activities threaten independent living of elderly when becoming problematic: inspiration for meaningful service robot functionality". In: *Disability and Rehabilitation: Assistive Technology* 9.6, pp. 445–452.

Bobick, Aaron F and James W Davis (2001). "The recognition of human movement using temporal templates". In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 3, pp. 257–267.

Broadbent, Elizabeth, Rebecca Stafford, and Bruce MacDonald (2009). "Acceptance of healthcare robots for the older population: Review and future directions". In: *International journal of social robotics* 1, pp. 319–330.

Brunelli, R. (2009). *Template Matching Techniques in Computer Vision: Theory and Practice*. Wiley.

Carreira, Joao, Eric Noland, Andras Banki-Horvath, et al. (2018). "A short note about kinetics-600". In: *arXiv preprint arXiv:1808.01340*.

Carreira, Joao, Eric Noland, Chloe Hillier, et al. (2019). "A short note on the kinetics-700 human action dataset". In: *arXiv preprint arXiv:1907.06987*.

Cheng, Zhongwei et al. (2012). "Human daily action analysis with multi-view and color-depth data". In: *European Conference on Computer Vision*. Springer, pp. 52–61.

Cover, Thomas M (1999). *Elements of information theory*. John Wiley & Sons.

Dalal, Navneet and Bill Triggs (2005). "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee, pp. 886–893.

Dalal, Navneet, Bill Triggs, and Cordelia Schmid (2006). "Human detection using oriented histograms of flow and appearance". In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part II 9*. Springer, pp. 428–441.

Dallel, Mejdi et al. (2020). "Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics". In: *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, pp. 1–6.

Damen, Dima, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, et al. (2018). "Scaling egocentric vision: The epic-kitchens dataset". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 720–736.

Damen, Dima, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, et al. (2022). "Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100". In: *International Journal of Computer Vision* 130.1, pp. 33–55.

Davidson, D. (1959). "Experimental tests of a stochastic decision theory using pairwise comparisons". In: *Journal of Mathematical Psychology* 6.1, pp. 1–20.

Donahue, Jeffrey et al. (2015). "Long-term recurrent convolutional networks for visual recognition and description". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634.

Ejaz, Naveed, Tayyab Bin Tariq, and Sung Wook Baik (2012). "Adaptive key frame extraction for video summarization using an aggregation mechanism". In: *Journal of Visual Communication and Image Representation* 23.7, pp. 1031–1040.

Feichtenhofer, Christoph (2020). "X3d: Expanding architectures for efficient video recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213.

Feichtenhofer, Christoph, Haoqi Fan, et al. (2019). "Slowfast networks for video recognition". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211.

Feichtenhofer, Christoph, Axel Pinz, and Richard P Wildes (2017). "Spatiotemporal multiplier networks for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4768–4777.

Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman (2016). "Convolutional two-stream network fusion for video action recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941.

Fischler, Martin A and Robert C Bolles (1981). "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *Communications of the ACM* 24.6, pp. 381–395.

Gao, Xinbo et al. (2009). "Shot-based video retrieval with optical flow tensor and HMMs". In: *Pattern Recognition Letters* 30.2, pp. 140–147.

Georgiadis, Dimosthenis et al. (2016). "A Robotic Cloud Ecosystem for Elderly Care and Ageing Well: The GrowMeUp Approach". In: *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*. Springer, pp. 919–924.

Gorelick, Lena et al. (2007). “Actions as space-time shapes”. In: *IEEE transactions on pattern analysis and machine intelligence* 29.12, pp. 2247–2253.

Goyal, Raghav et al. (2017). “The" something something" video database for learning and evaluating visual common sense”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 5842–5850.

Gu, Ye et al. (2020). “Multiple stream deep learning model for human action recognition”. In: *Image and Vision Computing* 93, p. 103818.

Guo, Yuejun et al. (2016). “Selecting video key frames based on relative entropy and the extreme studentized deviate test”. In: *Entropy* 18.3, p. 73.

Hannane, Rachida, Abdessamad Elboushaki, and Karim Afdel (2018). “MSKVS: Adaptive mean shift-based keyframe extraction for video summarization and a new objective verification approach”. In: *Journal of Visual Communication and Image Representation* 55, pp. 179–200.

Harris, Christopher G, Mike Stephens, et al. (1988). “A combined corner and edge detector.” In: *Alvey vision conference*. Vol. 15. 50. Citeseer, pp. 10–5244.

He, Dongliang et al. (2019). “Stnet: Local and global spatial-temporal modeling for action recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33, pp. 8401–8408.

Herath, Samitha, Mehrtash Harandi, and Fatih Porikli (2017). “Going deeper into action recognition: A survey”. In: *Image and vision computing* 60, pp. 4–21.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). “Long short-term memory”. In: *Neural computation* 9.8, pp. 1735–1780.

Horn, Berthold KP and Brian G Schunck (1981). “Determining optical flow”. In: *Artificial intelligence* 17.1-3, pp. 185–203.

Hutchinson, Matthew S and Vijay N Gadepally (2021). “Video action understanding”. In: *IEEE Access* 9, pp. 134611–134637.

Ji, Shuiwang et al. (2012). “3D convolutional neural networks for human action recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.1, pp. 221–231.

Jia, Baoxiong et al. (2020). "Lemma: A multi-view dataset for learning multi-agent multi-task activities". In: *European Conference on Computer Vision*. Springer, pp. 767–786.

Karpathy, Andrej et al. (2014). "Large-scale video classification with convolutional neural networks". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732.

Kay, Will et al. (2017). "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950*.

Klaser, Alexander, Marcin Marszałek, and Cordelia Schmid (2008). "A spatio-temporal descriptor based on 3d-gradients". In.

