A Deep Learning System for Daily Activity Recognition in Smart Home Environments

Augusto Luvisa Dessanti

Dissertation presented to the School of Technology and Management of Bragança to obtain the Master Degree in Electrotechnical and Computer Engineering.

Work oriented by:

Prof. Dr. José Luís Sousa de Magalhães Lima

Prof. Dr. Gustavo Kunzel

Bragança

2025/2026

A Deep Learning System for Daily Activity Recognition in Smart Home Environments

Augusto Luvisa Dessanti

Dissertation presented to the School of Technology and Management of Bragança to obtain the Master Degree in Electrotechnical and Computer Engineering.

Work oriented by:

Prof. Dr. José Luís Sousa de Magalhães Lima

Prof. Dr. Gustavo Kunzel

Bragança

2025/2026

Dedication

I dedicate this work to my family, my pillars and greatest source of strength. To my parents and sister, for always supporting me and believing in my potential. To my dear grandmother and my uncles and aunts, for their affection, encouragement, and for celebrating every small victory by my side. To my friends and colleagues, indispensable companions who accompanied me through every stage of this journey, from beginning to end, sharing challenges, knowledge, and the joy of achievements. Without your support and presence, reaching this point would not have been possible.

Acknowledgement

The completion of this work represents the culmination of a cycle and would not have been possible without the support and contributions of numerous individuals and institutions to whom I am deeply grateful.

First and foremost, my sincerest gratitude goes to my advisor, Prof. Dr. José Luís Sousa de Magalhães Lima, for his guidance, patience, availability, and the knowledge he shared, all of which were essential to the development of this research.

To my co-advisor, Prof. Dr. Gustavo Kunzel, I extend my thanks for his support, insightful advice, and valuable contributions, which greatly enriched this study.

In a very special way, my deepest gratitude goes to Rebeca Baron Kalbermatter, a PhD student under Prof. José Luís. Her constant support, crucial exchange of ideas, the companionship during challenging moments, and unwavering willingness to assist were undoubtedly the greatest support throughout this journey. This work also owes much to her collaboration and daily encouragement.

I extend my thanks to the Federal Institute of Rio Grande do Sul (IFRS) – Campus Farroupilha, where I began my academic journey in Control and Automation Engineering. The knowledge and opportunities gained there have been instrumental in paving the way for this master's research at IPB.

To the Polytechnic Institute of Bragança (IPB), I express my gratitude for welcoming me and offering the academic environment and resources necessary to

carry out this dissertation.

To all who, directly or indirectly, contributed to the realization of this work, my heartfelt thanks.

Abstract

This work presents the implementation of a system for daily activity classification using 3D Convolutional Neural Networks (3D CNN) and the Toyota Smarthome Dataset. The system aims to generalize and correctly classify activities, even in the face of data limitations such as high class imbalance, ambient occlusions, and similarities between classes. To overcome these challenges, preprocessing techniques and data augmentation were applied, including spatio-temporal resizing and image enhancement, with the objective of optimizing learning and generalization capabilities of the model. The proposed approach proved to be effective compared to other models on the same dataset, achieving 85.7% accuracy, 0.8568 precision, and 0.8570 recall.

Keywords: Activities of Daily Living; Convolutional Neural Networks; Video Classification; Toyota Smarthome Dataset; Data Augmentation.

Resumo

Este trabalho apresenta a implementação de um sistema para classificação de atividades diárias utilizando redes neurais convolucionais 3D (3D CNN) e o Toyota Smarthome Dataset. O sistema visa generalizar e classificar corretamente as atividades, mesmo diante de limitações nos dados, como alto desbalanceamento de classes, oclusões ambientais e similaridades entre as classes. Para superar esses desafios, foram aplicadas técnicas de pré-processamento e melhoria de dados (data augmentation), incluindo redimensionamento espaço-temporal e aprimoramento de imagens, com o objetivo de otimizar o aprendizado e a capacidade de generalização do modelo. A abordagem proposta mostrou-se eficaz em comparação com outros modelos no mesmo conjunto de dados, alcançando 85,7% de acurácia, 0,8568 de precisão e 0,8570 de sensibilidade.

Palavras-chave: Atividades de Vida Diária; Redes Neurais Convolucionais; Classificação de Vídeos; Toyota Smarthome Dataset; Data Augmentation;

Contents

Acknowledgement **vii**

Abstract **ix**

Resumo **xi**

Acronyms **xix**

1 Introduction **1**

1.1 Problem Description 2

1.2 Motivation 3

1.3 Objectives 4

1.4 Methodology Approach 5

1.5 Document Structure 6

2 State of art **9**

2.1 Toyota Smarthome Dataset 10

2.2 Neural network models 11

3 System methodology **19**

3.1 The Toyota Smart Home Dataset and Kinect Sensor Technology . . 19

3.2 System Based on 3D Convolutional Neural Networks 20

3.2.1 Main Processing Pipeline 21

4 Development **23**

4.1 Data Preprocessing 23

4.1.1 Video Resizing 23

4.1.2 Temporal Segmentation 24

4.1.3 Padding 24

4.1.4 Data Augmentation 25

4.2 Convolutional Network Architecture 26

4.3 Training 28

4.3.1 Dataset Splitting 28

4.3.2 Hyperparameter Configuration 29

4.3.3 Monitored Metrics 29

4.3.4 PyTorch Lightning Integration 30

4.4 Integration with Third-Party Code 31

5 Results **35**

5.1 Evaluation Context 35

5.1.1 Evaluation Metrics 36

5.2 Quantitative Results 37

5.2.1 Overall Model Performance 37

5.2.2 Comparison with the State of the Art 39

5.3 Analysis by Class 40

5.4 Qualitative Results 45

5.5 Limitations Identified 46

6 Conclusion and Future Work **49**

6.1 Conclusions 49

6.2 Future Work 51

List of Tables

2.1 Summary of Neural Network Models for Activity Recognition 14

4.1 Used libraries 31

5.1 Performance Metrics Across Training, Validation, and Test Sets . . 37

5.2 Performance comparison between the proposed model and the mentioned related works for 31 classes. 40

5.3 Performance Metrics by Class on the Test Set 40

List of Figures

5.1 Training and validation accuracy throughout the training process. . 38
5.2 Training and validation loss curves throughout the training process. 38
5.3 Absolute confusion matrix. It displays the exact number of test samples for each class. 42
5.4 Normalized confusion matrix. The diagonal represents the recall percentage for each class, while off-diagonal elements show the misclassification patterns between activities. 43
5.5 Examples of activities correctly classified [16] 45
5.6 Examples of activities incorrectly classified [16] 46

Acronyms

π-ViT Pose Induced Video Transformer.

3D CNN 3D Convolutional Neural Networks.

ADL Activities of Daily Living.

AI Artificial Intelligence.

Att-GNN Attention-based GNN.

Bi-LSTM Bidirectional Long Short-Term Memory.

CNN Convolutional Neural Networks.

F1-Score Harmonic Mean of Precision and Recall.

FN False Negative.

FNN Feedforward Neural Networks.

FP False Positive.

fps frames per second.

GNN Graph Neural Networks.

GPU Graphics Processing Unit.

GRU Gated Recurrent Units.

I3D Inflated 3D ConvNet.

IoT Internet of Things.

IoU Intersection over Union.

JP-GNN Joint-Prediction Network.

LSTM Long Short-Term Memory.

MLP Multilayer Perceptron.

ReLU rectified linear unit.

RGB Red, Green, Blue.

RNN Recurrent Neural Networks.

S-GNN GNN with Split Prediction.

SMOTE synthetic minority oversampling technique.

STA spatio-temporal attention.

TN True Negative.

TP True Positive.

VRAM Video Random Access Memory.

Chapter 1

Introduction

The global trend of increasing life expectancy and the consequent aging of populations requires innovative technological solutions to support the well-being and independence of the elderly [1]. This demographic shift imposes growing demands on healthcare systems and informal caregivers, highlighting the need for automated monitoring systems that can assist in care provision and ensure safety [2]. The increase in the elderly population worldwide presents a significant social challenge, creating a strong motivation for research into assistive technologies like automated activity recognition. This trend implies a growing market and need for solutions that can support independent living and reduce the burden on caregivers.

Automatic classification of activities from sensor data is essential for the development of intelligent systems capable of detecting abnormal events, such as falls or prolonged inactivity, without continuous human supervision, thus offering a cost-effective and less intrusive monitoring solution [1]. Accurate activity classification can enable the personalization of care plans and interventions based on an individual's activity patterns and specific needs, leading to more effective and targeted support for the elderly. The ability to classify activities automatically is crucial for the creation of truly intelligent assistive systems. This automation

reduces reliance on manual monitoring, freeing up caregivers for more direct interaction and personalized care, and allows for continuous, 24/7 monitoring, which may not be feasible with human supervision alone.

In this context, the recognition of Activities of Daily Living (ADL) emerges as a research area of significant importance, with potential applications in health monitoring, security, and personalized care in smart home environments [3]. The accurate and reliable identification of these activities can contribute to the early detection of health problems, accident prevention, and the improvement of the quality of life for the elderly.

1.1 Problem Description

Despite advancements in sensor technology and machine learning, developing reliable and practical human activity recognition systems presents significant hurdles. Current solutions face several key limitations that hinder their widespread adoption and effectiveness, particularly in complex real-world scenarios.

Traditional approaches to human activity classification face critical limitations when monitoring the elderly in home environments. These models show low adaptability to the natural variability in task execution, influenced by factors like individual habits, physical conditions, and environmental contexts. The need for frequent manual adjustments to accommodate new users or changes in the environment makes them impractical for dynamic real-world scenarios, where unpredictable conditions require autonomous continuous adaptation systems.

