Recognition of Quotidian Activities in Support of Independent Living Using a Single Wrist-worn Inertial Measurement Unit

DARIO ORTEGA ANDEREZ

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of

Doctor of Philosophy

July, 2020

This thesis is dedicated to my father and mother. Dad, I know you would be very proud of this achievement. This thesis would have not been possible without everything you have given and taught me throughout my entire life.

Acknowledgements

The work described in this thesis was carried out at Nottingham Trent University between October 2015 and July 2019, while I was working as a full-time doctoral research student.

I owe my utmost sincere gratitude to my director of studies, Prof. Ahmad Lotfi, for his guidance throughout the last four years. Your consistent support, motivation and kindness will never be forgotten. Many thanks for making your office accessible at all times despite your always busy calendar. Also, I would like to thank the team of co-supervisors: Dr. Caroline Langensiepen, I am grateful for the creativity and good spirit you were bringing to the table every Friday afternoon. Thank you Dr. Kofi Appiah for the supervision given at the early stage of my research. Lastly, to Dr. Amir Pourabdollah, many thanks for the corrections you made to the publications we share.

My sincere gratitude to my mother and father for their unconditional love, care and support throughout my entire life. Dad, I know how happy you would be. I wish you could be here to witness this achievement and celebrate it together. Mum, thank you for always being there when I need you. I do not have enough words to describe what a blessing it is to be your son. I would also like to thank my girlfriend Leticia for her support and patience throughout my PhD. Thank you so much for giving me that extra dose of energy every time I needed it.

I would like to appreciate my "family" from Nottingham including my friends and my landlord, and of course to my beloved "Jonis". Having friends like you makes everything in life much easier. Thank you to my

family, specially to my uncle "Julito" for always being there. Finally, thank you to my colleagues at the Computational Intelligence and Applications research group for the time shared during the past few years. To everyone who in one way or the other contributed to the success of my research.

This work was supported by Nottingham Trent University's Vice Chancellors' Research Scheme award, for the duration 2016 - 2019.

Dario Ortega Anderez
July 2020

Abstract

The field of Ambient Assisted Living (AAL) is gaining increasing attention from the research community in recent years with the rapid present and future ageing of the population worldwide. This problem has been widely recognised as has the need for it to be addressed both from an economic and societal perspective. Assisted living environments incorporate technological solutions to create a better condition of life for older adults. However, in order to create a better condition of life, it is crucial to understand the specific needs of each individual. To this regard, self-assessment of daily activities has shown to be subjective and variable, presenting important discrepancies with those performed by clinicians.

The above challenges have fostered the search for alternative monitoring solutions, increasing the research efforts upon the field of Human Activity Recognition (HAR). A vast array of sensing devices, including ambient sensors, video cameras and wearable devices, has been employed for the automatic monitoring of a person in a home environment. However, the research focus is shifting towards wearable solutions, which avoid the privacy concerns related to the use of video cameras in a home environment while providing more intrinsic information about the user than ambient devices.

The focus of this research is the investigation of signal processing and machine learning techniques for the recognition of quotidian activities concerning self-neglect (a behavioural condition in which individuals, generally older people, disregard the attention, intentionally or unintentionally, of their basic needs). More precisely, the aimed group of activities include those concerning personal hygiene, namely hands

washing and teeth brushing, as well as those directly related to dietary behaviour, namely eating and drinking.

The work undertaken in this thesis is divided into three different stages. First, given the continuous quasi-periodic behaviour of hands washing and teeth brushing, these are studied alongside a group of other quotidian activities which also exhibit continuity during their performance. These studies include the investigation of informative features for activity recognition as well as relevant classification models and signal processing techniques. In addition, a novel multi-level refinement approach is proposed as a way to improve the classification rate of those activities with lower inter-activity classification rate.

Second, a novel framework for fluid and food intake gesture recognition is developed. As opposed to the above activities, the nature of eating and drinking activities is neither static nor quasi-periodic. Instead, they are composed of sparsely occurring motions or gestures in continuous data streams. Given this characteristic, a novel signal segmentation technique, namely the Crossings-based Adaptive Segmentation Technique (CAST), is proposed to identify potential eating and drinking gestures while filtering out the remaining unwanted segments of the signals. In addition, various feature descriptors, namely a Soft Dynamic Time Warping (DTW) gesture discrepancy measure and time series to image encoding techniques, as well as various deep learning architectures are explored to overcome the notable existing similarity between eating and drinking gestures.

The third stage of the work aims at the identification of meal periods through the analysis of the distribution of eating gestures along time using low-computational cost signal processing techniques, including a moving average and an entropy measure.

The novel computational solutions and the results presented in this thesis, demonstrate a significant contribution towards the recognition of quotidian activities in support of independent living.

Publications

Journal Papers

Ortega Anderez, D., Lotfi, A. and Pourabdollah, A., 2020. A Deep Learning Based Wearable System for Food and Drink Intake Recognition. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-020-02684-7

Anderez, D.O., Lotfi, A. and Pourabdollah, A., 2020. Eating and drinking gesture spotting and recognition using a novel adaptive segmentation technique and a gesture discrepancy measure. Expert Systems with Applications, 140, p.112888.

Ortega-Anderez, D., Lotfi, A., Langensiepen, C. and Appiah, K., 2019. A multi-level refinement approach towards the classification of quotidian activities using accelerometer data. Journal of Ambient Intelligence and Humanized Computing, 10(11), pp.4319-4330.

Conference Proceedings

Anderez, D.O., Lotfi, A. and Pourabdollah, A., 2019, June. Temporal convolution neural network for food and drink intake recognition. In Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (pp. 580-586).

Anderez, D.O., Lotfi, A. and Langensiepen, C., 2018, September. A novel crossings-

based segmentation approach for gesture recognition. In UK Workshop on Computational Intelligence (pp. 383-391). Springer, Cham.

Anderez, D.O., Lotfi, A. and Langensiepen, C., 2018, June. A hierarchical approach in food and drink intake recognition using wearable inertial sensors. In Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference (pp. 552-557).

Anderez, D.O., Appiah, K., Lotfi, A. and Langesiepen, C., 2017, June. A hierarchical approach towards activity recognition. In Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments (pp. 269-274).

Contents

Nomenclature

AAL	Ambient Assisted Living
ADL	Activity of Daily Living
AMI	Ambient Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ApEn	Approximate Entropy
CAST	Crossings-based adaptive Segmentation Technique
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CS	Computational Solution
DBM	Deep Believe Networks
DTW	Dynamic Time Warping
EMG	Electromyography
FFT	Fast Fourier Transform
FN	False Negatives
FP	False Positives
FS	Feature Set
FSS	Feature Similarity Search
GAF	Gramian Angular Field
GAK	Global Alignment Kernel
GMM	Gaussian Mixture Model
HAR	Human Activity Recognition
HMM	Hidden Markov Model
IMU	Inertial Measurement Unit

KNN	k Nearest Neighbours
LMT	Logistic Model Tree
MEMS	Micro-electromechanical system
MTF	Markov Transition Field
PCA	Principal Component Analysis
PIRs	Passive Infrared Sensors
PLR	Piecewise Linear Representation
RBS	Radial Basis Kernel
RF	Random Forest
RMS	Root Mean Square
SNR	Signal to Noise Ratio
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWAB	Sliding Window and Bottom-up
TN	True Negatives
TP	True Positives
WMA	Weighted Moving Average
WSN	Wireless Sensor Network

List of Figures

List of Tables

Chapter 1

Introduction

The world population is ageing rapidly. According to [1], by 2050, the number of older adults will exceed the number of children for the first time in history. Besides, the old-age dependency ratio, calculated as the ratio of older adults (65 or older) to the working-age population (15 to 64), is growing fast, particularly within developed countries [2]. Studies indicate that older adults normally prefer to stay at their own homes as long as possible [3]. As a result of this, more older people live alone as sole occupants of a dwelling than any other population group [4]. In addition, the number of older adults needing peripheral support during their quotidian activities follows a worrying upwards trend, and it is predicted to reach the 22% by 2050 [2].

This problem has been widely recognised as it has the urgency for it to be addressed both from an economic and societal perspective. An increase in the number of care providers is a potential solution. However, the fast growth seen on the old-dependency ratio makes this option rather unrealistic. It is suggested the use of smart technologies can mitigate the impact of this demographic problem [5]. The significant advances in mobile and ubiquitous computing have already translated into increasing attention towards emerging research fields such as Ambient Intelligence (AMI) and Ambient Assisted Living (AAL). The aim is to enable independent living while promoting a better condition of life employing different assistive systems. Nonetheless, to be able to support and assist individuals, it is first crucial to understand their specific needs by the deployment of accurate monitoring systems.

Recent research in AAL has investigated the use of different monitoring platforms including video cameras [6, 7], ambient sensors such as passive infrared sensors (PIRs), pressure mats or magnetic sensors [8], as well as that of wearable devices incorporating motion sensors such as accelerometers and gyroscopes [9]. Although various pros and cons can be found on the employment of each sensing technology, research efforts are currently shifting towards the use of wearable solutions based on three main factors. First, the information provided by ambient sensors (normally in the form of binary data) is rather basic, being thus insufficient to monitor complex behaviours. Second, even though computer vision has been proven to be an accurate means of monitoring humans, there exist major privacy concerns with its use in home environments [10]. In addition, the problem of occlusion caused by the frequent presence of an object between the video camera and the subject makes the use of this solution unsuitable in most home environments. Third, recent surveys regarding the acceptability of the use of wearable devices have shown positive results, not only in adults [11], but also within the elderly population [12].

The remainder of this chapter is organised as follows. Section 1.1 presents the motivation behind this thesis. Section 1.2 provides an overview of the research undertaken in this thesis. Section 1.3 presents the project aim and objectives. Section 1.4 discusses the major research challenges identified for the completion of this work. Section 1.5 presents the major contributions achieved throughout the undertaken work. Finally, Section 1.6 outlines the organisation of the remaining chapters of the thesis.

1.1 Motivation

Current wearable and portable technologies such as smart phones, smart watches and/or fitness trackers incorporate a great array of sensors (i.e. accelerometers, gyroscopes, magnetometers), allowing for human behaviour analysis in different applications. Examples include fitness [13–17], rehabilitation [18], security [19] and health care [20].

Predominant attention has been given to fitness applications, where typically quasi-periodic activities such as walking, running or climbing stairs are analysed.

The efforts of this work are focused on the search for means to identify activities related to self-neglect issues. This not only includes the identification of quasi-periodic activities such as teeth brushing and hand washing but also the recognition of activities composed of sparsely occurring gestures like the intake of meals.

Self-neglect is defined as a behavioural condition in which individuals, normally older people, intentionally or unintentionally, disregard the attention of their essential needs [21]. These include any form of lack of personal hygiene, appropriate feeding or any other aspect regarding self-care. Research statistics from the National Health Service (NHS) [22], show that neglect and omission are the main risks for safeguarding inquiries, with rising figures comparing to previous years. Besides, self-neglect has been empirically related to cognitive impairment and depressive symptoms [23], and more importantly, it has been proven to be on its own an independent risk factor for death [24], with some of the most common diagnoses being hypertension, diabetes mellitus, dementia and depression.

The above strongly suggest that there is a research need for exploring novel monitoring solutions to track quotidian activities concerning self-neglect issues. The lack or under-performance of these activities could potentially indicate the need for peripheral support or the inability for independent living. Such information not only could be used to alert the older person's relatives, carers or medical institutions but to directly remind the person to carry out the activities themselves or yet cooperate with assistive robots to aid the individuals in the procedure as well. To the best of my belief, a self-neglect behaviour tracking system can be a significant contribution towards independent living in the way that it ensures the acknowledgement of the well-being of the subject by their relatives while passively contributing to the well-being of the person itself. In line with this, all the efforts of this thesis are directed towards the development of computational solutions for the accurate recognition of different activities concerning self-neglect issues. These mainly include hands washing, teeth brushing, as well as food and drink intake.

1.2 Overview of the Research

This study aims to develop computational solutions to accurately recognise quotidian human activities based on wrist-motion information. Human Activity Recognition (HAR) can be understood as a problem whose aim is to identify patterns on sensory temporal sequences of data or time series to infer the activity being performed by a person at a specific point in time. Specifically, the efforts are given to the recognition of a group of four activities concerning self-neglect issues. These include hand washing, teeth brushing, eating and drinking.

Driven by the differing nature of the activities of interest, the investigation is divided into two main parts, whereby quasi-periodic activities (hands washing and teeth brushing) and sequential activities (eating and drinking) are studied separately. For clarification purposes, it should be noted this work defines quasi-periodic activities as those activities whose motion signals exhibit certain similarity to a periodic function when being performed and sequential activities as those activities that are composed of sequences of sporadic gestures. Intuitively, the former group is studied through the analysis of consecutive time windows, whereas the latter group is studied through the temporal analysis of the occurrence of relevant gestures.

Despite the above, the recognition of quasi-periodic and that of sequential activities share three common areas of investigation. These are defined as follows:

- Signal pre-processing: this area embodies the investigation of signal processing techniques to suitably accommodate the raw sensory signals for further analysis. These include the removal of unwanted components of the signals through signal filtering, the creation of approximated functions to capture important patterns within the signals as well as the segmentation of the signals to either break them down into segments which share a common characteristic or to filter out unwanted segments.

- Feature extraction: this involves the investigation of means of transforming the pre-processed signals into a reduced number of variables (features) to facilitate the subsequent learning and generalisation of the activities and gestures of interest. A vast array of hand-crafted features are explored in

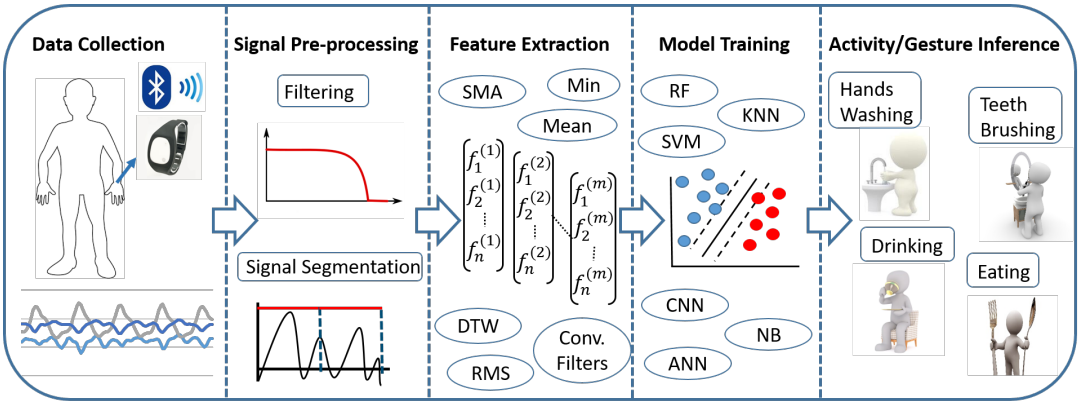

Figure 1.1: An illustration of the main steps of an activity or a gesture recognition system.

this work. Further feature descriptors are explored for the recognition of eating and drinking gestures based on their high degree of similarity in terms of wrist motion. These include the development of a gesture discrepancy measure based on Dynamic Time Warping (DTW) and the exploration of deep learning models as feature descriptors.

- Gesture/activity classification: this includes the investigation upon the performance of various supervised learning techniques for the detection of the activities and gestures of interest.

A further investigation was undertaken to identify a means of detecting meal periods based on the occurrence of eating gestures across the day. Low computational cost signal processing techniques were explored to study the distribution of eating gestures across time.

The above research areas were thoroughly investigated to address the main research question of this thesis - how can a single wrist-worn motion sensing unit be used to detect quotidian activities concerning self-neglect issues. For illustration purposes, Figure 1.1 depicts the main steps undertaken in an activity/gesture recognition problem.

1.3 Aim and Objectives

The aim of this research is to identify means of detecting quotidian activities concerning self-neglect using a single wrist-worn motion sensing unit, including a tri-axial accelerometer and a tri-axial gyroscope. The work undertaken embodies a thorough investigation of appropriate signal processing and computational intelligence techniques to be able to accurately recognise quotidian activities whose diminution or lack of performance can potentially be identified as jeopardy for the health and well-being of an individual, thus implying the need for peripheral support.

To achieve the above research aim, the following objectives were identified:

1. To conduct extensive research into existing methods for activity recognition using inertial sensors.

2. To propose a computational solution to recognise quotidian quasi-periodic activities in a home environment.

3. To extensively investigate signal processing, feature extraction and computational intelligence techniques to spot and recognise sporadic gestures from continuous motion data streams.

4. To explore deep neural network architectures for feature extraction and food and drink intake gesture recognition.

5. To develop a computational solution to accurately spot and recognise food and drink intake gestures from continuous wrist motion data streams.

6. To investigate the temporal occurrence of eating gestures to develop a computational solution for the recognition of meal periods under a free-living environment.

1.4 Research Challenges

In order to fulfil the above research aim and objectives, four main research challenges have been identified:

1. Sensing platform deployment: Various factors have to be taken into consideration for the selection of a suitable sensing platform for human monitoring. While the employment of numerous sensing devices can increase the sensory information obtained, unobtrusiveness and freedom of motion should be carefully taken into consideration. Another critical issue is that battery-life should be maximised utmost. Currently, wearable devices incorporate a great array of sensors, however, the overuse of resources can make a system inadequate for continuous monitoring. Therefore, it is a challenge to minimise the number of sensing units as well as the number of sensors within a sensing unit while maintaining an adequate recognition rate.

2. Structure of human activities: Human activity is normally referred to as a global term. However, it can be divided into different levels in a similar way when natural language is processed. For instance, a paragraph can be broken down into different words, a word can be divided into different syllables, and these can be further broken down into different letters. Likewise, activities can be broken down into smaller actions, and these can be further divided into basic movements. For instance, the activity 'eating pasta' could be further divided into smaller actions such as 'using the fork to take a bite', which at the same time could be broken down into 'lift up the hand' and 'put the hand back down to the rest position'. It is a challenge to accurately model the spatial and temporal relationship between the different outlined elements.

3. Differing nature of activities and signal segmentation: Although activity recognition is normally understood as a standalone problem, human activities differ significantly from each other in the way they are performed. For instance, while walking exhibits a continuous quasi-periodic temporal behaviour, eating is composed of sparsely occurring gestures. Signal segmentation is a crucial aspect for activity and gesture recognition which aims at either breaking down the signal into segments that share a common characteristic or to filter out unwanted segments of the signal. The differing nature of activities is, therefore, a key research challenge which demands

the investigation of adaptive signal segmentation techniques to comply with the differing characteristics of the activities themselves.

4. Gesture Similarity: This challenge refers to the high degree of similarity encountered on fluid and food intake activities in terms of the hand movements required to perform these two activities. It is crucial to explore discriminative feature descriptors so that discrepancies that will facilitate the learning of classification models can be found amongst such a high degree of similarity.

1.5 Major Contributions

The major contributions of this thesis are summarised as follows:

- To propose a computational solution for the recognition of quotidian activities concerning personal hygiene.
- Propose a novel multi-level refinement approach for activity recognition. As demonstrated in this work, the employment of this approach can achieve an improvement in the recognition rate of the activities which were originally lowering the performance of the whole system.
- Propose a novel adaptive signal segmentation technique (CAST) for spotting potential eating and drinking gestures within continuous motion data streams. This technique achieves a recall of 100%, therefore, it overcomes the main drawbacks encountered in previous attempts at developing segmentation techniques for spotting sporadic gestures. Given its outstanding results and its flexibility, CAST can be used in future activity and gesture recognition work.
- Propose the introduction of a DTW-based gesture discrepancy measure into long-established feature sets. As demonstrated in this work, the use of the gesture discrepancy measure consistently improves the gesture recognition rate across different experiments. This suggests the use of its employment as a feature descriptor in future activity and gesture recognition work.

- Propose a system for the recognition of eating and drinking gestures with the use of a novel adaptive segmentation technique and hand-crafted feature vector incorporating a personal gesture discrepancy measure.
- Propose a deep-learning-based domain knowledge free system for the recognition of eating and drinking gestures.
- Propose a novel approach for the detection of meal periods through the analysis of the occurrence of eating gestures across time.

1.6 Thesis Outline

In order to address the research question presented in Section 1.3 with regards to how a single wrist-worn motion sensing unit can be used to detect quotidian activities concerning self-neglect issues, Chapter 2 conducts an extensive literature review upon previous work in the field of activity recognition using wearable sensors to identify relevant research methods, as well as relevant research challenges and opportunities within the field. Chapter 3 presents a technical overview of the common methods employed for activity and gesture recognition and summarises how these are explored and employed throughout the work undertaken in the thesis. Chapter 4 presents a computational solution to recognise hygiene-related activities. Chapter 5 and Chapter 6 investigate the spotting and the recognition of sparsely occurring eating and drinking gestures from continuous data streams. Ultimately, making use of the knowledge gained in the previous two chapters, Chapter 7 investigates the recognition of meal periods based on the distribution of eating gestures across time. Putting all together, this thesis provides novel computational solutions to identify the main hygiene and nutrition-related activities concerning self-neglect issues using a single wrist-worn motion sensing unit.

In line with the above, The remainder of this thesis is organised as follows (see also Figure 1.2).

Chapter 2: Literature Review - This chapter provides an overview on previous work in the field of Ambient Assisted Living (AAL) as well as a detailed discussion on relevant literature in Human Activity Recognition (HAR) and its main crucial aspects.

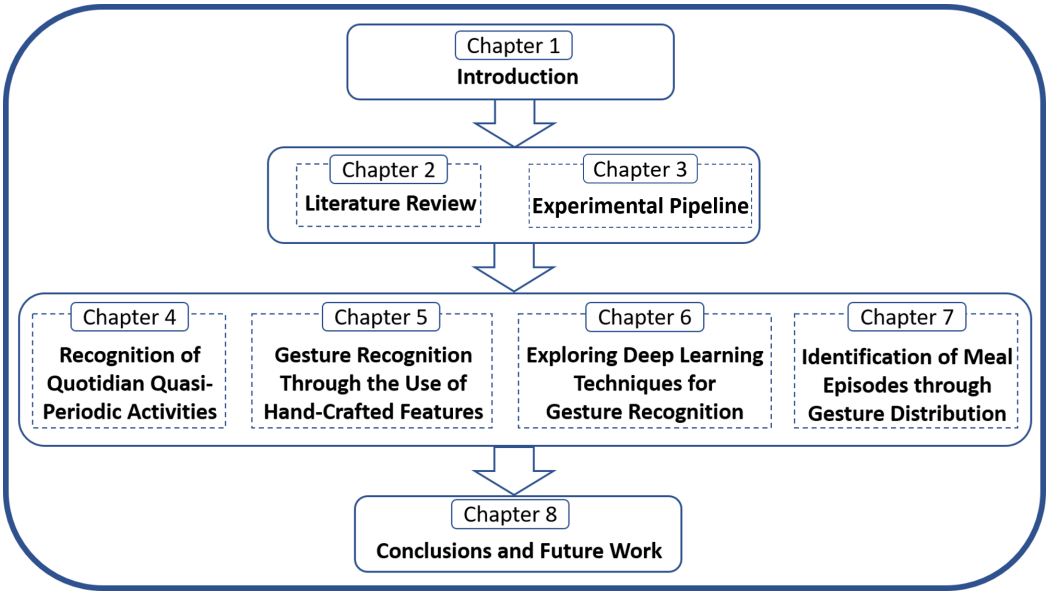

Figure 1.2: Thesis structure showing the organisation of the chapters.

Chapter 3: Experimental Pipeline - This chapter presents the common methods employed in HAR work with the use of wearable sensors. This includes the description of how data is collected, as well as the signal processing, feature extraction and computational intelligence techniques employed for gesture and activity recognition purposes.

Chapter 4: Recognition of Quotidian Quasi-Periodic Activities - This chapter describes in detail the proposed system for the recognition of the quotidian quasi-periodic activities of interest for this study. In addition, this chapter presents a novel multi-level refinement approach as an alternative to state-of-the-art classification approaches.

Chapter 5: Gesture Recognition Through the Use of Hand-Crafted Features and Gesture Discrepancy - This chapter presents a solution to spot and recognise eating and drinking gestures from continuous wrist motion data streams. This solution incorporates a novel signal segmentation approach (CAST) and the introduction of a gesture discrepancy measure into long-established feature sets. The experimental results achieved are presented and compared to those obtained by previous similar work.

Chapter 6: Exploring Deep Learning Techniques for Gesture Recognition -

This chapter explores the use of deep learning techniques for feature extraction and for the recognition of eating and drinking gestures. The performance of 1-Dimensional (1D) Convolutional Neural Networks (CNNs), 2D CNNs and multi-input networks are evaluated.

Chapter 7: Identification of Meal Periods Through Gesture Distribution - This chapter describes a novel approach to identify meal periods through the analysis of the occurrence of eating gestures across time.

Chapter 8: Conclusions and Future Work - This chapter presents the conclusions arisen from the thesis and propose directions for future work on human monitoring with the use of wearable sensors.

Chapter 2

Literature Review

2.1 Introduction

The increasing life expectancy alongside the decline in birth rates across the world is translating into a global ageing population structure [25], especially in developed nations such as the UK. Ageing is a phenomenon caused by the impact of a wide variety of cellular and molecular damage over time, which in turn lead to a gradual decrease in cognitive and physical abilities. As a result of this, the senior-age stage of a person's life is usually spent suffering from multiple disabilities [26], and as age increases, elderly individuals may lose the capacity to attend their basic needs (*i.e.* food or drink intake), therefore requiring peripheral support. There is thus a need to address the current demographic issue by providing means of sensing individual needs, through which the use of further resources (*i.e.* carers, care home spaces and assistive robotics) can be optimised. Given this, advances in sensing technologies can play a crucial role in preserving the wellbeing of older population groups while supporting their independent living.

In line with this, this chapter provides a comprehensive review of previous work in the fields of Ambient Assisted Living (AAL) and Human Activity Recognition (HAR). A particular emphasis is given to HAR work based on wearable motion sensors, where a thorough critical evaluation upon the different crucial aspects and open challenges found across the work in the field is undertaken. The aim of this chapter is, therefore, the justification of the overall experimental and

development work undertaken throughout this thesis through the identification, critical evaluation and discussion of the existing limitations found in HAR work in the context of AAL. Additionally, each experimental chapter is accompanied by its own specific review of related work and critical evaluation, justifying independently each of the experiments carried out and the methodology employed.

The remainder of this chapter is organised as follows: Section 2.2 discusses the rationale behind the field of AAL, the benefits it can bring to independent living and presents relevant work using alternative approaches to that of the use of wearable sensors. Section 2.3 provides a discussion upon the different crucial aspects of implementing activity recognition systems with the use of wearable sensors, including the sensor modality, the sensor placement, the sampling frequency, the signal pre-processing, the segmentation of the signals, the extraction of features and the use of classification models. Section 2.4 presents the conclusions drawn from the literature and the research opportunity identified through the analysis of previous work.

2.2 Assistive Technologies

The worrying upwards trend on global population ageing, alongside the current advances in ubiquitous computing, wireless technologies and robotics have made the development of assistive technologies become a crucial demand, leading to increasing research attention to the field of AAL and to the development of AmI technologies in recent years [27, 28]. Such technologies aim at the development of sensitive, responsive and adaptive environments to create a better condition of life for individuals with disabilities and older adults, while supporting their independent living. The effective use of sensing and monitoring devices is crucial to understand and anticipate the specific needs each individual may have. Likewise, gaining insights into such individual needs can improve the performance of assistive robots and optimise the use of human resources in terms of providing ad-hoc support to each individual. A broad categorisation of the sensing devices employed to monitor individuals in a home environment is generally made regarding the sensing modality employed. This includes three main categories, namely ambient sensors, video cameras (computer vision), and wearable sensors.

2.2.1 Ambient Sensor-based Technologies

Numerous research works have employed ambient devices for monitoring subjects in a home environment. The term 'ambient devices' embodies a broad array of sensing devices which are typically embedded within a home environment. The main examples include microphones [29], door magnetic switches [30, 31], RFID [32], Passive Infrared Sensors (PIRs) [30, 31], pressure mats [33–35] and infrared sensors [36]. Such devices are normally installed to form a multimodal collaborative sensing and intelligent Wireless Sensor Network (WSN) to be used in various application. For example, PIRs and magnetic door switches are normally utilised to track the movement of subjects across different areas of the house. Such information is then used, among other applications, to infer the activities being performed by the subject [37], to build personal behavioural patterns and identify abnormal behaviours or events [31, 38], as well as to infer single-occupancy and multi-occupancy scenarios in home environments [30]. Pressure mats are commonly used for the detection of sleeping periods and personal sleeping patterns [33–35], as well as for gait analysis and fall detection applications [39]. RFID technology is typically used to track the indoor location of individuals by the installation of various RFID tags around the home environment. Likewise, the use of RFID tags on everyday objects has been employed for activity recognition applications [40, 41]. Microphones are also employed as a means of indoor location tracking [42] and activity recognition [43]. However, the performances achieved at the latter application, ranging from 54% [44] to 80% [45] in terms of classification accuracy, are considerably low as compared to those achieved by the use of computer vision or wearable devices.

With regards to the recognition of Activities of Daily Living (ADLs) with the use of ambient sensors, various efforts are identified within the literature. For instance, the work in [46] proposes an audio-based system for the recognition of eating activities, achieving a classification recall of 76.3%. In [47] a pressure sensor matrix is used to detect the position of kitchen utensils on a table to infer eating related actions achieving a classification accuracy of 77%. The work in [40] makes use of a wrist-worn RFID reader and a network of RFID tags embedded into everyday objects to recognise a number of ADLs, including among

others, medication intake, food preparation related activities and teeth brushing, achieving an average per-class classification recall of 91%. A fluid intake recognition solution based on RFID technology is proposed in [48] by the attachment of four RFID tags to glass, and the analysis of the received signal strength (RSSI) between the RFID tags and a ceiling-mounted RFID antenna, achieving classification recalls in the range of 70.8% to 85.4% across different experiments.

2.2.2 Computer Vision-based Technologies

Computer vision-based technologies in the context of HAR make use of video cameras as a means of recognising and monitoring human activities, where typically the camera or combination of cameras employed are fixed at specific locations and angles within the home environment. Generally, the problem of activity recognition based on video sequences is tackled in three main steps. First, moving objects are segmented out from the video by the application of background subtraction image processing techniques such as Gaussian Mixture Models (GMM) [49] or by the use of threshold-based background modelling techniques based on individual pixel chromatic statistics such as the average [50] or the median [51]. A posteriori, a feature vector incorporating relevant characteristics of the human object such as the silhouette, the orientation, the change of the shape or the motion between consecutive frames is built [52]. Ultimately such feature vector is used to train a range of classification models to recognise the specific activity being performed across a sequence of video frames. Alternatively, deep learning techniques, especially CNN architectures, in which the feature learning takes place during the training phase, are widely employed recently for activity and gesture recognition applications [53–55].

As with ambient sensors, several attempts have been made to recognise ADLs in home environments with the use of computer vision. For instance, the work in [56] proposes a camera-based system to recognise eating and drinking activities from manually pre-segmented sequences of video frames achieving an overall classification rate of 93.3%. In [53] different CNN and hybrid CNN-LSTM networks are proposed to recognise food intake actions from video recordings achieving an F1 score of 85.8%. The work in [55] proposes a hybrid network-based bite

detection system achieving a classification recall of 91.71%.

2.2.3 Wearable Sensors-based technologies

Wearable sensors are being increasingly adopted for the development of activity recognition and behaviour analysis systems in view of their many applications in fitness, security and health care. Major examples of these applications include fall detection [57, 58], sleeping analysis [59], gait analysis [60], door security [19], activity recognition [14–16, 61] and gesture recognition [62, 63]. Within activity and gesture recognition work, predominant attention has been given to fitness applications in which quasi-periodic activities such as walking, running or climbing stairs are studied [14–17]. In contrast, more limited research has been reported concerning the recognition of quotidian daily activities, such as eating [9], drinking [64, 65] or hygiene-related activities [66], which in turn could potentially be used as an indicator of the mental health and physical well-being of older adults living independently. Besides, the use of wearable solutions can help to overcome the existing occlusion issues and the privacy concerns related to the use of video cameras in a home environment, while providing explicit personal-oriented data, as opposed to that provided by ambient sensors, which is rather object-oriented and simplistic (normally in the form of binary data). In addition, previous surveys regarding the acceptability of the use of wearable devices [11, 12] support the adoption of wearable devices as human monitoring systems, not only in adults but also within older population groups.

2.3 Activity and Gesture Recognition Using Wearable Sensors

A wearable-based HAR system generally embodies four main steps, namely data collection, signal pre-processing, feature extraction and activity classification. As a summary, the data collection deals with the gathering of informative data regarding the activities of concern for a specific application. This involves the adequate selection of the sensors employed and their placement on the human body, as well as the sampling frequency utilised to collect experimental data. At

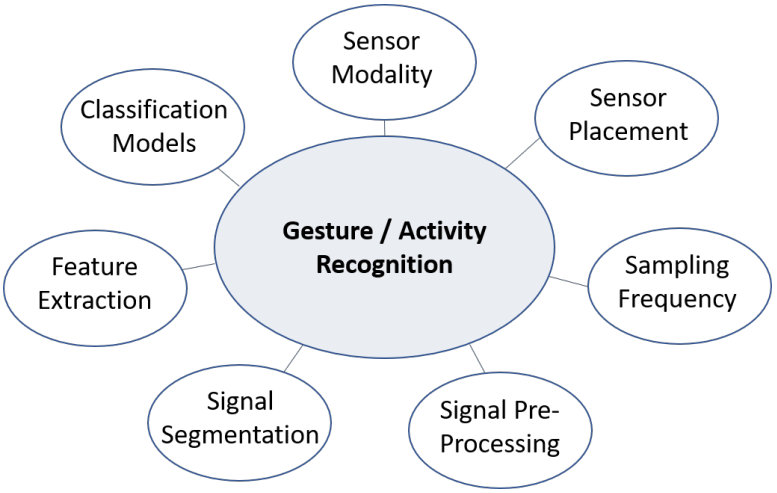

Figure 2.1: The main crucial aspects of an activity or gesture recognition system.

the pre-processing step, filtering techniques are employed to reduce the signal noise and enrich the information provided by the collected raw data. The pre-processed data is then transformed into a feature set or feature vector with a reduced number of variables. Ultimately, the resultant feature set is used to train a range of classification models, which throughout the training process, learn the relationship between the feature set and the activities included in the activity set.

The remainder of this section discusses the crucial aspects involved in recognition of gestures and activities using wearable sensors depicted in Figure 2.1.

2.3.1 Sensor Modality

The first question that arises in the context of activity recognition using wearable devices is which sensor/s should be employed. The continuous miniaturisation of electronics and Micro-Electro Mechanical Systems (MEMS) has enabled the possibility of embedding several sensors into a single wearable or portable device such as a smartphone, a smart band, a smart watch or smart clothes. A broad range of sensors has been employed either independently or through information fusion. Major examples include tri-axial accelerometers [67–69], tri-axial gyroscopes [69], tri-axial magnetometers [70], barometric sensors [71], light sensors [72], heart rate sensors [73] and wearable microphones [72, 74]. However, as more sensors are em-

bedded into a platform or a device, obtrusiveness and power consumption issues can compromise the acceptance of wearable devices as all-day monitoring mechanisms [75]. Although other resources like memory, bandwidth and processing power are constantly being improved to keep up with the increasing storage, communication and processing power demanded by novel applications like HAR, the battery capacity grows only at an approximate rate of 10% per year [76]. Therefore, it is crucial to keep a fair balance between classification performance and power consumption. Given this, most previous work in the field has opted for the employment of inertial-based motion sensors, especially that of accelerometers and gyroscopes. As demonstrated in previous work [15, 77], the employment of these two motion sensors as a source of data for activity recognition can lead to the achievement of a classification accuracy in the range of 90% to 99%. As suggested by the results reported in [78], they can complement each other when combined together. However, the same results suggest that when the employment of one of these two sensors individually achieves a satisfying classification performance, the addition of a second sensor may not lead to an improvement on the original performance. Therefore, this fact should be considered before the blind employment of several sensors. In line with this, most previous work in HAR makes use of stand-alone accelerometers [15, 67, 77, 79], which in addition to being able to provide data from which accurate activity recognition algorithms are built, their power consumption is approximately ten times lower than that of gyroscopes [9], therefore being more appropriate for all-day monitoring applications.