Kliper-Gross, Orit, Tal Hassner, and Lior Wolf (2011). "The action similarity labeling challenge". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.3, pp. 615–621.

Kong, Yu and Yun Fu (2018). "Human action recognition and prediction: A survey". In: *arXiv preprint arXiv:1806.11230*.

— (2022). "Human action recognition and prediction: A survey". In: *International Journal of Computer Vision* 130.5, pp. 1366–1401.

Kostavelis, Ioannis et al. (2019). "Understanding of human behavior with a robotic agent through daily activity analysis". In: *International Journal of Social Robotics*, pp. 1–26.

Kuehne, Hildegard et al. (2011). "HMDB: a large video database for human motion recognition". In: *2011 International Conference on Computer Vision*. IEEE, pp. 2556–2563.

Laptev, Ivan (2005). "On space-time interest points". In: *International journal of computer vision* 64.2-3, pp. 107–123.

Laptev, Ivan et al. (2008). "Learning realistic human actions from movies". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8.

Li, Ang et al. (2020). "The ava-kinetics localized human actions video dataset". In: *arXiv preprint arXiv:2005.00214*.

Li, WenLin et al. (2020). "Video summarization based on mutual information and entropy sliding window method". In: *Entropy* 22.11, p. 1285.

Liu, Jingen, Jiebo Luo, and Mubarak Shah (2009). "Recognizing realistic actions from videos "in the wild"". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1996–2003.

Liu, Jun et al. (2019). "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding". In: *IEEE transactions on pattern analysis and machine intelligence* 42.10, pp. 2684–2701.

Lowe, David G (1999). "Object recognition from local scale-invariant features". In: *Proceedings of the seventh IEEE international conference on computer vision*. Vol. 2. Ieee, pp. 1150–1157.

Lucas, Bruce D and Takeo Kanade (1981). "An iterative image registration technique with an application to stereo vision". In: *IJCAI'81: 7th international joint conference on Artificial intelligence*. Vol. 2, pp. 674–679.

Ma, L et al. (2018). "Image keyframe-based visual-depth map establishing method". In: *J. Harbin Inst. Technol* 50.11, pp. 23–31.

Mahdisoltani, Farzaneh et al. (2018). "On the effectiveness of task granularity for transfer learning". In: *arXiv preprint arXiv:1804.09235*.

Marszalek, Marcin, Ivan Laptev, and Cordelia Schmid (2009). "Actions in context". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 2929–2936.

Meghdadi, Amir H and Pourang Irani (2013). "Interactive exploration of surveillance video through action shot summarization and trajectory visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12, pp. 2119–2128.

Mei, Shaohui et al. (2015). "Video summarization via minimum sparse reconstruction". In: *Pattern Recognition* 48.2, pp. 522–533.

Meng, Ting-Chun et al. (2016). "Event-driven video summarization: A benchmark dataset and baseline results". In: *arXiv preprint arXiv:1606.04662*.

Nagendran, Arjun, Don Harper, and Mubarak Shah (2010). "New system performs persistent wide-area aerial surveillance". In: *SPIE Newsroom* 5, pp. 20–28.

Peng, Cheng et al. (2020). “Motion boundary emphasised optical flow method for human action recognition”. In: *IET Computer Vision* 14.6, pp. 378–390.

Pérez-Rodrıéguez, Rodrigo et al. (2019). “FriWalk robotic walker: usability, acceptance and UX evaluation after a pilot study in a real environment”. In: *Disability and Rehabilitation: Assistive Technology*, pp. 1–10.

Qiu, Renxi et al. (2012). “Towards robust personal assistant robots: Experience gained in the SRS project”. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 1651–1657.

Rai, Nishant et al. (2021). “Home action genome: Cooperative compositional action understanding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11184–11193.

Rao, Pamarthy Chenna and M Mariya Das (2012). “Keyframe extraction method using contourlet transform”. In: *Proceedings of the 2012 International Conference on Electronics, Communications and Control*, pp. 437–440.

Reddy, Kishore K and Mubarak Shah (2013). “Recognizing 50 human action categories of web videos”. In: *Machine Vision and Applications* 24.5, pp. 971–981.

Robinson, Anthony J and F Fallside (1988). “Static and dynamic error propagation networks with application to speech coding”. In: *Neural information processing systems*, pp. 632–641.

Rosebrock, Adrian (2022). “Deep Learning for Computer Vision with Python”. In.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). “Learning representations by back-propagating errors”. In: *nature* 323.6088, pp. 533–536.

Schuldt, Christian, Ivan Laptev, and Barbara Caputo (2004). “Recognizing human actions: a local SVM approach”. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* Vol. 3. IEEE, pp. 32–36.

Scovanner, Paul, Saad Ali, and Mubarak Shah (2007). “A 3-dimensional sift descriptor and its application to action recognition”. In: *Proceedings of the 15th ACM international conference on Multimedia*. ACM, pp. 357–360.

Shahabian Alashti, Mohamad Reza et al. (2023). "RH-HAR-SK: A Multi-view Dataset with Skeleton Data for Ambient Assisted Living Research". In: *ACHI 2023: The Sixteenth International Conference on Advances in Computer-Human Interactions*. IARIA.

Shahroudy, Amir et al. (2016). "Ntu rgb+ d: A large scale dataset for 3d human activity analysis". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019.

Shi, Qinfeng et al. (2011). "Human action segmentation and recognition using discriminative semi-markov models". In: *International journal of computer vision* 93.1, pp. 22–32.

Sigurdsson, Gunnar A et al. (2018). "Actor and observer: Joint modeling of first and third-person videos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7396–7404.

Simonyan, Karen and Andrew Zisserman (2014). "Two-stream convolutional networks for action recognition in videos". In: *Advances in neural information processing systems* 27.