Manual extraction of meaningful features from complex sensor data constitutes one of the main methodological bottlenecks. Temporal, multivariate, and noisy data - as captured by the Kinect - present additional challenges due to its multimodal nature (sequences of *frames*, multiple sensory channels) and the presence

of artifacts (occlusions, interference). This complexity makes manual *feature* engineering particularly challenging, requiring specialized expertise in activity recognition and signal processing. Consequently, relevant discriminative patterns are often overlooked in conventional analysis, limiting the effectiveness of traditional image processing techniques [4].

Many existing activity recognition models lack robustness and may not generalize well across different home environments due to variations in sensor placement, environmental conditions (such as lighting and furniture arrangement), and individual activity patterns, leading to a drop in performance when deployed in new and unseen environments [5]. Personalized models, trained on individual user data, have shown promising results in addressing this limitation, capturing user-specific activity characteristics and adapting to their unique environment, often outperforming universal models trained on a general population[6]. The variability across different home environments and individual behaviors highlights the need for models that can effectively generalize or be personalized for specific users and their surroundings. This suggests that a universal model trained on a diverse dataset can be a starting point, but fine-tuning or adapting the model to individual users may be necessary for optimal performance in real-world applications.

1.2 Motivation

This work stems from multiple interconnected factors emerging from the contemporary social and technological landscape. Firstly, the increasing global population aging and the consequent demands for more efficient and less intrusive health and care systems drive the search for innovative technological solutions that can promote the well-being and independence of the elderly.

In this context, automated monitoring systems in ambient assisted living emerge

as a promising approach. However, the effective implementation of these systems critically depends on the ability to autonomously classify the activities performed by individuals. Significant challenges reside here, as the natural variability of human behavior, the complexities of domestic environments, and the need for robustness in uncontrolled scenarios impose barriers to the creation of truly reliable and generalizable classification models.

Despite these challenges, the application opportunities are vast and impactful. Autonomous activity classification systems, based on advanced Artificial Intelligence (AI) models, have the potential to transform elderly care, allowing for continuous and discreet monitoring, increased safety and accident prevention, support for independent living, personalized care and interventions based on real data from daily activities.

Therefore, the central motivation of this work lies in the need to advance the state of the art in autonomous daily living activity recognition in realistic environments, overcoming existing technical challenges. By validating the potential of 3D CNN models (like the Inflated 3D ConvNet (I3D)) applied to video data from real domestic environments, this work seeks to contribute to the realization of the promising opportunities that AI-based activity classification offers for the care and support of the elderly.

1.3 Objectives

The main objective of this work is to validate the potential of Convolutional Neural Networks (CNN) to correctly classify activities of daily living in realistic home environment scenarios using video data. This approach leverages the ability of 3D CNN to capture both spatial and temporal correlations, which are critical for activity classification.

The specific objectives of this work are as follows:

- Implementation of the I3D model with modifications to accommodate the dataset used in the study.
- Application of data augmentation methods to enhance the model's generalization capabilities.
- Evaluate the performance of the implemented model in classifying the defined activities of daily living on the target dataset.
- Analyze the effectiveness of the proposed approach, including the applied modifications and data augmentation, for ADL recognition in realistic home environments.

1.4 Methodology Approach

The methodology in this work is based on the current objectives and implementation pipeline. It is structured as follows:

- **State of the art review**: Research for the recent studies of activity classification and its models implementations, study their data preparation, their models uses and its performance.
- **Dataset analysis**: Study the Smarthome dataset and its difert types of data.
- **Data pre-processing**: Pre-processing of the data for implementation on the model.
- **Data augmentation**: Applying Image Augmentations technics looking for a better generalization model.

- **Model development**: Develop the model and adjust it to the specifics parameters used on the data.

- **Train and evaluation**: Train the model and evaluate it with unseen data to check its capability and performance to generalize and classify the activities.

- **State of the art comparision**: Compare the model performance to the state of the art models that uses the same dataset.

1.5 Document Structure

This work is structured into chapters, organized as follows:

- **Chapter 1**: Introduction Presents the problem, motivation, objectives, and methodology of the work.

- **Chapter 2**: State of the Art Describes the research and analysis of the literature on artificial intelligence classification systems applied to activity monitoring and analysis. Includes information about the Toyota Smarthome dataset and Neural Network models for Activity Classification.

- **Chapter 3**: System Methodology Presents the complete methodology employed in the development of the daily activity classification system, based on 3D CNN.

- **Chapter 4**: Development Details the data preprocessing steps, the neural network architecture used, and the training process.

- **Chapter 5**: Results Presents and analyzes the quantitative and qualitative results obtained from the model evaluation.

- **Chapter 6**: Conclusion and Future Work Synthesizes the work performed, presents the main conclusions, contributions, and limitations, as well as suggesting directions for future work.

Chapter 2

State of art

The state of the art in this dissertation covers the study and analysis of literature on artificial intelligence classification systems. These are applied in the monitoring and analysis of activities, with the aim of detecting changes and deviations from expected behaviors. This analysis is crucial for understanding how emerging technologies can be used to identify alterations in daily activities and potentially predict and mitigate problems before they occur.

Advances in computer vision and increasing computational power have made video data a rich source of information for understanding human actions and their interactions with the environment [7]. In parallel, the evolution of deep learning [8], particularly CNN and Recurrent Neural Networks (RNN), has revolutionized the field of video-based activity recognition [9]. These deep learning models demonstrate the ability to automatically learn complex features from video data, resulting in substantial improvements in accuracy compared to traditional methods based on manual feature extraction (such as decision trees, support vector machines, and spatio-temporal local representation methods) [10], [11] or approaches based on predefined rules [12].

However, a critical limitation in early benchmarks was the reliance on young

subjects simulating elderly activities [13], which introduces a significant domain shift due to differences in kinematics and mobility patterns. To address this, the ETRI-Activity3D dataset [14] established a new standard by capturing synchronized RGB, depth, and skeleton data explicitly from elderly participants in a controlled robot-view setting, allowing for intergenerational motion analysis. Furthermore, to mitigate the scarcity of large-scale annotated data and the ethical risks of capturing dangerous events like falls, the field has seen the emergence of synthetic datasets. Platforms such as ElderSim [15] have enabled the creation of massive datasets like KIST SynADL [15], facilitating the pre-training of models on diverse environmental conditions and rare activity classes through Sim-to-Real transfer. Despite these methodological advances in controlled and synthetic environments, the validation of assistive systems requires exposure to the unstructured and unpredictable nature of genuine domestic life, which remains the primary bottleneck for deployment.

In this landscape, the Toyota Smarthome dataset [16] stands out as a valuable resource for research in ADL recognition in the elderly. It is a real-world dataset, specifically designed for this purpose, and includes RGB, depth, and 3D skeleton video streams [16]. The relevance of this dataset lies in its unscripted nature and the involvement of real elderly participants, which differentiates it from other datasets that may be recorded with actors [16]. This authenticity makes Toyota Smarthome an important benchmark for the development and evaluation of activity recognition models aimed at practical applications in elder care.

2.1 Toyota Smarthome Dataset

The Toyota Smarthome dataset was collected in an apartment equipped with seven Kinect v1 cameras. It records common daily living activities performed by 18

individuals, aged between 60 and 80 years. The videos have a resolution of 640x480 and offer three main data types: RGB, depth, and 3D skeleton, the latter being extracted from RGB information. A notable characteristic of the dataset is its realistic nature, as participants were not given specific instructions on how to perform the activities, resulting in spontaneous and diverse actions.

The dataset is available in two main versions: Toyota Smarthome Trimmed and Toyota Smarthome Untrimmed. The Trimmed version is designed for the activity classification task and contains 31 distinct activity classes, totaling 16,115 short video clips in RGB+D format, each containing a single activity. On the other hand, the Untrimmed version is aimed at activity detection in long, continuous videos. It comprises 536 videos with an average duration of 21 minutes, in which 51 activities are densely annotated.

The Toyota Smarthome dataset presents several challenges inherent to the uncontrolled nature of real-world data collection. High intra-class variation, for example, results from the fact that the same activity can be performed in different ways by different individuals. Class imbalance, where some activities have a significantly larger number of samples than others, is also a challenge to be considered. Furthermore, some activities involve similar movements, making it difficult to distinguish between them. Variation in the duration of activities within the same class also adds complexity to the recognition task. Other challenges include the occurrence of occlusions and variation in the distance between the camera and the subject, due to the multiple camera perspectives used in the collection.

2.2 Neural network models

Early approaches using neural networks for activity classification on the Toyota Smarthome dataset demonstrated the potential of this technique. An initial study

proposed a framework that automatically encodes correlations between various detected interactions through collected relational data [3]. By using conventional feedforward neural networks, the framework achieved an accuracy of 49% in identifying ADL, allowing for the classification of different daily activities. This initial demonstration established a basis for more in-depth investigations with more complex neural network architectures.

RNN and, in particular, Long Short-Term Memory (LSTM) networks, have proven effective in modeling time series, making them suitable for analyzing video data for activity recognition. One study explored different variants of LSTM to classify five daily activities (drinking from a glass, eating at a table, reading a book, using a phone, and walking) on the Toyota Smarthome dataset. Experimental results indicated a promising accuracy of 94% using an LSTM architecture. Furthermore, comparison with other architectures, such as simple RNN and Feedforward Neural Networks (FNN), revealed that LSTM showed the best classification accuracy, demonstrating its ability to capture the temporal dependencies inherent in human activities [17]. Another study delved deeper into this investigation, comparing the performance of RNN, LSTM, Bidirectional Long Short-Term Memory (Bi-LSTM), and Gated Recurrent Units (GRU) on the same task of classifying five activities [17]. The conventional LSTM demonstrated the best performance, achieving accuracies of 93.75% with 2D pose data, 97.08% with 3D pose data, and 94.58% when combining 2D and 3D pose data. These results reinforce the effectiveness of LSTM for activity recognition on the Toyota Smarthome dataset and highlight the potential of using 3D pose information to increase accuracy.