2.3.2 Sensor Placement

The sensor placement indicates the part or parts of the body where the sensing device(s) are worn or placed. The selection of the optimal sensor placement is a crucial aspect of achieving accurate recognition rates in activity recognition. However, there is still some debate over this research topic. Various attempts have been made to unravel this question by evaluating the recognition rate on certain activity sets with different sensor placements [80–83]. For instance, the work in [80] evaluated the recognition rate achieved by using seven different accelerometer placements on five groups of activities, where the activities were

grouped from very low level (e.g. laying down) to high level (e.g. running) according to the level of motion required at their performance. The results of the study suggest the waist is the optimal position for the recognition of low-level activities such as eating or drinking, whereas the ear is the preferred option for high-level activities such as cycling or running. The work in [81] evaluates the classification performance achieved by six different accelerometer placements on a group of seven activities, including both high level and low-level activities. In this case, the best results are obtained when the accelerometer is worn on the hip. In [83], three different sensor placements, namely right wrist, left ankle and chest, are evaluated using two different activity sets, including eight different activities. In this case, the ankle is shown to be the optimal placement, achieving the best overall activity recognition rate. However, the optimal placement varies across individual activities. In [82], the wrist is shown to outperform the hip and the thigh on the overall recognition rate of a set of fourteen activities.

Given the inconsistency found between the different attempts to define an ideal part of the body for the placement of a sensing device, HAR studies normally select the sensor placement based on their specific application. For instance, wrist-worn placements are convenient to recognise activities implying the use of the hands such as teeth brushing, smoking, eating or ironing [20, 61, 84, 85]. Foot-mounted or ankle-mounted sensors are preferred for step counting and gait analysis applications [86–88]. Thigh-mounted sensors can well reflect the leg motion involved in activities such as jogging or walking [82, 89, 90].

Other studies make use of multiple motion sensing devices placed at different parts of the body [91–95] or yet incorporate additional sensing devices such as surface Electromyography (EMG) electrodes or stethoscope microphones for the recognition of complex activities such as eating [74, 96]. As one would expect, the HAR systems incorporating multiple sensing devices generally outperform those which incorporate only a single sensing unit [81, 83]. However, such improvement in the classification performance comes along with the undesired extra obtrusiveness and lack of usability in real-life scenarios of multi-sensor setups.

2.3.3 Sampling Frequency

The sampling frequency refers to the frequency at which the sensory data is collected. The main concern regarding the selection of an adequate experimental sampling frequency is to comply with the Nyquist-Shannon sampling theorem [97] which establishes that the minimum sampling rate at which a continuous-time signal of finite bandwidth should be sampled to capture all the relevant information, must be at least twice the highest significant frequency in the signal. While complying with this is crucial, it is also important to consider the computational cost associated to the selection of an excessively high sampling frequency, since the computations required in further steps, namely in the signal pre-processing and feature extraction steps, would have to deal with an unnecessary excess of data, therefore increasing the overall computational cost and power consumption of the entire system. In this regard, the work in [98] studies the impact of the sampling frequency on the final classification performance across five benchmark datasets composed of different activities. According to the results obtained, the optimal sampling frequency, that is, the lowest sampling frequency achieving comparable performance to those achieved with higher sampling frequencies, varies significantly with regards to the activities being included in the activity set. In a similar effort, the work in [99] compares the classification performance achieved by five different sensor placement setups as a function of the sampling frequency using a unique activity set. The results indicate the classification rate increases marginally by just 1% above 20Hz and stabilise beyond 50Hz. According to [67], the fundamental frequencies of human activities do not exceed 20Hz. However, this statement is made with regards to the findings in [100] which concerns gait analysis only. Given the discrepancies found across the above experiments, the sampling frequency across the different works in the field has generally been arbitrarily selected by taking into consideration the nature of the activities to be classified, with sampling frequencies ranging from as low as 1Hz [101] to as high as 200Hz [99] and with the majority of studies ranging from 50Hz to 100Hz [13–15, 77, 102, 103].

2.3.4 Signal Pre-processing

Pre-processing techniques have been widely used to enrich the information provided by unprocessed experimental raw data. These include the application of digital filtering and smoothing techniques with the aim of reducing potential noise in the collected signals, the isolation of specific frequency components of the signal to be analysed individually, as well as the computation of additional time series from which to extract further features.

With regards to noise reduction, various studies have made use of Butterworth low pass filters to eliminate those frequency components which are not believed to be caused by human activities or human actions [104,105]. Beside Butterworth filters, smoothing techniques such as a moving average [79], a weighted moving average (WMA) [20], median filters [106, 107] or Gaussian-weighted windows [9] have also been employed as a means of noise reduction, while other works have opted for not to carry out noise reduction [16,77]. Concerning noise, an investigation upon the signal to noise ratio (SNR) offered by different filtering techniques on accelerometer data, is carried out in [67], with one of the conclusions which can be drawn from this study being that a median filter offers a good balance between computational cost and SNR as compared to other filtering techniques.

Filtering techniques have also been used to isolate different frequency components of the signal. This is generally done with the purpose of separating the low-frequency component of the signal caused by the gravitational force, from the high-frequency component due to the linear acceleration caused by the motion of the body part where the sensor is placed. For instance, the work in [15] makes use of a low pass filter and a high pass filter both with a cut-off frequency of 1Hz to separate the gravity component from the body acceleration component to then perform feature extraction from the two separated components independently. Similarly, in [14], the gravity and the acceleration components are separated using a digital filter with a cut-off frequency of 0.25Hz. However, only the high-frequency component is kept for further analysis in this case.

Further processing techniques have been used to compute additional time series from those provided by the sensory devices. A common approach within the work in the field [13, 15, 69] is to complement the tri-axial signal with a

fourth time series, namely the magnitude of the tri-dimensional vector, to provide orientation-free information. Even, in [108], although it did not lead to a good performance as compared to similar work, only the magnitude time series is used to train a CNN, discarding, therefore, the tri-dimensional accelerometer signal. Other studies [109,110] propose the use of the jerk (the rate at which acceleration changes) to provide additional time series for feature extraction.

2.3.5 Signal Segmentation

Once the signals are pre-processed, the next step is to divide the resultant data streams into shorter windows or segments to facilitate the later feature extraction and learning. Given its simplicity, a common approach throughout the field is to employ sliding windows, through which the data streams are divided into consecutive (often overlapping) time windows of equal length [13–17,77,111–116]. The intuition behind this approach is to identify a window length that incorporates the fundamental characteristics of an activity or an activity cycle. Window sizes have varied considerably between studies, with lengths varying from as low as 0.08 seconds [112] to as high as 30 seconds [113]. No reasoning is provided regarding the window length employed, therefore implying studies in the field commonly rely on an arbitrary window size or the success of a specific window length in previous similar work. Hereof, the impact of the window size on the classification performance of HAR systems has been investigated in various research works [117, 118]. The results reported in [118] with regards to the classification rate achieved as a function of the window size across different activities, obtaining differing optimal values for various activities, suggest that the optimal window size in a HAR system is dependent on the activity set studied. Nonetheless, given the optimal window size intervals provided in this same study, a safe option could be the selection of window size of around 1 second. A similar approach to sliding windows is proposed in [119]. In this study, the segmentation of the signals is done through the division of the data streams into extremely short fundamental movements to then through the clustering of such fundamentals, perform activity classification by studying the distribution of the fundamentals across each activity.

Despite the success achieved by the use of sliding windows on activity classification problems, adaptive segmentation techniques have been shown to offer better performance when tackling gesture recognition problems [120]. Within adaptive segmentation techniques, Piecewise Linear Representations (PLRs) are well-known techniques [121, 122]. In PLRs, segments of time series are approximated to a line either by the application of linear regression or interpolation, until a customised threshold error is exceeded. A posteriori, a Feature Similarity Search (FSS) is normally used to narrow down the number of segments [123]. In point of fact, the work in [123], employed a PLR, namely the Sliding Window and Bottom-up technique (SWAB), to spot a set of fluid and food intake gestures.

Besides PLRs, various adaptive segmentation approaches have been proposed for spotting sporadic gestures or actions from continuous inertial data streams. For instance, the work in [120] proposes an extendable Gaussian probability function-based window. In [61], a segmentation approach based on a re-adjustable resting position and a distance peak detector from the most current resting position is proposed. The work in [9] employs a wrist motion energy threshold-based segmentation approach. In [124] the sign changes on the accelerometer signal are used to divide it into different potential segments of interest.

2.3.6 Feature Extraction

Feature extraction is the process of computing abstractions from the raw or preprocessed segmented sensory signals to extract the characteristics that better describe the original signal [125]. Through such abstraction, a large set of data is transformed into a reduced representation, namely a feature vector, which includes relevant cues for the categorisation of the activities themselves. The resultant feature vector is a posteriori used as the input for classification models. In the context of HAR, features can be divided into two main categories regarding how they are extracted, namely hand-crafted features and automatically learned features.

Hand-crafted features are those features which are based on domain knowledge [25], being purposely computed for a specific application. The intuition behind the use of these features is that a sensory signal value at an instant point

in time does not provide sufficient information to infer which activity is being performed at the time. However, certain informative characteristics of the signals over a period of time or time window can provide valuable information. Generally, hand-crafted features are computed in the time and frequency domains, with the former domain being dominant across the different works in the field. The extraction of features in the wavelet domain has also been explored [17], however the extra computational cost did not translate into a better classification performance. Within the time domain, statistical features have been shown to include relevant characteristics of the signals which can help to distinguish between different activities. For instance, the standard deviation over a time window can provide valuable information to a classification model to make a decision upon whether a static activity such as 'standing' or a dynamic activity such as 'running' is being performed over the period of time delimited by that window. Main examples of statistical features include the mean [14, 15, 17, 72, 77], the standard deviation [14, 15, 17, 72, 77], the signal magnitude area [17], the root mean square (RMS) [14, 15, 17, 72], the inter-quartile range [17, 72] or the correlation between different signal axes [15, 77]. Main examples of features in the frequency domain include the bandwidth [13, 17] and the spectral energy [17, 126]. Dynamic Time Warping (DTW), which is a time series dissimilarity measurement algorithm, has also been widely employed as a feature descriptor [18, 127, 128].

In contrast to hand-crafted features, automatically learned features do not require specific domain knowledge since they are automatically learned throughout the training process of the corresponding deep learning classification algorithm. Although hand-crafted features were dominant in the field for a long period of time, deep learning is increasingly employed for both feature extraction and activity/gesture classification given the promising results achieved, especially with the use of Convolutional Neural Networks (CNNs) [90, 129–132].

2.3.7 Classification Models

The last step in the development of a gesture/activity recognition system is the classification of different gestures/activities. The aim of this step is to relate the information gained throughout the feature extraction to the different activities

or gestures using the different observations. Classification models are normally divided into conventional models and deep learning models.

Within conventional classification models, Support Vector Machines (SVM) [13,14,16,77,133,134], k-Nearest Neighbours (KNN) [77,79,133,135] and Random Forest (RF) [13–15, 133] have shown good classification performance across different feature sets and experiments. However, as with the different HAR-related aspects discussed in previous sections, a precise evaluation upon which classification model exhibits the best performance in HAR systems cannot be made, since their performance varies across different experiments and activities. For instance, the results in [14] show SVM offers better performance than that of RF at recognising walking related activities, whereas RF obtains a better performance at recognising dancing-related activities. Given this, the tendency amongst the works in the field is to evaluate the performance of the proposed systems using a range of classification models [14, 15, 77, 83]. In addition to the above classification models but to a lower extent, Hidden Markov Models (HMM) [65, 85] and Naive-Bayes [85] have also been employed in previous HAR work.

The recent advances in deep learning architectures are revolutionising the field of HAR. In contrast to conventional classification models, the majority of deep learning models do not rely on domain-specific knowledge to exhibit good classification performance, since the features are automatically learned throughout the training process. An increasing number of research works are recently employing deep learning models for activity recognition with the use of wearable sensors, specially CNNs [90, 129–132], which as shown in [130] can clearly outperform other deep learning approaches such as Artificial Neural Networks (ANNs) and Deep Believe Networks (DBN), as well as conventional classification models. As demonstrated by previous work [90, 130, 132, 136] CNNs can achieve a classification accuracy of over 90% across different human activity classification problems. Besides, CNNs have been used in combination with LSTM recurrent layers to exploit the temporal dynamics present in human activities [137]. Although LSTMs are normally used to make further predictions on given sequences of data, models exclusively formed by LSTM layers have also been employed by previous work in the field achieving satisfactory results in activity recognition problems [138, 139]. However, when compared to the performance of CNNs on a benchmark dataset

(UCI HAR [109]), CNNs [132] have shown a better classification performance than that shown by LSTMs [138, 139]. ANNs have also showed good performance in HAR studies [126], especially when combined with a thorough process of hyper-parameter tuning. However, in contrast to CNNs, which exhibit a good classification performance with the use of raw data, the results reported in [126] indicate ANNs should rather be fed with a small feature set of hand-crafted features than with raw data, with this implying the feature extraction efficiency of ANNs on raw time series is limited as compared to that of CNNs.

2.4 Discussion and Research Opportunity

From the analysis of the literature, it can be concluded that wearable devices show significant advantages as compared to alternative sensing solutions. Wearable solutions overcome the occlusion issues and privacy concerns of systems employing video cameras in a home environment [140]. Motion wearable sensors provide more intrinsic information about the subjects than systems using ambient sensors. Although ambient systems have shown good results at detecting simple activities such as sleeping or toileting, those results are significantly worsened when attempting the recognition of complex activities like eating [141].

Even though increasing efforts and subsequent achievements are being made in the field of HAR with the use of wearable sensors, most of these are being directed towards fitness applications. Also, the discrepancies found amongst the different studies in the field, suggest many of the crucial aspects for activity recognition depend on the activity set studied. Many attempts have been made to recognise quasi-periodic activities. However, there are still many open challenges in recognition of quotidian activities, especially of those composed of sparsely occurring gestures such as food and fluid intake.

The selection of the sensor/s placement is key to achieve successful activity recognition rates. However, as shown in the literature, there is still some debate as to what is the optimal placement for HAR systems. The analysis of previous work suggests the sensor placement should be decided based on the application itself. With regards to the work in this thesis, the active role of the arm, and especially of the hand, on the performance of the target activities, the literature

suggest that the wrist may potentially be the optimal placement for the sensing device.

Another crucial aspect to be taken into consideration is the overuse of power consumption since sensing devices are to be potentially employed as "all-day" monitoring mechanisms. This suggests that the employment of sensors should be considered carefully so that only sensors which contribute significantly to higher recognition rates are incorporated as a means of data collection. In this sense, tri-axial accelerometers are the preferred option among the different works in the field. The analysis of the literature also indicates tri-axial accelerometers may be employed alongside tri-axial gyroscopes in HAR applications. However, this should not be done blindly, since the power consumption of tri-axial gyroscopes is approximately ten times higher than that of tri-axial accelerometers. Further to the selection of the sensors, the sampling frequency employed for the collection of data also plays a crucial role on the power consumption of HAR systems, as an increase in the sampling frequency leads to an increase in the power consumption of the sensors, as well as in the amount of data that needs to be a posteriori processed. However, the issue of signal aliasing, which occurs when the sampling frequency is not at least twice the highest relevant component of the signal, should also be taken into consideration to appropriately select the sampling frequency of the different sensors employed.

The sliding window technique is an effective way of tackling the segmentation of sensory signals for the recognition of quasi-periodic activities. However, adaptive segmentation techniques are preferred when attempting the recognition of sparsely occurring gestures. Statistical features in the time domain and spectral information computed in the frequency domain have achieved good classification performance across numerous studies. Besides, automatically learned features through the use of deep learning techniques, and especially of CNNs, are increasingly adopted in recent work in the field, given their good performance and the advantage they provide in the sense that they are computed without the need for domain-specific knowledge. The classification model to be employed in HAR applications is another open area for discussion. This is mainly due to the variation with regards to the performance across different experiments of the main classification models employed for activity recognition. This variation indicates

the performance of the classification models depends on the specific application and feature set provided.

The above conclusions drawn from the analysis of relevant work in the field of HAR suggest that there are still many open challenges in activity recognition with the use of wearable sensors, especially in activities composed by sporadic occurring gestures. In line with this, the efforts of this thesis are focused on exploring computational solutions to provide alternative approaches to the current state-of-the-art.

Chapter 3

Experimental Pipeline

3.1 Introduction

This section presents a technical overview of the common methods employed for activity and gesture recognition for the justification of their exploration throughout the work undertaken in this thesis. Experiment-specific methods will be presented at the corresponding experimental chapters. Besides, the main datasets utilised in this thesis are also explained in this chapter. An activity/gesture recognition system typically involves five main steps, namely data collection, signal processing, feature extraction, gesture/activity classification and model evaluation. In view of this, the remainder of this chapter is organised as depicted in Figure 3.1.

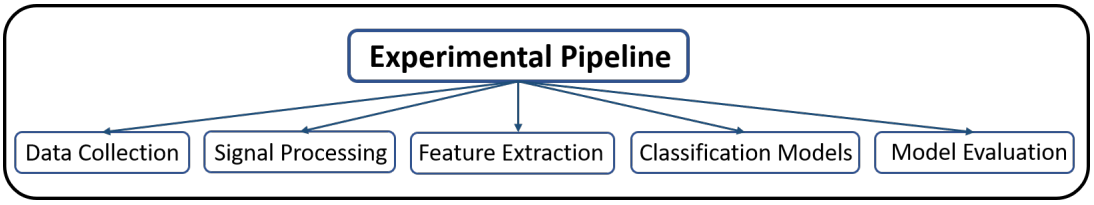

Figure 3.1: Block diagram representing the chapter organisation.

3.2 Data Collection

Data collection in the context of activity and gesture recognition is the process of gathering data by the use of different sensory devices to ultimately develop classification models to predict the activity or gesture being performed at a given time sequence. To do so, a range of experimental participants perform a series of activities while wearing one or various sensory devices. These recordings are annotated (labelled) so that the resultant wrist motion signals at any given time are associated with a specific class or label (activity or gesture in this case).

3.2.1 Sensory Device

A wrist-worn device composed of, among other sensors, a tri-axial accelerometer and a tri-axial gyroscope has been employed throughout the different experimental work undertaken in this thesis. As outlined in Section 2.3.1, accelerometers have been almost unanimously employed by previous work in the field [14, 15, 17, 68, 77, 79, 108, 127, 131, 142, 143] given the good balance between the performance shown by acceleration-based classification models and the power consumption demanded for the collection of the data. Besides, as demonstrated in [78], gyroscopes can in some circumstances improve the recognition performance achieved by classification models based solely on accelerometer data. However, the power consumption of gyroscopes is approximately ten times higher than that of accelerometers [9]. Based on this, the work in this thesis employs a tri-axial accelerometer, with occasional use of a tri-axial gyroscope, as a means of data collection for the recognition of the activities of interest.

The selection of the wrist as the location for the sensory device is based on two distinct factors. First, as discussed in Section 2.3.2, there exist major discrepancies between different studies aiming at the definition of optimal sensor placement for activity recognition, leading to a common tendency within the field for determining the sensor placement based on the application itself. In other words, the sensor placement is typically based on the nature of the target activities. In this case, the nature of the target activities 'Teeth brushing', 'Hands Washing', 'Eating' and 'Drinking', suggest that the wrist is potentially the most suitable sensor placement, given the higher number of degrees of freedom as compared

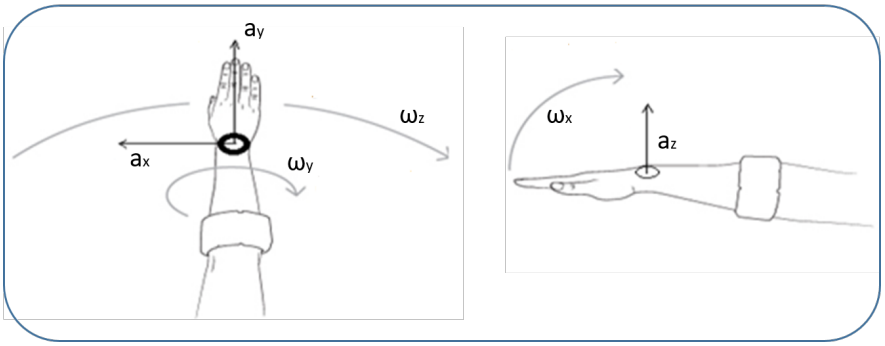

Figure 3.2: Visual representation of the signals measured by a wrist-worn tri-axial accelerometer. The data provided by the accelerometer is composed of three different time series a_x, a_y, a_z, which correspond to the medio-lateral, vertical and antero-posterior acceleration inputs respectively. The data provided by the gyroscope is composed of $\omega_x, \omega_y, \omega_z$, which correspond to the angular velocities around the x, y and z axes respectively.

to other parts of the arm and its evident connection with the target activities. Besides, the wrist is a natural place for instrumentation which minimises undesired obtrusiveness while increasing social acceptance due to the resemblance of a wrist-worn device to a common watch.

At present, motion sensors come in the form of Micro-Electro Mechanical Systems (MEMS) which embody both mechanical and electronic components of very small size. These range from a few micrometres to one millimetre. In particular, the datasets used for the work undertaken throughout this thesis were collected with a Mbientlab Meta Motion R [144], which comprises a BMI160 Inertial Measurement Unit (IMU) consisting of a state-of-the-art 3-axis low-g accelerometer and a low power 3-axis gyroscope both with 16bit resolution and a sample rate of up to 800Hz.

A detailed description of the working principle of tri-axial accelerometers and tri-axial gyroscopes are provided in Sections 3.2.1.1 and 3.2.1.2 respectively.

3.2.1.1 Tri-axial Accelerometer

A tri-axial accelerometer is a sensing device which measures linear acceleration along three axes, namely x, y and z. Given that during the undertaken experimental work, the device is worn on the wrist, the accelerometer, therefore,

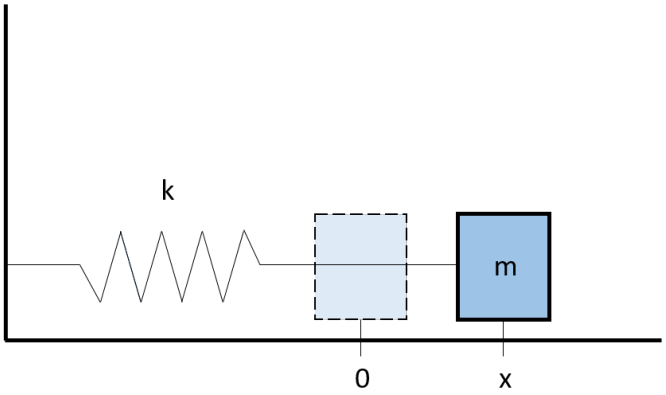

Figure 3.3: Visual representation of Hooke's law, where the dashed square represents the mass at its original position, and the solid square represents the same mass after a displacement x has occurred.

measures the linear motion of the wrist as this moves about in space. This can be visualised in Figure 3.2.

The working principle of an accelerometer is based on the Hooke's law (see Figure 3.3), which states that for relatively small displacements of an object (a mass), such displacement is directly proportional to the force or load that causes the displacement itself. Mathematically, this can be expressed as:

$$F = -kx \tag{3.1}$$

where F is the force that causes the displacement, k is a constant factor characteristic of the spring and x is the displacement of the mass from its equilibrium position.

Besides, Newton's Second Law states that the acceleration of an object is directly proportional to the magnitude of the force applied to it and inversely proportional to the mass m of the object. Assuming that the mass m remains constant, this can be mathematically expressed as:

$$a = \frac{F}{m} \tag{3.2}$$

where a is the acceleration of the object, F is the force applied to the object, and m is the mass of the object.

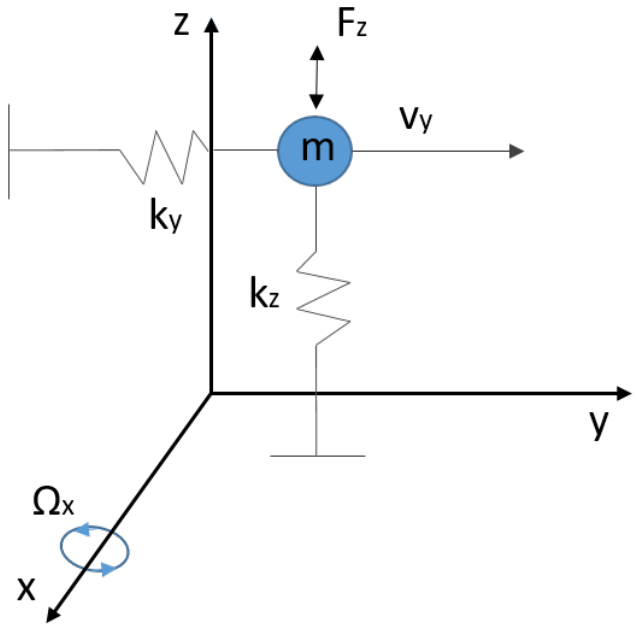

Figure 3.4: Visual representation of the Coriolis effect on a mass-spring model of a MEMS gyroscope.

3.2.1.2 Tri-axial Gyroscope

A tri-axial gyroscope is a sensing device which measures angular velocity around the x, y and z axes (see Figure 3.2). These are commonly known as pitch, roll and yaw respectively. MEMS gyroscopes make use of a vibrating mechanical element for detecting the angular velocity of a rotating mass. The measuring process is based on the Coriolis effect, an apparent acceleration proportional and perpendicular to the linear and angular velocities, which is observed in the rotating mass. The working principle of a gyroscope considering a mass m moving along the y-axis with a velocity v is depicted in Figure 3.4.

Given the scenario presented in Figure 3.4, the Coriolis force is calculated as:

$$F_z = |2m\boldsymbol{\Omega} \times \boldsymbol{v}| \tag{3.3}$$

where F_z is the Coriolis force along the z axis in this case, $\boldsymbol{\Omega}$ is the angular velocity and $\boldsymbol{v}$ is the linear velocity of the mass relative to the reference frame.

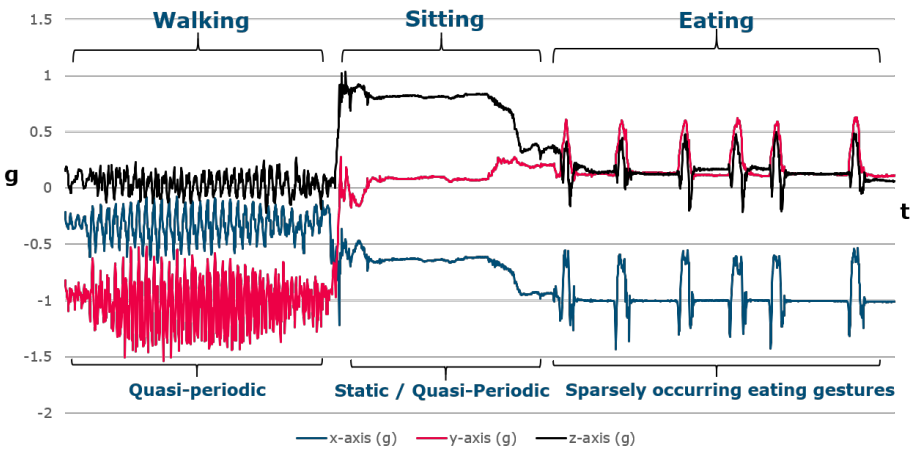

Figure 3.5: Example of accelerometer data from different quotidian activities. The x axis represents time and the y axis represents the corresponding acceleration measured in g.

3.2.2 Datasets

To conduct the research in this thesis, three distinct datasets are used. A brief description of the datasets is provided below. The first, namely Dataset 1, embodies data from a range of quotidian quasi-periodic activities to support the development of techniques to enable the recognition of hygiene-related activities among other quotidian activities. The second dataset, namely Dataset 2, incorporates data from a range of free-living actions as well as from meal periods to support the development of computational solutions to spot and recognise food and fluid intake gestures from continuous data streams. Ultimately, Dataset 3, embodies free-living recordings, to enable the development of computational solutions to detect meal periods from the distribution of food intake gestures across time. A pictorial example of the resultant accelerometer signal collected from the performance of various quotidian activities is provided in Figure 3.5. Further details about the three datasets are provided below.

3.2.2.1 Dataset 1 - Recognition of Quotidian Quasi-periodic Activities. Activities

Six subjects, two female and four male, including three undergraduate and three postgraduate students, participated in this research experiment. The participants were asked to perform a set of seven quotidian daily activities while wearing the employed IMU [144] on their dominant hand. The experiment was run at the Crime Scene Training Facility at Nottingham Trent University, Clifton Campus. The activities performed are listed below:

1. Hand Washing,
2. Teeth Brushing,
3. Standing,
4. Sitting,
5. Picking up an object from the floor[1],
6. Walking Upstairs,
7. Walking Downstairs.

No instructions were given as to how to perform the activities. This ensured the participants could perform the activities naturally, therefore providing reality to the resultant data set. To avoid undesired aliasing effect in the collected signals, a sampling frequency of 100 Hz was used.

3.2.2.2 Dataset 2. Spotting and Recognition of Eating and Drinking Gestures

This experiment embodied six volunteers, five male and one female (five postgraduate students and a senior member of staff), having a meal which included crisps, soup, chicken breast and cake. Given the food provided, the experiment included the use of diverse utensils. Moreover, the utensils provided differed

[1]This action was repeatedly performed, thus becoming a quasi-periodic activity.

between different participants (i.e. various participants used a mug to drink water while others used a glass), therefore incorporating inter-utensil variability. Furthermore, one left-handed person took part in the experiment, thus adding extra variability to the dataset. The resultant data set embodied the following gestures (labels):

1. Null (irrelevant gestures not related to eating or drinking),
2. Drinking (using a glass or mug to drink water),
3. Hand (Using the hand to take a bite of crisps or cake),
4. Spoon (Using the spoon to eat soup),
5. Fork (Using the fork to eat chicken).

As with the previous data set, the participants were asked to wear the IMU on their dominant hand, and no instructions were given as to how to carry out any of the actions during the recordings. Before the meal took place, the participants were asked to act freely around the house for an unlimited period of time. For this experiment, a sampling frequency of 25 Hz was used, given the lower frequency components present in meal intake gestures as compared to those present in other quotidian activities such as teeth brushing. The experiment was carried out at the new Crime Scene Training Facility at Nottingham Trent University, Clifton Campus.

3.2.2.3 Dataset 3. Recognition of Meal Periods

This experiment embodied four male volunteers (three postgraduate students and a senior member of staff), who were asked to wear the IMU on their dominant hand before, during and after their meals (breakfast, lunch and dinner) for various days. The participants were trained on how to use the sensory device and they were asked to annotate the beginning and the end of each meal with a minute precision. The purpose of this experiment was to collect sufficient pre-meal, during a meal and post-meal wrist motion data to identify their meal periods along the day. Consequently, the resultant data set embodied the two following classes:

1. Null (everything but meal periods),

2. Meal Period.

As in Dataset 2, a sampling frequency of 25 Hz was used for the recordings. As per the settings, the experiments were run under free-living conditions. A total of 41 meals with an average duration of 9.02 minutes were reported by the participants of the experiment. The total duration of the recordings is 3774 minutes $\approx$ 63 hours.

3.2.2.4 Datasets Remarks and Inter-Subject Variability

The work in this thesis has been primarily based on the use of the above mentioned datasets. Chapter 4 is based on a dataset composed of seven different quotidian quasi-periodic activities (Dataset 1). Chapter 5 and Chapter 6 exploit a dataset with different embedded food and drink intake gestures with the aim of developing computational solutions to accurately spot and recognise the target gestures from continuous data streams (Dataset 2). The work in Chapter 7 is based on longer-term recordings, aiming at the recognition of meal periods (e.g. lunch) based on the distribution of non-annotated eating and drinking gestures across time (Dataset 3).

The experimental setups for the collection of the above datasets were, to the extent possible, designed to incorporate the variability one would expect to encounter in real life. Although Dataset 1 and Dataset 2 were collected under semi-controlled environments, the experimental participants were told to perform the different activities freely. To the extent possible, participants from different cultures, nationalities, sex and age groups were recruited for the different experiments. This was done with the aim of ensuring the activity and gesture recognition systems proposed throughout the work in this thesis had to deal with the expected intra-subject and inter-subject variability present in the performance of the different activities. Ultimately, Dataset 3 was collected under free-living conditions. In this case, the experimental participants were given a sensing device, whereby they were able to collect data under different scenarios (e.g. their respective home environments).

It should be mentioned that the variation in the experimental participants across the different experiments was caused by the different availability of subjects at the time when the experiments were conducted.

3.3 Signal Processing

To remove unwanted components and enrich the collected motion signals for further analysis, various signal processing techniques are employed in HAR work. This section provides an overview of the main techniques employed to smooth, filter and segment accelerometer and gyroscope signals.

3.3.1 Filtering and Smoothing

Motion sensing devices such as the tri-axial accelerometer and the tri-axial gyroscope employed in this project are subject to undesired instrumentation, random and electric/electronic noise. To minimise the impact of such noise on the sensor measurements, various factors are taken into account. First, a digital low pass filter can be used to remove the frequency components above a specific frequency threshold from which human activity is not expected to happen. The filtering of the undesired high frequency components can be achieved by the employment of a Butterworth low pass filter [145] with the desired cut off frequency. Besides, according to the results in [67], median filters can offer a good balance between the computational cost required and the Signal to NoiseRatio (SNR) when used with motion signals.

3.3.2 Gravity vs. Linear Motion

A common step in HAR employing accelerometers is to separate the acceleration caused by the linear motion of the selected body part where the sensory device is placed, from that caused by the gravitational force [14, 15]. The gravitational component is associated with the low frequency component of the signal. In contrast, the acceleration caused by human motion is associated with the high frequency component. Generally, low-pass filters with cut-off frequencies of up to 1Hz are employed to separate these two components. In this work, a Butterworth

filter with a cut-off frequency of 1Hz is employed to isolate the above-mentioned components in Chapter 4. A posteriori, features can be extracted from the resultant two components. This step can be of crucial value to differentiate between dynamic and static activities since the existence of an acceleration in a static position can be further miss-interpreted by the employed classification models. The extraction of the gravity component from the overall accelerometer signal leads to the following resultant signals:

1. Linear acceleration due to wrist motion: a_x, a_y, a_z
2. Gravity component: $a_{g_x}, a_{g_y}, a_{g_z}$

with the computation of the above components, two further time series from which features can be extracted are obtained. This step can be crucial to obtain features regarding the orientation of the wrist using the gravity component.

3.3.3 Computation of Additional Signal Time Series

As explained in the section above, accelerometer signals from each independent axis can be split into two components, namely the gravity component and the linear acceleration component. Further to these, the computation of additional time series from which to extract features can also benefit the ultimate classification performance [146–148]. For instance, the computation of the magnitude, calculated over the tri-dimensional accelerometer and gyroscope signals, can mitigate the device orientation dependency when utilising each independent axis [147]. By doing so, each sensory signal is now represented by 4 different time series, instead of the original three provided by the sensory devices. The incorporation of the jerk (rate at which the acceleration changes) into the set of time series can also contribute to higher classification rates [146]. Thereby, the following time series can be computed for the further feature extraction:

1. Accelerometer signal: $a_x, a_y, a_z, |a|$
2. Linear acceleration due to wrist motion: $a_{x_m}, a_{y_m}, a_{z_m}, |a_m|$
3. Gravity component: $a_{x_g}, a_{y_g}, a_{z_g}, |a_g|$

4. Jerk: $j_x, j_y, j_z, |j|$

5. Angular Velocity: $\omega_x, \omega_y, \omega_z, |\omega|$

where $|a|$ is the magnitude of the corresponding tri-dimensional vector $[a_x, a_y, a_z]$ calculated as:

$$|a| = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{3.4}$$

and the acceleration jerk j at time t is given by:

$$j[t] = \frac{a[t] - a[t-1]}{\Delta t} \tag{3.5}$$

the computation of the above time series enables the extraction of further features which account for the rate of change of the acceleration along the different axes, as well as the extraction of orientation independent features which provide a means of accounting for the magnitude of the acceleration and angular velocity of the wrist as it moves about in space.

3.3.4 Signal Segmentation

Signal segmentation is the process of dividing the collected time series into smaller segments from which a posteriori the feature vector is calculated. As discussed in Section 2.3.5, there are two main ways of segmenting motion signals, namely through artificial segmentation and through adaptive segmentation. These are presented below in the context of the work undertaken in this thesis.