Singh, Meghna, Anup Basu, and Mrinal Kr Mandal (2008). "Human activity recognition based on silhouette directionality". In: *IEEE transactions on circuits and systems for video technology* 18.9, pp. 1280–1292.

Singh, Sanchit, Sergio A Velastin, and Hossein Ragheb (2010). "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods". In: *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*. IEEE, pp. 48–55.

Smaira, Lucas et al. (2020). "A short note on the kinetics-700-2020 human action dataset". In: *arXiv preprint arXiv:2010.10864*.

Soomro, Khurram and Amir R Zamir (2014). "Action recognition in realistic sports videos". In: *Computer vision in sports*. Springer, pp. 181–208.

Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah (2012). "UCF101: A dataset of 101 human actions classes from videos in the wild". In: *arXiv preprint arXiv:1212.0402*.

Stauffer, Chris and W Eric L Grimson (1999). "Adaptive background mixture models for real-time tracking". In: *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. Vol. 2. IEEE, pp. 246–252.

Sun, Deqing, Stefan Roth, and Michael J Black (2010). "Secrets of optical flow estimation and their principles". In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. 2432–2439.

Szeliski, Richard (2022). *Computer vision: algorithms and applications*. Springer Nature.

Tran, Du, Lubomir Bourdev, et al. (2015). "Learning spatiotemporal features with 3d convolutional networks". In: *Proceedings of the IEEE international conference on computer vision*, pp. 4489–4497.

Tran, Du and Alexander Sorokin (2008). "Human activity recognition with metric learning". In: *European conference on computer vision*. Springer, pp. 548–561.

Tran, Du, Heng Wang, et al. (2018). "A closer look at spatiotemporal convolutions for action recognition". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459.

Van Oost, Ellen and Darren Reed (2010). "Towards a sociological understanding of robots as companions". In: *International Conference on Human-Robot Personal Relationship*. Springer, pp. 11–18.

Varol, Gül, Ivan Laptev, and Cordelia Schmid (2017). "Long-term temporal convolutions for action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 40.6, pp. 1510–1517.

Voorhees, Ellen M (n.d.). "National Institute of Standards and Technology Gaithersburg, MD 20899". In: ().

Wang, Heng, Alexander Kläser, et al. (2013). "Dense trajectories and motion boundary descriptors for action recognition". In: *International journal of computer vision* 103, pp. 60–79.

Wang, Heng and Cordelia Schmid (2013). "Action recognition with improved trajectories". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3551–3558.

Wang, Limin, Zhe Wang, et al. (2015). "CUHK&SIAT submission for thumos15 action recognition challenge". In: *THUMOS Action Recognition challenge*, pp. 1–3.

Wang, Limin, Yuanjun Xiong, et al. (2015). "Towards good practices for very deep two-stream convnets". In: *arXiv preprint arXiv:1507.02159*.

Wang, Wenguan et al. (2015). "Robust video object cosegmentation". In: *IEEE Transactions on Image Processing* 24.10, pp. 3137–3148.

Wang, Yunbo et al. (2017). "Spatiotemporal pyramid network for video action recognition". In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1529–1538.

Weinland, Daniel, Remi Ronfard, and Edmond Boyer (2006). "Free viewpoint action recognition using motion history volumes". In: *Computer vision and image understanding* 104.2-3, pp. 249–257.

Wolf, Christian et al. (2012). "The LIRIS Human activities dataset and the ICPR 2012 human activities recognition and localization competition". In: *LIRIS Umr* 5205.

Xu, Zhongwen, Yi Yang, and Alex G Hauptmann (2015). "A discriminative CNN video representation for event detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1798–1807.

Yan, Haibin, Marcelo H Ang, and Aun Neow Poo (2014). "A survey on perception methods for human–robot interaction in social robots". In: *International Journal of Social Robotics* 6.1, pp. 85–119.

Yao, Ping (2022). "Key frame extraction method of music and dance video based on multicore learning feature fusion". In: *Scientific Programming* 2022.1, p. 9735392.

Yin, Yifang, Roshan Thapliya, and Roger Zimmermann (2016). "Encoded semantic tree for automatic user profiling applied to personalized video summarization". In: *IEEE Transactions on Circuits and Systems for Video Technology* 28.1, pp. 181–192.

Yu, Linchen et al. (2018). "Key frame extraction scheme based on sliding window and features". In: *Peer-to-Peer Networking and Applications* 11, pp. 1141–1152.

Zhang, Lanshan et al. (2016). "KaaS: A standard framework proposal on video skimming". In: *IEEE Internet Computing* 20.4, pp. 54–59.

Zhang, Zufan et al. (2020). "Human action recognition using convolutional LSTM and fully-connected LSTM with different attentions". In: *Neurocomputing* 410, pp. 304–316.

Zhou, Kaiyang, Yu Qiao, and Tao Xiang (2018). "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Zhu, Jianqing et al. (2014). "High-performance video condensation system". In: *IEEE Transactions on Circuits and Systems for Video Technology* 25.7, pp. 1113–1124.

Zhu, Yanming, Kun Li, and Jianmin Jiang (2014). "Video super-resolution based on automatic key-frame selection and feature-guided variational optical flow". In: *Signal Processing: Image Communication* 29.8, pp. 875–886.

Zhu, Yi et al. (2019). "Hidden two-stream convolutional networks for action recognition". In: *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, pp. 363–378.

