SIMILARITY SEGMENTATION APPROACH FOR SENSOR-BASED HUMAN ACTIVITY RECOGNITION

ABDULRAHMAN M A BARAKA

UNIVERSITI SAINS MALAYSIA

2024

SIMILARITY SEGMENTATION APPROACH FOR SENSOR-BASED HUMAN ACTIVITY RECOGNITION

by

ABDULRAHMAN M A BARAKA

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

January 2024

ACKNOWLEDGMENT

IN THE NAME OF ALLAH, THE ALL-COMPASSIONATE, ALL-MERCIFUL

I would like to express my gratitude and appreciation to my main supervisor, Dr. Mohd Halim Mohd Noor, for his real continuous support and valuable assistance. His constant guidance gives me motivation and strength that enables me to move through this challenge. May Allah bless him.

Also, I would like to express my gratitude to Prof. Azuraliza Abu Baakr, Prof. Umi Kalsom Yusof, Dr. Syaheerah Lebai Lufti, Dr. Ahmad Sufril Azlan Mohamed, and Dr. Azleena Mohd Kassim for their valuable comments that enriched the thesis and increased its value.

Special gratitude to my mother, my family, my wife, my brothers, my sisters, and every special person that supported me to complete this journey.

Finally, I want to thank all staff of the school of computer science and my university (USM) for helping and supporting me to complete this great journey.

Our last prayer is Alhamdulillah Lord of the Worlds.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ALGORITHEMS

Page

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
BA	Basic Activity
CNN	Convolution Neural Network
CPD	Change Point Detection
DBN	Deep Belief Network
DL	Deep Learning
DRL	Deep Reinforcement Learning
DSS	Deep Similarity Segmentation
DT	Decision Tree
E2E	End-To-End
HAR	Human Activity Recognition
IDE	Integrated Development Environment
IOT	Internet of Things
LDA	Linear Discriminant Analysis
LSTM	Short-Term Memory
ML	Machine Learning
SSA	Similarity Segmentation Approach
STD	Standard Deviation
SVM	Support Vector Machine
TA	Transitional Activity

PENDEKATAN SEGMENTASI PERSAMAAN UNTUK PENGECAMAN AKTIVITI MANUSIA BERASASKAN PENDERIA

ABSTRAK

Pengecaman aktiviti manusia berasaskan sensor memainkan peranan penting dalam banyak bidang, seperti pengawasan warga emas dan pengesanan jatuh dalam penjagaan kesihatan, penyelia senaman dan rumah pintar dalam aplikasi Internet of Things (IoT). Pengecaman aktiviti manusia (HAR) dilakukan melalui tiga peringkat: pembahagian isyarat, pengekstrakan ciri, dan peringkat pengelasan. Kaedah tetingkap gelongsor bersaiz tetap ialah kaedah yang paling banyak digunakan untuk pembahagian isyarat. Walau bagaimanapun, disebabkan tempoh masa aktiviti manusia yang berbeza-beza, tetingkap gelongsor bersaiz tetap mungkin tidak menghasilkan proses pembahagian yang optimum, terutamanya semasa aktiviti peralihan. Oleh itu, memilih saiz tetingkap yang optimum adalah tugas yang mencabar dan penting, terutamanya jika aktiviti peralihan dipertimbangkan. Para penyelidik cuba meningkatkan kaedah segmentasi dengan mencadangkan pelbagai teknik. Walau bagaimanapun, kebanyakan daripada mereka menumpukan pada setiap ciri tetingkap, dan sedikit yang menganggap hubungan temporal antara tetingkap bersebelahan. Oleh itu, analisis kesan saiz tetingkap terhadap prestasi pengecaman aktiviti asas dan peralihan dilakukan menggunakan model pembelajaran mendalam. Kemudian, dua pendekatan berasaskan persamaan dicadangkan. Pendekatan pertama ialah pendekatan segmentasi berasaskan persamaan yang mengeksploitasi struktur temporal isyarat aktiviti dengan membandingkan persamaan antara tingkap bersebelahan. Secara khusus, ciri dalaman diekstrak menggunakan kejuruteraan ciri untuk setiap tetingkap, dan kemudian

persamaan ciri antara tetingkap bersebelahan diukur menggunakan nilai ambang. Ini membolehkan peringkat pembahagian menjadi lebih berjaya dengan membenarkan model membezakan antara tetingkap peralihan dan bukan peralihan. Walaupun kejuruteraan ciri dan kaedah berasaskan ambang memerlukan beberapa eksperimen dan beberapa pakar, model pembahagian persamaan yang mendalam dicadangkan untuk meningkatkan pendekatan pembahagian persamaan dan mengelakkan kaedah kejuruteraan ciri dan berasaskan ambang. Model pembelajaran mendalam menganggap tugas pembahagian sebagai tugas pengelasan binari. Ia mengekstrak ciri tempatan setiap tetingkap dengan menggunakan lapisan rangkaian saraf konvolusi. Kemudian ia mengekstrak ciri temporal dengan mengukur persamaan antara ciri tingkap bersebelahan. Kedua-dua ciri digabungkan untuk membentangkan ciri akhir bagi setiap tetingkap. Pendekatan yang dicadangkan dinilai menggunakan dua set data awam dan dibandingkan dengan prestasi model HAR yang terkini. Keputusan eksperimen menunjukkan bahawa kaedah cadangan pertama mencapai ketepatan 92.71% dan 86.65% untuk kedua-dua set data, dan model cadangan kedua mencapai ketepatan 93.35% dan 84.96% untuk kedua-dua set data. Keputusan ini mengatasi hasil tetingkap gelongsor tetap serta mengatasi prestasi model terkini.