3.3.4.1 Artificial Segmentation

Artificial segmentation has almost unanimously employed by previous work in the field of HAR due to the good performance exhibited across a wide range of experiments concerning quasi-periodic activities [14,15,17,77,111,131]. The term artificial indicates the segmentation of the signals is not dependent on the signals themselves but on different parameters arbitrarily provided. Artificial segmentation is done through the use of sliding windows, whereby the signals are divided into windows of equal length. These windows are delimited by the window length of n, which defines the length of the windows from which a posteriori the features

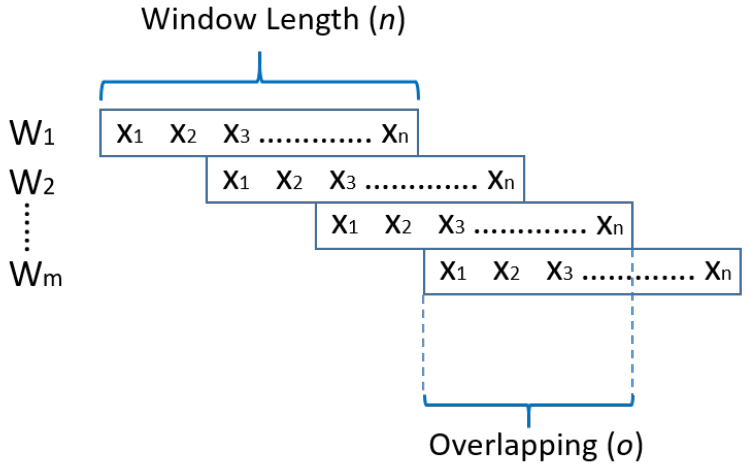

Figure 3.6: Example of the sliding window segmentation technique with a window length n and an overlapping o.

are calculated, and the overlapping o, which defines the overlapping between consecutive windows. The working principle of the sliding window technique with a window length n and an overlapping o is depicted in Figure 3.6.

3.3.4.2 Adaptive Segmentation

As suggested by the work in [120], adaptive segmentation techniques can outperform artificial segmentation techniques like the commonly employed sliding windows presented above. Especially, adaptive segmentation can be useful when tackling the spotting of sporadic gestures with different lengths such as food and drink intake gestures. Contrary to continuous quasi-periodic activities such as walking or teeth brushing, activities like eating and drinking are composed of sparsely occurring gestures embedded in continuous data streams. Motivated by the above, an adaptive segmentation technique is proposed to tackle the experimental work in this thesis with regards to the recognition of eating and drinking gestures. This makes use of characteristics of the signals themselves to identify potential segments of interest while filtering out unwanted segments of the signal which are not believed to be related to intake gestures. Two main constraints are identified on the segmentation of eating and drinking gestures. First, an eating or a drinking gesture can exhibit different lengths in time (*i.e.* a person may drink

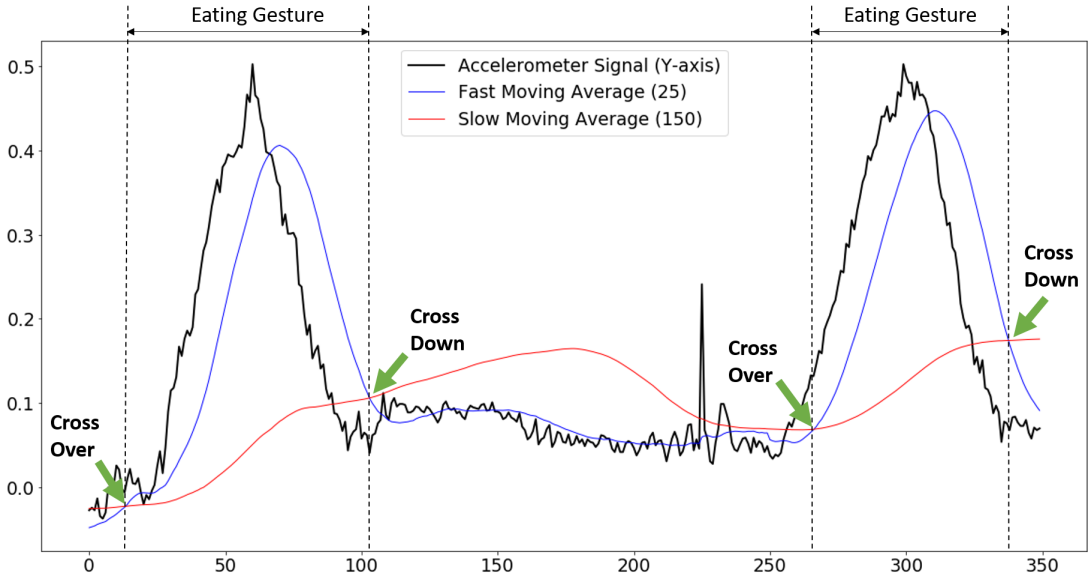

Figure 3.7: Crossings-based adaptive segmentation technique applied to a sample signal with two consecutive eating gestures.

for 1 second but may also do so for 10 seconds). This implies the segmentation has to adapt to such variability to extract the fundamental characteristics of each gesture. Second, segments need to be adjusted as new incoming data is received.

The crossings of two moving averages are explored to determine the potential segments containing eating or drinking gesture. Given its functionality, the technique is referred to as Crossings-based Adaptive Segmentation Technique (CAST). The intuition behind the CAST is the sequence of hand motions involved in an eating or a drinking gesture. First, the corresponding tool (*i.e.* a glass) is taken to the mouth. This is followed by a movement of the hand back to the rest position. Such a sequence of motions leads to a rapid increase on the fast moving average when food or a drink are taken to the mouth, crossing over the slow moving average. A hand movement to the rest position follows, producing a rapid decrease on the fast moving average and the consequent cross down of the slow moving average. This is illustrated in Figure 3.7, where the segmentation of two consecutive eating gestures using the CAST is shown.

The CAST can be explained as follows. Consider a signal $y[t]$. The moving

average $\bar{y}[t]$ of $y[t]$ is defined as:

$$\bar{y}[t] = \frac{1}{n} \sum_{i=0}^{n-1} y[t-i] \tag{3.6}$$

where n is the number of data points over which the moving average is calculated. Two moving averages $\bar{y}_1[t]$ and $\bar{y}_2[t]$ are calculated over the intervals T_1 and T_2 respectively, such that $T_2 > T_1$. If $y[t]$ increases, the CAST moving average $\bar{y}_1[t]$ will react faster to that increase on $y[t]$. Likewise, a faster reaction is also observed when a decrease is seen on $y[t]$.

Provided the higher power consumption of gyroscopes as compared to that of accelerometers, and the prospective use of HAR systems for all-day monitoring applications, the accelerometer signal is explored as a means of the segmentation of the signal.

The values for T_1 and T_2, as well as the accelerometer axis over which the moving averages are calculated, are experimentally evaluated by testing the following values for T_1 and T_2:

$\overline{T_1}$: [25, 50, 75, 100]

$\overline{T_2}$: [50, 100, 150, 200]

Given that more computational intensive tools are to be applied after the segmentation step, T_1 and T_2 are selected based on the gesture spotting recall. In other words, the aim of the segmentation technique is to maximise the true positives so that gestures of interest are not missed at the segmentation stage. The experiments show the optimal values for T_1 and T_2 are $n = 25$ and $n = 150$ respectively when used with the y-axis signal of the accelerometer. With these values, a 100% spotting recall is achieved. Considering a sampling frequency of 25 Hz, $\bar{y}_1[t]$ and $\bar{y}_2[t]$ are therefore the moving averages of the acceleration on the y-axis over 1 second and 6 seconds respectively.

Overall, the CAST overcomes the challenges exposed at the beginning of this section. First, it adapts to the nature of the signal, since both moving averages $\bar{y}_1[t]$ and $\bar{y}_2[t]$ react in consonance with the changes on $y[t]$. Second, it deals with different length of gestures successfully. For instance, in a long drinking gesture,

the decrease in the fast moving average $\bar{y}_1[t]$ after the glass has been taken to the mouth is slower than in a short gesture, since the hand movement that causes the decrease in $y[t]$ and therefore in $\bar{y}_1[t]$ is deferred. Third, CAST can be used real-time since it adapts to new incoming data adjusting the moving averages accordingly.

3.4 Feature Extraction

The computation of efficient representation of data in the form of a feature set is a crucial aspect in wearable HAR. In this process, data abstraction is computed upon each segment of the signal, so that the new representation is more relevant to the activity or gesture associated with the segment. The use of hand-crafted features has shown good recognition performance across numerous studies. The term hand-crafted means the features are calculated by leveraging specific domain knowledge. In the context of activity recognition with the use of wearable sensors, statistical features computed in the time domain and spectral-focused features computed in the frequency domain are widely employed. Based on the different works reviewed in Section 2.3.6, common hand-crafted features across relevant HAR work are presented below:

- Mean:

$$\overline{a} = \frac{1}{n} \sum_{t=0}^{n} a_t \tag{3.7}$$

  where a_t is the acceleration at time t and n is the window length expressed as the number of samples.

- Standard Deviation:

$$\sigma = \sqrt{\frac{\sum_{t=0}^{n} (a_t - \overline{a})^2}{n-1}} \tag{3.8}$$

  where a_t is the acceleration at time t, n is the window length expressed as the number of samples, and $\overline{a}$ the mean acceleration of the corresponding window.

- Signal Magnitude Area:

$$sma = \frac{1}{n}\int_{0}^{T=n}(|a_{x_t} - \overline{a_x}| + |a_{y_t} - \overline{a_y}| + |a_{z_t} - \overline{a_z}|)dt \tag{3.9}$$

where $a_{x_t}, a_{y_t}, a_{z_t}$ are the acceleration at time t on the x, y and z axes respectively, n is the window length expressed as a number of samples, and $\overline{a_x}$, $\overline{a_y}$, $\overline{a_z}$ the mean acceleration on the corresponding axis in the corresponding window.

- Signal Entropy:

$$H(a) = \sum_{t=0}^{Wl} |a_t - \overline{a}| \log_{10} |a_t - \overline{a}| \tag{3.10}$$

where a_t is the acceleration at time t, n is the window length expressed as the number of samples and $\overline{a}$ the mean acceleration in the corresponding window.

- Correlation:

$$r_{xy} = \frac{Cov(a_x, a_y)}{\sigma(a_x)\sigma(a_y)} \tag{3.11}$$

where $Cov(a_x, a_y)$ is the covariance of the acceleration on the axes x and y, and $\sigma(a_x) and \sigma(a_y)$ are the standard deviation for the acceleration on the axes x and y respectively.

- Skewness:

$$\gamma_1 = \frac{\frac{1}{n}\sum_{t=0}^{n}(a_t - \overline{a})^3}{(\sigma(a))^3} \tag{3.12}$$

where a_t is the acceleration at time t, n is the window length expressed as the number of samples and, $\overline{a}$ and $\sigma(a)$ are the mean acceleration and the standard deviation in the corresponding window respectively.

- Kurtosis:

$$\beta_2 = \frac{\frac{1}{n}\sum_{t=0}^{n}(a_t - \overline{a})^4}{(\sigma(a))^4} \tag{3.13}$$

where a_t is the acceleration at time t, n is the window length expressed as

the number of samples and, $\overline{a}$ and $\sigma(a)$ are the mean acceleration and the standard deviation in the corresponding window respectively.

- Root Mean Square

$$RMS = \sqrt{\frac{1}{n}\sum_{t=0}^{n}(a_t)^2} \tag{3.14}$$

where a_t is the acceleration at time t and n is the window length expressed as the number of samples.

The transformation from the time domain to the frequency domain has been computed using the Fast Fourier Transform:

$$A(k) = \sum_{t=0}^{n-1} a_t e^{(-i2\pi kt/n)} \tag{3.15}$$

where a_t is the acceleration at time t and n is the window length expressed as the number of samples, A(k) is the sequence of n complex-valued numbers given the sequence of data a(t).

- Energy

$$E = \frac{\sum_{k=1}^{n} |a_k|^2}{n} \tag{3.16}$$

where a1, a2, ... a_n are the FFT components of the corresponding window of length n.

3.5 Classification Models

The last stage in an activity or gesture recognition system is the training and testing of a machine learning classification model. Once trained, the classification model is used to predict the categorical class of new incoming instances. That is, given the set of features calculated over a time series segment, the classification model predicts the class (in this case, the activity or the gesture) of that segment. Therefore, the task of the predictive model is that of approximating a function f

from input variables X to a discrete output variable y. This section presents the most common supervised classification algorithms employed throughout previous work in the field, namely K-Nearest Neighbours (KNN), Support Vector Machine (SVM) and Random Forest (RF). Given their good performance across different studies (see Section 2.3.7), these are widely employed throughout the different experimental chapters of this thesis.

3.5.1 K-Nearest Neighbours

The KNN is a non-parametric instance-based supervised classification model which requires no learning process. Instead, through 'lazy learning', the class $\hat{f}(x_q)$ of a new incoming instance x_q is predicted based on the most common class among its k nearest neighbours estimated by their Euclidean distance to the current instance (see Figure 3.8). The KNN classification model can be defined as follows:

Given an instance x_q to be classified and the set of training samples $(x, f(x))$, let $x_1...x_k$ denote the k instances from the training examples nearest to x_q. The predicted class of x_q, denoted by $\hat{f}(x_q)$ is given by:

$$\hat{f}(x_q) \leftarrow \arg\max_{c \in C} \sum_{i=1}^{k} \delta(c, f(x_i)) \tag{3.17}$$

where $\delta(a, b)$=1 if a=b and $\delta(a, b) = 0$ otherwise.

3.5.2 Support Vector Machine

Support Vector Machines (SVM) is a classification model based on the margin maximisation principle. That is, during the training phase, a separating hyperplane that maximises the distance between the instances corresponding to different classes is estimated, with the support vectors being the closest points to such separating hyperplane. Commonly, before the estimation of the optimal separating hyperplane, a kernel function, typically a Radial Basis Function (RBF), is used to map the input space into a higher dimensional space where the distance between the instances of the different classes can be further maximised. With this, a linear classifier can be used to solve a non-linear problem. Considering a

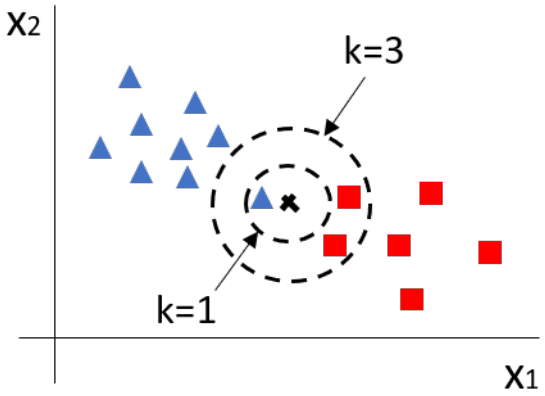

Figure 3.8: KNN classification model with k=1 and k=3 in a binary classification problem.

binary classification problem, the separating hyperplane can be defined as:

$$\mathbf{w}^{\intercal}\mathbf{x} + b = 0 \tag{3.18}$$

where $\mathbf{w}$ is a weight vector, $\mathbf{x}$ is the input feature vector and b is the bias.

The working principle of an SVM classification model on a binary classification problem is depicted in Figure 3.9. However, activity recognition is normally tackled as a multi-class classification problem, since generally, several activities are included in the activity sets. In this case, a One-vs-Rest classification strategy can be employed. With this, an N-class classification problem is tackled through the use of N binary classification models, where each binary classification model is aimed at estimating an optimal hyperplane that separates a specific class from the rest of the classes.

3.5.3 Random Forest

Random Forest (RF) is an ensemble supervised classification model based on the voting of a large number of randomly created and merged decision trees, where each of the decision trees can be thought of as a series of nodes with yes/no questions regarding the features that compose the feature vector, leading to a predicted class. At each node, decision trees search through the features for the

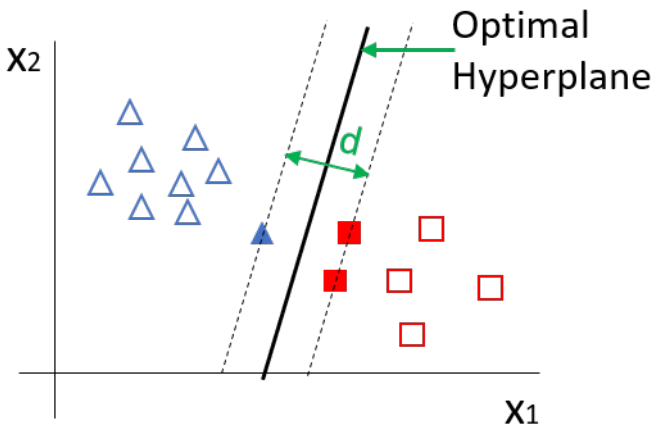

Figure 3.9: SVM classification model in a binary classification problem, where d is the maximum margin between the 2 classes.

value from which the split is made that results in the greatest reduction in the Gini Impurity. The Gini Impurity is given by:

$$I_G(n) = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{3.19}$$

where C is the total number of classes in the classification problem and p_i is the fraction of samples labelled with class i.

As shown in Figure 3.10, Random Forest makes use of n decision trees to make a prediction based on the majority voting across the trees. Besides, Random Forest de-correlates the different decision trees that form the forest by considering only a selection of features at each tree node split.

3.6 Model Evaluation

This section presents the evaluation measures and cross-validation strategies employed by the different activity and gesture recognition frameworks developed throughout this thesis.

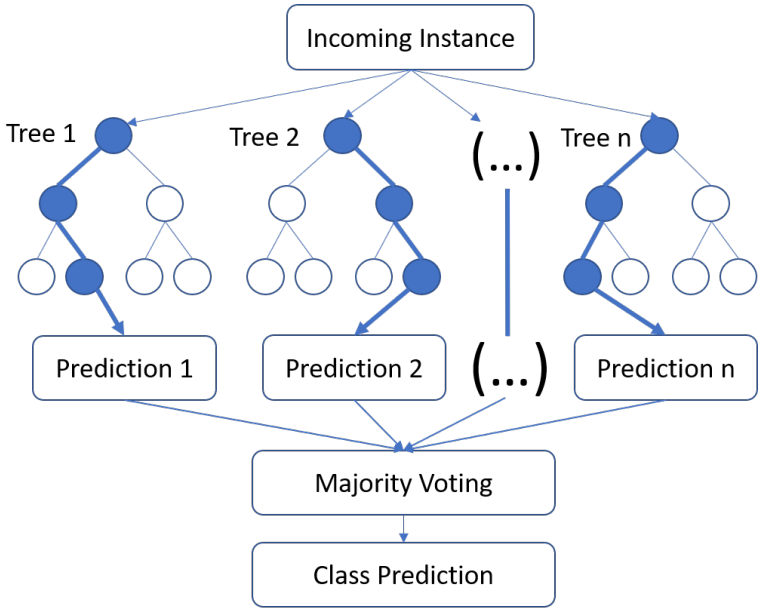

Figure 3.10: Random Forest working principle for predicting new incoming instances.

3.6.1 Evaluation Measures

To evaluate the performance of the classification models, three evaluation measures, namely classification accuracy, classification recall and classification precision, have been widely employed.

The classification accuracy provides an overview of the performance of the evaluated classification model calculated as the ratio of the number of correct predictions to the total number of predicted samples as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.20}$$

Alongside the classification accuracy, the classification recall and classification precision are used to mitigate the poor generalisation performance of the classification accuracy at evaluating unbalanced datasets. The classification recall, also known as the true positive ratio, is calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{3.21}$$

The classification precision is expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{3.22}$$

It should be noted that, in this thesis, the above measures are provided for each class within the classification problem, so that intuition on how the classification model performs on each specific activity or gesture can be obtained. These are referred to as per-class measures. On top of these, the average per-class accuracy, precision and recall are used to evaluate the overall performances of the implemented systems.

3.6.2 Evaluation Strategy

Further to selecting appropriate measures for the specific classification problems, the adoption of an adequate cross-validation strategy is crucial to guarantee the robustness of the results achieved. K-fold and Leave-One-Out cross-validation strategies are widely employed for evaluating HAR systems. By the employment of K-fold, the dataset is shuffled and split into K folds or groups. The model is then evaluated K times, where each time a different group is used as the test set and the remaining groups as the training set. A posteriori the average value across the K runs is provided as the final results. The Leave-One-Out cross-validation strategy follows a similar working principle. However, the split of the dataset is based on the number of participants, whereby each participant's instances are used once as the test set. Figure 3.11 illustrates the working principle of a 5-fold cross-validation strategy.

In the context of the experimental work carried out in this thesis, three different strategies are employed as follows. In Chapter 4, a 3-fold cross-validation where each of the three runs is divided into three sets, namely training, validation and test sets. The intuition behind the employment of this strategy comes from the need to have different sets for evaluating the robustness of the multi-refinement approach proposed in the chapter. In this context, the validation set is used to find refinement opportunities whereas the test set is used to test the improvement of such refinement.

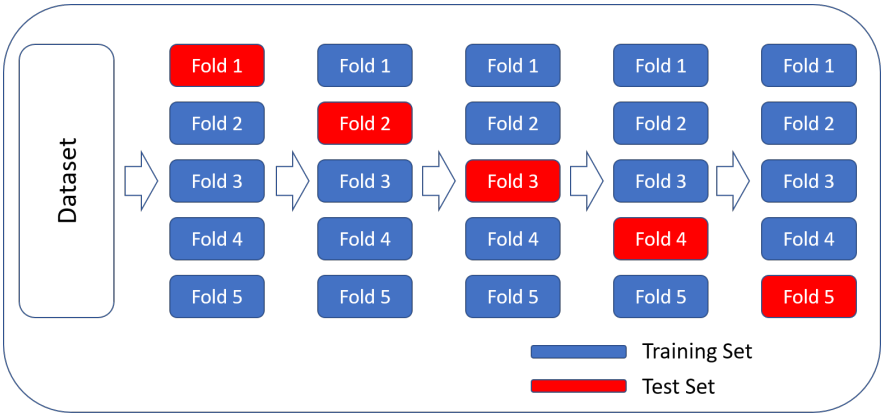

Figure 3.11: 5-fold cross-validation strategy

In Chapter 5 and Chapter 6 a Leave-One-Subject-Out strategy, whereby each of the experimental subject's data is used once as the test set, is adopted. The adoption of this strategy is mainly motivated by two main factors. First, the adaptive segmentation technique employed results in a reduced sample set as compared to problems tackled using artificial segmentation techniques where all the signal frames are considered for classification. Second, the dataset collected for these experiments includes data from a left-handed person, therefore potentially increasing the variations between participants. While a single train-test split may be used to evaluate the performance of classification models on larger datasets, the Leave-One-Subject-Out cross validation strategy is in this case considered the most appropriate strategy to evaluate the robustness of the gesture recognition approaches proposed in these two chapters given the reduced size of the dataset. In addition, the adoption of this strategy allows for the evaluation of the performance of the different classification models on a left-handed individual when these are trained on data from right-handed individuals only.

Lastly, in Chapter 7 a 10-fold cross validation strategy is employed. The adoption of this strategy is justified by the very limited data samples to be fed into the classification model, since samples are given in the form of whole meal periods rather than as short frames or segments of an activity. Given the limited size of the dataset, it is crucial to provide sufficient training data to the model for it to be able to generalise competently on unseen data. In this context, a 10-fold cross validation is considered an appropriate solution to evaluate the robustness of the

model while mitigating the impact of the limited size of the dataset.

3.7 Conclusions

This section has presented common methods employed in human activity recognition work, including signal processing techniques to reduce noise and to extract further information in the form of additional time series from those collected by the tri-axial accelerometer, signal segmentation techniques to break the sensory signals into smaller segments and to filter out unwanted segments of the signals, feature extraction of domain-specific hand-crafted features, the working principle of a range of classification models, as well as cross-validation strategies to evaluate the performance and robustness of the classification models. These methods, alongside further methods developed in this work with the aim of mitigating some of the limitations found in the state-of-the-art in the field, are explored in the following chapters.

Chapter 4

Recognition of Quotidian Quasi-Periodic Activities

4.1 Introduction

Many attempts have been made to develop HAR systems using wearable sensors. However, most of these efforts have been directed to the recognition of fitness-related activities such as walking or running [14–17]. By contrast, limited research has explored the recognition of hygiene-related activities [149, 150]. As suggested in [102], gaining insights into the daily behaviour of an older adult living independently can be valuable information for clinicians, who would be able to react in consonance with such information. Regarding hygiene issues, research reported in [151] suggests a big proportion of elderly individuals need assistance with their daily oral hygiene, however, only a very small percentage of those actually receives it. Hand washing is another crucial factor in personal hygiene, being widely recognised as a critical infection control mechanism [152]. As evidenced by the reiterative recommendations made by medical institutions during the current COVID-19 (Coronavirus Disease 2019) pandemic, preventing any form of infection through the maintenance of adequate hand hygiene habits can be critical for the health of individuals, especially for those in advanced age with weakened immune systems.

Motivated by the above, this chapter explores computational solutions to

recognise hygiene-related activities, namely teeth brushing and hand washing, among other common quotidian activities such as sitting, standing and walking-related activities. As mentioned in Section 2.3.6, the use of hand-crafted features have shown good classification performance across different studies concerning the recognition of continuous quasi-periodic activities. However, it is argued that activities differ diversely from each other, leading to the judgement that a specific set of features may be informative to classify a specific set of activities, but such informativeness should not necessarily be extended to a different activity set. In this context, a multi-level refinement approach, through which the selection of features is optimised for those activities which show lower classification performances as compared to that of the overall activity recognition system, is proposed. With this approach, after the classification takes place, information is extracted from the confusion matrix to focus the computational efforts on those activities with lower classification performances.

The remainder of this chapter is structured as follows: Section 4.2 presents a review of work on activity recognition. Section 4.3 presents the motivation behind the work in this chapter. Section 4.4 describes the methodology followed for the recognition of hygiene-related activities among other quotidian activities. Section 4.5 presents the results achieved. Ultimately, Section 4.6 draws the conclusions from this chapter based on the results obtained.

4.2 Review of Work on Activity Recognition

Numerous research works concerning activity recognition with the use of wearable sensors have been proposed in the last years, with these varying the type and number of sensors, their placement, the activities to be tracked, the pre-processing techniques, the feature extraction and selection methods, the classification approaches as well as the research purpose itself. This section describes some of the undertaken studies to put activity recognition into context. For instance, the work in [14] makes use of a smart-phone with a single tri-axial accelerometer to evaluate the activity recognition performance on a set of six activities, carrying the phone in the pocket and carrying the phone in hand. A maximum accuracy of 91.15% is achieved on the 'in-hand' experiment using a combination of differ-

ent classifiers. In [15], a group of five activities is studied using a single tri-axial accelerometer worn on the chest, achieving a maximum accuracy of 94% using a Random Forest classifier. In [77] data from a tri-axial accelerometer worn near the pelvic region is used to study eight different activities, obtaining a classification accuracy of over 99% combining different classifiers by plurality voting. The authors in [143] propose a Gaussian Continuous Hidden Markov Model (cHMM)-based sequential classifier using data from five bi-axial accelerometers to classify seven different activities, achieving a maximum accuracy of 98.4%. In [13], a circuit composed of eight fitness activities is studied using a system embodying a tri-axial accelerometer and a tri-axial gyroscope embedded in a vest and located at the upper trunk of the experimental participants, achieving a maximum accuracy of 92% using a Logistic Model Tree (LMT). The work in [17] makes use of data from a tri-axial accelerometer worn on the chest to study seven activities. Their approach includes an ensemble feature selection, which combined with a Random Forest classifier, obtains a maximum accuracy of 88%. In [102] data from two accelerometers worn on the sternum and the right thigh is used to classify sitting, standing and stepping with a classification accuracy of 98%. In [135], transition movements including sit-to-stand, stand-to-sit, lie-to-stand and stand-to-lie are studied using tri-axial accelerometer data, achieving an average classification accuracy of 84% with a KNN classifier.

Concerning hygiene-related activities, the work in [149] makes use of data from a wrist-worn bi-axial accelerometer to propose Gaussian Mixture Models (GMMs) combined with a majority voting system for the recognition of three early morning activities, including face washing, teeth brushing and shaving, achieving a classification rate of 83.9% with the use of a 16-dimensional hand-crafted feature vector. Using the same sensor but five more subjects, the work in [153] achieves a classification accuracy of 90.1% on the same activity set. In [150], teeth brushing is studied along with other seven quotidian activities with the use of data from a single wrist-worn tri-axial accelerometer. In this study, a 24-dimensional hand-crafted feature vector is used alongside a range of classification models, including KNN and ANNs. The results report the achievement of an average recognition accuracy of 95.24%.

In summary, a large number of research studies concerning activity recog-

nition with the use of wearable devices can be identified within the literature. Despite the existing differences between studies in terms of the experimental setups, the signal processing techniques and the classification models employed, a common factor can be identified across the different studies by analysing further the results achieved. Such an analysis reveals that some activities are less easily discriminated than others, exhibiting notable differences in terms of class-specific classification performances.

4.3 Motivation

The motivation behind the work in this chapter is to provide unobtrusive means of recognising hygiene-related activities among other quotidian activities, given the crucial impact personal hygiene can have on the health of older individuals and the limited studies found regarding this matter. Besides, the multi-level refinement approach proposed in this chapter is based on the analysis of different HAR research works providing class-specific evaluation measures. As mentioned above, from these studies, a common issue can be identified; some activities are less easily discriminated than others. For instance, authors in [67] struggle to classify 'sit' and 'fall'. In [18], it can be observed that the highest number of false detections occur in two specific activities; 'get up' and 'max-reach'. The authors in [74] find the groups 'spoon' and 'apple' to have considerably lower detection rate than others. In [119], the recognition rate of 'walking upstairs' and 'walking downstairs' is lower as compared to that of other walking-related activities.

4.4 Methods

This section presents the methodology proposed for the recognition of hygiene-related activities. An explanatory diagram with a summary of the different methodology steps can be seen in Figure 4.1.

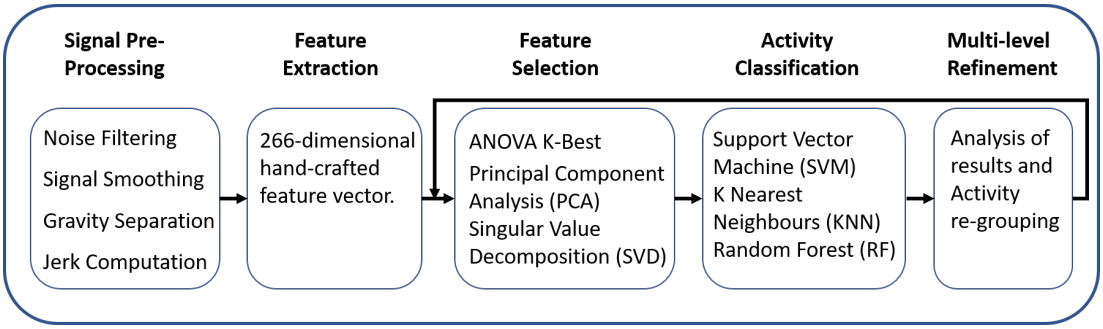

Figure 4.1: Steps of the proposed multi-level refinement approach. Pairs of activities which worsen the performance of the classification model, are grouped together for further inspection.

4.4.1 Experimental Setup

The work in this chapter makes use of the Dataset 1 (see Section 3.2.2.1) with tri-axial accelerometer data. The dataset embodies seven quotidian activities, including teeth brushing and hands washing.

4.4.2 Signal Pre-processing

The accelerometer data is composed of three different time series a_x, a_y, a_z, which correspond to the mediolateral, vertical and anteroposterior acceleration inputs respectively. A fourth time series, namely $|a|$, is computed as the magnitude of the tri-dimensional vector. In addition, the rate of change of the acceleration (jerk) is computed to obtain additional information from the accelerometer signals.

A median filter with a window length $n = 7$ is used for signal smoothing purposes. The frequency components of the signal above 20 Hz are filtered out with the use of a low-pass 20 Hz Butterworth filter. The gravity and motion components are separated through the use of a 1 Hz Butterworth filter.

The above signal processing steps result in a set of time series as follows:

- Accelerometer signal: $a_x, a_y, a_z, |a|$
- Linear acceleration due to wrist motion: $a_{x_m}, a_{y_m}, a_{z_m}, |a_m|$
- Gravity component: $a_{x_g}, a_{y_g}, a_{z_g}, |a_g|$

- Jerk: $j_x, j_y, j_z, |j|$

The segmentation of the signals is performed using sliding windows with a window length of 1 second and a 40% overlapping percentage.

4.4.3 Feature Extraction

A broad range of commonly employed hand-crafted features are computed for the construction of the feature vector (see Section 3.4). Within the time domain, different statistical features were explored. These include measures of central tendency like the mean and Root Mean Square (RMS), measures of statistical dispersion such as range, standard deviation and interquartile range, measures of distribution shape such as kurtosis or skewness, measures of dependence between different axes, such as Pearson's correlation and measures of the magnitude of varying quantity such as the signal magnitude area. On top of the statistical features, the number of peaks in the signal, the number of zero-crossings (number of times the signal crosses the 0 level) and signal entropy are computed to add additional information about the dynamics of the signals. After converting the signal to the frequency domain through the Fast Fourier Transform (FFT), the largest magnitude of the signal spectrum, the index of the spectrum component with the highest magnitude and the energy of the signal are also computed. Except for a few cases where it was not appropriate (*e.g.* correlation), the proposed features were calculated over all the time series exposed in Section 4.4.2. The dimensionality of the resultant feature vector is $n = 266$.

This feature set has been carefully selected to embody informative and discriminative information with regards to a wide array of signal characteristics, such as range, dispersion, central tendency, periodicity, frequency distribution, magnitude and changes in direction. A summary of the feature vector dimensionality, the total number of instances and the class distribution of the classification problem is given in Table 4.1.

Table 4.1: Post-segmentation data summary

Dim.	Inst	Hands. W	Teeth. B	Stand	Sit	P. Object	W. Down	W. Ups
266	8674	25.29%	9.52%	14.15%	18.38%	28.12%	2.32%	2.23%

4.4.4 Feature Selection and Reduction

As stated in [143], when the dimension of feature space is considerably high, learning the parameters for a classifier becomes a difficult and consuming task. In addition, feature selection/reduction can maintain or even increase the discriminative capability of the whole feature set. Three different methods for dimensionality reduction are explored. First, an Analysis of Variance (ANOVA) is conducted, where features are ranked according to their F measure, calculated as the ratio of the variance between classes and the variance within the class. Even though the use of ANOVA on non-normally distributed data can increase the chances of obtaining false positives, the F measure here is only used as a feature ranking mechanism regarding the dissimilarity between classes. After features are ranked, the subset that maximises the classification result is selected. Principal Component Analysis (PCA) and truncated Singular Value Decomposition (SVD) are also explored. These two approaches perform an orthogonal transformation of the data into a new coordinate system where the new coordinates are those which maximise the variance of the data, being the difference between the two approaches that PCA centres the data before computing the singular value decomposition.

4.4.5 Classification

As shown in the description of the dataset in Section 3.2.2.1, seven activities are investigated in this study. The performance of three different classification models, namely K-Nearest Neighbours (KNN), Random Forest (RF) and Support Vector Machine (SVM) using a Radial Basis Kernel (RBS), is evaluated. The optimal classifier is selected during the feature selection stage, along with the optimal number of features/components.

A 3-fold cross-validation method is used to test the robustness of the classification results. Each fold includes three different sets -the training set, the

validation set and the test set. The validation set is used to identify refinement opportunities and the test set to validate the performance improvement during the different refinement steps. In other words, once the model is trained, the test set is used to report the classification performance of the system and the validation set is used to identify those activities with lower classification performances. For clarification purposes, after a random shuffle, 60% of the dataset is used as the training set and the remaining 40% is split into the test set (20%) and the validation set (20%). With this, alongside the 3-fold cross-validation strategy employed, whereby each of the training, test and validation sets are split into three further respective subsets, the algorithm proposed is designed to ensure a competent robustness while dealing with the expected variability of the different actions or activities performed by the different experimental subjects.

4.4.6 Multi-Level Refinement

The proposed multi-level refinement can be defined as an algorithm that aims at optimising the classification accuracy of a group of classes by an improvement on the recognition rate of those classes which lower the classification rate of the whole group. Its implementation is justified by the fact that in a classification problem, a classification accuracy lower than 100% is normally caused by the difficulty to classify specific classes, unless the recognition rate is identical for all the classes, though this is not a common occurrence.

After the activity classification takes place, the confusion matrix is further analysed and activities are compared in pairs. If the classification accuracy between a pair of activities is lower than that on the whole model, those activities are grouped together for refinement. Activities which are found to lower the accuracy of the system due to their misclassification rate with an activity already pertaining to a refinement group are added to that same group; otherwise a new refinement group is constructed with these pair of activities. At this point, the feature selection is optimised for each group selecting the most informative feature set for each of them. This process is repeated until groups of two activities are reached.