Chapter 8

Appendix

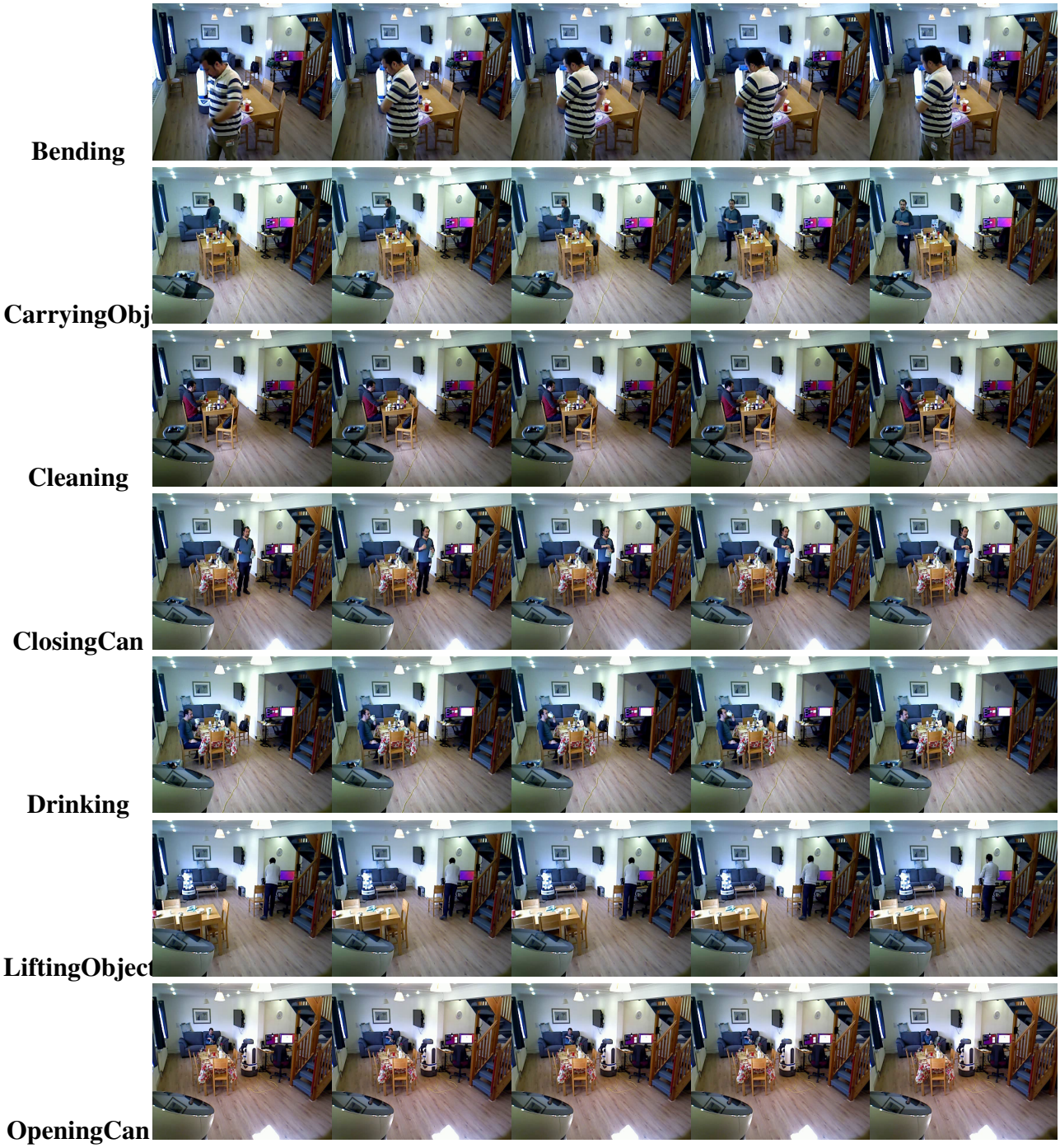
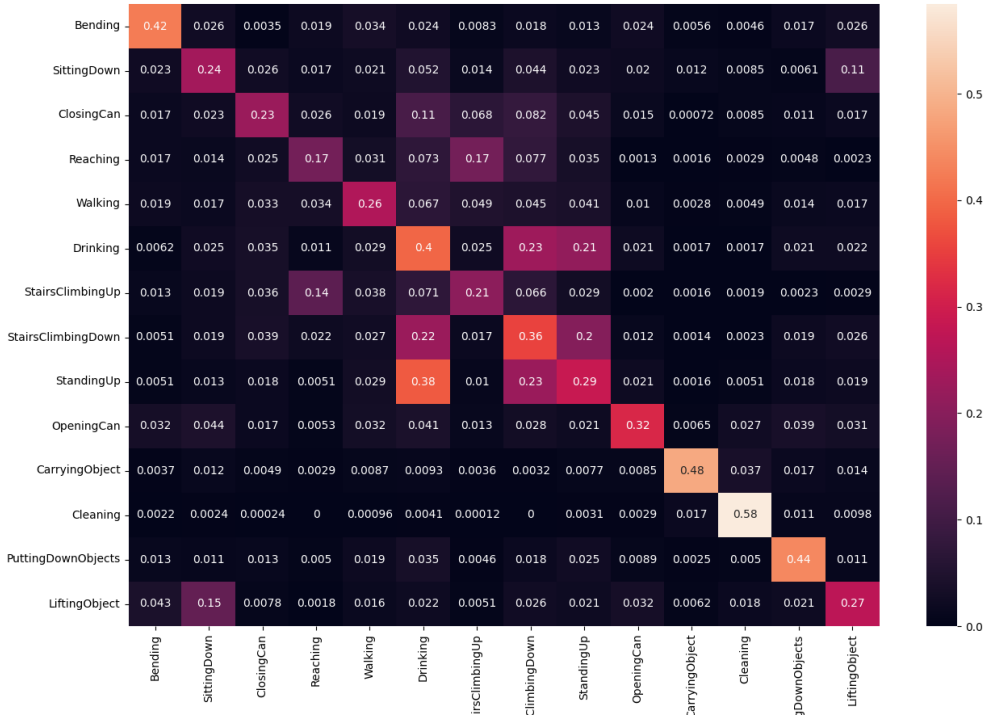
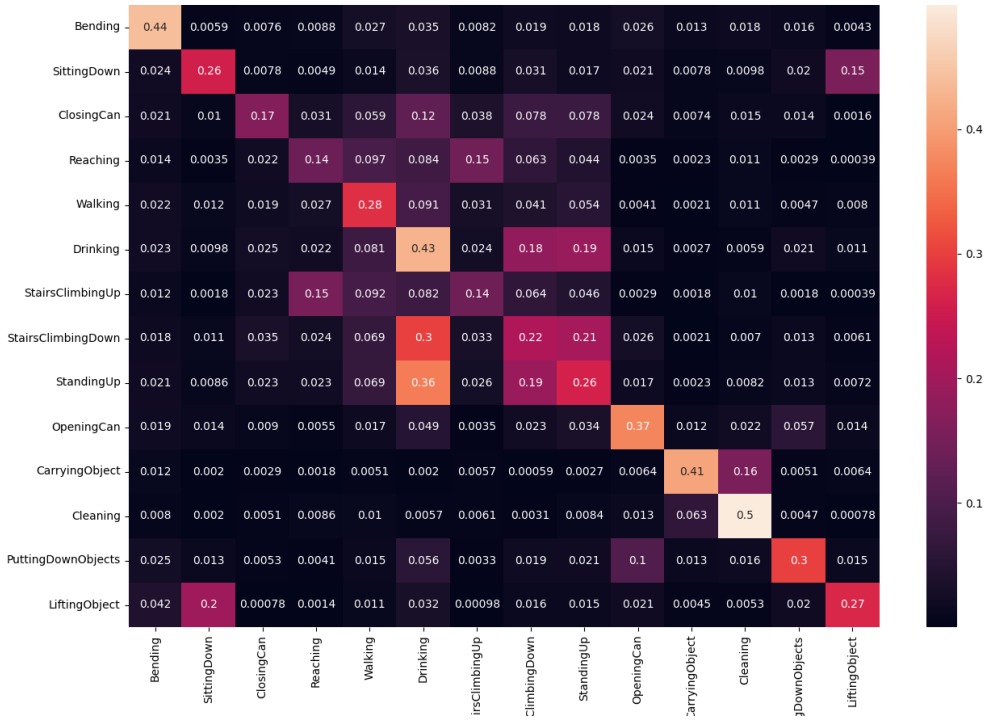