SIMILARITY SEGMENTATION APPROACH FOR SENSOR-BASED HUMAN ACTIVITY RECOGNITION

ABSTRACT

Sensor-based human activity recognition plays a significant role in many fields, such as elder surveillance and fall detection in healthcare, workout supervisor, and smart homes in Internet of Things (IoT) applications. Human activity recognition (HAR) is performed through three stages: signal segmentation, feature extraction, and classification stage. The fixed-size sliding window method is the most widely used method for signal segmentation. However, due to the varying duration times of human activities, the fixed-size sliding window may not produce an optimal segmentation process, particularly during transitional activity. Hence, selecting the optimal window size is a challenging and crucial task, especially if transitional activities are considered. The researchers attempted to enhance the segmentation method by proposing various techniques. However, most of them focus on each window's features, and few consider the temporal relationships between the adjacent windows. Therefore, an analysis of the impact of window size on the performance of basic and transitional activity recognition is performed using a deep learning model. Then, two similarity-based approaches are proposed. The first approach is a similarity-based segmentation approach that exploits the temporal structure of the activity signal by comparing the similarity between the adjacent windows. Specifically, the inner features are extracted using feature engineering for each window, and then the similarity of the features between the adjacent windows is measured using threshold values. This enables the segmentation stage to be more successful by allowing the model to distinguish between the transitional and non-transitional windows. While feature engineering and threshold-

based methods required some experiments and some experts, a deep similarity segmentation model is proposed to enhance the similarity segmentation approach and avoid feature engineering and threshold-based methods. The deep learning model treats the segmentation task as a binary classification task. It extracts the local features of each window by using convolutional neural network layers. Then it extracts the temporal features by measuring the similarity between the features of adjacent windows. Both features are combined to present the final features for each window. The proposed approaches are evaluated using two public datasets and compared with the performance of state-of-the-art HAR models. The experimental results show that the first proposed method achieved an accuracy of 92.71% and 86.65% for both datasets, and the second proposed model achieved an accuracy of 93.35% and 84.96% for both datasets. These results outperformed the fixed sliding window result as well as outperformed the performance of state-of-the-art models.

CHAPTER 1

INTRODUCTION

1.1 Background

Humans perform various tasks and activities in daily life. Improving an individual's life in various fields requires detecting and identifying these activities. Human Activity Recognition (HAR) is concerned with the ability to recognize human activities in order to understand human behaviors. Also, it is able to learn extensive information about humans. Understanding human activity helps improve human life. The potential application of HAR is widespread in various applications of the Internet of Things (IoT) (Acampora et al., 2020). IoT is a network of physical objects that are integrated with computing components and can transmit and receive data on their own to provide tailored services such as remote activity monitoring, identifying falls in the elderly (Ghahramani et al., 2020; Parvaneh et al., 2017), security (Ali et al., 2022), fitness and lifestyle (Zhang et al., 2022), surveillance systems (Islam et al., 2022; Liu et al., 2018), sport (Ghazali et al., 2018; Whitlock et al., 2018), entertainment systems (Islam et al., 2022), and smart home (Aminikhanghahi & Cook, 2019; Bermejo et al., 2021). Thus, over the past decade, human activity recognition (HAR) has become an important and vibrant field of research (K. Chen et al., 2021; Kumar & Hamirpur, 2021).

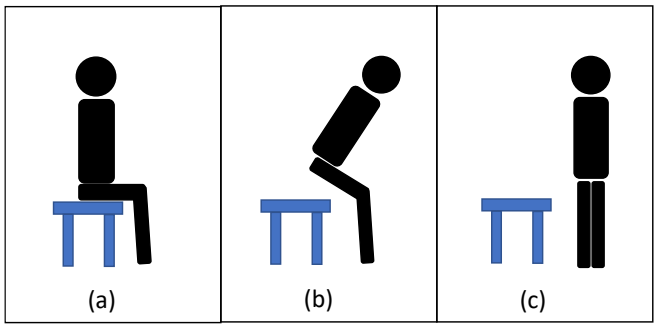

Figure 1.1. Example of activity types: a) basic activity (sitting). b) transition activity (sit to stand). c) basic activity (standing).

In HAR, generally, human physical activities can be categorized into basic and transitional activities. Basic activity (BA) is the activity performed by humans in daily life, such as running, standing, lying down, walking, and sitting, and the transitional activity occurs between two successive basic activities, such as standing-to-sit, sitting-to-stand, and sit-to-lie. Figure 1.1 shows examples of basic and transitional activities. Recognizing transitional activities is more challenging than recognizing basic activities because the basic activity lasts for a longer duration than the transitional activity. Although human activity is affected by human variability (Jimale & Mohd Noor, 2021), the main characteristics of transitional activities are that they are performed in a shorter time period and with a lower incidence rate compared with basic activities. Specifically, the average duration of basic activities is 20.1s, while the average duration of transitional activities is 3.7s (L. Chen et al., 2020). This low duration may lead to difficulty in recognizing transitional activities (Li et al., 2019)(Reyes-Ortiz et al., 2016). Thus, only a few researchers considered transitional activities in their HAR models. However, transitional activities are essential in various applications, such as pervasive healthcare and activity monitoring. The importance of detecting and recognizing transitional activities manifests in the following applications: identifying falls in the elderly (Ghahramani et al., 2020; Wairagkar et al., 2021), fitness and lifestyle (Zhang et al., 2022), and smart home (Aminikhanghahi and Cook, 2019; Bermejo et al., 2021).