The multi-level refinement, therefore, focuses the computational efforts on the

classification of those activities that are more difficult to classify in the first place. A pseudo-code of the multi-level refinement algorithm is shown in Algorithm 1.

Algorithm 1 Multi-level Refinement

```
top:
accuracy ← classification_accuracy
c_m ← confusion matrix
i ← rows confusion matrix
j ← columns confusion matrix
for n_rows do
    for n_columns do
        if (row ≠ column) then
            if ((c_m[j,j]+c_m[i,i]) / (c_m[j,j]+c_m[i,i]+c_m[i,j]+c_m[j,i]) < accuracy) then
                activity_pairs.append[(i, j)]
            end if
        end if
    end for
end for
for activity_pair ∈ activity_pairs do
    for activity ∈ activity_pair do
        if (activity belongs to a group) then (add its pair to the group)
        else(create new group and add both activities)
        end if
    end for
end for
if All activities belong to the same group then
(remove activity with the highest accuracy)
end if
for group ∈ groups do
    Feature Selection
    Run Classifier
    if (grouplength > 2) then
        goto top.
    end if
end for
```

It should be mentioned that the refinement of a specific pair or group of activities does not affect the classification of other classes since only the instances previously classified as belonging to any of the classes in that pair or group, are taken forward for refinement.

4.5 Results

In this section, the experimental results achieved are presented, explained and discussed. This section is divided as follows;

Section 4.5.1 examines the feature selection methods proposed for dimensionality reduction. Section 4.5.2 presents the classification results and the improvement achieved by the multi-level refinement algorithm. Ultimately, Section 4.5.3 discusses the results obtained.

4.5.1 Feature Reduction

ANOVA K-best, PCA and SVD are computed to find out the optimal subset of features/components for the description of the data set. To do so, the performance of the different feature selection methods is examined across all the possible number of features ranging from n=1 to n=266 (whole feature set). These three feature reduction techniques are examined on the three different classifiers proposed; RF, KNN and SVM. The best classification performance (average per-class classification accuracy = 99.15%) is achieved using ANOVA K-Best alongside a Random Forest classifier when n=149, with n being the number of features after being ranked according to their F ratio. The performance of the different feature selection methods with Random Forest across the number of dimensions of the feature vector can be observed in Figure 4.2.

4.5.2 Classification and Refinement

The first step of deploying the multi-level refinement algorithm is to train and evaluate a 7-class classification model representing the 7 studied activities using the train set and test set respectively. The classification, as reported by the test set, resulted in the confusion matrix and classification metrics presented in Figure 4.3 and Table 4.2 respectively.

The average per-class classification accuracy achieved by the model is 99.15%. However, there exist relevant differences in terms of precision and recall between different activities. At this point, the multi-level refinement algorithm is used to identify the activities lowering the performance of the model. This is per-

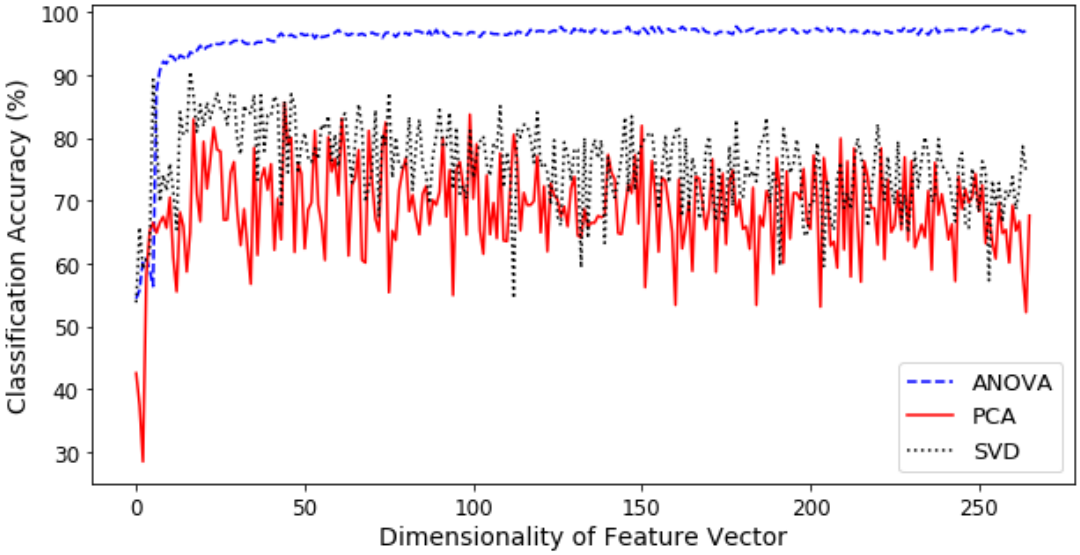

Figure 4.2: Classification performance of the feature selection methods based on a Random Forest classifier.

		Predicted Label							
	Activity	Hands Washing	Teeth Brushing	Standing	Sitting	Picking Object	Walking Downstairs	Walking Upstairs	Total
True Label	Hands Washing	1345	2	0	0	7	0	0	1354
	Teeth Brushing	2	447	1	6	7	0	1	464
	Standing	7	0	721	0	0	0	0	728
	Sitting	2	1	0	981	3	0	0	987
	Picking Object	37	8	0	1	1381	0	0	1427
	Walking Downstairs	0	11	0	0	0	88	26	125
	Walking Upstairs	0	21	0	0	0	17	82	120
	Total	1393	490	722	988	1398	105	109	5045

Figure 4.3: Confusion matrix before refinement using a Random Forest classification model.

formed using the validation set. "Teeth Brushing", "Walking Downstairs" and "Walking Upstairs" were identified to need refinement. For the benefit of reading convenience, this group of activities are referred to as "Group 1".

Taking into account the activities within Group 1, feature selection is per-

formed and a new 3-class model including the outlined set of activities was trained using the training set. The aim of this step is to optimise the feature vector for the classification of the identified set of activities, which was found to have a dimensionality of n=131. Such new 3-class model is now used to reclassify the samples which were previously predicted as to pertain to Group 1. The classification metrics and resultant confusion matrix after the first refinement step, as reported by the test set, can be seen in Table 4.3 and Figure 4.4 respectively.

After the first refinement step, the same process was repeated. In this case, the 3-class classification model alongside the validation set was used to identify activities which needed further refinement. The new set of activities found was formed by the activities "Walking Downstairs" and "Walking Upstairs". This group of activities are referred to as "Group 2". A new 2-class model was trained (using the training set) with an optimised feature vector for the classification of

Table 4.2: Classification metrics of the 7-class model before refinement.

	Accuracy	**Precision**	**Recall**
Hands Washing	98.88%	96.55%	99.34%
Teeth Brushing	99.12%	91.22%	99.55%
Stand	99.84%	99.86%	99.04%
Sit	99.74%	99.29%	99.39%
P. Object	98.77%	98.78%	96.78%
W. Downstairs	98.94%	83.81%	70.40%
W. Upstairs	98.73%	75.23%	68.33%
Average	**99.15%**	**92.11%**	**90.40%**

Table 4.3: Classification metrics after the first refinement step.

	Accuracy	**Precision**	**Recall**
Hands Washing	98.89%	96.55%	99.34%
Teeth Brushing	99.33%	93.32%	99.55%
Stand	99.84%	99.86%	99.04%
Sit	99.74%	99.29%	99.39%
P. Object	98.77%	98.78%	96.78%
W. Downstairs	99.20%	86.84%	79.20%
W. Upstairs	99.08%	82.88%	76.67%
Average	**99.26%**	**93.93%**	**92.85%**

True Label \ Predicted Label	Hands Washing	Teeth Brushing	Standing	Sitting	Picking Object	Walking Downstairs	Walking Upstairs	Total
Hands Washing	1345	2	0	0	7	0	0	1354
Teeth Brushing	2	447	1	6	7	0	1	464
Standing	7	0	721	0	0	0	0	728
Sitting	2	1	0	981	3	0	0	987
Picking Object	37	8	0	1	1381	0	0	1427
Walking Downstairs	0	8	0	0	0	99	18	125
Walking Upstairs	0	13	0	0	0	15	92	120
Total	1393	479	722	988	1398	114	111	5066

Figure 4.4: Confusion matrix after the first refinement step.

True Label \ Predicted Label	Hands Washing	Teeth Brushing	Standing	Sitting	Picking Object	Walking Downstairs	Walking Upstairs	Total
Hands Washing	1345	2	0	0	7	0	0	1354
Teeth Brushing	2	447	1	6	7	0	1	464
Standing	7	0	721	0	0	0	0	728
Sitting	2	1	0	981	3	0	0	987
Picking Object	37	8	0	1	1381	0	0	1427
Walking Downstairs	0	8	0	0	0	105	9	122
Walking Upstairs	0	13	0	0	0	21	89	123
Total	1393	479	722	988	1398	126	99	5069

Figure 4.5: Confusion matrix after the second refinement step.

the activities within Group 2. The dimensionality of the vector, in this case, was n=186. It can be noticed that the number of dimensions has now increased as compared to previous classifications. This may be due to the high similarity in terms of acceleration between walking downstairs and walking upstairs. The samples from the test set previously classified as to pertain to Group 2 were reclassified using the new 2-class model, leading to the classification metrics shown in Table 4.4 and the confusion matrix illustrated in Figure 4.5, as reported by the test set.

Table 4.4: Classification metrics after the second refinement step.

	Accucacy	**Precission**	**Recall**
Hands Washing	98.89%	96.55%	99.34%
Teeth Brushing	99.33%	93.32%	99.55%
Stand	99.84%	99.86%	99.04%
Sit	99.74%	99.29%	99.39%
P. Object	98.77%	98.78%	96.78%
W. Downstairs	99.26%	83.33%	86.07%
W. Upstairs	99.14%	89.90%	72.36%
Average	**99.28%**	**94.43%**	**93.22%**

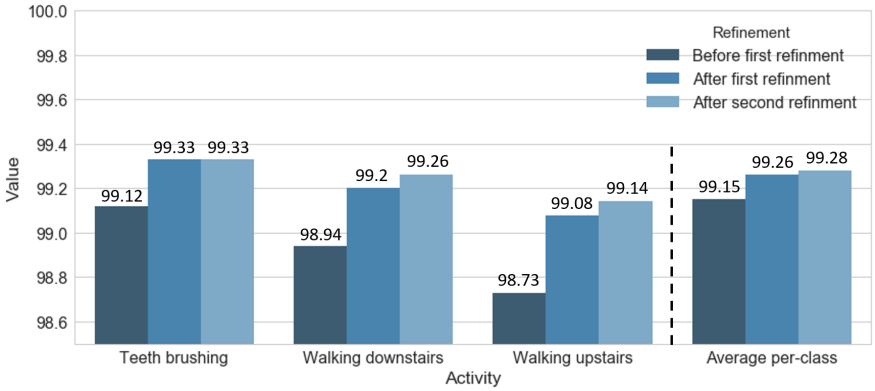

Figure 4.6: Comparison of activity classification accuracy before and after the different refinement steps.

The different refinement steps show an improvement in terms of per-class classification accuracy of the refined classes. This led to a better performance of the multi-level refinement approach proposed as compared to standard state-of-the-art classification approaches. Such improvement can be better visualised in Figure 4.6 where the per-class classification accuracy of the refined activities and the average per-class classification accuracy of the 7-class model across the different refinement steps are shown.

4.5.3 Validation and Discussion

To validate the multi-level refinement algorithm, a test is run on the [109] benchmark data set. The data set contains data collected from 30 volunteers performing

a group of six different Activities of Daily Living (ADLs) while wearing a tri-axial accelerometer and a tri-axial gyroscope on the waist. After the first classification, two groups of activities were suggested for refinement: 1) Sitting and Laying, 2) Walking Downstairs and Walking Upstairs. An improvement in the classification performance is observed in all activities after refinement. The per-class classification accuracy is improved from 98.29% to 100% for sitting, from 98.29% to 100% for laying, from 99.19% to 99.37% for walking downstairs, and from 99.39% to 99.56% for walking upstairs.

The results achieved in this chapter go in line with those achieved by the state-of-the-art. On the one hand, it can be observed that, as in the work in [109], higher classification recalls are obtained on static activities such as sitting or laying with classification recalls in the range of 95% to 100% as compared to those achieved on similar walking-related activities such as walking downstairs and walking upstairs, which are in the range of 72% to 86%. On the other hand, it is found that the overall classification results achieved on the utilised benchmark dataset tend to be high. For instance, in [154], a classification accuracy of 97.59% is achieved. In [132], the classification accuracy achieved is 95.75%.

With regards to hygiene related activities, the works in [149] and in [153] achieved classification accuracies of 83.9% and 90.1% on their respective datasets, which include teeth brushing among two other hygiene-related activities (shave and wash). Although a fair comparison cannot be made given the different datasets utilised, as compared to the results achieved in these two research works, a significantly higher classification accuracy is achieved in this work. In this regard, it should be outlined that the works in [149, 153] make use of 16 features only, whereas a feature vector with a significantly higher dimensionality (266) is proposed in this work.

Besides, the results obtained also suggest that the use of the proposed multi-level refinement can improve the classification accuracy on those activities that are more difficult to classify between when following a traditional classification approach. In addition, this method can benefit unbalanced experiments where data from specific activities are more difficult to collect as compared to others. After cross-validating the data-set, the amount of data from those specific activities may not be enough to classify them against similar activities. This problem

could be mitigated by applying the multi-level refinement approach proposed.

4.6 Conclusions

From the work undertaken in this chapter, it can be concluded that it is plausible to accurately recognise hygiene-related activities from other quotidian activities using tri-axial accelerometer data with long-established hand-crafted features and state-of-the-art classification models commonly employed for the recognition of other continuous quasi-periodic activities. In addition, the successful performance exhibited by the proposed multi-level refinement algorithm suggests that feature informativeness depends on the activity set chosen. Computational efforts should be given to particular groups of activities (or classes) with lower classification performance, in order to optimise the selection of features and consequently their classification rate. This approach could have a significant positive impact when the recognition of a specific class is crucial for the interest of the study, as well as when performing feature selection with imbalanced datasets. An example of this is a fall detection system.

Chapter 5

Gesture Recognition Through the Use of Hand-Crafted Features

5.1 Introduction

Following the recognition of hygiene-related activities, the remaining chapters of this thesis are aimed at the recognition of food and fluid intake gestures to further recognise meal periods from continuous sensory recordings. Recent statistics outline eating difficulties as a prevalent issue among the elderly population. For instance, the study carried out in [155], which includes 520 elderly patients in hospital rehabilitation, shows that 82% of them suffer some form of eating difficulty. The survey conducted in [156], including 3000 patients from 11 different hospitals, reveals that 21.1% of the patients younger than 80, and 36.4% of those aged 80 or older require some form of eating assistance. Eating difficulties are those that alone or in combination, hamper the intake or the preparation of food and/or beverages [157], with significant causes including poor appetite, cognitive impairment or feeding dependency. Incidentally, a poor diet can contribute to weight loss and malnutrition, leading to potential functional limitations, metabolic abnormalities and diminished immunity [158].

Additionally, maintaining an adequate hydration level is an essential aspect of dietary management [159]. Mainly, fluid intake is a severe issue in elderly care, where diminished thirst perception is frequently related to reduced cog-

nitive capabilities, leading concurrently to difficulty at remembering to drink enough [160]. Approximately 17 million people suffer a stroke yearly [161], with 77% of them enduring an upper extremity disability or a function loss of the limb upper motor [162]. Such function loss may lead stroke patients to difficulty at performing basic actions like eating or drinking, therefore limiting their own independence [140].

As opposed to quasi-periodic activities, which exhibit continuous behaviour in time, the difficulty of spotting gestures lies in their rather sparse nature. Further, spotting naturally learned gestures such as grasping a fork is harder than detecting gestures which have been purposely trained within a constrained environment, e.g. human-machine interaction gestures [123]. Given this, this chapter explores computational solutions to spot and recognise eating and drinking gestures from continuous sensory recordings.

The remainder of this chapter is organised as follows: Section 5.2 reviews relevant work on gesture spotting and recognition with the use of wearable sensors. Section 5.3 presents the motivation behind the work undertaken in this chapter. Section 5.4 presents the method proposed for the development of a fluid and food intake tracking system. Section 5.5 presents the results achieved and compares them to those of previous similar published works. Section 5.6 reports the conclusions drawn from the obtained results.

5.2 Review of Work on Gesture Recognition

Various solutions for spotting and recognising gestures have been proposed in recent years. In [124] a solution to recognise a set of seven basic hand gestures for human-machine interaction purposes using bi-axial data from a tri-axial accelerometer is proposed. In this work, a set of ten features is used to determine the gesture termination points. Once segments are found, three different models are proposed for the recognition of the gestures. Among the three models, the best results are achieved by a template matching model (95.6% classification accuracy). Similar work by [163] employs an LSTM network to recognise a set of six different hand gestures using tri-axial accelerometer and tri-axial gyroscope data from five users, achieving a classification accuracy of 95.85%. An adaptive segmentation

technique to spot a set of four transitional activities (sit-to-stand, stand-to-sit, sit-to-lie and lie-to-sit) is developed in [120] using data from a waist-worn accelerometer. First, a set of thirteen features is used on windows of fixed length to determine whether the different windows contain a transitional, a dynamic or a static activity. Windows classified as a transitional activity are extended until a decrease in likelihood for a particular transitional activity, given by the Gaussian probability density function, is identified. The results demonstrate an improvement in classification recall from 89.9% using an artificial segmentation approach to 93.0% with the adaptive segmentation technique. A solution for spotting and recognising smoking gestures using data from a wrist-worn quaternion is proposed by [61]. In this work, gestures are firstly spotted using a rest position tracking algorithm alongside a peak detector used to detect peaks on the distance between the most recent rest position and the current position. A posteriori, a feature vector from the extracted segments is calculated and used to train a Conditional Random Field (CRF) classifier. A classification precision of 91.0% and a classification recall of 81.0% are achieved by the proposed system.

Regarding the recognition of food and fluid intake gestures, the authors in [140] report a classification recall of 91.3% for the recognition of drinking gestures using a single wrist-worn IMU alongside an SVM classifier fed with a feature vector calculated over windows of 0.25 seconds. The work in [85] proposes a semi hierarchical approach for the recognition of recognition of eating and drinking in free-living conditions based on data collected from wrist-worn tri-axial accelerometers and gyroscopes. Through such approach, windows are firstly identified as 'eating' or 'not eating' using an artificial segmentation approach, namely a sliding window. A posteriori a dynamic segmentation technique is used to identify 'drinking' versus 'not drinking' actions using a threshold based adaptive segmentation technique using the magnitude of the gyroscope signal. To classify the different gestures, a feature vector composed of ten time-domain features is used as input to a range of six different classifiers, achieving a classification recall of 77% and 62% and a classification precision of 39% and 37% for eating and drinking respectively. In [65] a Gaussian Mixture Hidden Markov Models (GMM-HMMs) network is used for recognising drinking gestures. The experimental data is collected from 7 users following their usual daily activities while wearing a single

wrist-worn inertial sensor which includes a tri-axial accelerometer, a tri-axial gyroscope and a tri-axial magnetometer. An average classification precision and classification recall of 75.2% and 76.1% are achieved respectively. A drinking spotting solution based on a Feature Similarity Search (FSS) is proposed in [64]. In this work, the data is collected from six users wearing a single wrist-worn inertial unit containing a tri-axial accelerometer, a tri-axial gyroscope and a tri-axial compass while performing a set of various experimental scenarios. With this method, a classification recall of 84.0% is achieved. In [123], a solution for spotting and recognising a set of four dietary gestures (cutlery, drink, spoon and hand-held) using five inertial sensors (two on each arm and one on the trunk) is proposed. This solution is based on a two-stage gesture spotting approach through the combination of a sliding-window and bottom-up (SWAB) adaptive segmentation technique and an FSS. Once potential segments are identified by the two-stage gesture spotting technique, a Hidden Markov Model (HMM) is employed to classify the gestures of interest. This approach achieves a classification precision of 73.0% and a classification recall of 79.0%.

5.3 Motivation

The motivation behind the work in this chapter comes from the need to provide unobtrusive means of recognising eating and drinking gestures, as well as from the different limitations found in work concerning this matter. First, some studies rely on extremely constrained environments. For example, in [140] it is reported that on the recognition of drinking gestures, chairs are height-adjusted to individuals. In addition, individuals are told how to perform the drinking actions and the data is only composed of drinking gestures. The work by [164] on recognising door opening gestures makes no mention of a 'Null' class. The 'Null' class in a gesture recognition problem is the class composed by gestures outside the studied gesture set. This fact implies the experimental data set is built only with the gestures of interest. In the research work conducted by [124] on the recognition of a set of seven hand gestures, participants are told to hold the accelerometer horizontally during the experiments. Gesture spotting and recognition should be undertaken in realistic scenarios where participants perform the studied actions

freely. In addition, the resultant data sets should include a reasonable 'Null' class with a range of additional gestures outside the sought gesture set. Second, the classification performance of gesture spotting and recognition systems under unconstrained environments still lies far away from that in HAR systems. The main reason is that given the sparsity of gestures and the consequent difficulty of being accurately spotted from continuous data streams, a great number of true positives are missing at the segmentation (spotting) step. For example, the work presented by [123] results in a recall of 80% at the segmentation stage. The results in [64] indicate an 84% recall at spotting drinking gestures.

Besides, various fluid and food intake tracking solutions proposed are found to require the use of several sensor units [123, 165]. This could make such solutions be excessively intrusive for a daily use. Overall, the drawbacks above suggest there are still many open challenges in gesture spotting and recognition. The mitigation of the above drawbacks has motivated the work in this chapter, which aims at improving the performance achieved by previous work on the recognition of food and drink intake gestures while preserving unobtrusiveness and user comfort by the use of a single wrist-worn wearable device and the introduction of a novel adaptive segmentation technique able to accurately spot sparse eating and drinking gestures when these are perform freely, as well as of a novel feature descriptor based on the DTW distance that incorporates additional information to long-established feature sets.

5.4 Methods

This section presents the steps undertaken to develop the proposed fluid and food intake system based on hand-crafted features. The different stages of the proposed system are illustrated in Figure 5.1. First, potential segments containing eating or drinking gesture are identified using the Crossings-based Adaptive Segmentation Technique (CAST) proposed in this thesis and described in Section 3.3.4.2. A posteriori, four different Computational Solutions (CS) are proposed as follows:

CS1:- Dynamic Time Warping (DTW) Distance + K-Nearest Neighbours (KNN)

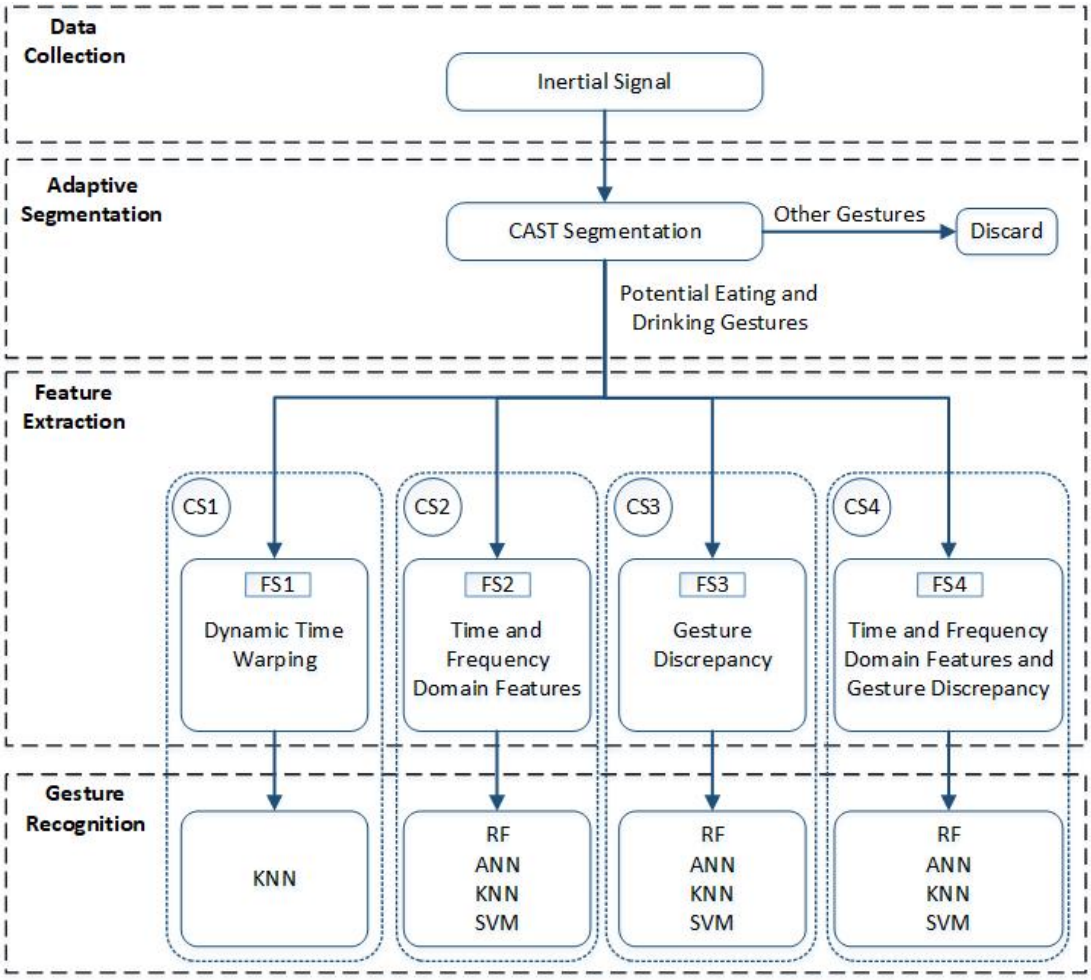

Figure 5.1: Schematic diagram of the proposed methodology to spot and recognise eating and drinking gestures.

CS2:- Feature set + range of state-of-the-art classification models

CS3:- Gesture discrepancy + range of state-of-the-art classification models

CS4:- Feature set+ gesture discrepancy + range of state-of-the-art classification models

The above computational solutions are used to methodically justify the addition of a gesture discrepancy measure to long-established features used in previous HAR work. In CS1, the use of Dynamic Time Warping is evaluated. Given the

challenging gesture set proposed, modest results are expected from CS1. However, this serves as a basis to justify the further exploration of DTW as a feature descriptor as well as to validate the CAST on the identification of eating and drinking gestures. In this context, such a justification can be based on the achievement of a fair classification performance by a computational solution employing DTW alongside a simple non-parametric distance-based classification model, such as KNN. CS2 explores the use of long-established features employed in previous HAR applications for the recognition of eating and drinking gestures. CS3 introduces the use of gesture discrepancy as a feature descriptor. Ultimately, CS4 evaluates the combination of the long-established range of features with the gesture discrepancy measure proposed. The achievement of an improvement on the classification performance of CS4 as compared to previous computational solutions will justify the addition of the gesture discrepancy measure in future activity and gesture recognition work. The performance of the proposed computational solutions is studied across three different gesture sets as follows:

2-Class: Null, Drinking or Eating

3-Class: Null, Drinking, Eating

5-Class: Null, Drinking, Spoon, Fork, Hand

where 'Null' refers to any gesture within the 'Null' class. That is any gesture which is not eating or drinking gesture.

5.4.1 Experimental Setup

The work in this chapter is based on the Dataset 2 presented in Section 3.2.2.2. This data comprises the tri-axial acceleration and the tri-axial angular velocity of the wrist of the different participants.

5.4.2 Signal Pre-processing

In order to minimise the computational cost of the system, a limited initial pre-processing is carried out on the raw inertial signals. The directions of the accelerometer y-axis and the gyroscope x and z axes are shifted 180° for the

left-handed participant, given the opposite orientation of these signals when the sensor unit is worn on the left hand.

5.4.3 Signal Segmentation and Gesture Spotting

The signal segmentation is carried out by the use of the proposed Crossings-based Adaptive Segmentation Technique (CAST) presented in Section 3.3.4.2. With this, the potential segments containing eating or drinking gestures are retrieved for further inspection.

5.4.4 Gesture Recognition

Once the potential segments containing an eating or drinking gesture are identified, gesture recognition is tackled as a classification problem. For the four proposed computational solutions (CS1, CS2, CS3, CS4), four different feature sets are employed as follows:

FS1:- Dynamic Time Warping

FS2:- Feature Vector

FS3:- Gesture Discrepancy

FS4:- Feature Vector and Gesture Discrepancy

More detail about these approaches is provided in the following sections.

5.4.4.1 Dynamic Time Warping

Let $q[t] = [q_1, q_2, ...q_n]$ and $s[t] = [s_1, s_2, ...s_n]$ be two temporal sequences with values at every time instant t=[1,2,...n]. The distance $d(q, s)$ is typically measured as their Euclidean distance:

$$d(q,s) = \sqrt{\sum_{t=1}^{n}(q[t] - s[t])^2} \tag{5.1}$$

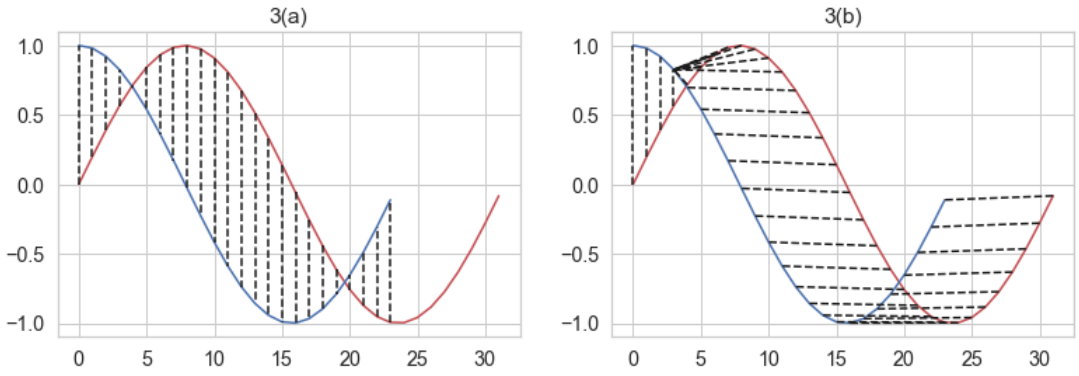

Figure 5.2: Difference between the Euclidean distance and the DTW distance of two signals; a) Euclidean distance, b) DTW distance: the distance between two points is calculated as their Euclidean distance (vertical distance) after alignment.

Two major constraints are found on the use of the Euclidean distance on time-dependent sequences: 1) the length of the sequences must be equal, i.e. $|q| = |s|$, 2) it does not consider the temporal distortion that may exist between q and s, since it measures the vertical distance between pairs of points according to their indexes at their respective sequences.

To overcome the above constraints, the optimal alignment between time-dependent sequences is calculated with the use of DTW [166]. The alignment can be explained as follows: Considering the two temporal sequences q and s of respective lengths $|q|$ and $|s|$, DTW finds a mapping path $\{(p_1, r_1), ..., (p_j, r_j)\}$ such that the distance on the mapping path $\sum_{i=1}^{j} |x(p_i) - y(r_i)|$ is minimised with the following two constraints:

$$\begin{cases} \text{Anchored beginning: } (p_1, r_1) = (1, 1) \\ \text{Anchored end: } (p_j, r_j) = (|q|, |s|) \end{cases} \tag{5.2}$$

The DTW distance between q and s is then calculated as the cost of the optimal alignment as follows:

$$D_{i,j} := D(q(i), s(j)) + min \left\{ \begin{array}{l} D(i-1, j) \\ D(i-1, j-1) \\ D(i, j-1) \end{array} \right\} \tag{5.3}$$

where $D(q(i) - s(j))$ is calculated as the Euclidean distance.

Figure 5.2 illustrates the use of the Euclidean distance and DTW to measure the similarity between temporal sequences. From the figure, it can be seen that DTW overcomes the drawbacks encountered when using the Euclidean distance. First, it can measure the distance between signals with different lengths, since one point of the sequence q can be aligned to more than one point of the sequence s and vice versa. Second, the alignment performed is able to capture the temporal distortion between the signals.

Ultimately, the DTW distance is used for gesture recognition. To do so, a K-Nearest Neighbours (KNN) classification model is employed, whereby unseen segments are assigned to the most common class among its k-nearest neighbours, with DTW being the distance measure between the different segments.

5.4.4.2 Feature Vector

This computational solution makes use of a long-established set of features used within the field of HAR [14, 15, 77, 167]. The feature vector has been conscientiously culled to provide a knowledgeable description of the data regarding a wide array of signal characteristics. These include measures of central tendency, periodicity, dispersion, changes in direction, frequency distribution and magnitude area. The range of features proposed was calculated over the mediolateral a_x, anteroposterior a_y and vertical a_z acceleration corresponding to the tri-axial accelerometer readings, as well as on the yaw ω_x, roll ω_y and pitch ω_z corresponding to the tri-axial gyroscope readings across the potential segments. On top of the above, the duration of each segment is also incorporated into the feature set. The resultant dimensionality of the feature vector proposed is $n = 85$.

5.4.4.3 Gesture Discrepancy

This computational solution introduces a gesture discrepancy measure as a signal descriptor. To do so, the Soft-DTW differentiable loss function proposed by [168] is employed to calculate a gesture barycenter for each of the gestures within the different proposed gesture sets through a minimisation problem. Further, the DTW distances to each of the calculated barycenters are used to build the feature

set.

Let's consider multivariate time series of varying length taking values in $\Omega \subset \mathbb{R}^{p}$, whereby they are represented as a matrix of p rows. Soft-DTW unifies the original DTW distance [166] and the Global Alignment Kernel (GAK) proposed by [169], both used to compare two time series $x[t] = [x_1, x_2, ..., x_n] \in \mathbb{R}^{p \times n}$ and $y[t] = [y_1, y_2, ..., y_m] \in \mathbb{R}^{p \times m}$.

Given the cost matrix $\Delta(x, y) := [\delta(x_i, y_j]_{ij} \in \mathbb{R}^{n \times m}$ and the set of binary alignments matrices $A_{n,m} \subset \{0, 1\}$, the inner product $\langle A, \Delta(x, y)\rangle$ of the cost matrix with an alignment matrix A in $A_{n,m}$ gives the score of A. DTW and GAK consider respectively the cost of all possible alignment matrices as follows:

$$DTW(x, y) := \min_{A \in A_{n,m}} \langle A, \Delta(x, y)\rangle, \tag{5.4}$$

$$\kappa_{GA}^{\gamma}(x, y) := \sum_{A \epsilon A_{n,m}} e^{-\langle A, \Delta(x,y)\rangle/\gamma} \tag{5.5}$$

From the equations above, a unified algorithm can be formulated as:

$$min^{\gamma}\{a_1, ..., a_n\} := \begin{cases} min_i \leq_n a_i, & \gamma = 0, \\ -\gamma \log \sum_{i=1}^{n} e^{-a_i/\gamma}, & \gamma > 0. \end{cases} \tag{5.6}$$

where γ is a smoothing parameter taking values in $\mathbb{R}_{\geq 0}$. Given the above, γ-soft-DTW can be defined as:

$$dtw_{\gamma}(x, y) := min^{\gamma}\{\langle A, \Delta(x, y)\rangle, A \in A_{n,m}.\} \tag{5.7}$$

Therefore, the original DTW score is recovered when γ is set to 0 and $dtw_{\gamma} = -\gamma log k_{GA}^{\gamma}$ when $\gamma > 0$.