Figure 8.1: Sequential Frames for Actions in the Front View of the RHM Dataset

Figure 8.1: Continue Sequential Frames for Actions in the Front View of the RHM Dataset

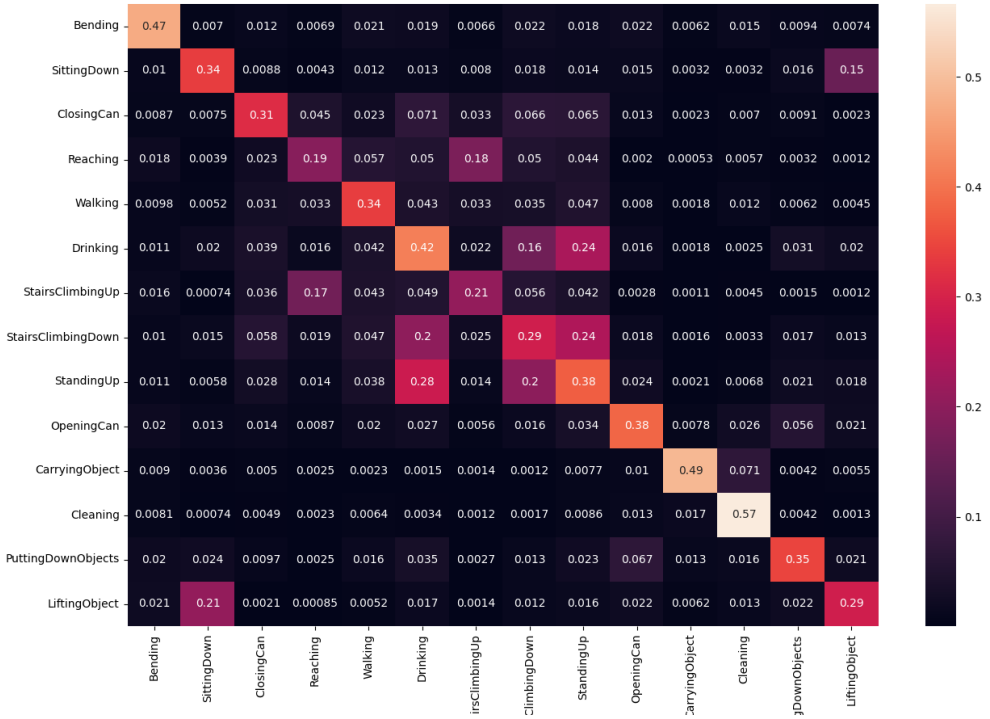


(a) RHM Confusion Matrix for Robot_Robot views with Dual-stream C3D Model

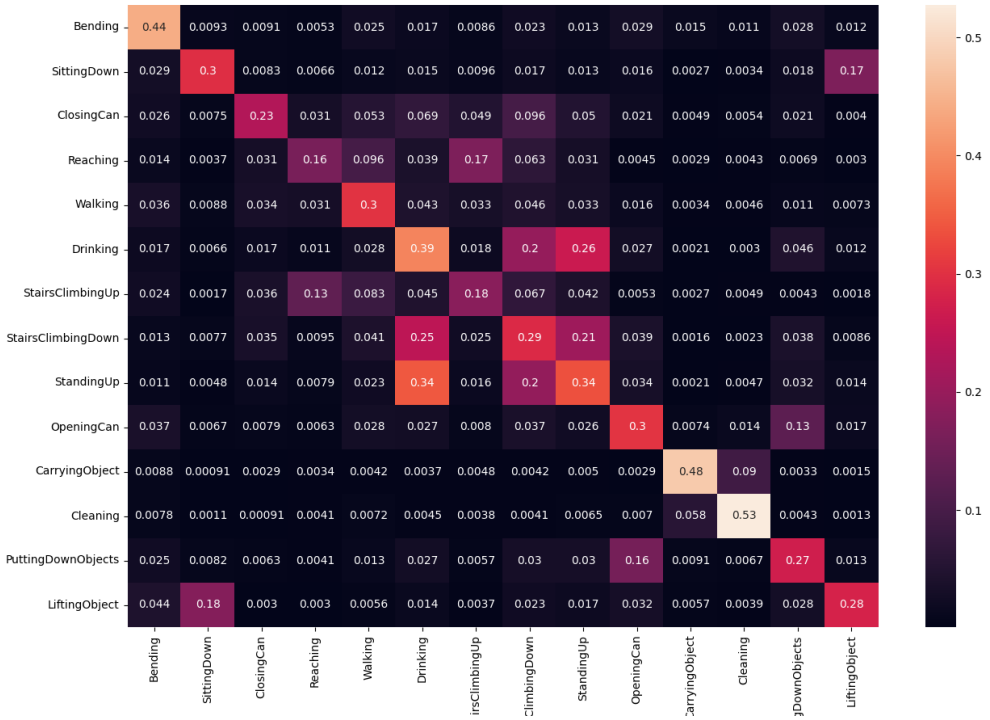


(b) RHM Confusion Matrix for Front_Front views with Dual-stream C3D Model