Human activity recognition can be performed through two sensing approaches: vision-based and sensor-based approaches (Minh Dang et al., 2020). Vision-based methods rely on visual features captured by cameras, while sensor-based approaches rely on signal features collected by sensors. Although vision-based approaches have some advantages (e.g., treating multiple users at the same time, treating entire body parts, and showing rich details for every activity), they also have several disadvantages

and constraints, such as the need for pre-installation, angle, location, illumination, potential obstruction, privacy, high computation, and complexity (Minh Dang et al., 2020). These constraints lead the researchers to go toward the sensors-based approach, especially since it has several advantages, including high sensitivity, location independence, low computation cost, simplicity of implementation, small data size, and suitability for real-time systems.

Wearable devices are the most widely used devices for HAR that embed accelerometer and gyroscope sensors, such as smart glasses, smartwatches, and smartphones (Abdel-Salam et al., 2021; Alves et al., 2020; Roobini & Fenila Naomi, 2019; Zhang et al., 2022). In addition, the accelerometer and gyroscope are the most commonly used devices in HAR (Demrozi et al., 2020; L. Zhou et al., 2020). Thus, using wearable devices to recognize human activity leads to a better understanding of human behavior in daily life.

1.2 Human Activity Recognition Stages

In general, sensor-based HAR goes through three basic stages: signal segmentation, feature extraction, and classification stages.

The signal segmentation stage separates the signals into several subsequent parts (segments) called windows (Li et al., 2019). This is typically done by using the sliding window approach. Each window contains a group of samples. Window size term refers to the number of samples inside the window. The fixed-size sliding window method is the most commonly used signal segmentation method. In this method, all windows have equal sizes (number of samples).

The feature extraction stage aims to reduce the input data size while preserving the description of the activities (Islam et al., 2022). Only significant information about the activity is retrieved.

The classification stage uses the extracted features to classify the activity type. The researchers utilize a variety of classifiers, including deep learning and machine learning models.

1.3 Research Problem

In the signal segmentation stage, the signal should be segmented into a sequence of multiple windows to describe the activity. The common technique for signal segmentation in HAR is the fixed-size sliding window, which divides the signals into windows of equal sizes (Zhang et al., 2022). In the signal segmentation stage, selecting an optimal window size is crucial for feature extraction and activity classification, especially the transitional activity, as it directly impacts the accuracy and efficiency of the classification models. However, selecting the optimal window size is a challenging task and not scalable due to the varying duration of human activities. (Atalaa et al., 2020; Ferrari et al., 2021; Qian et al., 2021). Selecting a small window size could split a particular activity over several windows and provide less information about the activity. On the other hand, a large window size might include data from different activities, which causes overlap between them and increases noise. Therefore a fixed window size for the signal segmentation stage is not the most effective way to perform activity recognition, especially if the transitional activity is considered (Li et al., 2019; Noor et al., 2017; Zhang et al., 2022).

Furthermore, the fixed sliding window does not consider the relationship between successive windows. It only allows the model to leverage local (current window) features.

Most state-of-the-art studies proposed an adaptive and dynamic window size to overcome the limitations of the fixed-size sliding window method. In the dynamic sliding window, the sensor data is segmented into different sizes of windows based on specific features (Alhammad & Al-Dossari, 2021; Noor et al., 2017). Most of these studies do not consider the transitional activities in their experiments, which could lead to low accuracy in recognizing the transitional activities. In addition, they do not consider the temporal relation between adjacent windows and the similarity concept in their studies.

Most signal segmentation studies rely on feature engineering methods for the feature extraction process, which involves extracting relevant features from signals by using some ML methods (Atalaa et al., 2021). These methods are often threshold-based, which requires predefined threshold values through experiments or designer experience. A convolutional neural network (CNN) is a deep learning (DL) framework that can extract features automatically through a learning-based pipeline without using or capturing any statistical information. CNN has done well in most areas due to its ability to automatically extract discriminative features, whether local features or the features that describe the relationship between adjacent windows. Although the deep learning technique has shown outperformance and efficiency in the processes of classification and feature extraction, few deep learning models are built for the signal segmentation stage (Islam et al., 2022).

1.4 Research Questions

1) How does the window size in the fixed sliding window method affect the performance of the human activity recognition model which obtains the best performance of detection both basic and transitional activities?
2) Does the similarity between the features of adjacent windows can be used to distinguish between basic and transitional windows to overcome the fixed-sliding window limitations?
3) Seeking for optimal deep learning model which able to measure the similarity between the features of successive windows automatically in order to distinguish between basic and transitional activities without using a threshold-based method.

1.5 Research Objectives

The aim of this thesis is to design a robust and effective signal segmentation approach for sensor-based human activity recognition. The main goal has been carried out through the following three research objectives:

1) To analyze the impact of the window size on the fixed-size sliding window method on the performance of activity recognition models that consider both basic and transitional activities.
2) To design a novel signal segmentation approach which extract the inner features of each window through measures the signal similarity inside each window, and then measures the dissimilarity of these inner features between adjacent windows., In order to distinguishes between transitional and basic activity to overcome the limitations of the fixed sliding window method.