Ultimately, given a group of N time series y_1, ..., y_N, that is, N matrices of p rows and varying number of columns, m_1, ..., m_N, the interest is to define a single barycenter time series x for that group under a set of normalised weights $\lambda_1, ..., \lambda_N \in \mathbb{R}_+$ such that $\sum_{i=1}^{N} \lambda_i = 1$. Thus, the barycenter is calculated by

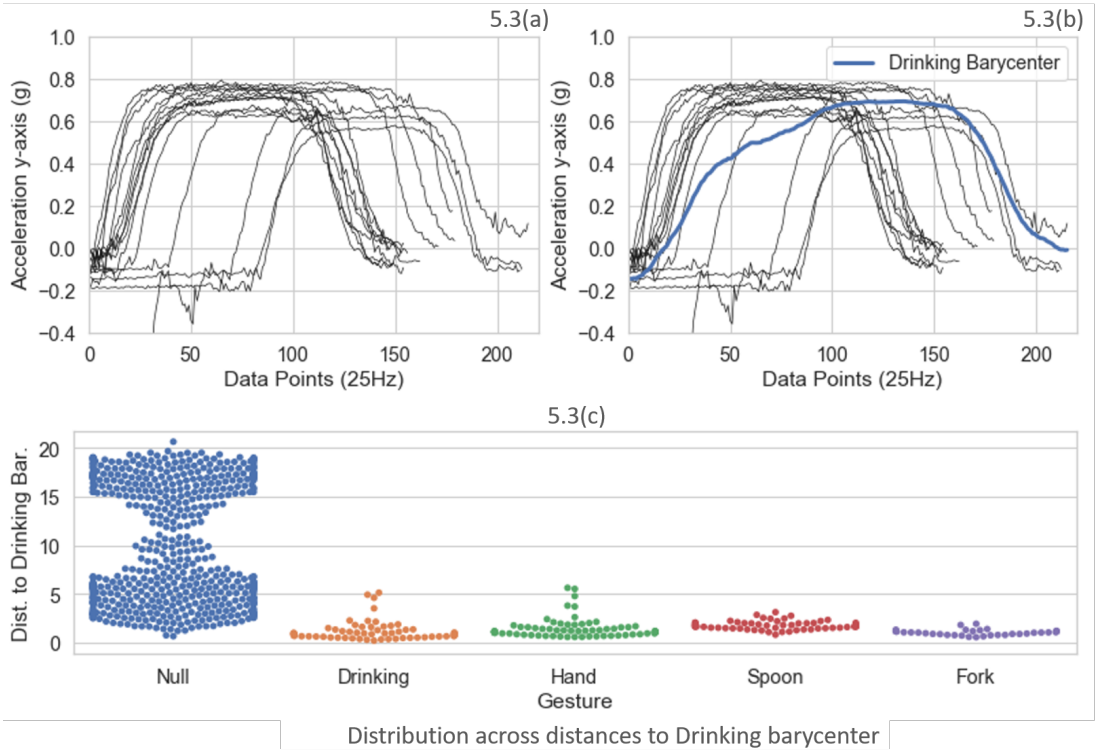

Figure 5.3: Distance to the drinking barycenter (accelerometer y-axis) of one of the experiment participants; a) Different drinking gestures from the participant, b) Calculation of the participant's drinking barycenter, c) Distribution of distances to the barycenter in (b) across the gestures from the rest of the participants.

approximately solving the following problem:

$$\min_{x \in \mathbb{R}^{p \times n}} \sum_{i=1}^{N} \frac{\lambda_i}{m_i} dtw_\gamma(x, y_i) \tag{5.8}$$

where it is assumed that x has fixed length n. Given the proposed gesture sets G_1, G_2, G_3 of respective lengths $|G_1|$, $|G_2|$, $|G_3|$, a barycenter was calculated for each of the gestures different from the 'Null' class $g_1, ..., g_{|G_i|-1}$ within $G_1, ...G_3$, for each of the experiment participants $P_1, ..., P_6$, for each of the time series in $a_x, a_y, a_z, \omega_x, \omega_y, \omega_z$, corresponding to the tri-axial accelerometer and the tri-axial gyroscope readings. A posteriori, the DTW distances to the set of calculated barycenters were computed and further used as feature descriptors.

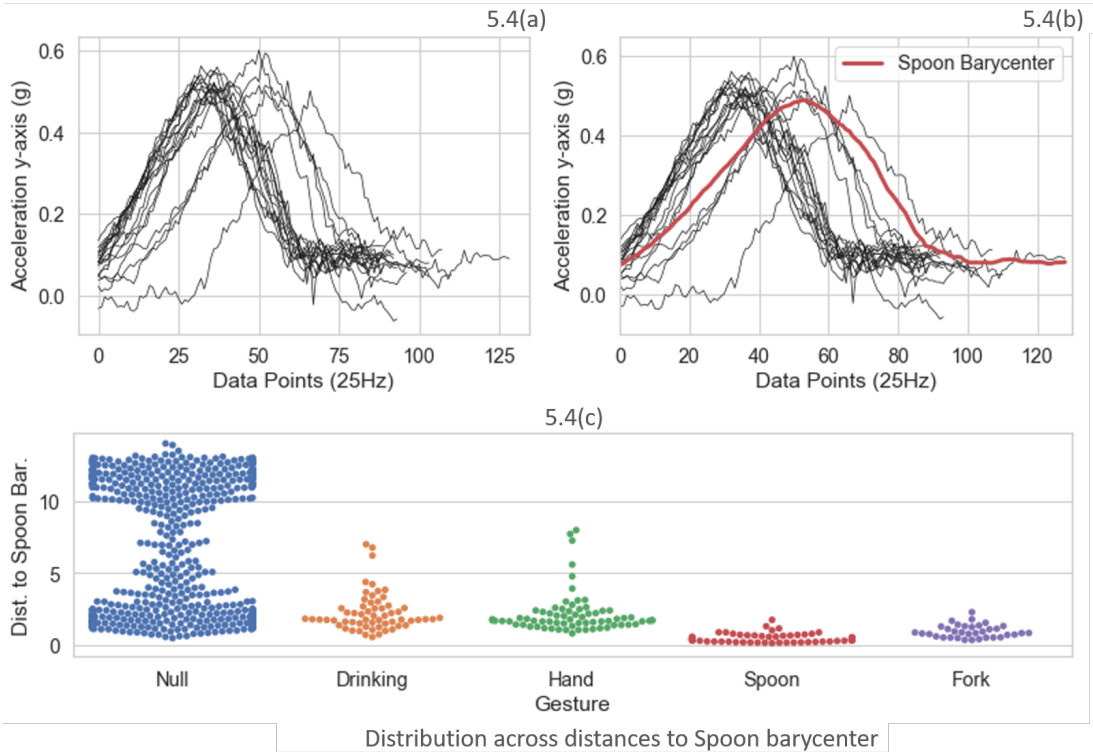

Figure 5.4: Distance to the spoon barycenter (accelerometer y-axis) of one of the experiment participants; a) Different spoon gestures from the participant, b) Calculation of the participant's spoon barycenter, c) Distribution of distances to the barycenter in (b) across the gestures from the rest of the participants.

Two pictorial examples of the calculation of a gesture barycenter and the distribution of the DTW distances to the calculated gesture barycenter across the different gestures are shown in Figure 5.3 and Figure 5.4. Further, the bi-dimensional distribution of the DTW distances to the barycenters exposed in Figure 5.3 and Figure 5.4 across the different gestures is shown in Figure 5.5 for illustration purposes. As a result of the above distance computations, the resultant dimensionality of the feature vector is $n = 36$ for the 2-class classification problem, $n = 72$ for the 3-class classification problem and $n = 144$ for the 5-class classification problem.

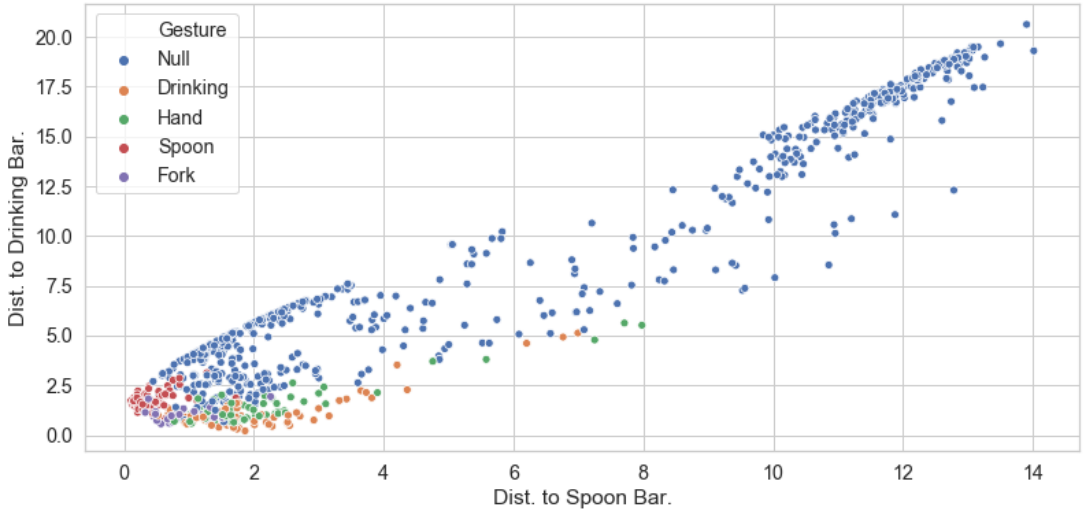

Figure 5.5: Bi-dimensional distribution of the DTW distances to the drinking and spoon barycenters of one of the participants across the gestures from the rest of the participants.

Table 5.1: Post-segmentation data summary

Class. Problem	Dim.	Inst.	Class 1	Class 2	Class 3	Class 4	Class 5
2-class	121	813	Null (72.2%)	Eat and Drink (27.8%)	-	-	-
3-class	157	813	Null (72.2%)	Drink (7.3%)	Eat (20.5%)	-	-
5-class	229	813	Null (72.2%)	Drink (7.3%)	Hand (8.7%)	Spoon (7.5%)	Fork (4.3%)

5.4.4.4 Feature Vector and Gesture Discrepancy

Feature set FS4 is a combination of the features introduced in FS2 and FS3 to evaluate whether the addition of a gesture discrepancy measure to long-established feature vectors improves the recognition rate of the system. The combination of both feature sets gives a resultant dimensionality of $n = 121$ for the 2-class classification problem, $n = 157$ for the 3-class classification problem and $n = 229$ for the 5-class classification problem. A summary of the feature vector dimensionality, the total number of instances and the class distribution for each of the classifications problems is given in Table 5.1.

5.4.5 Evaluation

A leave-one-out cross-validation strategy is employed for evaluating the different computational solutions. That is, a different participant was used as the test set on each of the cross-validation steps. For the feature sets including the gesture discrepancy measure (FS3 and FS4), the distances to the barycenters of the participant used as the test set on each cross-validation cycle were removed from the feature set.

Given the special structure of the feature set FS1 proposed in CS1, its performance was evaluated by the employment of a KNN classifier. The rest of the computational solutions were evaluated across a range of state-of-the-art classification models, including ANN, SVM, RF and KNN.

5.5 Results

This section presents the results obtained by the implementation of the presented methodology. Section 5.5.1 shows the performance of the proposed CAST segmentation technique at spotting potential eating and drinking gestures. Section 5.5.2 presents the results achieved by the different computational solutions proposed for gesture recognition. A discussion upon the results obtained is given in Section 5.5.3.

5.5.1 Gesture Spotting

As explained in Section 5.4, the first step on the development of the proposed fluid and food intake recognition system is to spot potential segments containing an eating or a drinking gesture. This is tackled by the implementation of CAST, which uses the crosses between two moving averages to spot those potential segments. A pictorial example for one of the experiment participants is shown in Figure 5.6.

Given that more computational intensive tools are to be applied at the gesture recognition step, the aim at this preliminary spotting step was to optimise the classification recall, that is, minimising the number of 'False Negatives', in this case eating or drinking gestures classified as pertaining to the 'Null' class. The

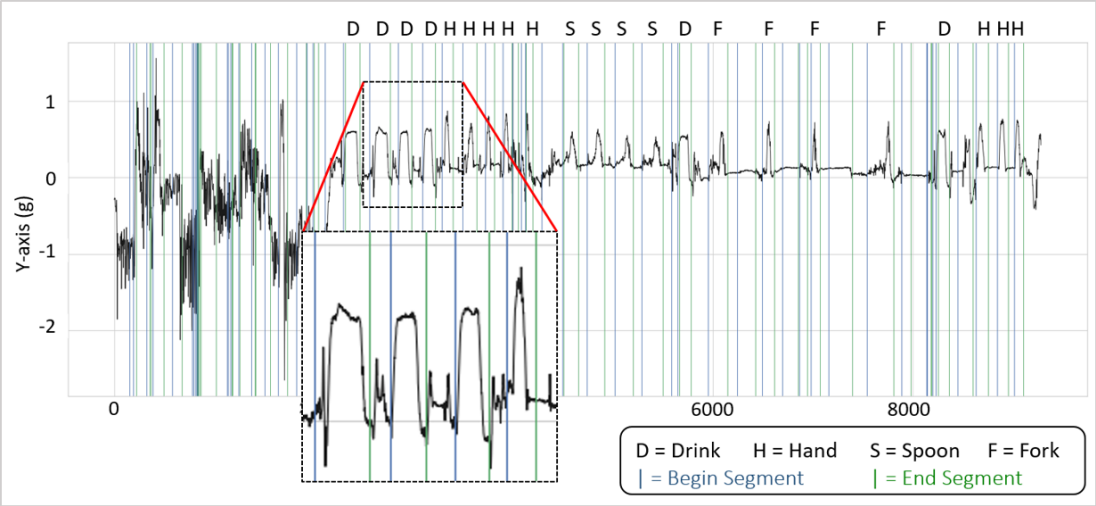

Figure 5.6: Performance of the Crossings-based Adaptive Segmentation Technique for one of the experiment participants.

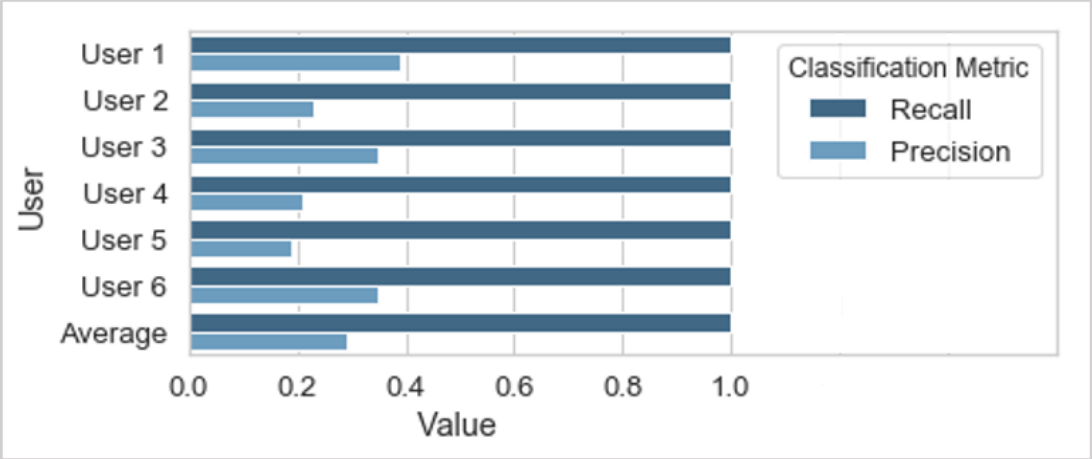

Figure 5.7: Spotting performance of CAST.

achieved spotting results shown in Figure 5.7 outline an average precision of 29% and an average recall of 100%, showing that this is successfully achieved by the segmentation technique proposed.

5.5.2 Gesture Recognition

After the segments potentially containing an eating or a drinking gesture are identified, gesture recognition comes into place. Four different computational

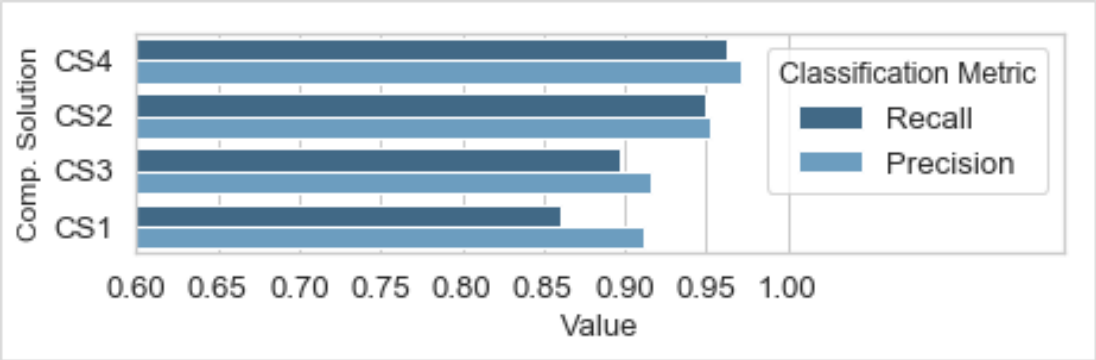

Figure 5.8: Classification performance of the four computational solutions proposed on the 2-class classification problem.

solutions were proposed across three different experiments. A comprehensive study upon the performance of the implemented computational solutions was performed and the best results obtained in each of the three experiments are presented below:

5.5.2.1 Experiment 1: 2-Class Classification Problem

In this experiment, eating and drinking gestures are grouped together and classified against the 'Null' class. The results presented in Table 5.2 outline an average per-class classification accuracy of 97.4%, a precision of 97.2% and a recall of 96.3% using a Random Forest Classifier on the feature set composed of the proposed range of features alongside the gesture discrepancy measure (FS4). Figure 5.8 shows the performance of the four computational solutions proposed.

Table 5.2: Classification metrics for the 2-class classification problem using CS4 with RF.

	Accuracy (%)	Precision (%)	Recall (%)
Null	97.4	97.6	98.8
Eating or Drinking	97.4	96.8	93.8
Average per-class	97.4	97.2	96.3

5.5.2.2 Experiment 2: 3-Class Classification Problem

This experiment aims at the recognition of eating and drinking gestures separately. This is therefore tackled as a 3-Class classification problem, with the classes being 'Null', 'Drinking' and 'Eating'. The classification metrics shown in Table 5.3 report an average per-class classification accuracy of 98.2%, a precision of 95.7% and a recall of 95.0%. The reported results are achieved using an Artificial Neural Network (ANN) on the feature set (FS4). The classification performance achieved by each of the computational solutions proposed are shown in Figure 5.9.

Table 5.3: Classification metrics for the 3-class classification problem using CS4 with an ANN

	Accuracy (%)	Precision (%)	Recall (%)
Null	97.9	98.1	99.0
Drinking	99.0	93.3	93.3
Eating	97.7	95.7	92.8
Average per-class	98.2	95.7	95.0

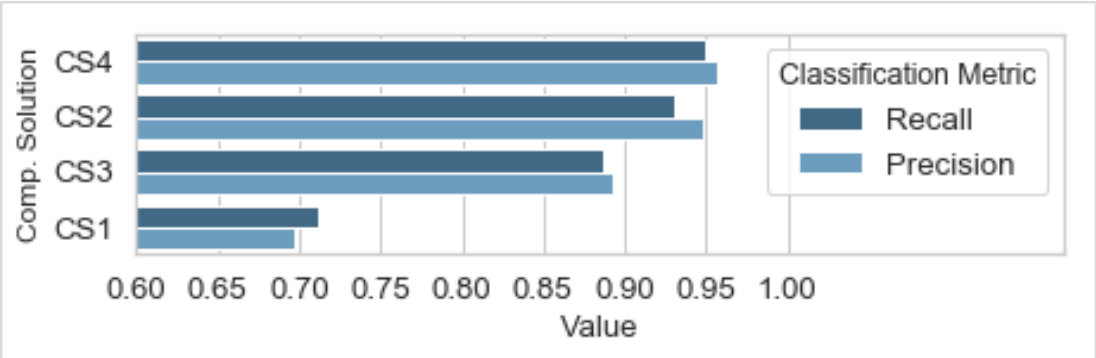

Figure 5.9: Classification performance of the four computational solutions proposed on the 3-class classification problem.

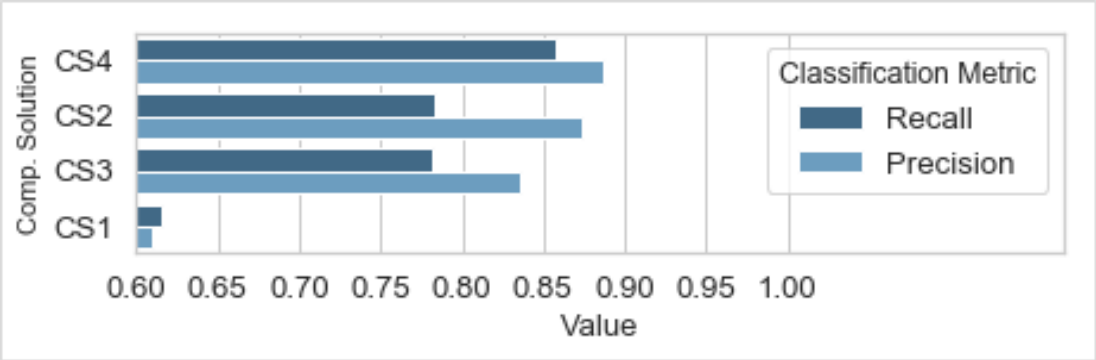

Figure 5.10: Classification performance of the four computational solutions proposed on the 5-class classification problem.

5.5.2.3 Experiment 3: 5-Class Classification Problem

In this experiment, the 'Eating' class is further divided into 3 different classes ('Spoon', 'Fork' and 'Hand'), leading to a 5-class classification problem, with the classes being 'Null', 'Drinking', 'Spoon', 'Fork' and 'Hand'. The classification metrics in Table 5.4 report an average per-class classification accuracy of 97.8%, a precision of 88.7% and a recall of 85.8%, using an ANN on the feature set (FS4). The classification performance of the four computational solutions are shown in Figure 5.10.

Table 5.4: Classification metrics for the 5-class classification problem using CS4 with an ANN.

	Accuracy (%)	Precision (%)	Recall (%)
Null	97.0	97.2	98.8
Drinking	98.6	90.2	91.7
Spoon	99.0	96.5	90.2
Fork	97.6	75.0	68.6
Hand	97.0	84.4	80.0
Average per-class	97.8	88.7	85.8

Table 5.5: Comparison of the proposed approach with previous work on the recognition of drinking gestures.

Method	Sensor Units	Spot.	Recog.	Accuracy	Precision (%)	Recall (%)
Junker *et al* (2008) [123]	5	✓	✓	-	88.0	83.0
Amft *et al* (2010) [64]	1	✓	✓	-	84.0	90.0
Chen *et al* (2017) [140]	1	X	✓	-	96.5	91.3
Serrano *et al* (2017) [170]	4	✓	✓	-	82.28	84.42
Ramos-Garcia *et al* (2014) [171]	1	✓	✓	86.5	-	-
Proposed Approach (3-class)	1	✓	✓	99.0	93.3	93.3
Proposed Approach (5-class)	1	✓	✓	98.6	90.2	91.7

5.5.3 Discussion

The methodology proposed addressed the problem of spotting and recognising fluid and food intake gestures with the use of a single wrist-worn inertial unit. At the spotting step, the aim was to minimise the number of false negatives. This was based on the fact that more computational intensive tools, namely classification models, were to be applied at the recognition step. The novel adaptive segmentation technique proposed (CAST) correctly identified all the drinking and eating gestures. Although the average precision is considerably low (29%), a 100% recall is achieved, indicating the aim proposed is successfully accomplished. Further, a range of four different feature sets was proposed for gesture recognition. As expected, the addition of the gesture discrepancy measure as a feature descriptor consistently improves the classification performance of the system across the three experiments proposed. This can be explained by the fact that the signal alignment performed through the use of DTW accounts for the gestures intra-person and inter-person temporal distortion, thus adding crucial information to long-established feature sets used in previous activity or gesture recognition problems.

Given the great variety of gestures involved in an eating activity, previous research has varied the way of tackling its recognition. To fairly evaluate the proposed methodology against previous similar work, the performance of the recognition of drinking gestures is considered. For fairness, and given the lack of a benchmark dataset concerning eating and drinking (each of the works in Table

5.5 were evaluated using different experimental datasets), the comparison is made with similar controlled or semi-controlled experiments undertaken in lab settings. As shown in Table 5.5, the proposed methodology shows competent results when compared to other drinking gestures recognition work, including both the spotting and recognition stages. Only the classification precision achieved in [140] shows a higher value. However, in [140], a spotting step was not included since the data set lacked a 'Null' class as well as other gestures different from drinking gestures. As a result of this, the precision and recall metrics were clearly boosted, since the experiment proposed was evidently biased towards the recognition of drinking gestures.

5.6 Conclusions

This chapter has proposed a novel approach for the spotting and recognition of eating and drinking gestures in a semi-controlled experimental setting using a single wrist-worn inertial unit as a means of data collection. Two major conclusions can be drawn from the results achieved. First, the CAST is shown to overcome the two major drawbacks observed in previous similar work. On the one hand, as contrary to previous adaptive segmentation techniques in the field, the CAST achieves a 100% spotting recall, thus preventing the system from having false negatives at the preliminary spotting phase. This is crucial since the errors at the spotting phase will propagate to the recognition phase, therefore limiting the performance of the whole system. Second, while long-established feature sets only incorporate shallow (normally statistical) characteristics of the signals, the Soft-DTW based gesture discrepancy measure proposed accounts for the intra and inter-personal temporal distortion at performing eating and drinking gestures. As shown by the results obtained, this clearly offers an advantage to the system, which has seen a consistent improvement across the three experiments proposed. In terms of the average per-class classification recall, the addition of the gesture discrepancy measure improves the performance of the system from 0.950 to 0.963, from 0.931 to 0.95 and from 0.783 to 0.858 for the 2-class, 3-class and 5-class classifications problems respectively. Regarding the average per-class classification precision, the performance improvements seen are from 0.952 to

0.972, from 0.948 to 0.957 and from 0.874 to 0.887 for the 2-class, 3-class and 5-class classifications problems respectively.

Chapter 6

Exploring Deep Learning Techniques for Gesture Recognition

6.1 Introduction

This chapter explores the use of deep learning techniques, and in particular, that of Convolutional Neural Networks (CNNs) for the recognition of eating and drinking gestures. As outlined in Section 2.3.7, CNNs are increasingly employed for activity recognition purposes given their good performance across different studies. In contrast to traditional classification models such as KNN, RF or SVM, CNNs do not require specific domain-knowledge, since the features are automatically learned throughout the training phase of the network. Provided the effectiveness of the CAST at spotting potential eating and drinking gestures, this technique is further made use of for the work undertaken in this chapter. With this, a study upon the complexity of the CNN architecture, through which the optimal hyper-parameters of the CNN for the recognition of eating and drinking gestures are identified, is carried out. These hyper-parameters include the number of layers, the number of filters or kernels, and the size of these filters. In addition, multi-input architectures are explored through the use of three time series to image encoding techniques, namely the signal spectrogram, the Markov Transition

Field (MTF) and the Gramian Angular Field (GAF), and the extraction of a low-dimensional feature vector.

The remainder of this chapter is organised as follows: Section 6.2 reviews relevant work on Deep Learning for activity recognition. Section 6.3 presents the motivation behind the work undertaken in this chapter. Section 6.4 presents the CNN-based method proposed for the development of a fluid and food intake tracking system. Section 6.5 presents the results achieved and compares them to those of previous similar published works. Section 6.6 reports the conclusions drawn from the obtained results.

6.2 Review of Work on Deep Learning for Activity Recognition

The use of deep learning, and especially that of CNNs has revolutionised the state-of-the-art of challenging problems such as speech recognition and image classification [90]. Likewise, CNNs are gaining increasing attention within the field of HAR due to the numerous advantages they provide as compared to traditional state-of-the-art HAR feature extraction and classification methods. First, conventional HAR solutions typically require the computation of hand-crafted or self-engineered features, thus relying on human domain knowledge. Second, according to human expertise, only shallow features, such as basic signal statistics, can be learned through the use of conventional hand-crafted feature extraction methods [130]. Despite the good performance exhibited by the use of shallow features on the recognition of low-level activities such as walking, sitting or jogging, gaining insights into context-aware activities such as using the toilet or having lunch, may require more complex computations [172]. Third, in contrast to traditional HAR approaches, CNNs are able to exploit the translation invariant nature of human gestures/activities as well as the local dependency attribute of temporal sequences [90].

The advantages presented above have recently deviated the attention of human activity/gesture recognition research work towards the implementation of CNN frameworks, which as shown by recent work in the field [129–131], can out-

perform traditional state-of-the-art approaches such as Random Forest, Support Vector Machines or K-Nearest Neighbours. However, despite the good performance exhibited by CNNs, major discrepancies are found among the literature.

One of such discrepancies is found on the segmentation of the sensory signals, which is mainly due to the differing duration of different gestures or activity cycles. While excessively short segments would miss fundamental characteristics of a gesture/activity, long sequences may retrieve characteristics from multiple gestures/activities, thus lowering the ultimate classification performance. Generally, the length of the segments is either roughly estimated based on the characteristics of the gesture or activity set studied [90, 132], or calculated as a hyper-parameter of the classification problem itself [108, 131].

Different approaches are also found on the pre-processing of the signals. Typically, 1D filters are directly used on the raw sensor data [90, 130–132]. However, alternative solutions have also been proposed. In [108], the accelerometer signals are unified into the magnitude of the tri-dimensional vector. While this approach can reduce the computational cost of the network, a poor performance (classification accuracy = 92.95%) at recognising a basic set of three high-level activities, suggests that crucial information is thrown away at such unification step. Various studies employing multiple sensor nodes for data collection [136, 154], propose time series to image encoding frameworks to capture the spatial dependency between the different sensors, as well as the local dependency over time. A posteriori 2D CNNs are used for feature learning and classification. As proven in [136], 2D CNNs can outperform 1D CNNs on time series classification tasks; however, the exhibited improvement is considerably low.

The network architecture has also varied considerably between different HAR works. While some studies propose shallow networks with only one convolutional layer [108, 131], other studies have opted for the employment of networks with two convolutional layers [136, 154] or yet deeper architectures [90, 130]. In theory, increasing the number of convolutional layers allows for the computation of more complex features, which as shown in [90], can lead to better classification performance. However, employing deep architectures may also lead to network overfitting and consequently to a worse classification performance [131].

Ultimately, the learning rate employed during the training phase of the net-

work is another factor open to discussion. While the majority of the studies employ an arbitrary learning rate [90, 130, 131], the work in [132] demonstrates the adequate tuning of the learning rate can have a significant positive effect on the overall classification performance of deep learning models aiming at activity recognition. To do so, the authors in [132] evaluate the classification performance of the network across different learning rates, ranging from 0.001 to 0.1, obtaining classification accuracies in the range (91.882% to 94.022%). This implies a relevant improvement of 2.3% in the classification accuracy of the model is achieved when comparing the worst and the best performing learning rates.

6.3 Motivation

As with the work undertaken in Chapter 5, the motivation of the work in this chapter comes from the many open challenges that exist on the implementation of systems for eating and drinking gesture recognition. In addition, the review of the literature concerning the use of CNNs for activity recognition, suggests there are still many unanswered questions with regards to the impact of the architecture of the network on the classification performance of CNN-based systems. Furthermore, the ability shown by CNNs to extract informative features and to accurately classify different activities, suggest that it is plausible to propose an accurate domain-knowledge independent eating and drinking recognition system.

6.4 Methods

This section presents the methodology employed to develop the proposed CNN-based fluid and food intake recognition system. The section is divided regarding the different methodology phases as follows. Section 6.4.1 presents the experimental setup, Section 6.4.2 describes the signal pre-processing step employed to correct the orientation of the accelerometer signal for left-handed participants, Section 6.4.3 presents the signal segmentation technique employed and the signal padding step applied to the accelerometer signals. Section 6.4.4 defines the time series to image encoding frameworks employed, ultimately Section 6.4.5 describes

the single-input and the multi-input multi-domain CNN-based frameworks proposed for gesture classification. The recognition of the gestures is tackled as a 3-class classification problem with the classes being 'Null', 'Eat', 'Drink'.

6.4.1 Experimental Setup

The work in this chapter is based on the Dataset 2 presented in Section 3.2.2.2. In this case, only the data incorporating the tri-axial acceleration of the wrist of the different participants are used for the development of the different CNN architectures. It must be noted that gyroscope data was also considered. However, the use of this data constantly led to network overfitting issues. Given this, the use of gyroscope data was disregarded.

6.4.2 Signal Pre-processing

To account for the difference in terms of sensor orientation found when the sensory device is worn on the left hand, the direction of the accelerometer y-axis is shifted 180° for the left-handed participant.

6.4.3 Signal Segmentation

As in the work presented in Chapter 5, an adaptive segmentation technique, namely the CAST (see Section 3.3.4.2), is employed to identify potential segments containing an eating or a drinking gesture. Contrary to traditional sliding-window approaches, the CAST adapts the segments of the signal to the duration of the gestures themselves, leading to a gesture set of signal segments with varying lengths. The segments are a posteriori padded to the length of the longest segment retrieved by the CAST ($n = 394$) to allow for network batch training. The GAF and the MTF time series to image encoding frameworks utilise such padded segments of length ($n = 394$). In the case of the signal spectrogram framework, n is rounded up to the nearest higher power of 2 ($n = 512$).

6.4.4 Time Series Imaging

Inspired by the work in [173, 174], three different frameworks are employed for encoding the accelerometer signal segments into images, namely the signal spectrogram, the Markov Transition Field (MTF) and the Gramian Angular Field (GAF). In this work, the image encoding is independently applied to the magnitude of the 3-dimensional accelerometer signal as well as to the y-axis signal (previously employed for signal segmentation). Examples of the time series to image encoding frameworks employed in this work are shown in Figure 6.1. It should be noticed that the different time series imaging-based frameworks are not employed independently. Instead, these are individually combined with the 1D CNN benchmark model at the fully connected layer to explore whether the classification performance achieved by the benchmark model fed with raw accelerometer signals can be further improved by the incorporation of additional features.

6.4.4.1 Signal Spectrogram

The signal spectrogram is a visual representation which depicts the strength spectrum of frequencies of a signal as it varies with time. Given a time series $X = \{x_1, x_2, ..., x_n\}$, X is first converted into the frequency domain using the Fast Fourier Transform (FFT) as follows:

$$FFT(X) = \frac{\sum_{k=1}^{n} |a_k|^2}{n} \tag{6.1}$$

where a1,a2,... a_n are the FFT components of the corresponding window of length n. In this case, a window length n of 32 samples with 50% overlapping is used across the padded segments ($N = 512$).
A posteriori, the signal spectrogram is calculated as follows:

$$spectrogram\{x(t)\}(\tau, \omega) = |X(\tau, \omega)|^2 \tag{6.2}$$

Eventually, the resulting signal spectrogram is encoded into a 2-dimensional (time and frequency) graph, with a third dimension (signal amplitude of a par-

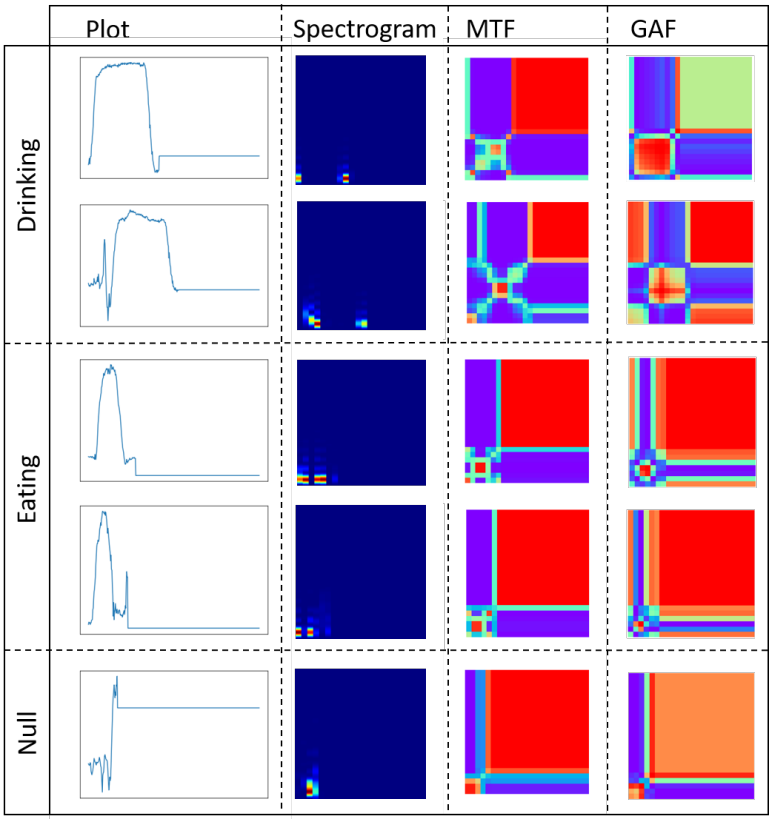

Figure 6.1: Examples of the employed imaging techniques for each of the classes ('Drink', 'Eat', 'Null'). In the examples provided, the plot and the corresponding spectrogram, MTF and GAF, are visual representations of the y-axis of the accelerometer signal.

ticular frequency at a specific time) represented by a colour scale.