Graph Neural Networks (GNN) have emerged as an interesting approach to modeling interactions between humans and objects, which are often crucial for ADL recognition. A GNN-based framework was proposed to analyze these interactions on the Toyota Smarthome dataset, aiming to classify eight daily activities

(cleaning, cooking, watching TV, eating, reading books, using the phone, using a laptop, and drinking water) [3]. The model constructed relational data from video frames, where objects were represented as nodes and interactions as edges, defined by temporal logic specifications based on Intersection over Union (IoU). The results demonstrated a TOP-1 accuracy of 88% for activity inference and 77% for object inference, surpassing baseline methods such as Multilayer Perceptron (MLP), GNN with Split Prediction (S-GNN), Attention-based GNN (Att-GNN), and Joint-Prediction Network (JP-GNN) [3]. This approach suggests that modeling the relationships between humans and objects can be fundamental for ADL recognition, and GNN offer a suitable architecture for this purpose.

More recently, Transformer-based architectures have gained prominence in the field of video activity recognition, including applications on the Toyota Smarthome dataset. Pose Induced Video Transformer (π-ViT) represents a state-of-the-art approach that enhances RGB representations learned by video transformers with 2D and 3D pose information [18]. This architecture achieved superior performance on the Toyota Smarthome dataset compared to other methods, including those that use both RGB and 3D pose data during inference. The ability of π-ViT to discard pose induction modules during inference is a valuable feature for real-time applications. These results indicate a growing trend in the use of Transformer architectures for video-based ADL recognition, with pose information playing a significant role in improving accuracy.

Hybrid approaches combining different neural network architectures and attention mechanisms have also been explored to address the challenges of the Toyota Smarthome dataset. A notable method proposed a separable spatio-temporal attention (STA) mechanism guided by pose, built upon 3D CNN, specifically I3D [19]. This model, capable of classifying the 31 activities of the Toyota Smarthome dataset, demonstrated superior performance to existing methods at the time of

its publication, achieving an average per-class accuracy of 54.2% (cross-subject), 35.2% (cross-view 1), and 50.3% (cross-view 2). The effectiveness of combining CNN for spatio-temporal feature extraction with attention mechanisms, especially when guided by pose information, was evidenced. A subsequent study improved this approach, focusing on enhancing action recognition through a separable spatio-temporal attention network with distinct branches for skeletal data (LSTM) and RGB data (I3D) [20]. By incorporating preprocessing techniques, such as view-invariant normalization of skeletal pose data and the use of full activity clips for RGB data, the model achieved even better results, with a mean per-class accuracy of 63.7% in cross-subject experiments. These advancements underscore the importance of data preprocessing techniques for optimizing the performance of deep learning models on the Toyota Smarthome dataset.

Table 2.1: Summary of Neural Network Models for Activity Recognition

Study	**Neural Network Model**	**Number of Classes**	**Performance Metric**	**Reported Value**	**Evaluation Protocol**
Relational data-based framework[3]	FNN	8	TOP-1 Accuracy	49%	Not specified
GNN based framework[3]	GNN	8	TOP-1 Accuracy	88%	Not specified
LSTM-based approach[17]	LSTM	5	Accuracy	94%	Not specified

Continued on next page

Table 2.1: Summary of Neural Network Models for Activity Recognition (Continued)

Study	**Neural Network Model**	**Classified Activities**	**Performance Metric**	**Reported Value**	**Evaluation Protocol**
Comparative study of RNNs[17]	LSTM	5	Accuracy (2D pose)	93.75%	Not specified
	LSTM	5	Accuracy (3D pose)	97.08%	Not specified
	LSTM	5	Accuracy (fused)	94.58%	Not specified
Spatio-temporal attention method[16]	I3D + pose-guided attention	31	Mean per-class acc.	54.2%	Cross-Subject
	I3D + pose-guided attention	19	Mean per-class acc.	35.2%	Cross-View 1
	I3D + pose-guided attention	19	Mean per-class acc.	50.3%	Cross-View 2
Spatio-Temporal attention network with preprocessing[20]	I3D + LSTM + attention	31	Mean per-class acc.	63.7%	Cross-Subject
	I3D + LSTM + attention	31	Overall accuracy	77.1%	Cross-Subject
	I3D + LSTM + attention	31	Mean per-class acc.	53.6%	Cross-View (joint)
	I3D + LSTM + attention	31	Overall accuracy	65.6%	Cross-View (joint)

Continued on next page

Table 2.1: Summary of Neural Network Models for Activity Recognition (Continued)

Study	Neural Network Model	Classified Activities	Performance Metric	Reported Value	Evaluation Protocol
π-ViT (Pose Induced Video Transformer)[18]	Video Transformer	31	Mean class accuracy.	72.9%	Cross-Subject
	Video Transformer	31	Mean class accuracy.	72.9%	Cross-Subject
	Video Transformer	19	Mean class accuracy.	55.2%	Cross-View 1
	Video Transformer	19	Mean class accuracy.	64.8%	Cross-View 2
	Pose induced Transformer	31	Mean class accuracy.	73.1%	Cross-Subject
	Pose induced Transformer	19	Mean class accuracy.	55.6%	Cross-View 1
	Pose induced Transformer	19	Mean class accuracy.	65.0%	Cross-View 2

The performance analysis reveals that different neural network architectures yield distinct results in the activity classification task on the Toyota Smarthome dataset. LSTM-based models demonstrated effectiveness in capturing temporal dependencies, achieving high accuracies in classifying a subset of activities [17]. GNN showed potential in modeling human-object interactions, which is relevant for recognizing various ADL [3]. Hybrid approaches combining CNN for spatio-temporal feature extraction with attention mechanisms, especially when guided

by pose information, showed significant performance improvements, being able to classify a larger number of activities with considerable accuracies [16], [20]. More recently, the introduction of Transformer-based architectures, such as π-ViT, represents an advancement in the state of the art, achieving superior results across different evaluation protocols [18].

The use of different data types also influences the performance of the models. Studies have shown that incorporating 3D pose data can lead to higher accuracy compared to using only RGB or 2D pose data. The combination of different modalities, such as RGB and skeleton, allows models to capture complementary information about visual appearance and body movement, resulting in more robust activity recognition.

Evaluation protocols, such as cross-subject and cross-view, also play a crucial role in assessing model performance. Results can vary significantly depending on the protocol used, reflecting the difficulty of models generalizing to new subjects or camera perspectives not seen during training.

This state-of-the-art review demonstrates the crucial role of the Toyota Smarthome dataset as a valuable resource for research in the recognition of ADL in elderly individuals using neural networks. The analysis of research papers reveals a progression from initial approaches with feedforward neural networks and RNN/LSTM to more sophisticated architectures such as GNN and Transformers, as well as hybrid methods employing attention mechanisms. The performance results achieved in different studies highlight the potential of neural networks for the task of activity classification on the dataset, although the challenges inherent to the real-world nature of the data, such as intra-class variation and class imbalance, continue to motivate research.

The comparison of the Toyota Smarthome with other activity recognition datasets underscores its unique characteristics, such as its focus on the elderly,

the recording of unscripted activities in a smart home environment, and the availability of multiple data types. These features make the Toyota Smarthome an important benchmark for the development of models aimed at practical applications in elderly care.

Chapter 3

System methodology

This chapter presents the complete methodology employed in the development of the daily activity classification system for elderly monitoring, using video data captured by the Kinect sensor.

The methodological approach is based on the application of deep learning techniques, specifically 3D CNN, to process visual data captured in a home environment. The addressed problem, the architecture of the proposed system, the implementation procedures, and the relevant technical considerations for its understanding and reproducibility are described in detail. The methodology is structured to provide a transparent and reproducible account of the research process, aligned with the study's objectives of enhancing elderly care through automated activity monitoring.

3.1 The Toyota Smart Home Dataset and Kinect Sensor Technology

For this work, the Toyota Smart Home dataset was selected. This dataset, captured using the Kinect sensor, provides a rich source of multimodal data (Red,

Green, Blue (RGB), depth, and skeleton), making it highly suitable for investigating activity recognition in realistic home environments [16]. The Kinect sensor's ability to capture diverse data types enables flexibility in exploring multiple approaches for activity classification, including the effectiveness of individual data streams (e.g., RGB-only) or their combinations [21].

The Toyota Smart Home dataset is particularly valuable for this research due to its multimodal nature and emphasis on activities relevant to elderly care in home settings. While this study focuses on RGB data, the availability of depth and skeleton data acknowledges the potential for future exploration of multimodal fusion, demonstrating a comprehensive understanding of the dataset's versatility.

3.2 System Based on 3D Convolutional Neural Networks

The proposed solution for the problem of automatic elderly activity classification consists of a system based on 3D CNN, specifically an adapted version of the I3D architecture to classify daily activities of the elderly using RGB video data from the Kinect sensor obtained from the ToyotaSmartHome dataset [16]. The system leverages the I3D architecture's ability to simultaneously learn spatial and temporal features from video sequences, making it suitable for analyzing human activities that naturally unfold over time and involve movement and interaction with the environment. The choice of an I3D network directly addresses the limitations of traditional methods by employing a deep learning approach capable of automatic feature extraction and modeling the spatial and temporal aspects of activities, which are crucial for accurate recognition in video data.

The I3D architecture, an extension of 2D convolutional networks to 3D, is specifically designed to capture both the spatial appearance of objects and people

within video frames and the temporal dynamics of actions by applying 3D convolutional filters across successive frames. Its ability to learn spatio-temporal features directly from video data eliminates the need for manual feature engineering, addressing a fundamental limitation of traditional rule-based methods and allowing the model to discover complex patterns that may be difficult for humans to define. Convolutional models pre-trained on large-scale video datasets, such as Kinetics [22], have demonstrated satisfactory performance on several action recognition tasks, suggesting their potential for this application. The proven success of the I3D architecture in action recognition and its capability to learn spatio-temporal features automatically make it a strong candidate for this research, offering a significant advantage over traditional rule-based methods.