3) To improve the signal segmentation approach based on a deep learning technique by building a deep similarity segmentation model able to extract the local and similarity features automatically without using feature engineering and threshold-based methods.

1.6 Research Scope

This research focuses on human activity recognition using sensor-based wearable devices, considering the following aspects:

- This research recognizes simple basic (e.g., running and walking) or transitional activities (e.g., sit-to-stand and sit-to-lying). As a result, there is no treatment for the complex activity that takes a long time to perform (e.g., eating) or the activity that consists of the consequence of multiple basic activities (e.g., cooking).

- The accelerometer and gyroscope sensors are used in the experiments regardless of other sensors.

1.7 Research Contributions

This research aims to improve the signal segmentation task by designing two ways to distinguish between basic and transitional activity windows. Both methods rely on the similarity feature of the adjacent windows. The first approach used the features engineering method, whereas the other used the deep learning method. The research contributions can be summarized by importance as follows:

A. Analysis of the impact of window size on the performance of human activity recognition models that considered basic and transitional activities.

B. Build a similarity segmentation approach (SSA) that can distinguish between the basic and transitional activity windows by using feature engineering to measure the similarity between adjacent windows.
C. Build a deep similarity segmentation (DSS) model that can distinguish between the basic and transitional activity windows based on deep learning techniques, in order to enhance the SSA by extracting the local and similarity features by using deep learning techniques to avoid feature engineering and threshold-based methods.

1.8 Thesis Organization

The thesis consists of seven chapters. The rest of the thesis is organized as follows:

The second chapter discusses the literature review of human activity recognition, as well as explaining the most important state-of-the-art of related works with their contributions. Chapter Three gives our general framework and the research methodology used. As well as illustrating the evaluation benchmarks and metrics used in the research. Chapter Four analyzes the impact of the performance of the human activity recognition model for both basic and transitional activities. In Chapter Five, the similarity segmentation approach is explained with experimental results and compared with the results of state-of-the-art models. In Chapter Six, the deep similarity segmentation model is explained with a discussion of the experimental results and a comparison of the results with the results of state-of-the-art models. The last Chapter shows the conclusion and future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter explains the fundamentals of human activity recognition and covers related concepts. Also, the related works are explained in the research field, sorted by HAR stages, with their contributions.

The chapter is organized as follows: the definition of human activity recognition in Section 2.2. The sensor types are described in Section 2.3. The human activity types are explained in Section 2.4. Section 2.5 explains HAR stages and provides some contributions from state-of-the-art approaches for each stage. In Section 2.6, the state-of-the-art signal segmentation approaches are discussed. The gap analysis is discussed in Section 2.7. Finally, chapter summarization is provided in Section 2.8.

2.2 Human Activity Recognition

In daily life, humans perform many different types of activities, including running, standing, lying down, walking, and sitting. Human Activity Recognition (HAR) is the ability to recognize human daily activities in order to understand human behaviors and learn extensive information about human activity. Understanding human activity helps to improve human life. However, different types of sensors capture diverse human activities. The following sections provide details on the different kinds of sensors and activities.

2.3 Sensor Types

Human activity recognition can be classified into two categories: vision-based and sensor-based (Minh Dang et al., 2020). Vision-based methods rely on visual characteristics and features captured by a camera to identify human activity. This type offered several advantages, including the ability to treat multiple users at the same time, treat entire body parts, and show rich details for every activity. On the other hand, some points limit dependence on it for recognizing human actions, the most important of which are:

- Affected by factors of camera constraints, such as angle, location, illumination, potential obstruction, and privacy.
- Camera constraints such as angle, location, illumination, potential obstruction, and privacy all have an impact.
- Affected by environmental factors, such as lighting.
- Problems of capture angle, such as occlusions.
- The large size of the treated data.
- High costs of processing (time and effort).
- High computation and complexity.
- It is difficult to use in low-power real-time applications.

Sensor-based methods capture and collect the signals from sensors, which is considered a type of time series data. This kind offers the following benefits: High sensitivity, location independence, low cost (less memory and calculation time), simplicity of implementation, small data size, and suitability for real-time systems.

The limitations of vision-based sensors and the advantages of sensor-based recognition led researchers to turn to sensor-based human activity recognition.

However, there are three common types of sensors: wearable, dense, and hybrid. The wearable sensors are worn or put in any human body position, as shown in Figure 2.1. There are two types of wearable sensors: physiological and inertial. Physiological sensors (body-worn) detect involuntary physiological signals such as skin conductance, heartbeat, and electrical muscle activity. Usually, these sensors require special pre-installation on the body before use. Inertial sensors are widely used as embedded sensors in wearable devices called Inertial Measurement Units (IMU) (L. Zhou et al., 2020). The main advantages of this type are that it is lightweight and low-cost. There are three sensors that are most commonly used in this type: the accelerometer, gyroscope, and magnetometer. The accelerometer measures the speed, displacement, and acceleration along three axes (i.e., x, y, and z). The magnetometer measures the earth's magnetic field (Roobini & Fenila Naomi, 2019; Zhang et al., 2022). All sensors measure among three linear axes: x, y, and z.

Dense sensors are usually used to infer user-object interaction. There are two types of this sensor. The first type is environmental sensors, which monitor how the environment changes when the user engages in an activity. e.g., temperature, humidity, and particulate matter. Based on the activities, the selection of appropriate environmental sensors needs to be carefully made (Minh Dang et al., 2020). The second type is object sensors, which are attached to a piece of equipment used for a particular task. For instance, a sensor might be integrated into a smart cup to track consumption and examine drinking patterns. The primary issues with this type of sensor are setup and high expenses (Minh Dang et al., 2020).