6.4.4.2 Markov Transition Field

The Markov Transition Field (MTF) framework is employed to encode dynamical transition statistics of the signal. To preserve the sequential information enclosed within the signal, the framework proposed by [173] is employed, whereby the Markov transition probabilities are represented sequentially, thus preserving information in the time domain. Given a time series $X = \{x_1, x_2, ..., x_n\}$, Q quantile bins are identified and each x_i is assigned to the corresponding bins q_j ($j \in [1, Q]$). A posteriori a QxQ weighted adjacency matrix W is constructed with the count of the transitions among quantile bins in the form of a first order Markov chain

along the time axis. $w_{i,j}$ is then estimated as the frequency at which a point in the quantile q_j is followed by a point in the quantile q_i. This, after normalisation $\sum_j w_{i,j} = 1$ gives as a result the Markov transition matrix W. However, W is insensitive to the distribution of X and the temporal dependency on time steps t_i.

To overcome the loss of the temporal dependency, the Markov Transition Field (MTF) matrix M is defined as follows:

$$M = \begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ w_{ij|x_2 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_2 \in q_i, x_n \in q_j} \\ \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix} \tag{6.3}$$

The QxQ Markov transition matrix (W) is computed by dividing the data into Q quantile bins, where the quantile bins that contain the data at time stamp i and j are q_i and q_j respectively ($q \in [1, Q]$). M_{ij} in MTF denotes the transition probability of $q_i \rightarrow q_j$. That is, the matrix W is spread out into the MTF matrix M by considering temporal position.

6.4.4.3 Gramian Angular Field

The Gramian Angular Field (GAF) is a time series to image encoding technique by which a time series is encoded into a polar coordinate system from its original Cartesian coordinates while preserving the temporal correlation. Given a time series $X = \{x_1, x_2, ..., x_n\}$ where each x_i is normalised as:

$$\tilde{x_i} = \frac{(x_i - max(X) + (x_i - min(X))}{max(X) - min(X)} \tag{6.4}$$

$\tilde{X}$ can be represented in polar coordinates as follows:

$$\begin{cases} \phi = \arccos(\tilde{x_i}), -1 \leq \tilde{x_i} \geq 1, \tilde{x_i} \in \tilde{X} \\ r = \frac{t_i}{N}, t_i \in N \end{cases} \tag{6.5}$$

where t_i is the time stamp and N is a constant regularisation factor of the polar coordinate system.

The above encoding offers two major advantages. First, the function is bijective. That is, each value in the original signal correspond to one value in the polar coordinate representation and vice versa. Second, the absolute temporal relations are preserved through the r coordinate.

Further to the conversion, the angular perspective can be easily exploited by considering the trigonometric sum between each pair of points. Thusly, the GAF is defined as:

$$G = \begin{bmatrix} cos(\phi_1 + \phi_1) & \dots & cos(\phi_1 + \phi_n) \\ cos(\phi_2 + \phi_1) & \dots & cos(\phi_2 + \phi_n) \\ \vdots & \ddots & \vdots \\ cos(\phi_n + \phi_1) & \dots & cos(\phi_n + \phi_n) \end{bmatrix} \tag{6.6}$$

Taken the definition of the inner product of two vectors x and y as:

$$< x, y >= x \cdot y - \sqrt{1 - \tilde{X}^2} \cdot \sqrt{1 - \tilde{X}^2} \tag{6.7}$$

G is therefore a Gramian matrix as shown in Equation 6.8:

$$G = \begin{bmatrix} < \tilde{x}_1, \tilde{x}_1 > & \dots & < \tilde{x}_1, \tilde{x}_n > \\ < \tilde{x}_2, \tilde{x}_1 > & \dots & < \tilde{x}_2, \tilde{x}_n > \\ \vdots & \ddots & \vdots \\ < \tilde{x}_n, \tilde{x}_1 > & \dots & < \tilde{x}_n, \tilde{x}_n > \end{bmatrix} \tag{6.8}$$

6.4.5 Network Architectures

This work proposes a range of single-input and multi-input multi-domain CNNs for the recognition of eating and drinking gestures from continuous accelerometer readings (see Figure 6.2). The intuition behind this is the great potential of CNNs to identify the relevant patterns from accelerometer temporal sequences given the translation invariant nature of gestures. In addition, CNNs are domain-knowledge independent since the features are automatically learned through the training step. Such feature learning takes place following a hierarchical structure, whereby the most elementary patterns are captured at the left-most layers, and more complex patterns are learned at the right-most ones. It should be mentioned

that the use of LSTMs was also considered. However, the significantly worse performance seen through initial experiments where LSTMs were employed alone and in combination with convolutional layers as compared to the performance achieved by networks using only convolutional layers, led to focusing the research efforts on optimising CNN architectures.

6.4.5.1 Benchmark Model - 1D CNN

A 1D CNN fed with raw accelerometer data is proposed as a benchmark model. Given the accelerometer time series $x_i^0 = [x_1, ..., x_N]$, where N is the length of the accelerometer segments (in this case, N=394 samples), the output of the convolutional layers is given by:

$$c_i^{l,j} = \sigma\left(b_j^l + \sum_{m=1}^{M} w_m^{l,j} x_{i+m-1}^{l-1,j}\right), \tag{6.9}$$

where l is the layer index, M is the kernel size, w_m^j is the weight for the j^{th} map and m^{th} filter index, b_j^l is the bias term for the j^{th} filter at layer l, and σ is the activation function.

In this case, the activation function employed is the rectified linear unit (ReLU):

$$\sigma(z) = max(0, z) \tag{6.10}$$

Following the convolutional layer, a pooling layer performs a non-linear down-sampling by retrieving the maximum value among a set of nearby inputs. This is given by:

$$p_i^{l,j} = \max_{r \in R}\left(C_{ixT+r}^{l.j}\right) \tag{6.11}$$

where T is the pooling stride and R the pooling size (in this study, 1 and 2 respectively).

Several convolutional and pooling layers can be stacked to form deeper network architectures. The output from the stacked convolutional and pooling layers is flattened to form the feature vector $f^I = [f_1, ..., f_I]$, where I is the number of units in the last pooling layer. f^I is then used as input to the fully-connected

layer:

$$h_i^l = \sum_j w_{ji}^{l-1}\left(\sigma(f_i^{l-1}) + b_i^{l-1}\right) \tag{6.12}$$

where w_{ji}^{l-1} is the connection weight term from the i^{th} node on layer $l-1$ to the j^{th} node on layer l, σ is the activation function (ReLU) and b_i^{l-1} is the bias term.

The output from the fully connected layer is then used as input to the softmax function, by which the gesture classification is computed as:

$$P(c|p) = \underset{c \in C}{argmax} \frac{e^{(f^{l-1}w^L + b^L)}}{\sum_{k=1}^{N_C} e^{(f^{L-1}w_k)}} \tag{6.13}$$

where L is the index of the last layer, c is the gesture class and N_C is the total number of gesture classes.

The network training is conducted using the Adaptive Moment Estimation (Adam) optimiser on batches of 32 accelerometer segments for a total of 30 epochs. Categorical cross-entropy is used as the loss function. A dropout rate of 0.5 is used on the fully connected layer to mitigate overfitting issues.

6.4.5.2 Benchmark Network Optimisation

The performance of the 1D CNN is studied across various key network parameters. These include the number of layers (l), the number of filters within a layer (j) and the filter size (M) as follows:

- l = [1,2,3]
- j = [16,32,64,128,256]
- M = [6,12,25,50,75,100,125,150]

Given the sampling frequency employed for data collection (25 Hz), the filter size ranges from $M = 0.24$ seconds to $M = 6$ seconds. The learning rate employed is 0.001.

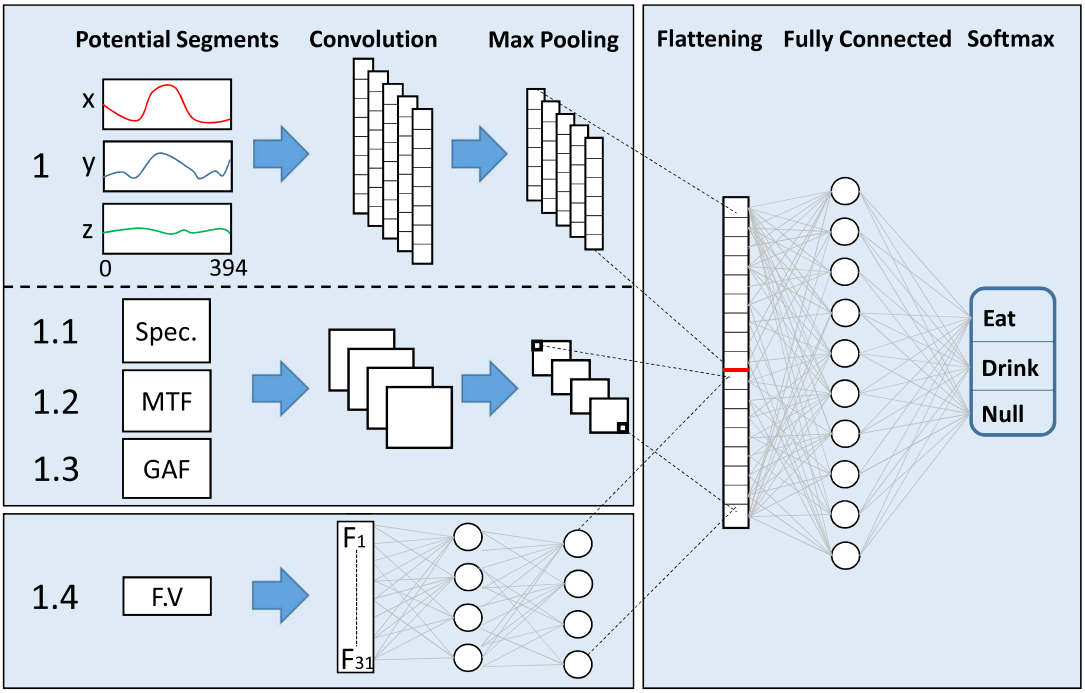

Figure 6.2: Diagram showing the different single-input and multi-input multi-domain networks proposed. It should be noticed that the top part (1) is a common factor on all the proposed networks. The rest of the models are built on top of that one by combining the respective learned features at a common fully connected layer. That is, the features learned by Model 1 in the figure are combined independently at the fully connected layer with the features learned by each of the 1.1, 1.2, 1.3 and 1.4 models after flattening.

6.4.5.3 CNN Frameworks Description

Once the 1D benchmark network is optimised, various multi-input multi-domain networks are built on top to evaluate whether a further improvement on the classification performance can be achieved. The different proposed CNN frameworks are described below (see also Figure 6.2):

- 1. 1D CNN: Optimised 1D CNN benchmark network fed with raw accelerometer data.
- 1.1.1. Spec(Mag): Optimised 1D CNN benchmark network fed with raw ac-

celerometer data combined with a 2-layered 2D CNN fed with spectrogram images of the magnitude of the tri-dimensional accelerometer signal.

- 1.1.2. Spec(y): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with spectrogram images of the y-axis of the accelerometer signal.
- 1.2.1. MTF(Mag): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with MTF images of the magnitude of the tri-dimensional accelerometer signal.
- 1.2.2. MTF(y): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with MTF images of the y-axis of the accelerometer signal.
- 1.3.1. GAF(Mag): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with GAF images of the magnitude of the tri-dimensional accelerometer signal.
- 1.3.2. GAF(y): Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered 2D CNN fed with GAF images of the y-axis of the accelerometer signal.
- 1.4. F.V: Optimised 1D CNN benchmark network fed with raw accelerometer data combined with a 2-layered NN fed with a 31-dimensional hand-crafted feature vector.

The architecture of the 2D CNNs employed for the feature learning of the resultant spectrogram, MTF and GAF images is defined by $l = 2$, $j = 5x5$ and $M = 16$. The framework, including the hand-crafted feature vector (F.V), employs a 2-layered Neural Network (NN) with 16 neurons on each layer. Such a F.V includes a wide range of descriptive signal statistics as well as the duration of the different gestures.

A posteriori, the classification performance of each of the frameworks is evaluated by adopting a leave-one-out cross-validation strategy, whereby on each validation step one of the experiment participants is used as the test set and

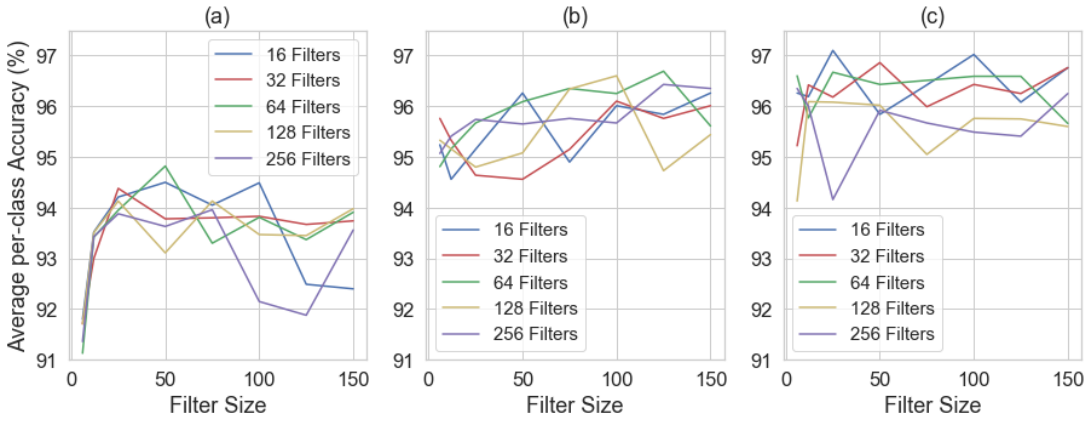

Figure 6.3: Classification performance of the 1D CNN across the parameters l, j and M, where (a) depicts the average per-class classification accuracy of the 1-layered CNN, (b) of the 2-layered CNN and (c) of the 3-layered CNN.

the remaining subjects as the training set. Ultimately, making use of the best performing network architecture, the impact of the learning rate is studied across the following values: [0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1].

6.5 Results

The results achieved by the different CNN-based frameworks for the recognition of eating and drinking gestures are presented in this section. The problem has been tackled as a 3-class classification problem, with the classes being 'Drink', 'Eat' and 'Null'. The 'Null' class embodies all the irrelevant gestures retrieved by the segmentation technique. That is, all the gestures which are not an eating or a drinking gesture.

A parametrically optimised 1D CNN fed with raw accelerometer data is first proposed as a benchmark classification model. Such optimisation is achieved by studying the performance of the network across the number of layers l, the number of filters j and the filter size M.

This can be better observed in Figure 6.3 where the average per-class classification accuracy of the networks is plotted against the different studied parameters. The optimisation process is performed layer by layer. That is, once the values

j and M are optimised for the 1-layered CNN, a second convolutional layer is added to that optimised network. This process is then repeated for the implementation of the 3-layered CNN. From Figure 6.3, it can be seen that while an increase on the classification performance is achieved by increasing the number of layers (this is confirmed by an ANOVA-Tukey HSD test), no direct relationship can be observed between the classification performance and the number of filters or the filter size. Despite the improvement seen on the classification performance achieved by the increase made to the number of layers, further analysis is made by analysing the distribution of the average per-class classification accuracy across the different configurations. As it can be seen in Figure 6.4, the performance distribution exhibited by the 1-layered and the 3-layered CNNs exhibit a negative skewness. This indicates the use of a 1-layered network and that of a 3-layered network for this specific problem can lead to underfitting and overfitting issues respectively. Therefore, a 2-layered network would be recommended as the more conservative architecture for future similar problems where the execution of network optimisation is not possible.

In this case, as shown in Table 6.1, the best average performance across j and M is achieved by the 3-layered CNN with an average per-class classification accuracy of 96.06%. The best classification performance is also achieved using a 3-layered CNN (the configuration is described in the table). Such network achieves an average per-class classification accuracy of 97.10%, an average per-class classification precision of 93.01% and an average per-class classification recall of 93.96%. The results achieved for each specific class are shown in Table 6.2.

After the optimisation of a 1D CNN, the different frameworks proposed in Section 6.4.5.3 are evaluated. The classification performance achieved by each of

Table 6.1: Summary of results. The Avg. perform. (%) column reports the mean of the average per-class classification accuracy across j and M. Acc. (%), Prec. (%) and Rec. (%) report the respective values achieved by the best network configurations described in the Best Configuration column.

1D CNN	Avg. perform. (%)	Best Configuration	Acc. (%)	Prec. (%)	Rec. (%)
1 Layer	93.36	j^1=64 filters, M^1=50	94.82	86.46	90.23
2 Layers	95.59	j^2=64 filters, M^2=125	96.69	91.40	94.28
3 Layers	96.06	j^3=16 filters, M^3=25	97.10	93.01	93.96

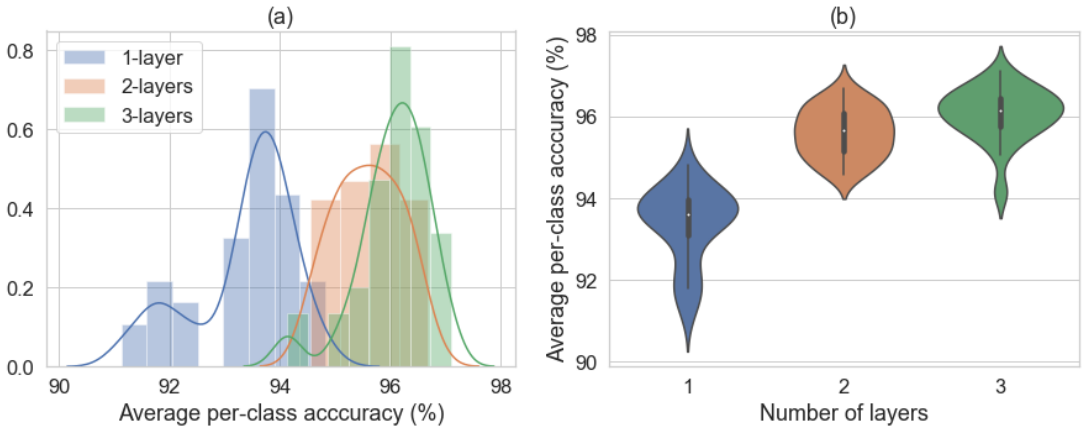

Figure 6.4: Study upon network architecture (number of layers). (a) shows the distribution of the classification accuracies achieved by the 1-layered, 2-layered and 3-layered CNNs. (b) shows the corresponding violin plot.

Table 6.2: Classification metrics for the CNN showing the best classification performance.

	Accuracy (%)	Precision (%)	Recall (%)
Null	96.29	98.60	96.25
Drinking	98.73	92.98	89.83
Eating	96.29	87.43	95.81
Average per-class	97.10	93.01	93.96

the frameworks can be seen in Figure 6.5. The results indicate the benchmark 1D CNN model outperforms the rest of the proposed frameworks, with only the F.V framework obtaining comparable results. Despite the additional implicit information provided by the rest of the frameworks, the required additional complexity of the network led to overfitting issues.

Using the best performing architecture, further hyper-parameter tuning is performed to evaluate the impact of the learning rate on the classification performance of the network. To do so, the classification performance is studied across the following learning rate values [0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1], starting from the highest value (0.1). With this, it is observed that high learning rates, including 0.1, 0.05, 0.01 and 0.005 lead the network to

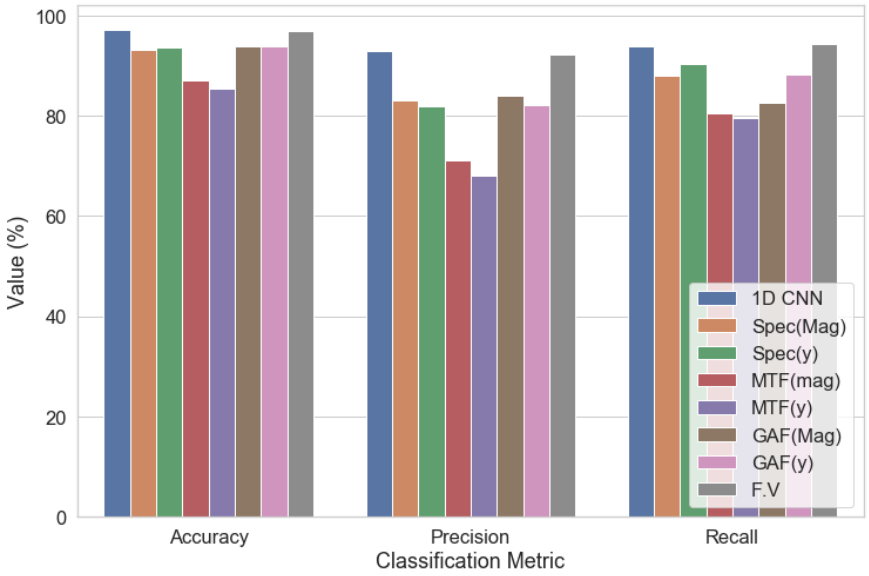

Figure 6.5: Classification performance achieved by the proposed CNN-based frameworks

Table 6.3: Classification performance comparison of the different learning rates.

Performance/ Learning Rate	**0.00001**	**0.00005**	**0.0001**	**0.0005**	**0.001**
Max. Accuracy (%)	86.64	91.78	93.43	94.17	**94.63**
Mean Accuracy (%)	83.44	87.40	89.62	**91.37**	90.78

converge too quickly to a sub-optimal solution whereby all the instances are classified as the dominating class ('Null'). The maximum and the mean classification performances achieved by each of the remaining learning rates are reported in Table 6.3. As it can be seen in the table, the maximum classification accuracy (94.63%) is achieved using the default learning rate (Lr) of 0.001. The classification performance of this network across the different epochs can be seen in Figure 6.6.

6.5.1 Discussion

The CNN-based system proposed addressed the problem of spotting and recognising eating and drinking gestures with the use of a single wrist-worn tri-axial

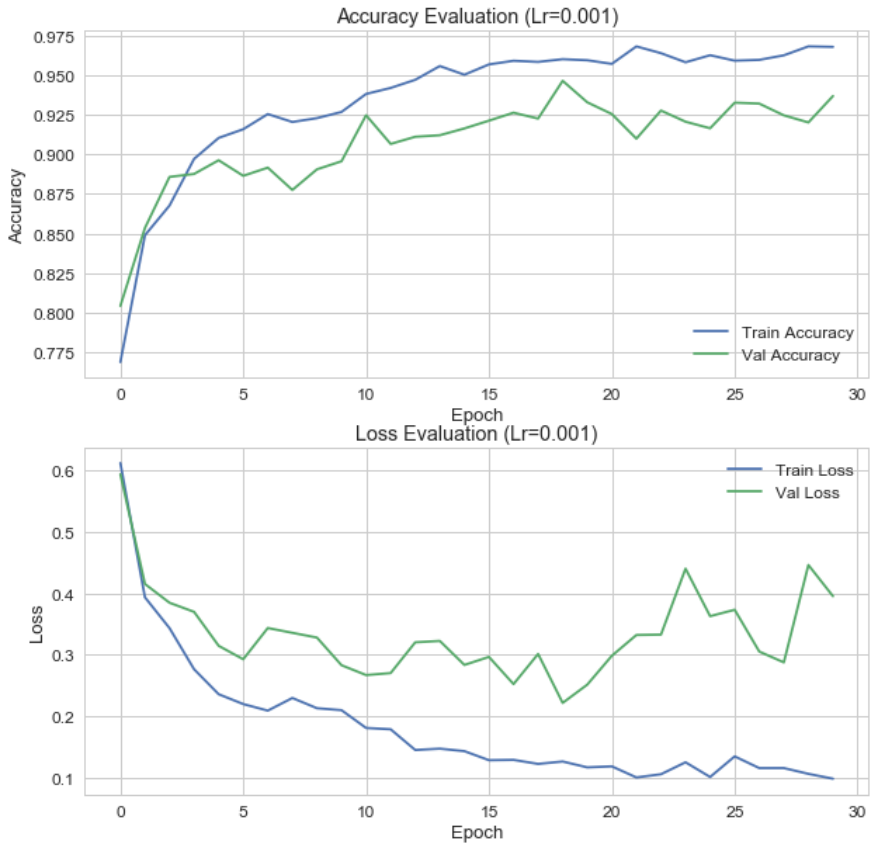

Figure 6.6: Classification performance of the benchmark architecture using the best performing learning rate (Lr=0.001).

accelerometer. As demonstrated in Chapter 5, the adaptive segmentation technique employed (CAST), correctly spots all the eating and drinking gestures embedded in the accelerometer readings. This overcomes the drawback found in previous work at trying to estimate a suitable segment length for the specific classification problem [108, 131].

Despite the efforts given to improve the classification performance of the 1D CNN fed with raw accelerometer data, these mostly led to overfitting issues. However, the satisfactory results achieved outline the suitability of CNNs for gesture recognition. As compared to the system proposed in Chapter 5, the CNN-based system presented in this chapter achieves slightly worse classification performance. However, it must be noted that despite the lower classification performance, two major advantages are identified. First, the CNN-based system

Table 6.4: Comparison of the proposed system to previous work on the recognition of drinking gestures.

Method	Sensor Units	Spot.	Recog.	Accuracy	Precision (%)	Recall (%)
Chapter 5	1	✓	✓	99.0	93.3	93.3
Proposed System	1	✓	✓	98.73	92.98	89.83
Junker *et al* (2008) [123]	5	✓	✓	-	88.0	83.0
Amft *et al* (2010) [64]	1	✓	✓	-	84.0	90.0
Chen *et al* (2017) [140]	1	X	✓	-	96.5	91.3
Serrano *et al* (2017) [170]	4	✓	✓	-	82.28	84.42
Ramos-Garcia *et al* (2014) [171]	1	✓	✓	86.5	-	-

is domain-knowledge independent. Second, the presented system only makes use of accelerometer data, whereas the system proposed in Chapter 5 makes use of both accelerometer and gyroscope data. Table 6.4 shows the comparison of the system proposed with similar work based on controlled or semi-controlled experiments run in laboratory settings.

6.6 Conclusions

This chapter has presented a CNN-based system to address gesture recognition with a case study on eating and drinking. First, an adaptive segmentation technique, namely the CAST, was employed for spotting potential eating and drinking gestures within the continuous accelerometer readings. This technique exhibits a 100% spotting recall, therefore overcoming the drawbacks found in previous literature, where true positives are missing at this preliminary step. This is crucial since the errors taking place at this step propagate to the classification step, therefore affecting the overall performance of the system.

Further to signal segmentation, a thorough study on CNNs for eating and drinking gesture recognition was undertaken. A 1D CNN fed with raw accelerometer data was parametrically optimised and proposed as a benchmark classification model. The best classification results were obtained with a network architecture composed of 3 convolutional layers with an overall per-class classification accuracy of 97.10%. However, certain architectural configurations of the 3-layered CNN,

show symptoms of model overfitting. Thus, it is crucial not to assume complex networks will perform better and keep an adequate balance between the complexity of the network, the data available and the complexity of the classification problem itself.

Further to defining a 1D CNN benchmark classification model, various attempts were made to enrich the feature learning process performed through such a benchmark model. These included the use of various 2D CNNs fed with the resultant images obtained by the employment of three different time series to image encoding frameworks, as well as a NN fed with a 31-dimensional handcrafted feature vector. A posteriori, the above feature learning techniques were combined with the resultant features of the benchmark network at a common fully connected layer. Despite the good performance exhibited by the employed time series to image encoding frameworks in different applications such as audio analysis [175] or EEG-based sentiment classification [173], in this case, their use did not lead to a better classification performance when added to the 1D benchmark network. The model incorporating the 31-dimensional feature vector did not improve the classification performance of the benchmark model either. This suggests problems of model overfitting may occur when fitting excessive information into the network. Thus, it can be concluded that raw accelerometer data alongside the use of a 1D CNN is the preferred solution, since it offers an adequate balance between underfitting and overfitting, leading to a better classification performance when unseen data is fed into the network. In addition, as shown by the differing results achieved by the employment of different learning rates, the tuning of this hyper-parameter should be carefully considered in future activity classification work using deep learning models, since the employment of an excessively high learning rate may lead to a quick convergence at a sub-optimal solution, and the employment of a excessively low learning rate may lead the network to get stuck at a sub-optimal local minimum.

Overall, the results obtained suggest that the use of domain-knowledge independent CNNs for the recognition of eating and drinking gesture is plausible. However, attention should be given to overfitting issues, since these can significantly compromise the performance of the network.

Chapter 7

Identification of Meals Intake Through Gesture Distribution

7.1 Introduction

In previous chapters, the recognition of food and fluid intake gestures across periods of eating was studied. Following this, this chapter exploits the good performance exhibited by the previously proposed algorithms at recognising eating and drinking gestures in a semi-controlled environment, to investigate whether that can be further taken advantage of to propose a means of identifying the intake of the main daily meals, namely breakfast, lunch and dinner, under free-living conditions.

Results reported by nutrition research work [176] suggest that bad eating habits such as eating before bedtime are risk factors for new-onset hypertension in older seniors. In addition, as reported in [177], regular eating habits reduce the incidence of mental illness in older population groups. This means that adequate and balanced nutrition is decisive not only to maintain good physical health but also to avoid undesired psychological problems. Dietary behaviour thus plays a vital role in our day to day lives and health. While obesity is a significant risk factor for heart diseases, stroke, high blood pressure or diabetes [178], malnutrition is considered as a confounding factor for developing chronic diseases [74]. As of now, dietary behaviour is usually tracked in the form of

self-assessment questionnaires. However, two major drawbacks are found in the use of conventional dietary tracking approaches. First, the data entry process may result cumbersome, since questionnaires have to be filled manually by the subjects typically. Second, numerous studies indicate self-reported estimates of daily activities are subjective and variable [179, 180].

Given this, the investigation of unobtrusive alternative ways of monitoring personal dietary behaviour would be highly beneficial to understand the dietary needs of older adults living independently. In line with this, this chapter studies the distribution of eating gestures across time by investigating the suitability of low computational cost signal processing techniques, namely an entropy measure and a moving average, to identify the intake of the main meals across continuous free-living recordings. The intuition behind the employment of these two signal processing techniques is that the number of eating gestures is expected to vary and be higher when a meal takes place. In other words, it is expected that outside meal periods, eating gestures will occur less often, therefore leading to a potentially lower variation in the number of eating gestures as well as to a lower number of them. By contrast, when a meal takes place, a greater number of eating gestures caused by the intake of the meal is expected. This should potentially lead to an increase in both the variation (or unpredictability) as well as in the count of the gestures which at the same time should be reflected by the entropy measure and the moving average respectively.

Two different approaches are proposed to determine whether those periods actually correspond to the intake of a meal; 1) A threshold-based approach. 2) A 2-class classification problem ('Meal' Vs 'Non-meal') for which a low-dimensional feature vector incorporating two features is proposed.

The remainder of this chapter is organised as follows: Section 7.2 reviews relevant work on the detection of meal intake and eating periods using inertial sensors. Section 7.3 presents the motivation behind the work undertaken in this chapter. Section 7.4 presents the proposed methodology for the detection of meals. Section 7.5 presents the results achieved and discuss the performance achieved by the proposed approach. Ultimately, Section 7.6 reports the conclusions drawn from the obtained results.

7.2 Review of Work on the Recognition of Eating Periods

As compared to the work undertaken on the recognition of other daily activities, a much narrower body of research has studied the recognition of periods of eating based on the signals provided by wearable motion sensors.

In [9], a two-stage approach for spotting periods of eating using data from a single wrist-worn inertial sensor is proposed. First, a custom-peak algorithm based on wrist motion energy is used as a means of identifying potential periods of eating. The intuition behind this approach is that periods of eating are preceded and followed by periods of larger wrist motion energy caused by pre-meal food preparation-related actions and post-meal tidying up-related actions. Once the potential periods of eating are identified, an array of four features extracted across those periods is used to train a Naive Bayes classifier, by which a classification recall of 86.2% and a classification precision of 20.9% are achieved.

The work in [85] combines a traditional activity recognition approach (a sliding window segmentation technique with a window length of 10 seconds, a range of 10 hand-crafted features and a range of classification models) with a set of ad-hoc restrictions to identify eating windows of 1 minute across free-living recordings of wrist tri-axial acceleration and angular velocity data. Among the ad-hoc restrictions proposed to enhance the performance given by the classification models, it is assumed, as in [9], that periods of eating are preceded by periods of higher wrist motion energy. The authors also assume the duration of each meal should be at least 5 minutes. In addition, a time-based probability function with regards to the time at which each instance occurs is used to classify the eating gestures. The classification is assessed in 1 minute periods, whereby an eating period is only considered when three or more out of the six windows within that minute are classified as eating. By following this approach, the authors achieve a classification precision of 64% and a classification recall of 69%.

In [84], the recognition of eating periods is attempted in a two-fold process using data from a wrist-worn tri-axial accelerometer. With this approach, eating gestures are firstly classified using a sliding window segmentation alongside a set of five hand-crafted features and a Random Forest classification model. A posteriori,

the density of those gestures are studied across time using the Density-based Spatial Clustering (DBSCAN) to predict whether a meal has taken place over a window of time. To do so, different window sizes are employed, with windows of 60 minutes (50% overlap) achieving the best performance. The results using such window size report a classification precision of 66.7% and a classification recall of 88.8%.

7.3 Motivation

As mentioned above, only a narrow body of literature has concerned the recognition of periods of eating using inertial sensors under free-living conditions. In addition, within those works, various limitations can be outlined. For instance, in work by [9] it is assumed that periods of eating are preceded and followed by periods of larger wrist motion energy caused by the tasks related to the preparation of the food and the tiding and cleaning related post-meal tasks. This idea is also incorporated in most recent work by [85] to prevent false positives. While this may be the natural behaviour for individuals with some specific cultural background, this approach may lack generality when applied to societies in which a meal may be simply composed by a ready-to-eat food item (*e.g.* a sandwich). In addition, the performance exhibited by the research works attempting the recognition of periods of eating [9, 84] still lies far away from those achieved by works attempting the recognition of other quotidian activities. Specially, very moderate values for the classification precision are reported across the different works studying the recognition of eating periods.

This confirms the challenges exposed in Section 1.4 with respect to the differing nature and structure of activities. Unlike continuous quasi-periodic activities, an eating activity can incorporate a large variety of gestures and actions, which result in a significant challenge when attempting the modelling of the activity as a whole. In addition, the moderate classification precision reported by previous work suggests that, in free-living conditions, gestures which share some similarity with eating gestures such as face touching or smoking gestures, can be easily confounded with eating gestures.

Overall, the above suggests that there are still many opportunities for further

research and the exploration of novel approaches to recognise eating periods from inertial signals. In line with this, the remainder of this chapter exploits the good performance exhibited by the gesture recognition approach proposed in Chapter 5 in a semi-controlled lab-based experiment to explore alternative means of tackling the challenging task of recognising the intake of meals in free-living conditions.

7.4 Methods

This section presents the steps undertaken to develop the proposed periods of eating detection system based on the distribution of the occurrence of eating gestures across time. To do so, the gesture recognition system proposed in Section 5.4.4.4 is employed to tackle the recognition of eating gestures embedded in the experimental data streams. The predictions made by the recognition system are used to transform the collected signals into binary time series where a non-eating gesture is represented by a '0' and an eating gesture is represented by a '1'. With this, a moving average and an entropy measure are employed as a means of detecting potential segments containing a meal period by accounting for the dissimilarity in the number of eating gestures identified across consecutive minutes. A posteriori, a threshold-based approach and a classification based-approach proposed to predict whether each potential segment retrieved by the signal processing techniques contains a meal period. The intuition behind this approach is that unlike other quotidian activities of continuous nature, periods of eating can embody a variety of actions and movements which do not correspond to the gestures of interest. That is, between two consecutive eating gestures, an individual may perform other actions unrelated to the activity of interest. The different stages of the proposed system are illustrated in Figure 7.1 and further described in detail below.