While the Kinect sensor provides depth and skeleton data, this research focuses on utilizing the RGB data stream to train the I3D model as the primary input data type. RGB data offers rich visual information about the activities and environment, including details about the objects being used, interactions between the person and their surroundings, and the general context of the activity, which can be effectively processed by the I3D network to learn discriminative features. Using only RGB data simplifies input requirements and focuses the model on learning the visual appearance and motion patterns of activities, providing a clear baseline for performance before potentially incorporating more complex multimodal inputs. Starting with RGB data provides a focused approach and allows for a straightforward application of the I3D architecture.

3.2.1 Main Processing Pipeline

The main pipeline of the proposed system consists of three main steps.

The first step consists of pre-processing the RGB video data to a fixed resized image resolution, adjust the temporal window to a predetermined quantity of

frames and applying data augmented methods to increase model robustness and generalization, this ensures consistency and prepare it for the model's train

The second step is train the modified I3D model using the pre-processed data to learn activity patterns, and activity classification.

For the third step the trained I3D model then classifies the pre-processed video segments into predefined activity categories (e.g., cooking, walking, watchTV).

This clear outline of the processing pipeline provides a roadmap for the subsequent detailed sections, ensuring that the reader understands the overall information flow and the key steps involved in the activity recognition process.

Chapter 4

Development

This chapter covers the implementation of the system in detail, as mentioned in Chapter 3. It sequentially covers the data preprocessing, the architecture of the model used and its modifications, the training configurations, and the metrics used to evaluate its performance.

4.1 Data Preprocessing

Before feeding the video data into the 3D CNN, several preprocessing steps are applied to ensure the data is in a suitable format and to enhance the training process. These steps are crucial for standardizing the input and improving the model's ability to learn effectively.

4.1.1 Video Resizing

All video frames from the Toyota Smart Home dataset are resized to a spatial dimension of 298x224 pixels, preparing the data for a subsequent transformation to 224x224 via random cropping. This standardization ensures a consistent input size for the I3D network, which typically requires a fixed input resolution for

its convolutional and fully connected layers. Resizing to a common resolution is a standard preprocessing step in image and video analysis for deep learning, ensuring compatibility with the model architecture and potentially leveraging benefits from pre-trained weights if transfer learning is employed.

4.1.2 Temporal Segmentation

Each video is segmented into non-overlapping temporal windows of 128 frames. At a frame rate of 30 frames per second (fps), a 128-frame window corresponds to approximately 4.3 seconds of activity. This fixed-length segmentation allows the I3D network's 3D convolutional filters to effectively learn spatio-temporal patterns within a consistent temporal context, enabling the model to recognize activities based on both the sequence of movements and static poses within that period. The selection of a 128-frame window was primarily dictated by hardware constraints, as the available GPU memory limited the processing of longer sequences. Conversely, reducing the window size was avoided, as the approximately 4-second duration was deemed necessary to sufficiently capture the full temporal dynamics of the daily activities.

4.1.3 Padding

Videos with a total number of frames less than 128 are padded with zero frames at the end to ensure a uniform input length. Zero-padding is employed as the standard technique for this architecture, selected over alternatives such as frame repetition or video looping. While frame repetition was considered, it risks introducing artificial static patterns that the network could misinterpret as a stationary activity (e.g., freezing in a pose). Similarly, video looping was avoided as it alters the temporal structure of the event and introduces sharp discontinuities at the

loop boundary, potentially generating misleading motion cues. In contrast, zero-padding inserts values that produce zero activations during convolution, thereby effectively neutralizing the signal in the padded region without introducing spurious motion or static artifacts. This approach ensures that all input sequences meet the fixed temporal dimension required by the I3D network while preserving the integrity of the original activity patterns.

4.1.4 Data Augmentation

To enhance the robustness and generalization capability of the I3D model, two data augmentation techniques are applied to the training data: Random Crop and Random Horizontal Flip.

- **Random Crop**: During training, random 224x224 pixel regions are cropped from the resized video frames. This technique helps the model learn to recognize activities from different spatial perspectives and improves its invariance to minor variations in subject position or camera viewpoint within the scene, making the model more robust to variations in how activities are framed in the video. Random cropping simulates different viewpoints and subject placements, forcing the model to focus on the essential features of the activity rather than specific spatial contexts, thus improving its ability to generalize to new and unseen scenarios. This resolution is a common input size for many deep learning models.

- **Random Horizontal Flip**: With a 50% probability, video frames are randomly flipped horizontally. This augmentation technique increases the model's generalization capability by reducing its sensitivity to the left-right

orientation of activities, ensuring the model can recognize an activity regardless of whether it is performed on the left or right side of the frame. Horizontal flipping helps the model learn symmetric patterns in human movements and reduces its reliance on the specific orientation of the activity within the video frame, improving its ability to recognize activities performed from different perspectives.

The entire preprocessing pipeline was implemented using the Torchvision library, ensuring direct compatibility with the PyTorch framework and enabling efficient Graphics Processing Unit (GPU) operations when available. The implementation also includes optimizations for parallel data loading, utilizing multiple workers to minimize delays between training epochs.

4.2 Convolutional Network Architecture

This section details the architecture of the convolutional neural network employed for activity recognition. We describe the base architecture and the specific modifications implemented for this study.

The proposed system is based on the I3D architecture, which extends the principles of 2D convolutional networks (such as Inception [23]) to the spatio-temporal domain for video analysis by "inflating" 2D convolutional filters and pooling kernels to 3D, adding a temporal dimension. The I3D network uses 3D convolutional filters and pooling layers that operate on both the spatial and temporal dimensions of the video input, allowing them to capture motion patterns and temporal relationships between visual features across the frame sequence, which is crucial for distinguishing between activities with similar static poses but different movements. The architecture consists of multiple layers of 3D convolutions organized into inception modules, the rectified linear unit (ReLU) activation function, and

3D pooling operations to reduce dimensionality until reaching the classification layer for the predefined activity categories. The core advantage of the I3D architecture lies in its ability to process video as a 3D data sequence, enabling it to inherently learn both the spatial appearance of the person and the temporal evolution of their movements, making it highly suitable for activity recognition from video.

- **Input Layer Adaptation**: The input layer of the I3D model was adapted to accept tensors of dimension [3, 128, 224, 224], corresponding to the RGB channels, 128 frames, and 224x224 spatial dimensions of the pre-processed video segments from the Toyota Smart Home dataset. Adapting the input layer is a crucial step when using convolutional models on new datasets with different input characteristics, ensuring the data is correctly fed into the network.

- **Dropout Layers**: Dropout layers with a rate of 0.5 are included in the modified I3D architecture after certain convolutional layers or Inception modules. Dropout is a regularization technique that randomly zeros out a fraction (in this case, 50%) of the neuron outputs during training, preventing the network from becoming overly reliant on specific neurons and thus reducing overfitting on the training data. This helps prevent overfitting by reducing the model's reliance on specific neurons and encouraging it to learn more robust and generalizable features that can perform better on unseen data. A dropout rate of 0.5 is a commonly used value that generally provides a good balance between reducing overfitting and maintaining the model's learning capacity, although this value can be further tuned through experimentation on the validation dataset. The strategic placement of dropout layers within the I3D architecture helps improve the model's generalization performance

by reducing overfitting, which is particularly important when training complex deep learning models on relatively limited datasets.

4.3 Training

This section describes the methodology and configuration used to train the 3D CNN model. It covers the data splitting strategy, hyperparameter settings, monitored performance metrics, and the integration of PyTorch Lightning for efficient training management.

4.3.1 Dataset Splitting

The Toyota Smart Home dataset was divided into three subsets: a training dataset (90% of the data), a validation dataset (5%), and a test dataset (5%). Random stratified sampling was employed to ensure that the proportion of each activity class remains approximately consistent across all subsets.

Unlike standard Cross-Subject or Cross-View protocols, which isolate specific individuals or camera angles to test generalization to unseen domains, this random split strategy was deliberately chosen to maximize the model's exposure to the full diversity of the dataset. By including samples from all subjects and viewpoints in the training phase, the goal is to train the I3D model to handle the comprehensive variability of poses, lighting, and perspectives inherent in the smart home environment.

The training dataset allows the model to learn model parameters by minimizing the error between predictions and true labels. The validation dataset is used to tune hyperparameters and prevent overfitting. Finally, the test dataset provides a representative evaluation of the model's performance across the learned distribution of activities, subjects, and views. This approach ensures that the evaluation

reflects the model's capability to recognize activities within the diverse conditions present in the entire dataset.

4.3.2 Hyperparameter Configuration

The I3D model was trained with a physical batch size of 5 due to memory limitations of the NVIDIA Titan V GPU with 12GB of Video Random Access Memory (VRAM), which restricts the number of video segments processed simultaneously. To overcome this constraint and ensure stable convergence, a gradient accumulation strategy was employed with an accumulation step of 6. This technique aggregates gradients over 6 consecutive forward and backward passes before performing a single weight update.

Consequently, the model operates with an effective batch size of 30 (5×6). This approach combines the memory efficiency required by the hardware with the statistical stability of a larger batch size, ensuring accurate gradient estimation.

The Adam optimizer is used to train the model with a learning rate of 0.001. The Adam optimizer is known for its efficiency and ability to adapt the learning rate for each parameter. Combined with the effective batch size of 30, this setup facilitates smooth convergence. A learning rate of 0.001 acts as a standard starting point, providing a balance between speed and stability. The selection of these hyperparameters, particularly the use of gradient accumulation, reflects an optimized strategy to train the I3D network effectively despite physical hardware limitations.

4.3.3 Monitored Metrics

The performance of the I3D model is monitored and evaluated using two key metrics: Accuracy and Loss Function Value.

- **Accuracy**: Measures the overall percentage of correctly classified activity instances relative to the total number of instances in the test dataset. This metric offers a global view of the model's performance across all classes, indicating the proportion of correct predictions regardless of the activity type.

- **Loss Function Value**: Reflects the average error of the model's predictions during training and validation, calculated by a loss function (e.g., Cross-Entropy Loss). Lower values indicate that the model is generating predictions close to the true values, while high values suggest significant discrepancies. Monitoring this metric is essential for identifying issues like overfitting (when training loss is much lower than validation loss) or underfitting (when both values remain high), allowing for adjustments to the model's architecture or hyperparameters.