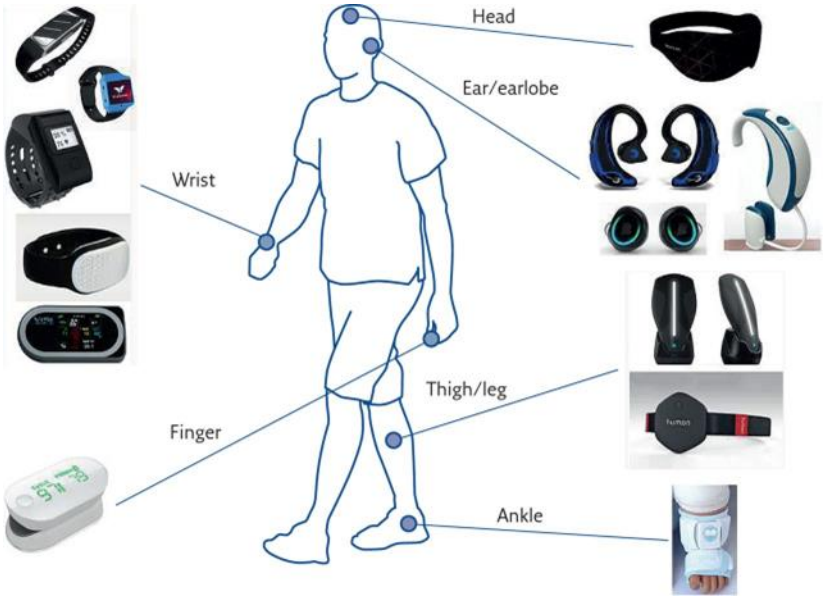

Figure 2.1. Wearable devices (Aliverti, 2017).

Hybrid sensors combine different sensor types to provide a precise human activity recognition system. The fusion stage is often challenging as there are multiple sensor types involved (Demrozi et al., 2020). To enhance the model performance, several researchers fuse multiple inertial sensors in their models (K. Chen et al., 2021; J. Wang et al., 2021). Despite integrating multiple sensors to give more information about motion than using a single sensor, sensor fusion approaches encounter several difficulties, including the choice of sensors, locations, and sync (Rahn et al., 2021).

2.4 Human Activity Types

Human activities can be divided into many categories according to a variety of factors, including subjects, duration time, motion, and the number of actions.

2.4.1 Based on Duration Time

The amount of time needed to perform any activity is referred to as duration time. Human activities are divided into two groups based on the duration of time: basic

activities (BA) and transitional activities (TA). The basic activity is what people do every day (e.g., walking, running, and standing), while the transitional activity occurs between two basic activities, such as sit-to-stand. There are two categories of TA: long and short. The long activity is carried out over an extended period of time. It often includes many supporting actions (e.g., walking activity, which consists of standing action). The short activity requires less time to complete (e.g., standing). The second group includes transitional activities (or posture activities), which are activities that are in between two basic activities.

The key characteristics of TA are that it has a low duration time and a low incidence rate. These characteristics led some researchers to neglect the transitional activities in their HAR model. They assume that these characteristics do not significantly affect the model's performance, especially when they are interested in recognizing the basic activity only (Abid et al., 2021; Alhammad & Al-Dossari, 2021; Oluwalade et al., 2021; Pilario et al., 2020; Y. Tang et al., 2021). Of course, this neglect leads to the non-application of these models in many applications that need to detect transitional activities, as mentioned earlier. Experimentally, some researchers prove that recognition of the TA affects classification accuracy. Specifically, the accuracy of the classification models that neglected the TA is higher than that considered it (Acampora et al., 2020; Hessfeld et al., 2021; Mohd Noor et al., 2022).

Furthermore, in deep learning models, the training stage requires a sufficient amount of labeled data to obtain the optimal training. Since the number of transitional activities data within the dataset is lower than the number of basic activities data, due to its characteristics, the problem of data imbalance occurs. The imbalance of data is considered one of the HAR DL model challenges (K. Chen et al., 2021). When using

an imbalanced dataset, conventional models tend to predict the class with the majority number of training samples while ignoring the class with few available training samples (K. Chen et al., 2021).

2.4.2 Based on Motion

The activities can be divided into static and dynamic categories (Lawal & Bano, 2020). Standing and sitting activities are examples of static activities because they don't involve any movement while being performed. Walking, leaping, and running are examples of dynamic activities that include motion while being performed. As a result, dynamic activities involve more motion than static ones. Figure 2.2 shows the differences in motion sensor signals between dynamic and static activities.

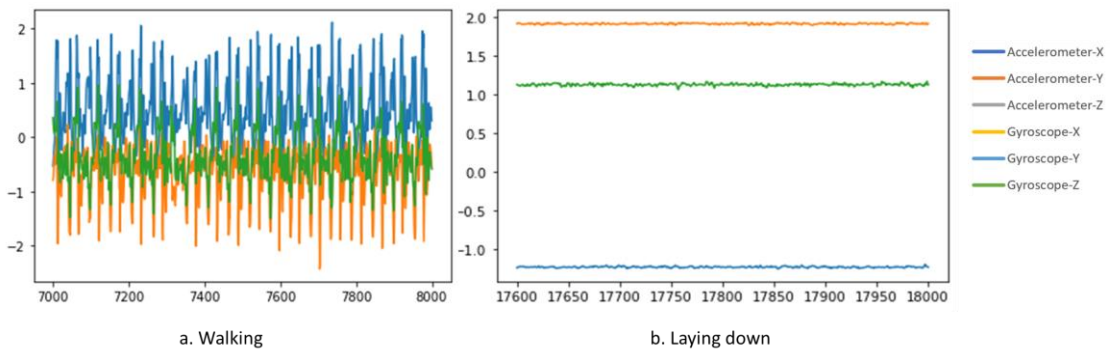

Figure 2.2 The pattern difference between static activity (lying down) and dynamic activity (walking).