Figure 8.2: Confusion Matrix for same views - Chapter 5 - Section 5.4.2

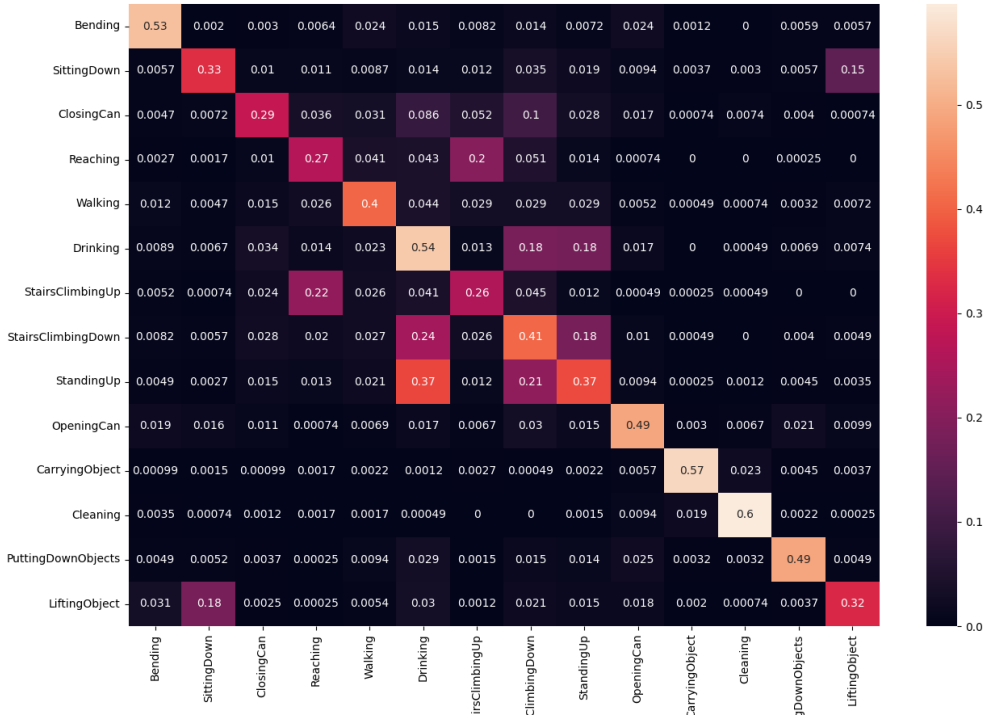


(c) RHM Confusion Matrix for Back_Back views with Dual-stream C3D Model

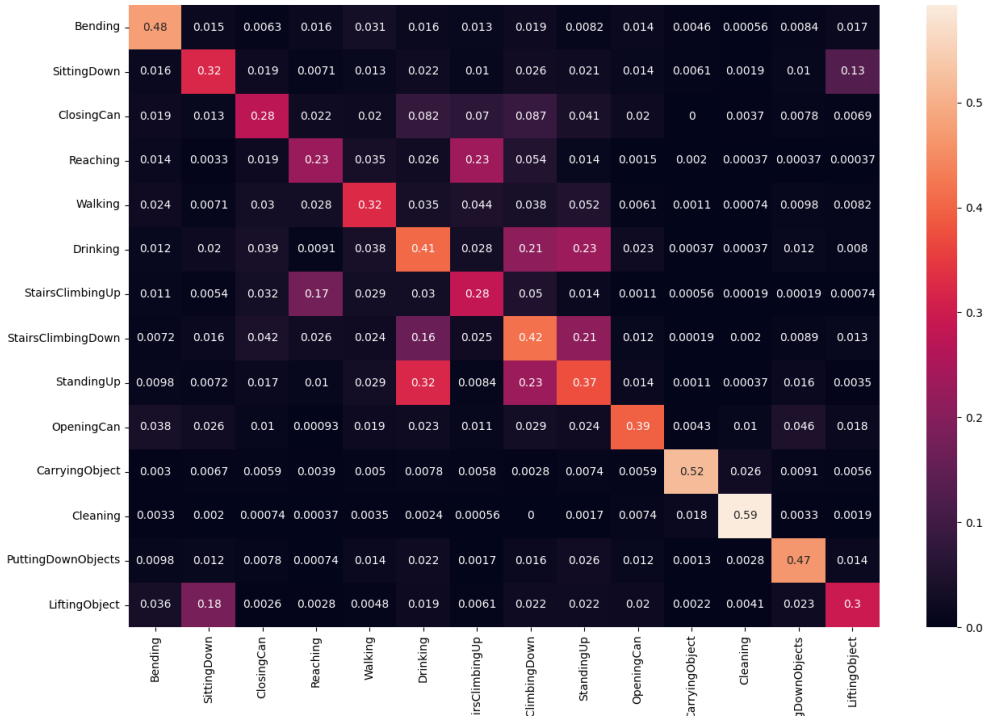


(d) RHM Confusion Matrix for Top_Top views with Dual-stream C3D Model

Figure 8.2: Continue Confusion Matrix for same views - Chapter 5 - Section 5.4.2

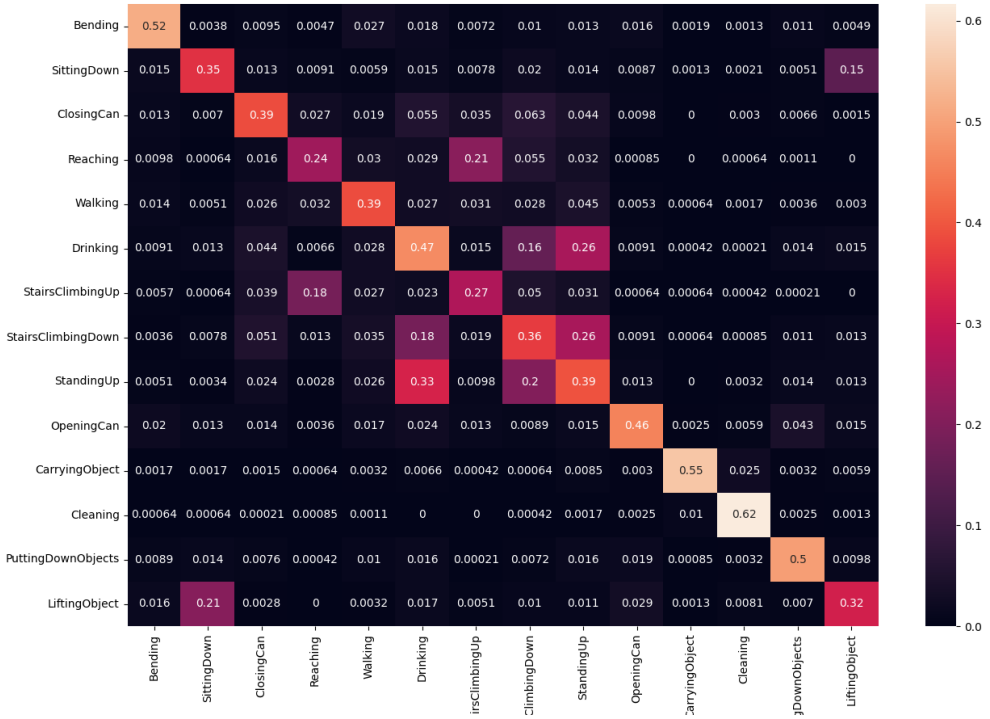