These complementary metrics offer a dual view: while accuracy shows practical classification performance, the loss value reveals the internal quality of the model's optimization during learning.

These metrics are calculated using the PyTorch Lightning and TorchMetrics libraries, ensuring efficient and standardized evaluation during the training, validation, and testing phases. The use of these metrics provides a comprehensive evaluation of the model's performance, covering both its predictive capability and its generalization ability.

4.3.4 PyTorch Lightning Integration

PyTorch Lightning is integrated into the training pipeline to automate many of the repetitive aspects of deep learning model training, such as managing the training loop, handling data loading, and providing support for GPUs. This framework

promotes reproducible research by enforcing a structured approach to organizing training and evaluation code, making it easier for other researchers to replicate experiments and build upon the results. PyTorch Lightning also facilitates the potential future scaling of the training process to multiple GPUs should more computational resources become available, which could significantly reduce training time for larger datasets or more complex models. The addition of PyTorch Lightning simplifies the development process, enhances reproducibility, and offers flexibility for future scalability, reflecting best practices in deep learning research.

4.4 Integration with Third-Party Code

Table 4.1 presents the libraries and tools used in this research, their versions, and their main functionalities. This detailed table underscores the dependence of modern machine learning research on a rich ecosystem of open-source libraries, each contributing specific functionalities to the development pipeline, from model building and training to data manipulation and evaluation.

Table 4.1: Used libraries

Library	**Version**	**Main Functionality**
PyTorch	2.6.0+cu126	CNN construction and automatic differentiation
PyTorch Lightning	2.5.0.post0	Training automation and multi-GPU support
Torchvision	0.21.0+cu126	Image preprocessing and data augmentation
Torchmetrics	1.6.1	Computation of evaluation metrics
Scikit-learn	1.6.1	Stratified dataset splitting
Pandas	2.2.3	Tabular data manipulation
NumPy	2.2.2	Numerical tensor processing
Matplotlib	3.10.0	Results visualization

Continued on next page

Table 4.1: Used libraries (Continued)

Library	**Version**	**Main Functionality**
tqdm	4.67.1	Progress bar for monitoring

While building upon existing architectural concepts and standard preprocessing techniques, this work includes specific implementation choices and adaptations tailored to the problem and dataset at hand. These original aspects are detailed below.

The input layers of the pre-trained I3D model were adapted to ensure compatibility with the specific tensor dimensions of the pre-processed RGB video data from the Toyota Smart Home dataset, as managed within the PyTorch Lightning structure. This involved adjusting the expected input shape of the first convolutional layer to match the required format for processing the 128-frame RGB video segments, which may differ from the input shape the model was originally trained on. This adaptation highlights the need to tailor deep learning models to the specific characteristics of the new dataset being used, particularly input dimensions, to ensure proper data flow and effective learning.

The choice and implementation of a fixed temporal window size of 128 frames for segmenting the video data was a specific design decision made in this research, considering the balance between capturing sufficient temporal context to accurately recognize the activities of interest and maintaining computational efficiency for model training and inference. This strategy was tailored to the characteristics of the activities in the Toyota Smart Home dataset, aiming to capture the typical duration of these activities, and to the capabilities of the I3D model to process sequences of this length effectively. The temporal segmentation strategy reflects a fundamental design choice in the methodology, balancing the need for adequate temporal information with computational constraints and the nature of

the activities being studied.

This research made the deliberate decision to focus exclusively on the RGB data stream from the Kinect sensor and not utilize the depth or skeleton information provided by the dataset for this initial implementation. This decision was motivated by the initial focus on leveraging the I3D architecture, which is primarily designed for RGB video input and has shown strong performance on action recognition tasks using this modality, and to establish a clear performance baseline using a single data source. Future work can explore the potential benefits of incorporating the other sensor modalities for multimodal activity recognition, acknowledging that depth and skeleton data can provide complementary information that enhances the system's robustness and accuracy. The decision to exclude depth and skeleton data at this stage of the research represents a focused approach to simplify the problem and establish a baseline using RGB data with the I3D model. The explicit mention of future work with other modalities indicates an understanding of the potential for improvement through multimodal fusion.

Chapter 5

Results

This chapter presents and analyzes the results obtained from the evaluation of the adapted I3D model for the classification of ADL on the Toyota Smart Home dataset. The evaluation structure, the metrics used, the quantitative and qualitative performance of the model, as well as a critical analysis of its limitations and contributions, are detailed in the subsequent sections.

5.1 Evaluation Context

The experimental evaluation conducted in this work had the central objective of validating the effectiveness of the proposed ADL classification system, as outlined in Chapter 3. Specifically, the tests aimed to verify whether the I3D model, trained exclusively with RGB data, is capable of classifying the activities present in the Toyota Smart Home dataset with a level of accuracy and robustness that supports its potential application in residential environments for elderly monitoring. The intention, therefore, was to measure the model's ability to learn discriminative spatio-temporal representations from 128-frame temporal windows and generalize this learning to unseen data, constituting a fundamental step towards enabling

more intelligent and proactive assistance systems.

5.1.1 Evaluation Metrics

For a comprehensive quantitative evaluation of the model's performance, the following metrics were selected: accuracy, precision, recall, and Harmonic Mean of Precision and Recall (F1-Score). The choice of these metrics is justified by their ability to provide a detailed view of the classifier's behavior, especially in scenarios with multiple classes and potential data imbalance:

- **Precision**: Measures the proportion of instances classified as positive that are truly positive. It represents the frequency with which the model is correct when predicting a class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.1}$$

- **Recall**: Measures the proportion of actual positive instances that are correctly identified. It indicates the model's ability to detect occurrences of a class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.2}$$

- **F1-Score**: The harmonic mean of precision and recall. It provides a balance between the two metrics, being useful for scenarios with imbalanced classes.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

- **Accuracy**: Measures the overall proportion of correct classifications relative to the total number of instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{5.4}$$

5.2 Quantitative Results

This section presents the quantitative results obtained from the experimental evaluation of the proposed activity recognition system. The performance metrics measured during the training and testing phases are detailed here.

5.2.1 Overall Model Performance

The I3D model was trained and evaluated using the data split (90% training, 5% validation, 5% test) and the hyperparameters detailed in Methodology 3. Table 5.1 presents a summary of the performance metrics obtained in the training, validation, and test phases.

Figures 5.1 and 5.2 display the learning curves, showing the evolution of the loss function and accuracy over training epochs for both the training and validation sets.

Table 5.1: Performance Metrics Across Training, Validation, and Test Sets

Metric	**Training**	**Validation**	**Test**
Accuracy	0.9043	0.8645	0.8570
Precision	0.9051	0.8596	0.8568
Recall	0.9043	0.8645	0.8570
F1-Score	0.9014	0.8584	0.8495

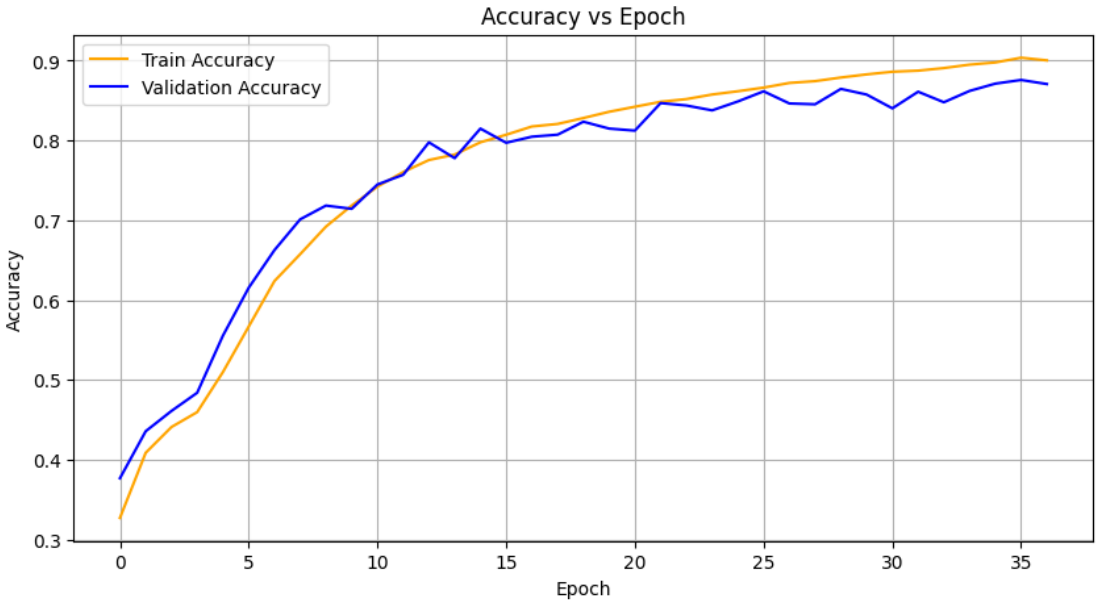


Figure 5.1: Training and validation accuracy throughout the training process.