2.4.3 Based on Number of Actions

Based on the number of actions, the activities are divided into simple and complex activities (Ferrari et al., 2021). Simple activities are performed on one stage, such as walking, sitting, and jumping. A complex activity consists of more than one stage, such as cooking and house cleaning.

2.5 Human Activity Recognition (HAR) Stages

Sensor-based human activity recognition goes through several stages (Ferrari et al., 2021), as shown in Figure 2.3.

2.5.1 Data Pre-processing

The captured signals from sensors usually have some noise and need some processing to make them suiTable for recognition tasks. The following are the most common preprocesses used on signals:

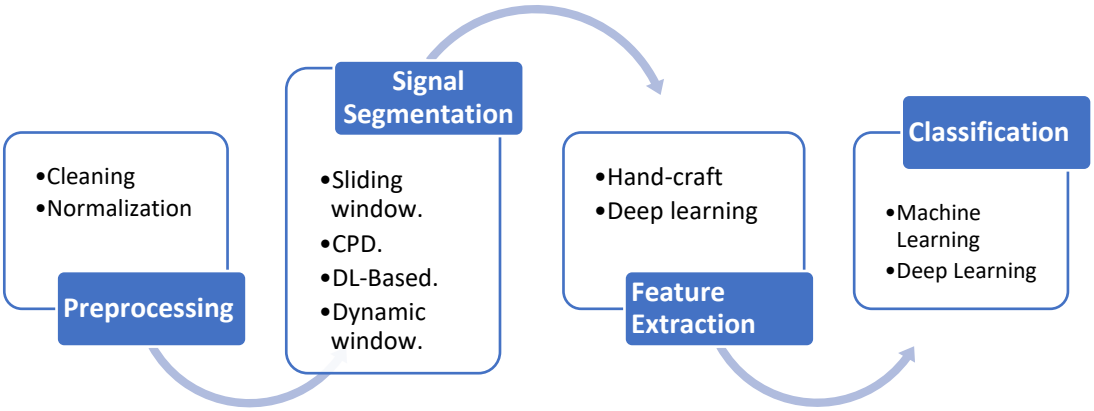

Figure 2.3. Main stages of sensor-based human activity recognition.

Cleaning: this process is used to remove the noise and undesired data from signals that affect recognition performance negatively.

Normalization: this process aims to scale the signal data, which makes the training process of the classification network more robust and less sensitive to covariate or distribution shifts (Lu et al., 2021).

2.5.2 Signal Segmentation

The wearable sensors read the signals over time. To reduce the size of the read data, a sampling process is performed. The sampling process represents the time dependence of the signal by a discrete set of samples (Weik, 2000). A single sample extracted from the sensors at a specific time does not give sufficient data to identify an activity (Minh Dang et al., 2020). Thus, the activity is represented by a group of samples. The number of samples taken per second is called the sample rate. Siirtola et al. (2011) demonstrated that 10Hz (i.e., 10 samples per second) is sufficient to recognize daily activities, and others adopt 30Hz (Alhammad & Al-Dossari, 2021). As a result, the signals should be split into small fragments to allow the model to recognize the activity through the stages of feature extraction and classification.

The goal of the signal segmentation process is to partition the signals into small continuous fragments. Each fragment is called a window (or segment). Each window contains a sequence of data samples from the sensors, whether equal or unequal in length. Window size refers to the number of samples inside the window. So, the number of samples for long activities is larger than for short activities. The optimal window size contains all samples of the specific activity.

The essential mistake is that some researchers believe that the signal segmentation stage is a pre-processing task and is done in advance. In fact, the errors in the signal segmentation stage may propagate to later steps (Lima et al., 2019; Qian et al., 2021). Thus, the accuracy of the signal segmentation method affects the accuracy of the recognition process, so it is considered a challenging task (Akbari et al., 2018; Atalaa et al., 2020).

Previous studies have indicated that the signal segmentation process affects the performance of HAR stages, whether model accuracy, quality of features, or training time (Atalaa et al., 2020). Atalaa et al. (2020) investigated the effects of window size on the HAR model accuracy by using an ANN classifier. They find that the performance of the HAR model is different due to the difference in window size (Atalaa et al., 2020). Ghazali et al. (2018) performed the same investigation and got the same results using Decision Trees, SVM and KNN (Ghazali et al., 2018).

The sliding window is the most popular method used for the signal segmentation process. The sliding window method slides within the signals to generate the temporal windows. Two issues are related to the window: window size and the overlapping between windows. In the sliding window method, two consecutive windows can be overlapped such that a certain number of samples from the previous window are included in the current window. The degree of overlap is also known as window shift. There are two types of windows (Jordao et al., 2018)**:** Non-Overlapping Windows: no overlap between windows. A lower number of samples are generated. This approach is widely used in HAR (Banos et al., 2014). Overlapping Windows: which presents 50% overlap between temporal windows, where the half data of the first window is the same as the half data of the next window, causing an increased sampling number. However, while window size has a high effect on the recognition performance, the overlapping between windows also affects the recognition model performance (Li et al., 2019).