(a) RHM Confusion Matrix for Robot_Front views with Dual-stream C3D Model

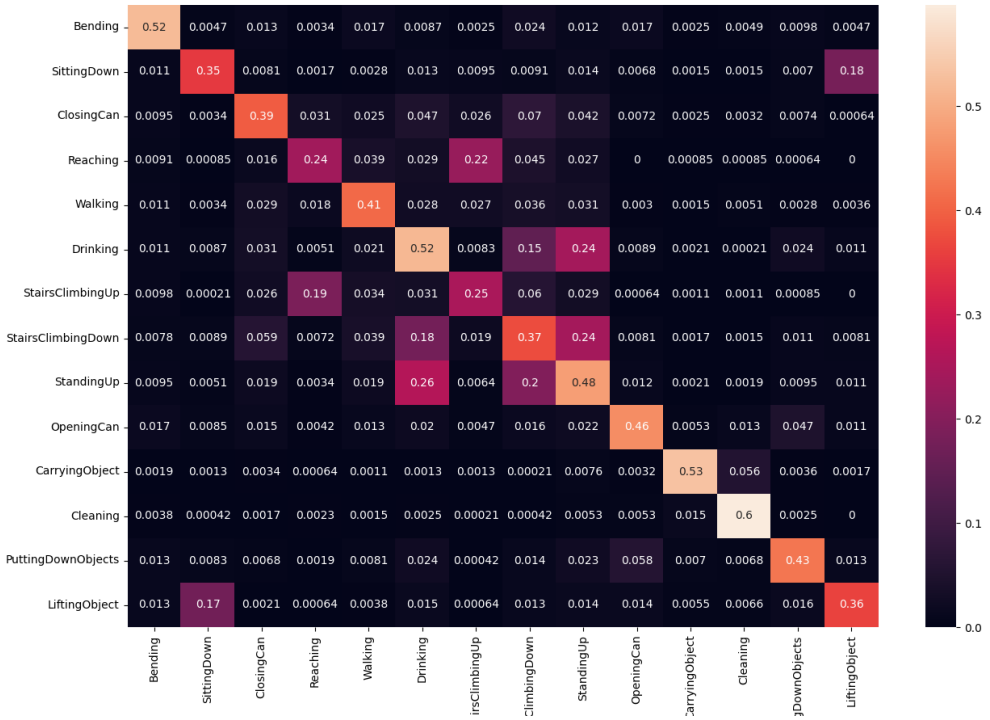


(b) RHM Confusion Matrix for Front_Robot views with Dual-stream C3D Model

Figure 8.3: Confusion Matrix for same views - Chapter 5 - Section 5.4.2

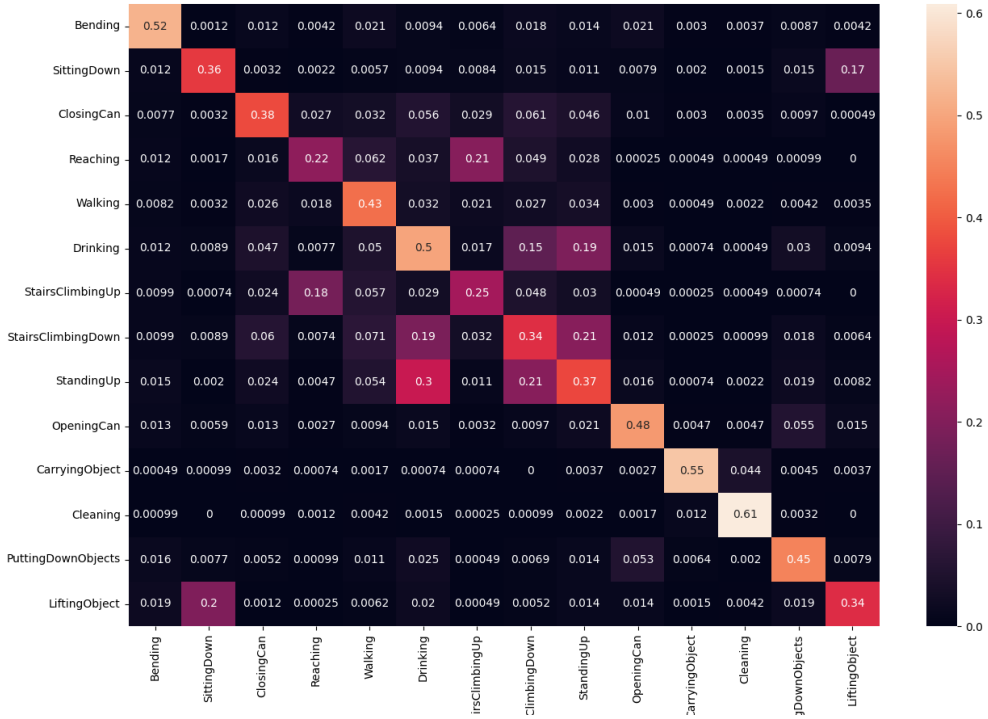


(c) RHM Confusion Matrix for Robot_Back views with Dual-stream C3D Model