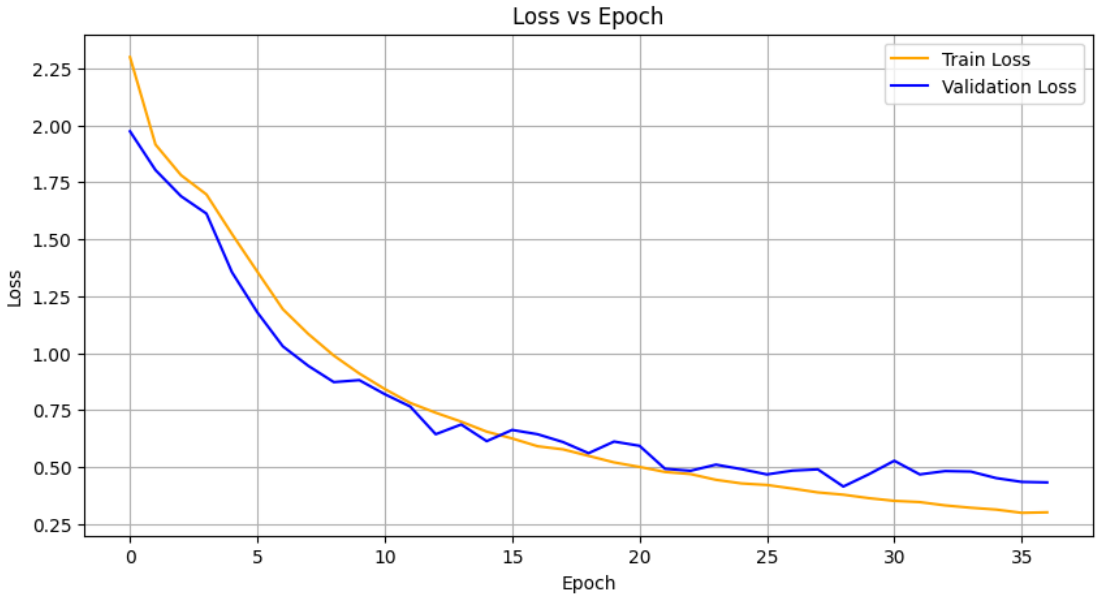


Figure 5.2: Training and validation loss curves throughout the training process.

5.2.2 Comparison with the State of the Art

To contextualize the performance of the proposed model, Table 2.1 presents the performance of other works that used the Toyota Smart Home dataset. It is important to note that direct comparisons are challenging due to variations in the number of classes considered, the metrics reported, evaluation protocols (e.g., cross-subject, cross-view), and the data modalities used (e.g., RGB, 2D/3D pose, multimodal data).

Analyzing Table 2.1, it is observed that the I3D model proposed in this work, using exclusively RGB data for 31 classes, achieved an overall accuracy of 85.70%. This result is competitive, especially when compared with approaches that also consider a high number of classes (31). As a comparision, the work by [20], which uses a more complex architecture (I3D + LSTM + attention), reported an overall accuracy of 77.1% under a Cross-Subject protocol for 31 classes. The π-ViT model [18], based on Transformers and pose information, achieved an average class accuracy of 73.1%.

It is crucial to highlight that many of the listed works use pose data (2D or 3D) or hybrid architectures that can capture complementary information. The current model, focusing only on RGB with a standard (modified) I3D architecture, demonstrates strong performance. Studies using LSTM [17] report higher accuracies (>90%), but for a significantly smaller number of classes (5 activities), which represents a less complex classification task. The absence of an explicitly defined "Cross-Subject" or "Cross-View" evaluation protocol for the model in this dissertation (beyond stratified splitting) in the table should be considered, as these more rigorous protocols tend to present additional generalization challenges. However, the performance obtained suggests that the I3D approach with the adopted preprocessing and training configurations is promising for the task of recognizing a diverse set of ADL.

Table 5.2: Performance comparison between the proposed model and the mentioned related works for 31 classes.

Architecture	**Evaluation Method**	**Accuracy**
I3D (Proposed Adapted Model)	Overall accuracy with random stratified splitting	85.70%
I3D + LSTM + Attention [20]	Overall accuracy with cross-subject protocol	77.1%
Transformers + Pose [18]	Class-Average accuracy with cross-subject protocol	73.1%

Table 5.2 presents the performance comparison between the proposed model and the other works analyzed in this section, highlighting the architectures used and the respective accuracies.

5.3 Analysis by Class

For a more granular understanding of the model's performance, its performance on each of the 31 activity classes was analyzed. Figures 5.3 and 5.4 present the confusion matrix and the normalized confusion matrix for the test set, allowing visualization of the classes where the model demonstrates higher assertiveness and those that are frequently confused. Table 5.3 details the precision, recall, and F1-Score by class.

Table 5.3: Performance Metrics by Class on the Test Set

Activity	**Precision**	**Recall**	**F1-Score**	**Samples**
Cook.CleanDishes	0.76	0.93	0.84	88
Cook.Cleanup	0.81	0.83	0.82	70
Cook.Cut	0.96	0.94	0.95	49
Cook.Stir	0.85	1.00	0.92	57
Cook.Usestove	0.86	0.75	0.80	8

Continued on next page

Table 5.3: Performance Metrics by Class on the Test Set (Continued)

Activity	**Precision**	**Recall**	**F1-Score**	**Samples**
CutBread	0.86	0.86	0.86	7
Drink.FromBottle	0.75	0.33	0.46	27
Drink.FromCan	1.00	0.33	0.50	24
Drink.FromCup	0.84	0.79	0.81	131
Drink.FromGlass	0.00	0.00	0.00	4
Eat.AtTable	0.86	0.89	0.88	36
Eat.Snack	0.67	0.15	0.25	13
Enter	0.74	0.76	0.75	34
GetUp	0.97	0.90	0.94	42
LayDown	1.00	0.73	0.84	11
Leave	0.52	0.88	0.65	16
MakeCoffee.PourGrains	0.86	1.00	0.92	6
MakeCoffee.PourWater	1.00	0.43	0.60	7
MakeTea.BoilWater	0.50	0.12	0.20	8
MakeTea.InsertTeabag	0.50	0.75	0.60	4
Pour.FromBottle	0.55	0.63	0.59	19
Pour.FromCan	0.00	0.00	0.00	4
Pour.FromKettle	0.36	0.40	0.38	10
ReadBook	0.93	0.97	0.95	497
SitDown	0.86	0.91	0.89	56
TakePills	0.61	0.63	0.62	35
UseLaptop	0.96	0.92	0.94	172
UseTablet	1.00	0.80	0.89	20
UseTelephone	0.80	0.62	0.70	97
Walk	0.83	0.91	0.86	254
WatchTV	0.84	0.87	0.86	166

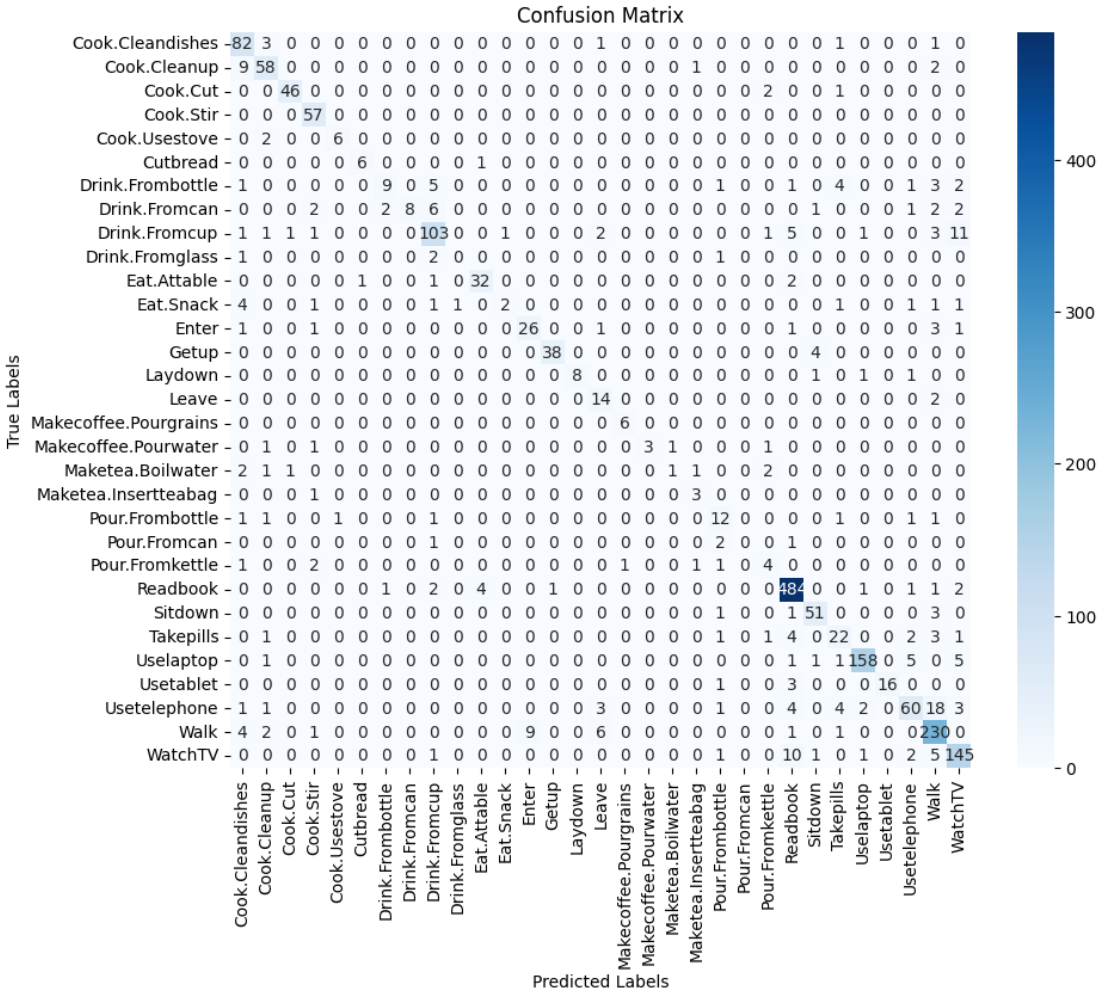


Figure 5.3: Absolute confusion matrix. It displays the exact number of test samples for each class.