2.5.3 Features Extraction

The method of reducing the amount of raw data input while keeping the activity descriptions is known as feature extraction. This procedure is essential since the raw data input from the sensors is significant and can include duplicate data. The process of

extracting features from unstructured data and converting them into forms compatible with machine learning models is known as feature engineering (Z. Alice & C. Amanda, 2018). The robust features represent the activities well and do not lose any important information, which leads to an accurate activity classification. However, there are two types of features: features engineering (statistical) or learning-based features.

The statistical features are extracted through handcrafting and feature engineering. This type is usually used in classical machine learning (ML) models. ML methods perform the feature extraction step by experts or by calculating a set of statistical functions, which is time-consuming. These methods are called features engineering features (e.g., mean, minimum, maximum, standard deviation (STD), autoregressive (AR), and root mean square) (Acampora et al., 2020; Alhammad & Al-Dossari, 2021; Alves et al., 2020; Issa et al., 2022; Lone et al., 2021; Ni et al., 2018). Furthermore, the kernel methods are able to learn non-linear transformations of input data as implicit features (Qian et al., 2019; Yulita & Saori, 2019).

L. Chen et al. (2020) used Fisher-Score, Relief-F, and Chi-Square to select several features to obtain a relatively good feature set. Various feature selection algorithms are used to select the higher-scoring features according to the specific classification. A variety of machine learning methods are used to classify and select the one with the highest classification accuracy. A support vector machine is used to classify the posture activity. Lone et al. (2021) used different methods for feature extraction and selection methods (PCA, Chi-squared, Relief, RFE, Boruta), and five different types of machine learning classification algorithms (SVM-L, Adaboost.M1, Stochastic GBM, XGBoost, AvNNET) to classify the basic activities and postural transitions (Lone et al., 2021).

Qian et al. (2021) proposed a weakly-supervised feature extraction method based on the technique of kernel embedding of distributions. This technique jointly segments sensor streams and then extracts statistical features from each segment (Qian et al., 2021). Acampora et al. (2020) proposed a feature buffer unit to improve the classification accuracy considering TA activity. Specifically, the extracted feature space is augmented with features coming from previous classification steps, then passed to the neural network (Acampora et al., 2020).

Furthermore, the number of features is a crucial factor that faces hand-crafted features. The model won't be able to complete the task at hand if there aren't enough informative characteristics. The model cost is high and is harder to train if there are too many features or if the majority of them are useless. The performance of the model might be impacted if the training procedure goes wrong in any way (Z. Alice and C. Amanda, 2018).

On the other hand, the learning-based features have been automatically extracted through deep learning (DL) models, which do not capture any statistical information. Temporal features are one of these types. Recently, DL has done well in most areas due to its ability to extract discriminative features automatically rather than using feature engineering methods. Therefore, most researchers have lately adopted DL methods to build their models and enhance their performance. The Convolution Neural Network (CNN) is the most widely used framework for DL (Mohd Noor et al., 2022). Mohd Noor (2021) proposed an unsupervised feature learning method in order to automatically extract and select the features. The proposed method jointly trains a convolutional denoising autoencoder with a convolutional neural network to learn the underlying features and produce a compact feature representation of the data (Mohd

Noor, 2021). Whitlock et al. (2014) proposed a model based on multi-task recurrent neural networks (RNN) to segment and recognize activities and cycles. The local context features are extracted by CNN and captured by RNN for longer-range local dependencies. This allows a longer-range transfer of knowledge from previous time steps to the current time step (Whitlock et al., 2014).

2.5.4 Classification Stage

For a decade year, researchers used many classifier techniques to classify human activity recognition. These techniques can be classified into two types: classical ML or DL techniques.

2.5.4(a) Machine Learning Models

Classical ML models usually adopt feature engineering for extracting and selecting the relevant features and modeling the features using machine learning techniques. The most commonly used ML algorithms are: Support Vector Machine (SVM) (L. Chen et al., 2020; Lawal & Bano, 2019; Ni et al., 2018; Pamplona-Beron et al., 2021; Yulita & Saori, 2019), Random Forest (Erdaş et al., 2016; Li et al., 2019), Naïve Bayes(Noor et al., 2017), k-NN(Ni et al., 2018; Noor et al., 2017), or decision tree (DT) (L. Chen et al., 2020; Ni et al., 2018; Noor et al., 2017).

Shi et al. (2020) distinguished transitional activity windows from basic activity windows through the standard deviation trend analysis (STD-TA) method. The signal is segmented by the sliding window method, and the window size is determined by the duration of a single activity. A group of statistical features is extracted, including mean, variance, and standard deviation (STD). Then it is passed into the SVM classifier for the classification task. While the accuracy of activity recognition depends on the sensor

placements (Rahn et al., 2021), the data is collected by putting the smartphone on the right leg and using an accelerometer with barometer sensors to try to obtain the best accuracy of classification (Shi et al., 2020).

Acampora et al. (2020) proposed a memory-based Artificial Neural Network (MANN) architecture to improve classification accuracy by considering TA activity. The MANN architecture extends the neural network with short-term memory information about the previous activities' features, where a memory buffer is used to store the information about features related to previous states. Specifically, the extracted feature space is augmented with features derived from previous classification steps. The extracted features include the mean, standard deviation, median, maximum, and minimum. An ANN classifier is used for the classification process, and it got 95.48% overall accuracy (Acampora et al., 2020).