7.4.1 Experimental Setup

The work in this chapter is based on the Dataset 3 presented in Section 3.2.2.3. This dataset, collected in free-living conditions, comprises the tri-axial acceleration and tri-axial angular velocity of the wrist of the different participants who

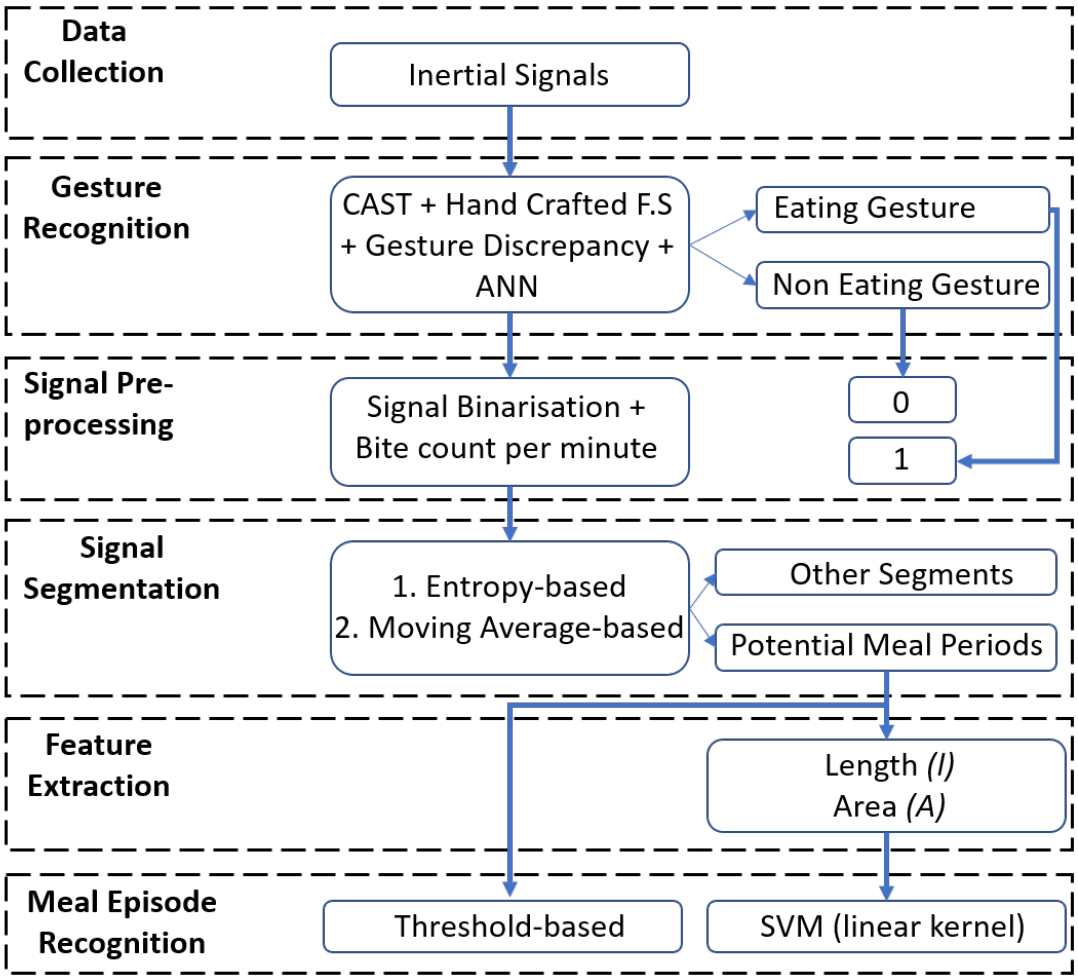

Figure 7.1: Schematic diagram of the proposed methodology to recognise periods of eating.

were asked to annotate the beginning and the end of their main periods of eating during the day, namely breakfast, lunch and dinner.

7.4.2 Gesture Recognition

The first step of the proposed approach to recognise periods of eating from the collected wrist-worn tri-axial accelerometer and tri-axial gyroscope signals is the recognition of eating gestures from the continuous data streams. To do so, the computational solution for gesture recognition proposed in Section 5.4.4.4 is em-

ployed. This is based on the a better classification performance on the recognition of eating gestures exhibited by this approach as compared to the rest of the gesture recognition approaches proposed in Chapters 5 and 6. In particular, it is important to notice the significantly higher classification precision (95.7%) achieved by this approach as compared to that achieved by the best performing CNN-based model (87.43%) for the 'Eating' class. As the analysis of the results achieved by previous work [9, 84, 85] suggest, the classification precision is a key challenge for identifying eating periods in free-living conditions. Therefore, this should be considered carefully at this preliminary step.

As a remainder, the selected computational solution incorporates the CAST as a signal segmentation mechanism, a range of hand crafted features and the proposed gesture discrepancy measure as the feature set, and an ANN as the classification model. The performance achieved by this computational solution at recognising eating gestures during a meal period can be seen in Table 5.3

With this, the collected time series, corresponding to the wrist tri-axial acceleration and angular velocity of the wrist of the experimental participants, are converted into sequences of labels embodying the following classes: 'Null', 'Drinking' and 'Eating', which correspond to the gestures predicted by the employed gesture recognition approach.

7.4.3 Signal Pre-processing

The different predicted gestures are utilised to build a binary sequence of gestures $G[n]$ as follows:

$$G[n] = \begin{cases} 1, \text{if } n = \text{'eating'} \\ 0, \text{otherwise} \end{cases} \tag{7.1}$$

that is, a '0' is assigned to the gestures not predicted as being an eating gesture, and a '1' is assigned to those gestures predicted as being an eating gesture. For illustration purposes, a visual example of a binarised signal corresponding to a lunch period can be seen in Figure 7.2.

A binary sequence of eating gestures $G[n]$ corresponding to the predictions given by the gesture recognition system can then be encoded into a time series $S = [s_i : 1 \leq i \leq N]$ where each s_i is the number of predicted eating gestures

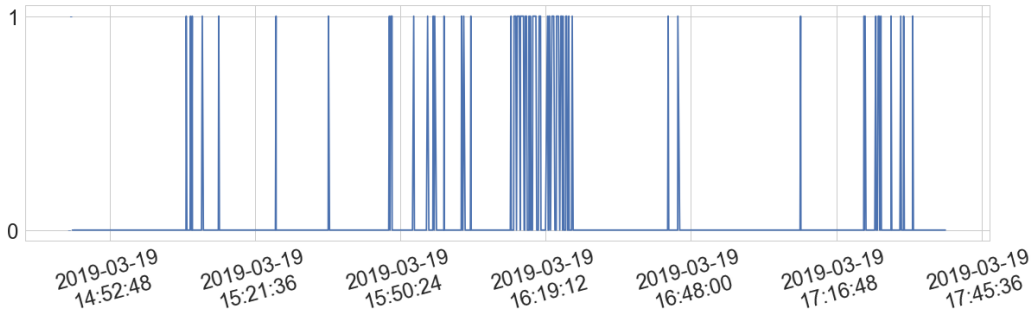

Figure 7.2: A binarised segment of a signal corresponding to the classifications made by the gesture recognition system, where a '1' indicates an eating gesture has been identified. The binarised signal corresponds to a reported lunch from 16:14 to 16:26

at a period of time i. For the purpose of the application studied in this work, periods of 1 minute were considered. With this, the binary signals are encoded so that values in S are equally spaced in time, representing, therefore the number of eating gestures identified per minute. The intuition behind the employed encoding method is that, as outlined in Section 1.4, an activity can be broken down into smaller actions or gestures which exhibit a clear connection with the activity itself. For instance, the activity eating a sandwich can be divided into multiple repetitive gestures which involve taking the sandwich towards the mouth to take a bite. In this context, the continuous occurrence of eating gestures across time can be associated with the occurrence of an eating activity. An example of a resultant encoded time series can be seen in Figure 7.3.

7.4.4 Identification of Potential Meals

As mentioned above, it is hypothesised that the number of gestures is more variable and higher during meal periods than outside such periods. Given these hypotheses, two distinct methods, namely a moving average and an entropy measure, are explored as a means of identifying the hypothesised higher number of gestures and higher variation respectively. This is ultimately aimed at the translation of these variables into the recognition of segments containing meal periods within the resultant time series S. As suggested in [85] an eating activity is not expected to last less than 5 minutes. Based on this, a window size of 5 was used

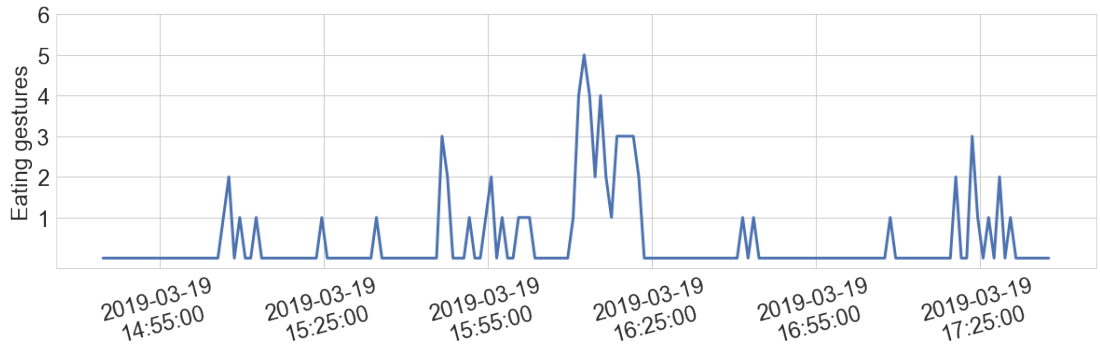

Figure 7.3: Encoded time series representing the number of eating gestures per minute predicted by the gesture recognition system. The plotted time series corresponds to a reported lunch from 16:14 to 16:26.

for the calculation of both the entropy measure and the moving average. Further details about these two techniques are presented below.

7.4.4.1 Approximate Entropy

Entropy can be defined as a measure that reflects the lack of order or predictability of physical property or a signal, which can be employed to estimate the uncertainty or degree of randomness in a system. The intuition behind the use of entropy to spot potential periods of eating is that it is expected that before, during and after the areas of interest, eating and non-eating gestures will alternate, resulting in an increase in the unpredictability of S. By contrast, S is expected to remain more stable outside periods of eating due to the lack or limited occurrence of eating gestures. In this context, the Approximate Entropy (ApEn) [181], is employed to identify a potential dissimilarity between periods of eating and non-eating. The selection of ApEn to measure the disorder in the sequences of gestures is motivated by the good performance exhibited by this entropy measure at reflecting the uncertainty present in binary signals pre-processed in a similar way [30].

Given a time series $S = [s_i : 1 \leq i \leq N]$ with N samples, a non-negative integer $m \leq N$ which represents the length of the blocks of data to be compared, and a positive real number r which represents the tolerance or filtering level, the

sequences or blocks of the vector S_i^m can be expressed as:

$$S_i^m = [s_i, s_{i+1}, ..., s_{i+m-1}], \text{for i = 1, 2, ...,(}N\text{-}m\text{+1)} \tag{7.2}$$

The distance between two sequences S_i^m and S_j^m is calculated as the maximum difference between their components given by:

$$d[S_i^m, S_j^M] = \max_{k=0,1,...,m-1}(|s_{i+k} - s_{j+k})|) \tag{7.3}$$

For each S_i^m, the number of $j \leq N - m + 1$ such that $d[S_i^m, S_j^m] \leq r$, expressed as $N_i^m(r)$, is used to calculate the parameters $C_i^m(r)$ as:

$$C_i^m(r) = \frac{1}{N - m + 1} N_i^m(r) \tag{7.4}$$

The mean value of the parameters $C_i^m(r)$ is given by:

$$\phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln C_i^m(r) \tag{7.5}$$

ultimately, with $\phi^m(r)$ and $\phi^{m+1}(r)$, the ApEn is calculated as:

$$\text{ApEn}(m, r, N) = \phi^m(r) - \phi^{m+1}(r) \tag{7.6}$$

where $N = 5$ in this case, which corresponds to the length of the 5 minutes window proposed. An example of the use of ApEn to identify the variations in S is shown in Figure 7.4.

7.4.4.2 Moving Average

As an alternative to ApEn, a moving average is employed to identify segments of the signal containing a meal period. The intuition behind the use of a moving average is the expected increase in the number of identified eating gestures during meal periods. Given this, the moving average is expected to reflect such increase while filtering out false positives, that is, non-eating gestures classified as being eating gestures. The moving average of S across windows of 5 minutes is given

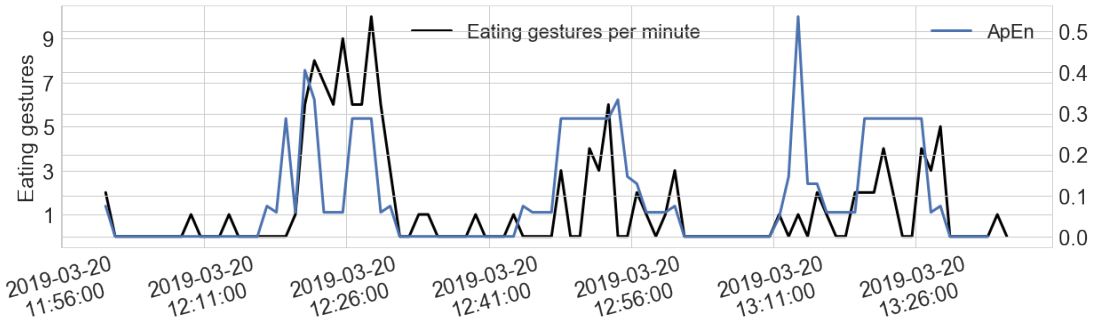

Figure 7.4: Encoded time series representing the number of eating gestures per minute predicted by the gesture recognition system. The plotted time series corresponds to a reported lunch from 12:23 to 12:30.

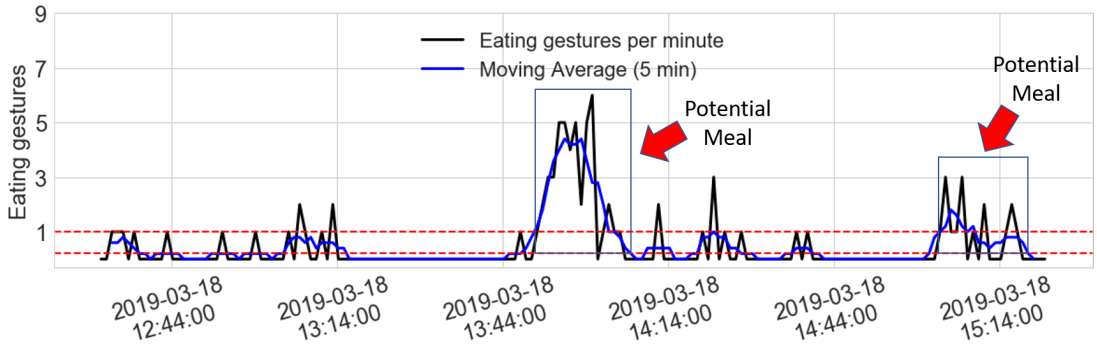

Figure 7.5: Encoded time series representing the number of eating gestures per minute predicted by the gesture recognition system and the moving average calculated across windows of 5 minutes. The plotted time series corresponds to a reported lunch from 13:53 to 14:02.

by:

$$\bar{S}[t] = \frac{1}{n}\sum_{i=0}^{n-1} S[t+i] \tag{7.7}$$

where in this case, $n = 5$.

7.4.4.3 Observations

After preliminary observations with different configurations for m and r, it is noticed the ApEn is excessively sensitive to the expected moderate gesture recognition precision in free-living conditions, showing a counterproductive behaviour

to the interests of the application of this work (see Figure 7.4). In contrast, by looking at Figure 7.5, it can be observed that the moving average $\overline{S}$ complies with two key aspects for it to be used as a means of identifying potential meal periods. On the one hand, $\overline{S}$ acts as a filter to sparse false positives. On the other hand, $\overline{S}$ reacts in consonance with the expected increase of eating gestures during meal periods. Based on the above, it was decided to explore S further while dismissing the use of ApEn.

7.4.5 Recognition of Meal Periods

Two different approaches are proposed to recognise meal periods based on the values of $\overline{S}$, namely a threshold-based approach and a classification-based approach. These are presented below.

7.4.5.1 Threshold-based approach

The threshold-based approach studies the performance of the system across the following ingestion rates given in bites (eating gestures) per minute $[1, 2, 3]$. In this context, a segment is predicted as a meal period if the moving average $\overline{S}$ crosses over the corresponding threshold ingestion rate and remains above for at least two minutes. The labelling of each segment above the threshold was done considering the span of $\overline{S}$. A segment is therefore labelled as a 'meal' if there is an overlap between the area where $\overline{S} > threshold$ with the area delimited by the reported beginning and end times of the meal by the corresponding experimental participant, considering the span of $\overline{S}$.

7.4.5.2 Classification-based Approach

Following the same labelling criteria, a classification-based approach is proposed. To do so, a 2-dimensional feature vector is computed to train a range of three state-of-the-art classification models. In this case, a unique ingestion rate of '1' is used to delimit the segments to evaluate the performance of the system. This is based on the higher recall obtained as compared to the corresponding metrics using higher ingestion rates. In order to account for possible meal interruptions, it seems reasonable to state that 5 minutes without the occurrence of an eating

gesture would mean the end of a meal. Given this, a threshold of $\overline{S} < 0.2$ was used to consider the end of a segment. This means, in the case of two consecutive segments where $\overline{S} > 1$, these are joined together.

Feature Extraction - The 2-dimensional feature vector proposed incorporates the segment length (l) and the segment area (a) given by the sum of all values of $\overline{S}$ where $\overline{S} > 1$. The intuition behind the use of these features is that meal periods are expected to have a longer duration and a higher number of eating gestures as compared to those segments containing false positives or the intake of little snacks. Based on this, both l and a are expected to show greater values within meal periods than outside these.

Classification - The feature set above is used to train an SVM classification model with a linear kernel. The intuition behind this is that the boundary between eating and non-eating periods is expected to be linear, potentially classifying those with longer lengths and higher areas as meal periods. The problem is tackled as a binary classification problem, where each area over the threshold is classified as either being a 'meal period' or a 'non-meal period'. A 10-fold cross-validation strategy is adopted to evaluate the performance of the classification model.

7.5 Results

The results of the threshold-based and the classification-based approaches for the recognition of eating periods are presented in Sections 7.5.1 and 7.5.2 respectively. A posteriori, a discussion upon the results are provided in Section 7.5.3.

7.5.1 Threshold-based Approach

As aforementioned, the threshold-based approach is evaluated across a range of different ingestion rates. The results obtained by the different the use of the different thresholds is provided in Table 7.5.1.

7.5.2 Classification-based Approach

As aforementioned, an SVM classification model with a linear kernel is trained with the proposed 2-dimensional feature vector to predict whether segments above the threshold (ingestion rate = 1 bite per minute) are caused by eating gestures within meal periods. The false negatives caused by the employment of the threshold are added up to the ones resulting from the evaluation of the classification model. With this, adopting a 10-fold cross-validation strategy, a classification precision of 93.10% and a classification recall of 65.85.% are achieved.

7.5.3 Discussion

The results achieved in this chapter indicate that the approach proposed to recognise the intake of meals through the study of the distribution of eating moments across time using a signal processing technique with low computational cost, namely a moving average, has a great potential to be employed as means of monitoring dietary behaviour across the day.

A commonly occurring problem was found across the literature with regards to showing certain imbalance between the classification recall and the classification precision, with the latter showing rather moderate values. This suggests that 1) gestures alike eating gestures, such as face touching gestures or smoking gestures may be easily confounded with eating gestures, 2) It is a challenge to build datasets in lab-based semi-controlled environments to later make predictions in free-living conditions with unseen environments.

As demonstrated by the evaluation made with the threshold-based approach with different ingestion rates, the classification precision and the classification recall can be reasonably balanced. This indicates two main points: 1) the gesture

Table 7.1: Classification metrics for the recognition of meal periods across the different ingestion rate-based threshold values.

Threshold	Precision (%)	Recall (%)
1	56.10	90.2
2	78.04	71.1
3	92.60	61.98

recognition model trained in a semi-controlled environment was able to generalise competently on the challenging task of spotting and recognising sparse gestures in differing free-living conditions. 2) The moving average-based approach proposed to model eating periods as a distribution of eating gestures across time was able to filter out to a high degree the false positives, potentially coming in the form of eating-alike gestures such as face touching gestures. According to work in this area [182], the frequency of face touching can be as high as 153 times per hour implying, therefore, face touching can occur as often as more than two times per minute.

The results obtained with the linear SVM classification model trained with the length and the area of the segments above the lower ingestion rate-based threshold (1 bite per minute), with a high classification precision as compared to the classification recall, suggest that the varying duration and abundance of different meals can play an important role on the imbalance shown by these two metrics. This is rather coherent, since for instance, the amount of food ingested at different meals may differ significantly between individuals, as well as between meals. Here, it should be noted that the experimental participants were only asked to annotate the beginning and the end of the main meal periods, namely breakfast, lunch and dinner. This implies that other traditional eating periods, such as the intake of pre-meal snacks or post-meal coffee/tea and sweets, if took place, were not annotated.

Overall, the results achieved suggest that the proposed approach alongside the previously proposed gesture recognition system signifies a great contribution towards the development of dietary behaviour monitoring with the use of a single inertial sensor.

7.6 Conclusions

In this chapter, a novel approach to recognise the intake of meals under free-living conditions has been proposed. Such an approach has been based on the distribution of eating gestures across time, which was expected to show higher and more variable values across the periods of time containing an intake of a meal. The system proposed can be divided into two different steps. First, the food and fluid

intake gesture recognition system proposed in Section 5.4.4.4 was employed to recognise the different eating gestures across the wrist inertial recordings collected in free-living conditions. Second, the count and the variation of the gestures across time was investigated by the employment of a moving average and an entropy measure respectively. This was based on the intuition that eating gestures should occur more often inside meal periods, therefore causing a potential increase in both the count and the variation and consequently in the moving average and the entropy measure. While the moving average based on the count of gestures per minute was found to be a suitable mechanism to identify potential periods of eating, the entropy measure was shown to be too sensitive to the expected moderate precision at recognising eating gestures in free-living conditions. Thus, further efforts were given to the recognition of meal periods based on the use of the moving average. With this, two distinct methods, namely a threshold-based approach and a classification-based approach, were explored.

From the promising results achieved, it can be concluded that the recognition of the intake of meals across free-living scenarios is plausible with the use of a single wrist-worn inertial sensor. In this context, a set of recommendations for further work are provided in the future work Section 8.4.

Chapter 8

Conclusions and Future Work

8.1 Thesis Summary

The work undertaken in this thesis has presented a range of novel frameworks for human activity and gesture recognition with the aim of enhancing the usability of wearable technologies in the field of Ambient Assisted Living. The motivation behind this work is the ageing population structure and the worrying upwards trend shown by the number of older adults needing peripheral support during their quotidian activities. Consequently, the efforts in the thesis have been directed to the recognition of quotidian activities concerning essential personal needs, namely hygiene and nutrition.

To answer the research question made in Chapter 1: "How can a single wrist-worn motion sensing unit be used to recognise quotidian activities concerning self-neglect issues?", efforts throughout this work have been given to investigate how the use of signal processing and Computational Intelligence techniques can relate the acceleration and angular velocity of the dominant wrist of individuals with the activities of interest. As a summary, in Chapter 4, the recognition of hygiene-related activities was explored. Chapters 5 and 6 explored the recognition of food and fluid intake gestures. Ultimately, Chapter 7 studied the recognition of the main meals across the day, namely breakfast, lunch and dinner, based on the distribution of food intake gestures across time.

8.2 Concluding Remarks

The work in this thesis has demonstrated the plausibility of the development of accurate frameworks for the recognition of quotidian daily activities based on the use of a single wrist-worn inertial measurement unit. The efforts made to explore novel ways of segmenting inertial sensory signals and the extraction of novel relevant features have led to the improvement on the classification performance achieved by some of the existing methods across different activity classification problems. Given the recognition rates achieved throughout the different experimental chapters, and the previously reported acceptability of wrist-worn sensory devices to monitor human activities in older population groups, it can be concluded that these devices exhibit a great potential to be employed as "all-day" monitoring mechanisms for older adults living independently. The conclusions for the various aspects covered in this project are presented below.

8.2.1 Recognition of Quotidian Quasi-Periodic Activities

The results reported in Chapter 4 suggest that quotidian quasi-periodic activities can be accurately recognised from accelerometer data with the use of an artificial segmentation technique, namely the sliding window approach, alongside long-established hand-crafted features and state-of-the-art classification models. In addition, form the multi-level refinement approach proposed, it can be concluded that feature informativeness depends on the activity set chosen to be studied. Given this, feature refinement may be used in activity recognition problems, especially in applications where the recognition of a specific activity is crucial for the interest of the application (*e.g.* fall detection).

8.2.2 Gesture Recognition Through the Use of Hand-Crafted Features

The results reported in Chapter 5 suggest that fluid and food intake gestures can be accurately recognised from continuous wrist inertial recordings in a meal context. Such results also indicate that the use of the proposed Soft-DTW gesture discrepancy measure can lead to a notable gain in classification performance in

experiments embodying the recognition of similar gestures. This outlines the reliability of the Soft-DTW based gesture discrepancy measure as a feature descriptor in human activity and gesture recognition problems. In addition, the good performance exhibited by the proposed framework outlines the crucial role of the development of accurate problem-specific adaptive segmentation techniques in the spotting of sparsely occurring gestures in continuous data streams. Although widely-used sliding windows exhibit good performance on the recognition of continuous quasi-periodic activities, their employment for spotting sporadic gestures may lead to the loss of fundamental characteristics of the gestures or to the incorporation of unwanted signal fragments which can potentially compromise the recognition performance of a gesture recognition system.

8.2.3 Exploring Deep Learning Techniques for Gesture Recognition

This work was undertaken to explore alternative ways of recognising eating and drinking gestures from inertial recordings to those proposed in Chapter 5. The results reported in Chapter 6, with a slightly worse classification performance to that achieved in Chapter 5, corroborate the potential of the employment of convolutional neural networks to undertake human activity and gesture recognition problems without the need of previous domain-specific knowledge. The undertaken study upon network complexity and the incorporation of additional features to those automatically extracted from raw recordings indicates the use of deep complex architectures or that of further features may not necessarily lead to better classification performance. Instead, problems of overfitting may occur. Given this, a recommendation to future work in the field using CNNs is first to explore the performance of shallow networks with two or three convolutional layers fed with raw data before trying to tackle the problem with complex network architectures or further data inputs.

8.2.4 Identification of Meals Intake through Gesture Distribution

The recognition of eating periods or that of the intake of meals is a very challenging task due to various reasons. First, the intake of a meal can take place in distinct environments (e.g. on dining tables of different heights, no dining table, etc.). Second, as opposed to continuous quasi-periodic activities such as walking or running, the intake of a meal is composed of sparsely occurring gestures. Therefore, to model the intake of a meal is crucial first to be able to spot and recognise those gestures from continuous data streams. The results reported in Chapter 7 outline the plausibility of recognising eating periods in free-living conditions by the analysis of the distribution of eating gestures across time. In addition, it was demonstrated that through the use of a low-cost signal processing technique, a balance between the classification precision and the classification recall could be achieved. Two main conclusions can be drawn: 1) Despite the difficulty of training accurate models based on datasets collected in controlled or semi-controlled lab-based environments to making predictions in free-living conditions, the gesture recognition model was able to generalise competently on the challenging task of spotting and recognising sparse gestures in differing free-living conditions. 2) The moving average-based approach proposed to model eating periods as a distribution of eating gestures across time was able to filter out to a high degree the false positives, potentially coming in the form of eating-alike gestures such as face touching gestures.

Overall, it has been demonstrated that the recognition of the intake of meals based on the distribution of eating gestures across time is feasible. To achieve further advances in the recognition of complex activities such as meal intakes, getting classification performances closer to those achieved on the recognition of continuous quasi-periodic activities, a set of recommendations are provided in Section 8.4. This will allow the potential of both the gesture recognition and the meal intake detection approaches to be further exploited.

8.3 Contributions

The major contributions of this thesis are outlined below:

- Propose accurate means of recognising hygiene-related activities among other quotidian activities by the use of an artificial segmentation technique alongside hand-crafted features and state-of-the-art classification models.
- Propose a novel multi-level refinement approach to optimise the selection of features for activities which exhibit a lower classification performance as compared to that presented by the overall activity recognition system. As demonstrated, the employment of this approach can lead to an improvement in the activity recognition system's classification rate.
- Propose a novel adaptive signal segmentation technique (CAST) for spotting potential eating and drinking gestures within continuous motion data streams. This technique, with a 100% classification recall, overcomes the main drawbacks encountered in attempting the spotting of sparse and duration varying gestures with artificial signal segmentation techniques which divide the continuous data streams into windows of equal length. Given the outstanding performance and the flexibility of this technique in terms of adjusting the moving averages according to the application needs, the CAST may be used in future activity and gesture recognition work.
- Propose a novel DTW-based gesture discrepancy measure as a feature descriptor to enrich the information gained through the extraction of long-established hand-crafted feature vectors. As demonstrated in this work, the use of the gesture discrepancy measure consistently improves the gesture recognition rate across different experiments. This supports its employment as a feature descriptor in future activity and gesture recognition work.
- Undertake a thorough investigation upon the use of convolutional neural networks for gesture recognition by which the impact of the network complexity and the use of additional features from those automatically extracted

by the network were studied. With this, the use of CNNs can be recommended for future gesture and activity recognition work, especially when a lack of specific domain-knowledge exists.

- Propose novel frameworks for the recognition of similar eating and drinking gestures during an eating period which show a great performance at such a challenging task.

- Propose novel means of identifying meal periods in free-living conditions through the analysis of the occurrence of eating gestures across time by the use of a low-cost signal processing technique.

8.4 Future Work

Following the work undertaken in this thesis, this section outlines the main directions for future work:

- The field of HAR with the use of wearable sensors continues to receive increasing efforts by the research society. However, as outlined in Section 2.3, most of these efforts are directed towards fitness application. Further work concerning quotidian activities which can be somehow associated with the mental and the physical health of older individuals living independently is still needed in order to employ wearable inertial sensors as standalone mechanisms for continuous monitoring. In addition, it is proposed the exploration of action variability. This may help to identify common patterns in the data, allowing for the optimisation of the signal-processing techniques employed in an activity recognition work.

- Research work concerning the recognition of quasi-periodic activities has often reported classification accuracies in the range of 95%-100%. On the other hand, the performance reported by previous work aiming at the recognition of sparse gestures such as food and fluid intake gestures or at the recognition of meal intakes still lies far from these figures, specially, when experiments are run in free-living conditions. An effort to improve the performance shown in previous studies have been made in this work. Nonethe-

less, there still exist areas for continued development. In line with this, the further exploration and development of signal processing, feature extraction and gesture classification techniques for food and fluid intake recognition is recommended for future work.

In this context, the exploration of data augmentation techniques such as the Synthetic Minority Oversampling Technique (SMOTE) on hand-crafted feature vectors and the computation of image transformations on encoded time series is recommended to mitigate the imbalance present on datasets embodying the recognition of sparse gestures such as eating and drinking gestures. In addition, the use of pre-trained deep learning models on larger activity recognition datasets could benefit the training phase of deep learning models aiming at the recognition of sparse gestures. Based on this, an investigation upon the use of transfer learning for the recognition of eating and drinking gestures is proposed for future work.

- The way eating and drinking gestures are performed may not vary significantly between healthy individuals. However, the performance of systems aiming at the recognition of these gestures on participants with functional limitations such as patients suffering from Parkinson's disease or stroke patients could potentially be compromised. In line with this, further experiments embodying these social groups are recommended.

- A novel method to detect periods of eating is proposed in Chapter 7, achieving promising results. Gesture labelling can be a very cumbersome work since, unlike continuous activities which can be labelled in chunks of data, gestures have to be independently annotated. Nonetheless, such an effort would be highly beneficial to achieve further improvements in recognition of eating gestures and consequently on the identification of the intake of meals. Given this, the collection of two extensive datasets is proposed as follows: 1) A dataset incorporating a vast array of hand to face gestures. 2) A dataset incorporating gestures from the intake of meals in different environments and positions. As shown in Chapter 5, the gesture recognition model proposed shows very good performance at recognising similar food and drink intake gestures. The incorporation of new classes to the

model should translate into further improvement of its performance and consequently into potential further advances in dietary monitoring.

- Alongside the wrist-worn inertial sensor, low-cost ambient sensors such as PIRs or strategically installed pressure sensors (mats) on specific chairs can offer dietary monitoring systems the opportunity to get valuable context awareness (*i.e.* the individual is sat at the dining table or in the kitchen). With this, probabilistic functions could be used on the output given by the food intake recognition system.

- A significant contribution towards the field of AAL will be the combination of the above with the exploitation of trend analysis techniques to develop intelligent systems able to identify anomalies on the dietary behaviour of individuals so that cases in which eating assistance is required can be identified.