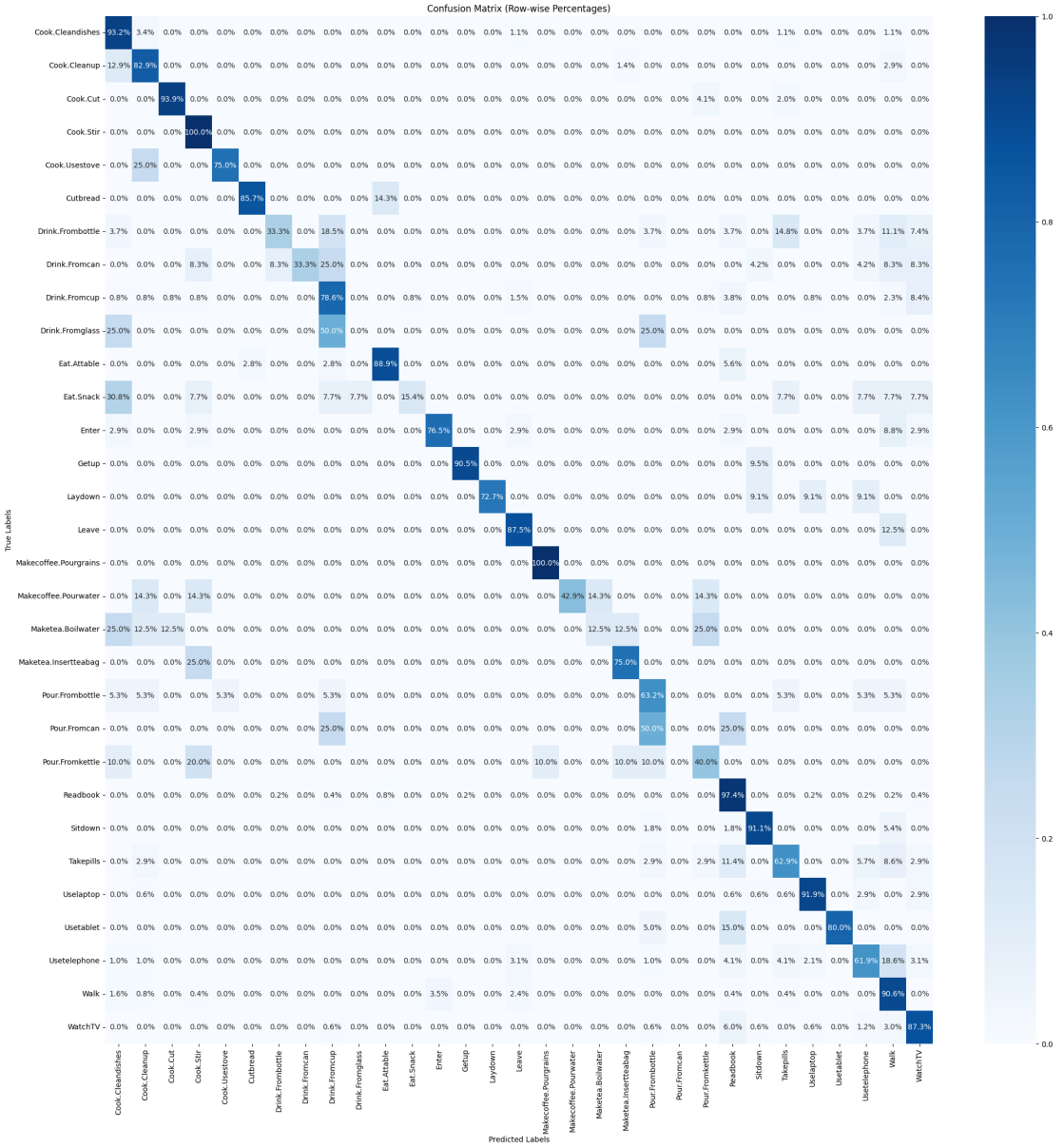
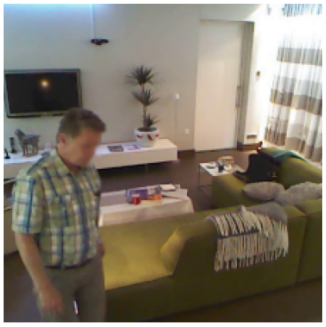


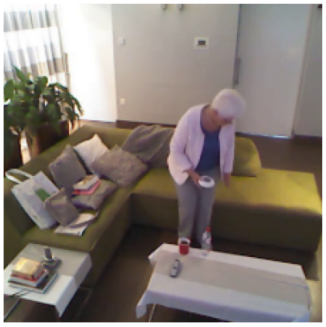
Figure 5.4: Normalized confusion matrix. The diagonal represents the recall percentage for each class, while off-diagonal elements show the misclassification patterns between activities.

The analysis of Table 5.3, the confusion matrices (Figures 5.3 and 5.4) reveals important peculiarities. Activities with a high number of samples and distinct visual characteristics, such as Readbook (F1-Score: 0.95), Uselaptop (F1-Score: 0.94), Walk (F1-Score: 0.86), and WatchTV (F1-Score: 0.86), were classified with high effectiveness. Similarly, more specific cooking activities like Cook.Cut (F1-Score: 0.95) and Cook.Stir (F1-Score: 0.92) also showed excellent performance.

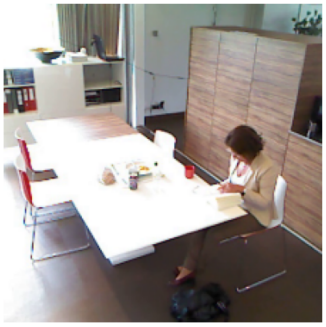
On the other hand, classes with few test samples and/or high visual similarity to other classes showed inferior performance. Drink.Fromglass and Pour.Fromcan stand out negatively, both with an F1-Score of 0.00, indicating that the model could not correctly identify them, possibly due to the very low number of samples (4 for both in the test set and 62 and 59 total samples respectively in the training dataset). Classes like Eat.Snack (F1-Score: 0.25) and Maketea.Boilwater (F1-Score: 0.20) also had low performance.



(a) Walk activity correctly classified.



(b) Sitdown activity correctly classified.



(c) Readbook activity correctly classified.

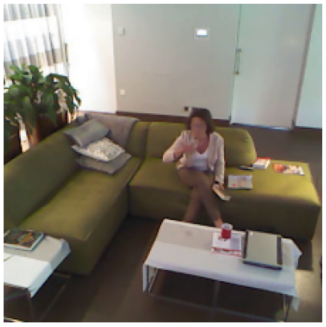
Figure 5.5: Examples of activities correctly classified (Walk, Sitdown, Readbook) [16].

5.4 Qualitative Results

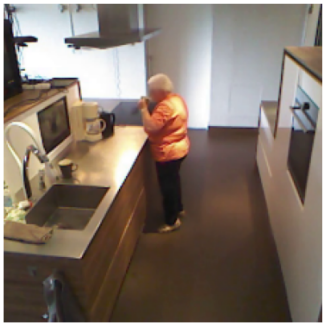
Qualitative analysis complements the quantitative results, offering insights into the model's behavior through the inspection of specific examples and the evolution of training.

Inspecting frames from videos (Figure 5.5 and Figure 5.6) that were classified correctly and incorrectly allows for a better understanding of the model's difficulties. It was observed that the model generally succeeds in classifying activities with large and contextually distinct body movements (e.g., "Walk", "Sitdown", "Readbook").

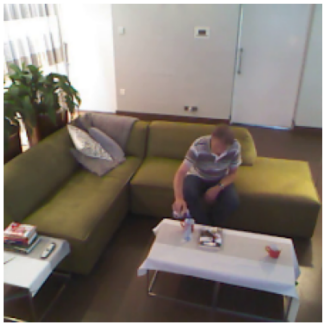
However, the most notable classification failures occurred in activities with high visual similarity and, frequently, a low number of samples. For instance, the "Drink" group activities (Drink.Fromglass, Drink.Frombottle, Drink.Fromcan) were consistently confused, with many instances being erroneously assigned to the Drink.Fromcup class. The latter has a considerably higher number of samples, which may have biased the model's learning. The percentage confusion matrix 5.4 shows, for example, that Drink.Frombottle was classified as Drink.Fromcup 18.5% of the time and as Walk 11.1% of the time, despite having a 33.3% accuracy



(a) Drink.Fromglass activity incorrectly classified as Drink.Fromcup.



(b) Drink.Fromglass activity incorrectly classified as Readbook.



(c) Pour.Fromcan activity incorrectly classified as Pour.Frombottle

Figure 5.6: Examples of activities incorrectly classified (Drink.Fromglass, Drink.Fromglass, Pour.Fromcan) [16].

within its own class. The Drink.Fromglass activity had 50% of its (few) instances classified as Drink.Fromcup and 25% as Pour.Frombottle, which can be attributed to the similarity in object usage and movement patterns between the activities.

Similarly, Pour.Fromcan, which is the class with the fewest training examples (59) and only 4 in the test set, was confused 50% of the time with Pour.Frombottle and 25% with Readbook (a less intuitive confusion, perhaps due to similar objects in the scene or even the environment itself). The difficulty lies in the intrinsic similarity of gestures and manipulated objects, a considerable challenge when using only RGB information and having few examples available. Even for a human observer, distinguishing these sub-categories of "Drink" or "Pour" can be complex without additional context or visualization of very specific details.

5.5 Limitations Identified

The analysis of the results and the development process allowed for the identification of some important limitations.

Model training was performed on a single NVIDIA Titan V graphics card with

12GB of VRAM. This hardware limitation restricted the physical batch size to 5 samples. However, the potential issue of noisy gradient estimates usually associated with small batches was effectively mitigated through gradient accumulation. By aggregating gradients to achieve an effective batch size of 30, the model benefited from stable convergence and accurate optimization steps. Nevertheless, the limited computational throughput of a single GPU remained a constraint regarding processing speed, resulting in a considerable training duration of approximately 83 hours.

The Toyota Smart Home dataset, while valuable, presents challenges. It was observed that several activities are visually similar (e.g., different forms of "Drink" or "Pour") and are often performed in the same room, making discrimination difficult based solely on RGB visual features. Furthermore, the dataset is significantly imbalanced, with some classes having a very small number of samples compared to others. This imbalance directly impacts the model's ability to learn robust representations for minority classes, as evidenced by the low performance in activities such as Drink.Fromglass and Pour.Fromcan.