Lone et al. (2021) investigated and analyzed the performance of different machine learning algorithms with various dimensionality reduction techniques. They used five different types of machine learning classification algorithms (SVM-L, Adaboost.M1, Stochastic GBM, XGBoost, AvNNET) to classify the basic activities and postural transitions. Different methods are used for feature extraction and selection, including PCA, Chi-squared, Relief, RFE, Boruta (Lone et al., 2021).

2.5.4(b) Deep Learning Models

The biggest drawbacks of feature engineering and ML techniques are that they can be laborious, time-consuming, difficult to estimate how many features there are, and prone to error. Deep learning techniques come to overcome the limitations of machine learning techniques. Recently, the adoption of deep learning for classification

has garnered the interest of various researchers and has become extremely popular in the HAR field in recent years (K. Chen et al., 2021; Ige & Mohd Noor, 2022; Kumar & Hamirpur, 2021; Nafea et al., 2021). With deep learning, optimal features can be extracted from the activity signals automatically. Thus, DL outperforms machine learning models in the recognition task in terms of classification performance measures such as accuracy, precision, recall, and F1-score (Islam et al., 2022; Nafea et al., 2021; Zhang et al., 2022).

There are several DL techniques that adopted by HAR studies such as Short-Term Memory (LSTM) (Whitlock et al., 2014), Convolutional Neural Network (CNN) (Ali et al., 2022), Autoencoder, Deep Belief Network (DBN), RNN, Deep Reinforcement Learning (DRL), Convolutional Long Short-Term Memory (Conv-LSTM) (Moreira et al., 2021), and Hybrid DL Models (Oluwalade et al., 2021; H. Wang et al., 2020).

Mohd Noor (2021) proposed an unsupervised feature learning method based on a denoising autoencoder, which aims to extract and select the discriminative features for the activity recognition task. The data is segmented by using an adaptive sliding window method presented in (Noor et al., 2017). Then the features are represented by denoising the autoencoder using 1D convolutional and 1D max-pooling layers. Also, proposed a joint training approach to enhance the reconstruction task (Mohd Noor, 2021).

Furthermore, some studies adopted long short-term memory (LSTM) to model the temporal sequences between windows. Xia et al. (2020) built a classification model consisting of two layers of LSTM followed by CNN layers. A fixed sliding window is

used for the segmentation task. The model achieved a significant performance; however, transitional activities were not considered in the study (Xia et al., 2020).

Irfan et al. (2021) proposed a hybrid multi-model activity recognition using by utilizing multiple deep learning models simultaneously. The proposed model integrates three deep learning models to classify the basic activities and transitional activity windows. The feature data is passed into the models simultaneously. The fusion of the results is done using the class probabilities of each model (Irfan et al., 2021). Mohd Noor et al. (2022) proposed a hybrid deep learning model which uses the deep temporal Conv-LSTM architecture. The proposed model consists of concurrent feature learning pipelines using CNN to extract the features from the windows. extraction. The model is integrated with a sequence learning module to learn the temporal features from the concatenated window features. This utilizes both temporal features and the relationship of windows (Mohd Noor et al., 2022). Dirgová Luptáková et al. (2022) proposed a transformer model for activity recognition using a deep learning model with an attention mechanism (Vaswani et al., 2017). The self-attention mechanism is inherent in the transformer, which expresses individual dependencies between stream signals (Dirgová Luptáková et al., 2022).

In order to achieve optimal performance for the DL model, it requires a large amount of data for the training stage. As a result, the accuracy of recognizing the basic activities is clearly higher than the accuracy of recognizing the transitional activities (Irfan et al., 2021; Jimale & Mohd Noor, 2021; Lone et al., 2021). This is due to the imbalance of data within the dataset, where the number of basic activity instances is much higher than transitional activities, which makes it difficult for the model to adequately learn the features of the transitional activities. In addition, most studies

treated each segment separately regardless of any relation between segments, whether the previous or the next of the current window.

2.5.4(c) Hybrid Models

Some studies combine ML and DL methods to leverage both feature engineering and learning-based features. Abid et al. (2021) proposed combining ML and DL methods through three pipelines. The first encompasses feature engineering-based classifiers, where ReliefF for feature selection is combined with an SVM. The second pipeline encompasses feature learning-based classifiers, which use linear discriminant analysis (LDA) as an automatic feature extractor. The last pipeline used the CNN architecture for both feature extraction and classification (Abid et al., 2021). Furthermore, such a task can be considered a sequence-to-sequence problem rather than a classification problem (Whitlock et al., 2014).

However, all previous studies focus on supervised learning. Supervised learning refers to the use of labeled data in the training stage, which helps the model learn more about the features of each label (or class). Labeling the data requires more human effort and is time-consuming. Some studies adopted other types of learning to avoid the labeling process or less use of labeled data, such as semi-supervised, weakly-supervised, unsupervised, or reinforcement learning.

Semi-supervised learning is halfway between supervised and unsupervised learning, where a small amount of data in the dataset is labeled and a lot is unlabeled (Devgan et al., 2020). C. I. Tang et al. (2021) proposed a training pipeline that combines self-training and self-supervised learning techniques. This allows deep learning models to learn more generalizable features by leveraging unlabeled data (C. I. Tang et al.,