References

[1] Department of Economic United Nations, "World ageing population", Tech. Rep., United Nations Publications, 2009. 1

[2] Department of Economic United Nations, "World population ageing 2009", Tech. Rep., United Nations Publications, 2010. 1

[3] Christina Harrefors, Stefan Sävenstedt, and Karin Axelsson, "Elderly peoples perceptions of how they want to be cared for: an interview study with healthy elderly couples in northern sweden", *Scandinavian Journal of Caring Sciences*, vol. 23, no. 2, pp. 353–360, 2009. 1

[4] Chun Zhu, *Hand gesture and activity recognition in assisted living through wearable sensing and computing*, PhD thesis, Oklahoma State University, 2011. 1

[5] Mukhtiar Memon, Stefan Wagner, Christian Pedersen, Femina Beevi, and Finn Hansen, "Ambient assisted living healthcare frameworks, platforms, standards, and quality attributes", *Sensors*, vol. 14, no. 3, pp. 4312–4341, 2014. 1

[6] Kofi Appiah, Andrew Hunter, Ahmad Lotfi, Christopher Waltham, and Patrick Dickinson, "Human behavioural analysis with self-organizing map for ambient assisted living", in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2014, pp. 2430–2437. 2

[7] Sandipan Pal and Charith Abhayaratne, "Video-based activity level recognition for assisted living using motion features", in *Proceedings of the 9th International Conference on Distributed Smart Cameras*. ACM, 2015, pp. 62–67. 2

[8] Nagender K Suryadevara and Subhas C Mukhopadhyay, "Determining wellness through an ambient assisted living environment", *IEEE Intelligent Systems*, vol. 29, no. 3, pp. 30–37, 2014. 2

[9] Yujie Dong, Jenna Scisco, Mike Wilson, Eric Muth, and Adam Hoover, "Detecting periods of eating during free-living by tracking wrist motion", *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1253–1260, 2014. 2, 16, 18, 21, 23, 30, 114, 115, 118

[10] Kelly E Caine, Arthur D Fisk, and Wendy A Rogers, "Benefits and privacy concerns of a home equipped with a visual sensing system: A perspective from older adults", in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Sage Publications Sage CA: Los Angeles, CA, 2006, vol. 50, pp. 180–184. 2

[11] Elizabeth C Nelson, Tibert Verhagen, and Matthijs L Noordzij, "Health empowerment through activity trackers: An empirical smart wristband study", *Computers in Human Behavior*, vol. 62, pp. 364–374, 2016. 2, 16

[12] Tara O'brien, Meredith Troutman-Jordan, Donna Hathaway, Shannon Armstrong, and Michael Moore, "Acceptability of wristband activity trackers among community dwelling older adults", *Geriatric Nursing*, vol. 36, no. 2, pp. S21–S25, 2015. 2, 16

[13] Daniel WT Wundersitz, Casey Josman, Ritu Gupta, Kevin J Netto, Paul B Gastin, and Sam Robertson, "Classification of team sport activities using a single wearable tracking device", *Journal of Biomechanics*, vol. 48, no. 15, pp. 3975–3981, 2015. 2, 20, 21, 22, 24, 25, 56

[14] Akram Bayat, Marc Pomplun, and Duc A Tran, "A study on human activity recognition using accelerometer data from smartphones", *Procedia Computer Science*, vol. 34, pp. 450–457, 2014. 2, 16, 20, 21, 22, 24, 25, 30, 38, 40, 54, 55, 79

[15] Pierluigi Casale, Oriol Pujol, and Petia Radeva, "Human activity recognition from accelerometer data using a wearable device", *Pattern Recognition*

and Image Analysis, pp. 289–296, 2011. 2, 16, 18, 20, 21, 22, 24, 25, 30, 38, 40, 54, 56, 79

[16] Nicole A Capela, Edward D Lemaire, and Natalie Baddour, "Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients", *PloS One*, vol. 10, no. 4, pp. e0124414, 2015. 2, 16, 21, 22, 25, 54

[17] Ç Berke Erdaş, Işıl Atasoy, Koray Açıcı, and Hasan Oğul, "Integrating features for accelerometer-based activity recognition", *Procedia Computer Science*, vol. 98, pp. 522–527, 2016. 2, 16, 22, 24, 30, 40, 54, 56

[18] Lieven Billiet, Thijs Willem Swinnen, Rene Westhovens, Kurt de Vlam, and Sabine Van Huffel, "Accelerometry-based activity recognition and assessment in rheumatic and musculoskeletal diseases", *Sensors*, vol. 16, no. 12, pp. 2151, 2016. 2, 24, 57

[19] Michael A Mahler, Qinghua Li, and Ang Li, "Securehouse: A home security system based on smartphone sensors", in *IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2017, pp. 11–20. 2, 16

[20] Saisakul Chernbumroong, Shuang Cang, Anthony Atkins, and Hongnian Yu, "Elderly activities recognition and classification for applications in assisted living", *Expert Systems with Applications*, vol. 40, no. 5, pp. 1662–1674, 2013. 2, 19, 21

[21] Aanand D Naik, Jason Burnett, Sabrina Pickens-Pace, and Carmel B Dyer, "Impairment in instrumental activities of daily living and the geriatric syndrome of self-neglect", *The Gerontologist*, vol. 48, no. 3, pp. 388–393, 2008. 3

[22] NHS Digital, "Safeguarding adults collection (sac). england 2016-2017 experimental statistics", 2017, [Online], Accessed: 2018-01-15. 3

[23] Robert C Abrams, Mark Lachs, Gail McAvay, Denis J Keohane, and Martha L Bruce, "Predictors of self-neglect in community-dwelling elders", *American Journal of Psychiatry*, vol. 159, no. 10, pp. 1724–1730, 2002. 3

[24] Mark S Lachs, Christianna S Williams, Shelley O'brien, Karl A Pillemer, and Mary E Charlson, "The mortality of elder mistreatment", *JAMA*, vol. 280, no. 5, pp. 428–432, 1998. 3

[25] Yan Wang, Shuang Cang, and Hongnian Yu, "A survey on wearable sensor modality centred human activity recognition in health care", *Expert Systems with Applications*, 2019. 12, 23

[26] Davide Giacalone, Karin Wendin, Stefanie Kremer, Michael Bom Frøst, Wender LP Bredie, Viktoria Olsson, Marie H Otto, Signe Skjoldborg, Ulla Lindberg, and Einar Risvik, "Health and quality of life in an aging population–food and beyond", *Food Quality and Preference*, vol. 47, pp. 166–170, 2016. 12

[27] Ruijiao Li, Bowen Lu, and Klaus D McDonald-Maier, "Cognitive assisted living ambient system: a survey", *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015. 13

[28] Diane J Cook, Juan C Augusto, and Vikramaditya R Jakkula, "Ambient intelligence: Technologies, applications, and opportunities", *Pervasive and Mobile Computing*, vol. 5, no. 4, pp. 277–298, 2009. 13

[29] Yun Li, Zhiling Zeng, Mihail Popescu, and KC Ho, "Acoustic fall detection using a circular microphone array", in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 2242–2245. 14

[30] Aadel Howedi, Ahmad Lotfi, and Amir Pourabdollah, "Exploring entropy measurements to identify multi-occupancy in activities of daily living", *Entropy*, vol. 21, no. 4, pp. 416, 2019. 14, 120

[31] Abubaker Elbayoudi, Ahmad Lotfi, and Caroline Langensiepen, "The human behaviour indicator: A measure of behavioural evolution", *Expert Systems with Applications*, vol. 118, pp. 493–505, 2019. 14

[32] Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M Rehg, "A scalable approach to activity recognition based

on object use", in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8. 14

[33] Aung Aung Phyo Wai, Kow Yuan-Wei, Foo Siang Fook, Maniyeri Jayachandran, Jit Biswas, and John-John Cabibihan, "Sleeping patterns observation for bedsores and bed-side falls prevention", in *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2009, pp. 6087–6090. 14

[34] Vangelis Metsis, Georgios Galatas, Alexandros Papangelis, Dimitrios Kosmopoulos, and Fillia Makedon, "Recognition of sleep patterns using a bed pressure mat", in *Proceedings of the 4th International Conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2011, p. 9. 14

[35] M Baran Pouyan, Mehrdad Nourani, and Matthew Pompeo, "Sleep state classification using pressure sensor mats", in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 1207–1210. 14

[36] Seiichi Honda, Ken-ichi Fukui, Koichi Moriyama, Satoshi Kurihara, and Masayuki Numao, "Extracting human behaviors with infrared sensor network", in *2007 Fourth International Conference on Networked Sensing Systems*. IEEE, 2007, pp. 122–125. 14

[37] Gadelhag Mohmed, Ahmad Lotfi, and Amir Pourabdollah, "Human activities recognition based on neuro-fuzzy finite state machine", *Technologies*, vol. 6, no. 4, pp. 110, 2018. 14

[38] Salisu Wada Yahaya, Ahmad Lotfi, and Mufti Mahmud, "A consensus novelty detection ensemble approach for anomaly detection in activities of daily living", *Applied Soft Computing*, vol. 83, pp. 105613, 2019. 14

[39] KH Low, Jeffry William Tani, Teguh Chandra, and Ping Wang, "Initial home-based foot-mat design & analysis of bio-gait characteristics to prevent fall in elderly people", in *2009 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2009, pp. 759–764. 14

[40] Michael Buettner, Richa Prasad, Matthai Philipose, and David Wetherall, "Recognizing daily activities with rfid-based sensors", in *Proceedings of the 11th International Conference on Ubiquitous Computing*, 2009, pp. 51–60. 14

[41] Jaeyoung Yang, Joonwhan Lee, and Joongmin Choi, "Activity recognition based on rfid object usage for smart mobile devices", *Journal of Computer Science and Technology*, vol. 26, no. 2, pp. 239–246, 2011. 14

[42] Ish Rishabh, Don Kimber, and John Adcock, "Indoor localization using controlled ambient sounds", in *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2012, pp. 1–10. 14

[43] Koji Yatani and Khai N Truong, "Bodyscope: a wearable acoustic sensor for activity recognition", in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, 2012, pp. 341–350. 14

[44] Long-Van Nguyen-Dinh, Ulf Blanke, and Gerhard Tröster, "Towards scalable activity recognition: Adapting zero-effort crowdsourced acoustic models", in *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, 2013, pp. 1–10. 14

[45] Dawei Liang and Edison Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos", *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–18, 2019. 14

[46] Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D Abowd, "Inferring meal eating activities in real world settings from ambient sounds: A feasibility study", in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015, pp. 427–431. 14

[47] Jingyuan Cheng, Mathias Sundholm, Bo Zhou, Marco Hirsch, and Paul Lukowicz, "Smart-surface: Large scale textile pressure sensors arrays for activity recognition", *Pervasive and Mobile Computing*, vol. 30, pp. 97–112, 2016. 14

[48] Asangi Jayatilaka and Damith C Ranasinghe, "Real-time fluid intake gesture recognition based on batteryless uhf rfid technology", *Pervasive and Mobile Computing*, vol. 34, pp. 146–156, 2017. 15

[49] Chris Stauffer and W Eric L Grimson, "Adaptive background mixture models for real-time tracking", in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. IEEE, 1999, vol. 2, pp. 246–252. 15

[50] P Wayne Power and Johann A Schoonees, "Understanding background mixture models for foreground segmentation", in *Proceedings Image and Vision Computing New Zealand*, 2002, vol. 2002, pp. 10–11. 15

[51] Nigel JB McFarlane and C Paddy Schofield, "Segmentation and tracking of piglets in images", *Machine Vision and Applications*, vol. 8, no. 3, pp. 187–193, 1995. 15

[52] Suad Albawendi, Kofi Appiah, Heather Powell, and Ahmad Lotfi, "Video based fall detection with enhanced motion history images", in *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 2016, pp. 1–7. 15

[53] Philipp V Rouast and Marc TP Adam, "Learning deep representations for video-based intake gesture detection", *arXiv preprint arXiv:1909.10695*, 2019. 15

[54] Brais Martinez, Davide Modolo, Yuanjun Xiong, and Joseph Tighe, "Action recognition with spatial-temporal discriminative filter banks", in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5482–5491. 15

[55] Dimitrios Konstantinidis, Kosmas Dimitropoulos, Ioannis Ioakimidis, Billy Langlet, and Petros Daras, "A deep network for automatic video-based food bite detection", in *International Conference on Computer Vision Systems*. Springer, 2019, pp. 586–595. 15

[56] Alexandros Iosifidis, Ermioni Marami, Anastasios Tefas, and Ioannis Pitas, "Eating and drinking activity recognition based on discriminant analysis of

fuzzy distances and activity volumes", in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 2201–2204. 15

[57] Roberto Ugolotti, Federico Sassi, Monica Mordonini, and Stefano Cagnoni, "Multi-sensor system for detection and classification of human activities", *Journal of Ambient Intelligence and Humanized Computing*, vol. 4, no. 1, pp. 27–41, 2013. 16

[58] Daniela Micucci, Marco Mobilio, Paolo Napoletano, and Francesco Tisato, "Falls as anomalies? an experimental evaluation using smartphone accelerometer data", *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 1, pp. 87–99, 2017. 16

[59] Francisco de Arriba-Pérez, Manuel Caeiro-Rodríguez, and Juan M Santos-Gago, "How do you sleep? using off the shelf wrist wearables to estimate sleep quality, sleepiness level, chronotype and sleep regularity indicators", *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 4, pp. 897–917, 2018. 16

[60] Delaram Jarchi, James Pope, Tracey KM Lee, Larisa Tamjidi, Amirhosein Mirzaei, and Saeid Sanei, "A review on accelerometry-based gait analysis and emerging clinical applications", *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 177–194, 2018. 16

[61] Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, and Evangelos Kalogerakis, "Risq: Recognizing smoking gestures with inertial sensors on a wristband", in *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2014, pp. 149–161. 16, 19, 23, 72

[62] Hari Prabhat Gupta, Haresh S Chudgar, Siddhartha Mukherjee, Tanima Dutta, and Kulwant Sharma, "A continuous hand gestures recognition technique for human-machine interaction using accelerometer and gyroscope sensors", *IEEE Sensors Journal*, vol. 16, no. 16, pp. 6425–6432, 2016. 16

[63] Renqiang Xie and Juncheng Cao, "Accelerometer-based hand gesture recognition by neural network and similarity matching", *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4537–4545, 2016. 16

[64] Oliver Amft, David Bannach, Gerald Pirkl, Matthias Kreil, and Paul Lukowicz, "Towards wearable sensing-based assessment of fluid intake", in *8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM)*, 2010, pp. 298–303. 16, 73, 74, 89, 110

[65] Giovanni Schiboni and Oliver Amft, "Sparse natural gesture spotting in free living to monitor drinking with wrist-worn inertial sensors", in *Proceedings of the International Symposium on Wearable Computers*. ACM, 2018, pp. 140–147. 16, 25, 72

[66] Josh Cherian, Vijay Rajanna, Daniel Goldberg, and Tracy Hammond, "Did you remember to brush? a noninvasive wearable approach to recognizing brushing teeth for elderly care", in *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2017, pp. 48–57. 16

[67] Wei-zhong Wang, Yan-wei Guo, Bang-yu Huang, Guo-ru Zhao, Bo-qiang Liu, and Lei Wang, "Analysis of filtering methods for 3d acceleration signals in body sensor network", in *International Symposium on Bioelectronics and Bioinformatics (ISBB), 2011*. IEEE, 2011, pp. 263–266. 17, 18, 20, 21, 38, 57

[68] Mario Munoz-Organero and Ahmad Lotfi, "Human movement recognition based on the stochastic characterisation of acceleration data", *Sensors*, vol. 16, no. 9, pp. 1464, 2016. 17, 30

[69] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga, "Complex human activity recognition using smartphone and wrist-worn motion sensors", *Sensors*, vol. 16, no. 4, pp. 426, 2016. 17, 21

[70] Mi Zhang and Alexander A Sawchuk, "A preliminary study of sensing appliance usage for human activity recognition using mobile magnetometer", in

Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012, pp. 745–748. 17

[71] Adil Mehmood Khan, Ali Tufail, Asad Masood Khattak, and Teemu H Laine, "Activity recognition on smartphones via sensor-fusion and kda-based svms", *International Journal of Distributed Sensor Networks*, vol. 10, no. 5, pp. 503291, 2014. 17

[72] Uwe Maurer, Asim Smailagic, Daniel P Siewiorek, and Michael Deisher, "Activity recognition and monitoring using multiple sensors on different body positions", in *International Workshop on Wearable and Implantable Body Sensor Networks (BSN'06)*. IEEE, 2006, pp. 4–pp. 17, 24

[73] Jingyuan Cheng, Oliver Amft, and Paul Lukowicz, "Active capacitive sensing: Exploring a new wearable sensing modality for activity recognition", in *International Conference on Pervasive Computing*. Springer, 2010, pp. 319–336. 17

[74] Oliver Amft, Martin Kusserow, and Gerhard Tröster, "Probabilistic parsing of dietary activity events", in *4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*. Springer, 2007, pp. 242–247. 17, 19, 57, 112

[75] Edward L Mahoney and Diane F Mahoney, "Acceptance of wearable technology by people with alzheimers disease: Issues and accommodations", *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 25, no. 6, pp. 527–531, 2010. 18

[76] Gabriel Orsini, Dirk Bade, and Winfried Lamersdorf, "Cloudaware: A context-adaptive middleware for mobile edge and cloud computing applications", in *2016 IEEE 1st International Workshops on Foundations and Applications of Self* Systems (FAS* W)*. IEEE, 2016, pp. 216–221. 18

[77] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman, "Activity recognition from accelerometer data", in *American Association for Artificial Intelligence (AAAI)*, 2005, vol. 5, pp. 1541–1546. 18, 20, 21, 22, 24, 25, 30, 40, 56, 79

[78] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga, "Fusion of smartphone motion sensors for physical activity recognition", *Sensors*, vol. 14, no. 6, pp. 10146–10176, 2014. 18, 30

[79] Enrique Garcia-Ceja and Ramon Brena, "Long-term activity recognition from accelerometer data", *Procedia Technology*, vol. 7, pp. 248–256, 2013. 18, 21, 25, 30

[80] Louis Atallah, Benny Lo, Rachel King, and Guang-Zhong Yang, "Sensor positioning for activity recognition using wearable accelerometers", *IEEE Transactions on Biomedical Circuits and Systems*, vol. 5, no. 4, pp. 320–329, 2011. 18

[81] Ian Cleland, Basel Kikhia, Chris Nugent, Andrey Boytsov, Josef Hallberg, Kåre Synnes, Sally McClean, and Dewar Finlay, "Optimal placement of accelerometers for the detection of everyday activities", *Sensors*, vol. 13, no. 7, pp. 9183–9200, 2013. 18, 19

[82] Alexander HK Montoye, James M Pivarnik, Lanay M Mudd, Subir Biswas, and Karin A Pfeiffer, "Comparison of activity type classification accuracy from accelerometers worn on the hip, wrists, and thigh in young, apparently healthy adults", *Measurement in Physical Education and Exercise Science*, vol. 20, no. 3, pp. 173–183, 2016. 18, 19

[83] Alok Kumar Chowdhury, Dian Tjondronegoro, Vinod Chandran, and Stewart G Trost, "Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data", *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 678–685, 2017. 18, 19, 25

[84] Edison Thomaz, Irfan Essa, and Gregory D Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing", in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 1029–1040. 19, 114, 115, 118

[85] Diana Gomes, João Mendes-Moreira, Inês Sousa, and Joana Silva, "Eating and drinking recognition in free-living conditions for triggering smart

reminders", *Sensors*, vol. 19, no. 12, pp. 2803, 2019. 19, 25, 72, 114, 115, 118, 119

[86] Zebo Zhou, Shanhui Mo, Jin Wu, and Hassen Fourati, "Behaviors classification based distance measuring system for pedestrians via a foot-mounted inertial sensor", *Asian Journal of Control*, 2019. 19

[87] Stefan I Madansingh, Dennis H Murphree, Kenton R Kaufman, and Emma Fortune, "Assessment of gait kinetics in post-menopausal women using triaxial ankle accelerometers during barefoot walking", *Gait & Posture*, vol. 69, pp. 85–90, 2019. 19

[88] Robert LeMoyne and Timothy Mastroianni, "Portable wearable and wireless systems for gait and reflex response quantification", in *Wearable and Wireless Systems for Healthcare I*, pp. 59–71. Springer, 2018. 19

[89] A Moncada-Torres, K Leuenberger, R Gonzenbach, A Luft, and Roger Gassert, "Activity classification based on inertial and barometric pressure sensors at different anatomical locations", *Physiological Measurement*, vol. 35, no. 7, pp. 1245, 2014. 19

[90] Charissa Ann Ronao and Sung-Bae Cho, "Deep convolutional neural networks for human activity recognition with smartphone sensors", in *International Conference on Neural Information Processing*. Springer, 2015, pp. 46–53. 19, 24, 25, 93, 94, 95

[91] Ling Bao and Stephen S Intille, "Activity recognition from user-annotated acceleration data", in *International Conference on Pervasive Computing*. Springer, 2004, pp. 1–17. 19

[92] Narayanan C Krishnan and Sethuraman Panchanathan, "Analysis of low resolution accelerometer data for continuous human activity recognition", in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3337–3340. 19

[93] Chun Zhu and Weihua Sheng, "Human daily activity recognition in robot-assisted living using multi-sensor fusion", in *2009 IEEE International Conference on Robotics and Automation*. IEEE, 2009, pp. 2154–2159. 19

[94] Ramona Rednic, Elena Gaura, James Brusey, and John Kemp, "Wearable posture recognition systems: factors affecting performance", in *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2012, pp. 200–203. 19

[95] Timo Sztyler, Heiner Stuckenschmidt, and Wolfgang Petrich, "Position-aware activity recognition with wearable devices", *Pervasive and Mobile Computing*, vol. 38, pp. 281–295, 2017. 19

[96] Oliver Amft and Gerhard Tröster, "Recognition of dietary activity events using on-body sensors", *Artificial Intelligence in Medicine*, vol. 42, no. 2, pp. 121–136, 2008. 19

[97] Martin H. Weik, *Nyquist theorem*, pp. 1127–1127, Springer US, Boston, MA, 2001. 20

[98] Aftab Khan, Nils Hammerla, Sebastian Mellor, and Thomas Plötz, "Optimising sampling rates for accelerometer-based human activity recognition", *Pattern Recognition Letters*, vol. 73, pp. 33–40, 2016. 20

[99] Lei Gao, AK Bourke, and John Nelson, "Evaluation of accelerometer based multi-sensor versus single-sensor activity recognition systems", *Medical Engineering & Physics*, vol. 36, no. 6, pp. 779–785, 2014. 20

[100] Erik K Antonsson and Robert W Mann, "The frequency content of gait", *Journal of Biomechanics*, vol. 18, no. 1, pp. 39–47, 1985. 20

[101] Shumei Zhang, Paul McCullagh, and Vic Callaghan, "An efficient feature selection method for activity classification", in *2014 International Conference on Intelligent Environments*. IEEE, 2014, pp. 16–22. 20

[102] A Godfrey, KM Culhane, and GM Lyons, "Comparison of the performance of the activpal professional physical activity logger to a discrete accelerometer-based activity monitor", *Medical Engineering & Physics*, vol. 29, no. 8, pp. 930–934, 2007. 20, 54, 56

[103] MA Álvarez De La Concepción, LM Soria Morillo, Luis Gonzalez-Abril, and JA Ortega Ramírez, "Discrete techniques applied to low-energy mo-

bile human activity recognition. a new approach", *Expert Systems with Applications*, vol. 41, no. 14, pp. 6138–6146, 2014. 20

[104] Dapeng Qiao, Grantham KH Pang, Man-Kit Mui, and David CC Lam, "A single-axis low-cost accelerometer fabricated using printed-circuit-board techniques", *IEEE Electron Device Letters*, vol. 30, no. 12, pp. 1293–1295, 2009. 21

[105] Kuang-Hsuan Chen, Jing-Jung Yang, and Fu-Shan Jaw, "Accelerometer-based fall detection using feature extraction and support vector machine algorithms", *Instrumentation Science & Technology*, vol. 44, no. 4, pp. 333–342, 2016. 21

[106] MJ Mathie, Nigel H Lovell, ACF Coster, and BG Celler, "Determining activity using a triaxial accelerometer", in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society][Engineering in Medicine and Biology*. IEEE, 2002, vol. 3, pp. 2481–2482. 21

[107] Yi He, Ye Li, and Shu-Di Bao, "Fall detection by built-in tri-accelerometer of smartphone", in *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2012, pp. 184–187. 21

[108] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho, "Human activity recognition from accelerometer data using convolutional neural network", in *IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2017, pp. 131–134. 22, 30, 94, 109

[109] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz, "Energy efficient smartphone-based activity recognition using fixed-point arithmetic.", *Journal of Universal Computer Science*, vol. 19, no. 9, pp. 1295–1314, 2013. 22, 26, 67, 68

[110] Niall Twomey, Tom Diethe, Xenofon Fafoutis, Atis Elsts, Ryan McConville, Peter Flach, and Ian Craddock, "A comprehensive study of activity recognition using accelerometers", in *Informatics*. Multidisciplinary Digital Publishing Institute, 2018, vol. 5, p. 27. 22

[111] Jozsef Suto, Stefan Oniga, and Petrica Pop Sitar, "Feature analysis to human activity recognition", *International Journal of Computers Communications & Control*, vol. 12, no. 1, pp. 116–130, 2017. 22, 40

[112] Martin Berchtold, Matthias Budde, Hedda R Schmidtke, and Michael Beigl, "An extensible modular recognition concept that makes activity recognition practical", in *Annual Conference on Artificial Intelligence*. Springer, 2010, pp. 400–409. 22

[113] Shaopeng Liu, Robert X Gao, Dinesh John, John W Staudenmayer, and Patty S Freedson, "Multisensor data fusion for physical activity assessment", *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 3, pp. 687–696, 2011. 22

[114] Frédéric Li, Kimiaki Shirahama, Muhammad Nisar, Lukas Köping, and Marcin Grzegorzek, "Comparison of feature learning methods for human activity recognition using wearable sensors", *Sensors*, vol. 18, no. 2, pp. 679, 2018. 22

[115] Jian Wu and Roozbeh Jafari, "Orientation independent activity/gesture recognition using wearable motion sensors", *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1427–1437, 2018. 22

[116] Nils Y Hammerla, Shane Halloran, and Thomas Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables", *arXiv preprint arXiv:1604.08880*, 2016. 22

[117] Benish Fida, Ivan Bernabucci, Daniele Bibbo, Silvia Conforto, and Maurizio Schmid, "Varying behavior of different window sizes on the classification of static and dynamic physical activities from a single accelerometer", *Medical Engineering & Physics*, vol. 37, no. 7, pp. 705–711, 2015. 22

[118] Oresti Banos, Juan-Manuel Galvez, Miguel Damas, Hector Pomares, and Ignacio Rojas, "Window size impact in human activity recognition", *Sensors*, vol. 14, no. 4, pp. 6474–6499, 2014. 22

[119] Mi Zhang and Alexander A Sawchuk, "Motion primitive-based human activity recognition using a bag-of-features approach", in *Proceedings of*

the 2nd ACM SIGHIT International Health Informatics Symposium. ACM, 2012, pp. 631–640. 22, 57

[120] Mohd Halim Mohd Noor, Zoran Salcic, I Kevin, and Kai Wang, "Adaptive sliding window segmentation for physical activity recognition using a single tri-axial accelerometer", *Pervasive and Mobile Computing*, vol. 38, pp. 41–59, 2017. 23, 41, 72

[121] Eamonn Keogh, Selina Chu, David Hart, and Michael Pazzani, "Segmenting time series: A survey and novel approach", in *Data Mining in Time Series Databases*, pp. 1–21. World Scientific, 2004. 23

[122] Miodrag Lovrić, Marina Milanović, and Milan Stamenković, "Algoritmic methods for segmentation of time series: An overview", *Journal of Contemporary Economic and Business Issues*, vol. 1, no. 1, pp. 31–53, 2014. 23

[123] Holger Junker, Oliver Amft, Paul Lukowicz, and Gerhard Tröster, "Gesture spotting with body-worn inertial sensors to detect user activities", *Pattern Recognition*, vol. 41, no. 6, pp. 2010–2024, 2008. 23, 71, 73, 74, 89, 110

[124] Ruize Xu, Shengli Zhou, and Wen J Li, "Mems accelerometer based nonspecific-user hand gesture recognition", *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1166–1173, 2012. 23, 71, 73

[125] Narayanan C Krishnan, Colin Juillard, Dirk Colbry, and Sethuraman Panchanathan, "Recognition of hand movements using wearable accelerometers", *Journal of Ambient Intelligence and Smart Environments*, vol. 1, no. 2, pp. 143–155, 2009. 23

[126] Jozsef Suto and Stefan Oniga, "Efficiency investigation of artificial neural networks in human activity recognition", *Journal of Ambient Intelligence and Humanized Computing*, vol. 9, no. 4, pp. 1049–1060, 2018. 24, 26

[127] Barbara Bruno, Fulvio Mastrogiovanni, Antonio Sgorbissa, Tullio Vernazza, and Renato Zaccaria, "Analysis of human behavior recognition algorithms based on acceleration data", in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1602–1607. 24, 30

[128] Nimish Kale, Jaeseong Lee, Reza Lotfian, and Roozbeh Jafari, "Impact of sensor misplacement on dynamic time warping based human activity recognition using wearable computers", in *Proceedings of the Conference on Wireless Health*, 2012, pp. 1–8. 24

[129] Stefan Duffner, Samuel Berlemont, Grégoire Lefebvre, and Christophe Garcia, "3d gesture classification with convolutional neural networks", in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5432–5436. 24, 25, 93

[130] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition", in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015. 24, 25, 93, 94, 95

[131] Andrey Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks", *Applied Soft Computing*, vol. 62, pp. 915–922, 2018. 24, 25, 30, 40, 93, 94, 95, 109

[132] Charissa Ann Ronao and Sung-Bae Cho, "Human activity recognition with smartphone sensors using deep learning neural networks", *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016. 24, 25, 26, 68, 94, 95

[133] Li Liu, Yuxin Peng, Shu Wang, Ming Liu, and Zigang Huang, "Complex activity recognition using time series pattern dictionary learned from ubiquitous sensors", *Information Sciences*, vol. 340, pp. 41–57, 2016. 25

[134] John Paul Varkey, Dario Pompili, and Theodore A Walls, "Human motion recognition using a wireless sensor-based wearable system", *Personal and Ubiquitous Computing*, vol. 16, no. 7, pp. 897–910, 2012. 25

[135] Roozbeh Jafari, Wenchao Li, Ruzena Bajcsy, Steven Glaser, and Shankar Sastry, "Physical activity monitoring for assisted living at home", in *4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*. Springer, 2007, pp. 213–219. 25, 56

[136] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi, "Multi-modal convolutional neural networks for activity recognition", in *2015 IEEE International*

Conference on Systems, Man, and Cybernetics. IEEE, 2015, pp. 3017–3022. 25, 94

[137] Francisco Javier Ordóñez and Daniel Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition", *Sensors*, vol. 16, no. 1, pp. 115, 2016. 25

[138] Yu Zhao, Rennong Yang, Guillaume Chevalier, Ximeng Xu, and Zhenxing Zhang, "Deep residual bidir-lstm for human activity recognition using wearable sensors", *Mathematical Problems in Engineering*, vol. 2018, 2018. 25, 26

[139] Masaya Inoue, Sozo Inoue, and Takeshi Nishida, "Deep recurrent neural network for mobile human activity recognition with high throughput", *Artificial Life and Robotics*, vol. 23, no. 2, pp. 173–185, 2018. 25, 26

[140] Liu-Hsuan Chen, Kai-Chun Liu, Chia-Yeh Hsieh, and Chia-Tai Chan, "Drinking gesture spotting and identification using single wrist-worn inertial sensor", in *IEEE International Conference on Applied System Innovation (ICASI)*, 2017, pp. 299–302. 26, 71, 72, 73, 89, 90, 110

[141] Jiahui Wen and Mingyang Zhong, "Activity discovering and modelling with labelled and unlabelled data in smart environments", *Expert Systems with Applications*, vol. 42, no. 14, pp. 5800–5810, 2015. 26

[142] Cagatay Catal, Selin Tufekci, Elif Pirmit, and Guner Kocabag, "On the use of ensemble of classifiers for accelerometer-based activity recognition", *Applied Soft Computing*, vol. 37, pp. 1018–1022, 2015. 30

[143] Andrea Mannini and Angelo Maria Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers", *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010. 30, 56, 60

[144] MetaMotionR, "Mbientlab", 2020, Accessed: 2017-07-15. 31, 35

[145] Stephen Butterworth et al., "On the theory of filter amplifiers", *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930. 38

[146] Wilhelmiina Hamäläinen, Mikko Järvinen, Paula Martiskainen, and Jaakko Mononen, "Jerk-based feature extraction for robust activity recognition from acceleration data", in *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, 2011, pp. 831–836. 39

[147] Phong Nguyen, Takayuki Akiyama, Hiroki Ohashi, Goh Nakahara, Katsuya Yamasaki, and Saito Hikaru, "User-friendly activity recognition using svm classifier and informative features", in *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2015, pp. 1–8. 39

[148] Yufei Chen and Chao Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition", *Ieee Access*, vol. 5, pp. 3095–3110, 2017. 39

[149] Cheol-Hong Min, Nuri F Ince, and Ahmed H Tewfik, "Generalization capability of a wearable early morning activity detection system", in *2007 15th European Signal Processing Conference*. IEEE, 2007, pp. 1556–1560. 54, 56, 68

[150] Jhun-Ying Yang, Jeen-Shing Wang, and Yen-Ping Chen, "Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers", *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2213–2220, 2008. 54, 56

[151] Marianne Forsell, Petteri Sjögren, and Olle Johansson, "Need of assistance with daily oral hygiene measures among nursing home resident elderly versus the actual assistance received from the staff", *The Open Dentistry Journal*, vol. 3, pp. 241, 2009. 54

[152] Ikuko Takahashi and Sue Turale, "Evaluation of individual and facility factors that promote hand washing in aged-care facilities in japan", *Nursing & Health Sciences*, vol. 12, no. 1, pp. 127–134, 2010. 54

[153] Cheol-Hong Min, Nuri F Ince, and Ahmed H Tewfik, "Classification of continuously executed early morning activities using wearable wireless sensors", in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 5192–5195. 56, 68

[154] Wenchao Jiang and Zhaozheng Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks", in *Proceedings of the 23rd ACM International Conference on Multimedia*. Acm, 2015, pp. 1307–1310. 68, 94

[155] Albert Westergren, Mitra Unosson, Ola Ohlsson, Birgitta Lorefält, and Ingalill R Hallberg, "Eating difficulties, assisted eating and nutritional status in elderly (65 years) patients in hospital rehabilitation", *International Journal of Nursing Studies*, vol. 39, no. 3, pp. 341–351, 2002. 70

[156] Christa Lohrmann, Ate Dijkstra, and Theo Dassen, "The care dependency scale: an assessment instrument for elderly patients in german hospitals", *Geriatric Nursing*, vol. 24, no. 1, pp. 40–43, 2003. 70

[157] Albert Westergren, Siv Karlsson, Pia Andersson, Ola Ohlsson, and Ingalill R Hallberg, "Eating difficulties, need for assisted eating, nutritional status and pressure ulcers in patients admitted for stroke rehabilitation", *Journal of Clinical Nursing*, vol. 10, no. 2, pp. 257–269, 2001. 70

[158] Hélène Payette and Bryna Shatenstein, "Determinants of healthy eating in community-dwelling elderly people", *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique*, pp. S27–S31, 2005. 70

[159] Michael N Sawka, Samuel N Cheuvront, and Robert Carter III, "Human water needs", *Nutrition Reviews*, vol. 63, pp. S30–S39, 2005. 70

[160] W Larry Kenney and Percy Chiu, "Influence of age on thirst and fluid intake.", *Medicine and Science in Sports and Exercise*, vol. 33, no. 9, pp. 1524–1532, 2001. 71

[161] Judith Mackay, *The atlas of heart disease and stroke*, vol. 5, World Health Organization, 2004. 71

[162] Enas S Lawrence, Catherine Coshall, Ruth Dundas, Judy Stewart, Anthony G Rudd, Robin Howard, and Charles DA Wolfe, "Estimates of the prevalence of acute stroke impairments and disability in a multiethnic population", *Stroke*, vol. 32, no. 6, pp. 1279–1284, 2001. 71

[163] Tsung-Ming Tai, Yun-Jie Jhang, Zhen-Wei Liao, Kai-Chung Teng, and Wen-Jyi Hwang, "Sensor-based continuous hand gesture recognition by long short-term memory", *IEEE Sensors Letters*, vol. 2, no. 3, pp. 1–4, 2018. 71

[164] Mei-Chuan Tseng, Kai-Chun Liu, Chia-Yeh Hsieh, Steen J Hsu, and Chia-Tai Chan, "Gesture spotting algorithm for door opening using single wearable sensor", in *International Conference on Applied System Invention (ICASI)*. IEEE, 2018, pp. 854–856. 73

[165] Dario Ortega-Anderez, Ahmad Lotfi, and Caroline Langensiepen, "A hierarchical approach in food and drink intake recognition using wearable inertial sensors", in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*. ACM, 2018, pp. 552–557. 74

[166] Hiroaki Sakoe and Seibi Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978. 78, 80

[167] Dario Ortega-Anderez, Ahmad Lotfi, Caroline Langensiepen, and Kofi Appiah, "A multi-level refinement approach towards the classification of quotidian activities using accelerometer data", *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 11, pp. 4319–4330, 2019. 79

[168] Marco Cuturi and Mathieu Blondel, "Soft-DTW: a differentiable loss function for time-series", in *Proceedings of the 34th International Conference on Machine Learning*, Doina Precup and Yee Whye Teh, Eds., International Convention Centre, Sydney, Australia, 06–11 Aug 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 894–903, PMLR. 79

[169] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui, "A kernel for time series based on global alignments", in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2007, vol. 2, pp. II–413. 80

[170] J Ignacio Serrano, Stefan Lambrecht, M Dolores del Castillo, Juan P Romero, Julián Benito-León, and Eduardo Rocon, "Identification of ac-

tivities of daily living in tremorous patients using inertial sensors", *Expert Systems with Applications*, vol. 83, pp. 40–48, 2017. 89, 110

[171] Raul I Ramos-Garcia, Eric R Muth, John N Gowdy, and Adam W Hoover, "Improving the recognition of eating gestures using intergesture sequential dependencies", *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 825–831, 2014. 89, 110

[172] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu, "Deep learning for sensor-based activity recognition: A survey", *Pattern Recognition Letters*, 2018. 93

[173] Zhiguang Wang and Tim Oates, "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks", in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. 97, 98, 111

[174] Isah A Lawal and Sophia Bano, "Deep human activity recognition using wearable sensors", in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments-PETRA'19*. ACM, 2019, pp. 45–48. 97

[175] Guoshen Yu and Jean-Jacques Slotine, "Audio classification from time-frequency texture", in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 1677–1680. 111

[176] Takeyasu Kakamu, Tomoo Hidaka, Tomohiro Kumagai, Yusuke Masuishi, Hideaki Kasuga, Shota Endo, Sei Sato, Akiko Takeda, Makoto Koizumi, and Tetsuhito Fukushima, "Unhealthy changes in eating habits cause acute onset hypertension in the normotensive community-dwelling elderly3 years cohort study", *Medicine*, vol. 98, no. 15, 2019. 112

[177] Elizabeth A Gollub and Dian O Weddle, "Improvements in nutritional intake and quality of life among frail homebound older adults receiving home-delivered breakfast and lunch", *Journal of the American Dietetic Association*, vol. 104, no. 8, pp. 1227–1235, 2004. 112

[178] Nancy S Wellman and Barbara Friedberg, "Causes and consequences of adult obesity: health, social and economic impacts in the united states", *Asia Pacific Journal of Clinical Nutrition*, vol. 11, pp. S705–S709, 2002. 112

[179] Ben J Smith, Alison L Marshall, and Nancy Huang, "Screening for physical activity in family practice: evaluation of two brief assessment tools", *American Journal of Preventive Medicine*, vol. 29, no. 4, pp. 256–264, 2005. 113

[180] Elaine C. rush, Mauro E Valencia, and Lindsay D Plank, "Validation of a 7-day physical activity diary against doubly-labelled water", *Annals of Human Biology*, vol. 35, no. 4, pp. 416–421, 2008. 113

[181] Steven M Pincus, "Approximate entropy as a measure of system complexity.", *Proceedings of the National Academy of Sciences*, vol. 88, no. 6, pp. 2297–2301, 1991. 120

[182] Yen Lee Angela Kwok, Jan Gralton, and Mary-Louise McLaws, "Face touching: a frequent habit that has implications for hand hygiene", *American Journal of Infection Control*, vol. 43, no. 2, pp. 112–114, 2015. 126