Chapter 6

Conclusion and Future Work

This work addressed the relevant challenge of automatic ADL recognition through a deep learning approach based on 3D CNN and RGB data. The results obtained are encouraging and demonstrate the potential of computer vision to create tools that support active and safe aging.

6.1 Conclusions

The central objective was to develop and evaluate an automatic ADL recognition system, focusing on applying the I3D architecture to classify 31 distinct activities from the Toyota Smart Home dataset, using exclusively visual data (RGB) captured by the Kinect sensor. The adopted methodology involved preprocessing videos into 128-frame temporal segments, spatial resizing to 224x224 pixels, applying data augmentation techniques, and training the I3D model with the help of the PyTorch Lightning framework.

The quantitative results obtained demonstrated the feasibility of the proposed approach. The model achieved an overall accuracy of 85.70% on the test set, with weighted precision and recall values of 0.8568 and 0.8570, respectively. Detailed

analysis revealed that the system is particularly effective in classifying frequent activities with distinct spatio-temporal patterns, such as "Readbook", "Uselaptop", or "Walk". These results position the model's performance competitively relative to some state-of-the-art works addressing a similar number of classes, especially considering the exclusive use of RGB data.

However, the evaluation also exposed significant challenges. The model's performance was considerably lower for activities with few samples in the dataset or with high intrinsic visual similarity (e.g., different forms of "Drink" or "Pour"), highlighting the impact of class imbalance and the complexity of disambiguation based only on visual information.

It is concluded that 3D CNN models like I3D are powerful tools for learning spatio-temporal representations from RGB data in complex ADL classification tasks. However, their effectiveness is strongly influenced by the quality, quantity, and balance of the training data, as well as the intrinsic visual distinguishability of the activities. The study reinforces the idea that, while promising, the recognition of fine-grained and similar ADL remains an open challenge, especially with unimodal data.

A reflection on the process suggests that incorporating additional data, namely skeleton (pose) data, could have been a path to generating better positive results by improving the model. Pose information is inherently more robust to variations in lighting and appearance, and can provide crucial discriminative cues for activities with similar body movements. This finding represents fundamental acquired knowledge about the limitations of the adopted unimodal RGB approach and points to clear paths for optimization.

During development, it was necessary to go beyond the initial planning to address resource limitations. The unplanned implementation of mixed precision

training was a pragmatic solution that allowed mitigating GPU memory constraints, reducing training time, and enabling experimentation with the I3D architecture within the available hardware.

The main contribution of this dissertation lies in the empirical demonstration of the capability and limitations of a standard I3D model, trained only with RGB data, for classifying a large and diverse set (31 classes) of ADL from the Toyota Smart Home dataset. Achieving an overall accuracy of 85.70% validates the architecture and methodological choices (such as the 128-frame temporal window and data augmentation techniques) as a solid foundation for this complex task.

The work contributes to the body of knowledge in the field of computer vision applied to health and well-being, providing benchmarks and insights into the performance of 3D CNN models in this specific context. From a practical standpoint, the results suggest the potential for developing non-invasive monitoring systems based on common RGB cameras, which can assist in monitoring the elderly in a home environment, identifying routines and potential behavioral changes, particularly for the more common and well-classified activities by the model.

6.2 Future Work

The current study has inherent limitations, notably the exclusive reliance on RGB data, the dataset imbalance, and hardware constraints that limited the batch size during training. These limitations open up opportunities for future work.

The main aspects with potential for improved model performance are the integration of skeleton and/or depth data with RGB data for different pattern recognition, the evaluation of newer model architectures such as Video Transformer Networks, and the application of resampling techniques (e.g., synthetic minority

oversampling technique (SMOTE) for video) or loss function adjustments to mitigate the negative impact of minority classes on overall and per-class performance.

Opportunities for future work include implementing the model on a prototype system for testing in real home environments with different users and under varying conditions, such as ambient light and occlusions. The work could also extend to functionalities beyond ADL classification, incorporating modules capable of detecting anomalies and unexpected events like falls and deviations from established routines. Finally, integrating with Internet of Things (IoT) systems would enable continuous and real-time monitoring, opening opportunities for an alert generation system and report visualization for real-time performance evaluation.

Despite the identified limitations, the knowledge acquired and the directions pointed towards future work constitute a step forward in the pursuit of more intelligent, discreet, and effective monitoring systems capable of contributing to the quality of life of the elderly population.

Bibliography

[1] F. X. Gaya-Morey, C. Manresa-Yee, and J. M. Buades-Rubio, "Deep learning for computer vision based activity recognition and fall detection of the elderly: A systematic review," *Applied Intelligence*, vol. 54, no. 19, pp. 8982–9007, Jul. 2024, ISSN: 1573-7497. DOI: 10.1007/s10489-024-05645-1. [Online]. Available: http://dx.doi.org/10.1007/s10489-024-05645-1.

[2] Z. Liu, S. Zhang, H. Zhang, and X. Li, "A study on caregiver activity recognition for the elderly at home based on the xgboost model," *Mathematics*, vol. 12, no. 11, 2024, ISSN: 2227-7390. DOI: 10.3390/math12111700. [Online]. Available: https://www.mdpi.com/2227-7390/12/11/1700.

[3] P. Su and D. Chen, "Adopting graph neural networks to analyze human–object interactions for inferring activities of daily living," *Sensors*, vol. 24, no. 8, 2024, ISSN: 1424-8220. DOI: 10.3390/s24082567. [Online]. Available: https://www.mdpi.com/1424-8220/24/8/2567.

[4] A. Ravuri, "A systematic literature review on human activity recognition," *Journal of Electrical Systems*, vol. 20, pp. 1175–1191, Apr. 2024. DOI: 10.52783/jes.2848.

[5] D.-A. Nguyen and N.-A. Le-Khac, *Sok: Behind the accuracy of complex human activity recognition using deep learning*, 2024. arXiv: 2405.00712 [eess.SP]. [Online]. Available: https://arxiv.org/abs/2405.00712.

[6] J. W. Lockhart and G. M. Weiss, "Limitations with activity recognition methodology & data sets," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct, Seattle, Washington: Association for Computing Machinery, 2014, pp. 747–756, ISBN: 9781450330473. DOI: 10.1145/2638728.2641306. [Online]. Available: https://doi.org/10.1145/2638728.2641306.

[7] D. Bouchabou, S. M. Nguyen, C. Lohr, B. Leduc, and I. Kanellos, "A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning," *CoRR*, vol. abs/2111.04418, 2021. arXiv: 2111.04418. [Online]. Available: https://arxiv.org/abs/2111.04418.

[8] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, *Video-based human action recognition using deep learning: A review*, 2022. arXiv: 2208.03775 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2208.03775.

[9] K. Alomar, H. I. Aysel, and X. Cai, *Rnns, cnns and transformers in human action recognition: A survey and a hybrid model*, 2024. arXiv: 2407.06162 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2407.06162.

[10] D. J. H. Kim and I. S. Nakamura, "Reconhecimento de atividades humanas por aprendizado de máquina," M.S. thesis, Departamento de Engenharia Mecatrônica e de Sistemas Mecânicos, Escola Politécnica da Universidade de São Paulo (PMR), São Paulo, Brasil, 2020.

[11] A. B. Sargano, X. Gu, P. Angelov, and Z. Habib, "Human action recognition using deep rule-based classifier," *Multimedia Tools Appl.*, vol. 79, no. 41–42, pp. 30 653–30 667, Nov. 2020, ISSN: 1380-7501. DOI: 10.1007/s11042-020-

09381-9. [Online]. Available: https://doi.org/10.1007/s11042-020-09381-9.

[12] A. Sarabu and A. Santra, "Human action recognition in videos using convolution long short-term memory network with spatio-temporal networks," *Emerging Science Journal*, vol. 5, pp. 25–33, Feb. 2021. DOI: 10.28991/esj-2021-01254.

[13] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, Dec. 2014, ISSN: 0169-2607.

[14] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "Etri-activity3d: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly," *CoRR*, vol. abs/2003.01920, 2020. arXiv: 2003.01920. [Online]. Available: https://arxiv.org/abs/2003.01920.

[15] H. Hwang, C. Jang, G. Park, J. Cho, and I. Kim, "Eldersim: A synthetic data generation platform for human action recognition in eldercare applications," *CoRR*, vol. abs/2010.14742, 2020. arXiv: 2010.14742. [Online]. Available: https://arxiv.org/abs/2010.14742.

[16] S. Das, R. Dai, M. Koperski, *et al.*, "Toyota smarthome: Real-world activities of daily living," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019.

[17] L. C. Hong, C. Tee, and M. K. O. Goh, "Activities of daily living recognition using deep learning approaches," *Journal of Logistics, Informatics and Service Science*, vol. 9, no. 4, pp. 129–148, 2022, ISSN: 2409-2665.

[18] D. Reilly and S. Das, *Just add π! pose induced video transformers for understanding activities of daily living*, 2023. arXiv: 2311.18840 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2311.18840.

[19] J. Carreira and A. Zisserman, *Quo vadis, action recognition? a new model and the kinetics dataset*, 2018. arXiv: 1705.07750 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1705.07750.

[20] P. Climent i Pérez and F. Flórez-Revuelta, "Improved action recognition with separable spatio-temporal attention using alternative skeletal and video preprocessing," *Sensors*, vol. 21, p. 1005, Feb. 2021. DOI: 10.3390/s21031005.

[21] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and rgb data," *Pattern Recognition Letters*, vol. 131, Mar. 2020. DOI: 10.1016/j.patrec.2020.01.010.

[22] W. Kay, J. Carreira, K. Simonyan, *et al.*, *The kinetics human action video dataset*, 2017. arXiv: 1705.06950 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1705.06950.

[23] C. Szegedy, W. Liu, Y. Jia, *et al.*, *Going deeper with convolutions*, 2014. arXiv: 1409.4842 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1409.4842.