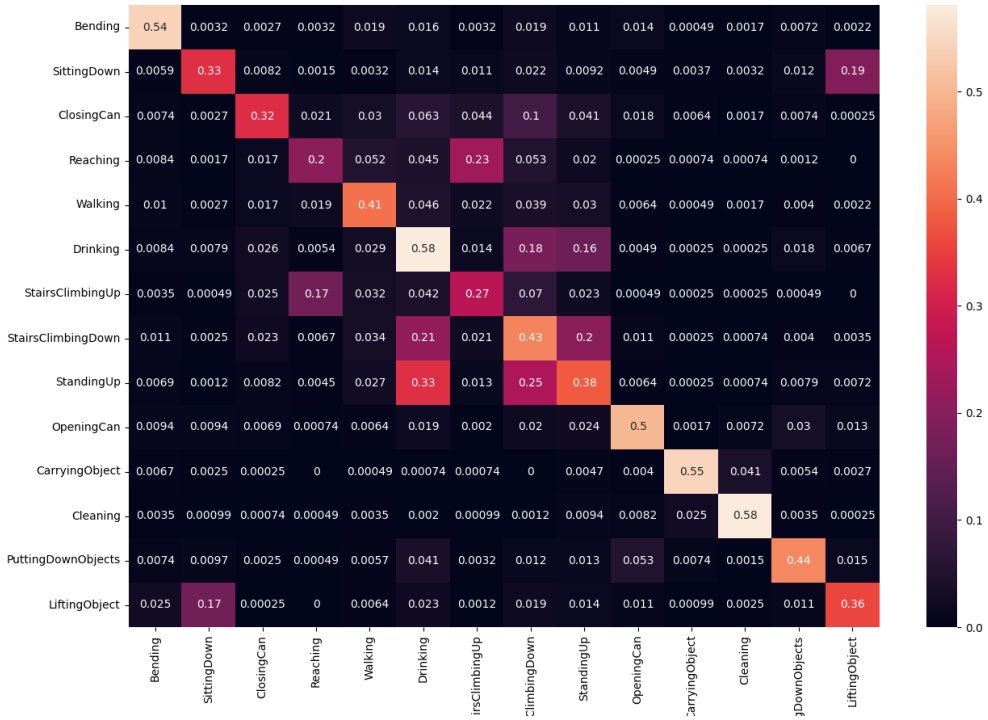


(d) RHM Confusion Matrix for Back_Robot views with Dual-stream C3D Model

Figure 8.3: Continue Confusion Matrix for same views - Chapter 5 - Section 5.4.2



(e) RHM Confusion Matrix for Robot_Top views with Dual-stream C3D Model



(f) RHM Confusion Matrix for Top_Robot views with Dual-stream C3D Model

Figure 8.3: Continue Confusion Matrix for same views - Chapter 5 - Section 5.4.2