Fuzzy Transfer Learning in Human Activity Recognition

David Ochoechi Ada Adama

School of Science and Technology

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

Doctor of Philosophy

May 2020

This thesis is dedicated to my father, I know you will be proud of this milestone accomplished.

Acknowledgements

The work described in this thesis was carried out at Nottingham Trent University between June 2016 and June 2019, while I was working as a full-time doctoral research student.

All glory to God Almighty for making this a reality.

My sincere gratitude to my director of studies, Prof. Ahmad Lotfi, for his guidance throughout the course of my research. Your consistent motivation and kind words will never be forgotten. Many thanks for making your office accessible despite your busy schedules. Also, I will like to thank the team of co-supervisors I had at different stages of my research: Dr. Caroline Langensiepen I am grateful for the creative ideas you always had to offer and for the many corrections you made to the publications leading to this thesis. I wish you were on my supervisory team till completion. Thank you Dr. Kevin Lee for the supervision given at the early stage, especially, for your invaluable corrections to the few publications we shared. Lastly, to Dr. Robert Ranson who came on board at a later stage, many thanks for the contributions made to the progress of my research.

My sincere appreciation to my family for their support through the course of my research. To my mother, Dr. Omeche Onoja, words fail me to describe what a blessing it is to be your son. Your love, constant prayers, words of wisdom and discipline have contributed in my journey to achieving this feat. I pray you live to enjoy the bountiful blessings of your labour. To all other family members, my step father, Prof. Sam Baba-Onoja, and siblings, I say God bless you all for your support. Dante and Ene Benita, I owe you guys a lot.

Thank you for the care, support and the home you provide here in UK.

Finally, I would like to appreciate my friends and colleagues at NTU for the time shared during the past few years of my research. Specifically, my fellow PhD researchers in the Computational Intelligence and Applications research group. To everyone who in one way or the other contributed to the success of my research, not forgetting my friend Pedro Trindade, I am most grateful.

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

David Ochoechi Ada Adama
May 2020

Abstract

Assisted living environments are incorporated with different technological solutions to improve the quality of life and well-being. In recent years, there has been a growing interest in the research community on how to develop evolving solutions to aid assisted living. Different techniques have been studied to address the need for technological systems which are intelligent enough to evolve their knowledge to solve tasks which have not been previously encountered. One such approach is Transfer Learning (TL), for example, between humans and robots.

Humans excel at dealing with everyday activities, learning and adapting to different activities. This comprises different complex techniques which enable the lifelong learning process from observation of our environment. To obtain similar learning in assistive agents, TL is needed. The aim of the research reported in this thesis is to address the challenge associated with learning and reuse of knowledge by assistive agents in an Ambient Assisted Living (AAL) environment. In this thesis, a novel approach to transfer learning of human activities through the combination of three methods; TL, Fuzzy Systems (FS) and Human Activity Recognition (HAR) is presented. Through the incorporation of FS into the proposed approach, uncertainty that is evident in the dynamic nature of human activities are embedded into the learning model.

This research is focused on applications in assistive robotics. This is with a purpose of enabling assistive robots in AAL environments to acquire knowledge of such activities as are performed by humans. To achieve this, an extensive investigation into existing learning

methods applied in human activities is conducted. The investigation encompasses current state-of-the-art of TL approaches employed in skill transfer across different but contextually related activities.

To address the research questions identified in the thesis, the contributions of the methodology employed are in three main categories; 1) Firstly, a novel framework for human activity learning from information observed. Experiments are conducted on selected human activities to acquire enough information for building the framework. From the acquired information, relevant features extracted are used in a learning model to recognise different activities. 2) Secondly, the sequence of occurrence(s) of tasks in an activity needs to be considered in the learning process. Therefore, in this research, a novel technique for adaptive learning of activity sequences from acquired information is developed. 3) Finally, from the sequence obtained, a novel technique for transfer of human activity across heterogeneous feature space existing between a human and an assistive robot is developed. These categories form the basis of the TL framework modelled in this research.

The framework proposed is applied to TL of human activity from data generated experimentally and benchmark datasets of various classes of human activities. The results presented in this thesis show that exploring the process of human activity learning is an important aspect in the TL framework. The features extracted sufficiently distinguish relevant patterns for each activity. Also, the results demonstrate the ability of the methodology to learn and predict human actions with a high degree of certainty. This encourages the use of TL in assisted living environments and other applications. This and many more applications of TL in technology would be a potential driver of the next revolution in artificial intelligence.

Publications

As a result of the research presented in this thesis, the following publications have been published:

David Ada Adama, Ahmad Lotfi and Robert Ranson. Adaptive Segmentation and Sequence Recognition of Human Activities from Skeleton Data. Expert Systems with Applications. March 2020 [Under review].

David Ada Adama, Ahmad Lotfi, Caroline Langensiepen, Kevin Lee and Pedro Trindade. Human Activity Learning for Assistive Robotics using A Classifier Ensemble. Soft Computing (2018) 22: 7027. DOI: 10.1007/s00500-018-3364-x.

Gadelhag Mohmed, David Ada Adama, and Ahmad Lotfi. (2019). Fuzzy Feature Representation with Bidirectional Long Short-Term Memory for Human Activity Modelling and Recognition. In: Advances in Computational Intelligence Systems (UKCI 2019), September 4-6, 2019.

David Ada Adama, Ahmad Lotfi, and Robert Ranson. (2019). Fuzzy Transfer Learning of Human Activities in Heterogeneous Feature Spaces. In Proceedings of 12th Conference on Pervasive Technology Related to Assistive Environments (PETRA'19), June 5-7, 2019. DOI: 10.1145/3316782.3322786.

David Ada Adama, Ahmad Lotfi, Robert Ranson, and Pedro Trindade. (2019). Transfer learning in Assistive Robotics: From human to robot domain. In: The 2nd UK-RAS Conference on Embedded Intelligence (UK-RAS19), pp. 6063.

David Ada Adama, Ahmad Lotfi and Caroline Langensiepen. (2019). Key Frame Extraction and Classification of Human Activities using Motion Energy. In: Advances in Computational Intelligence Systems. UKCI 2018. Advances in Intelligent Systems and Computing, vol 840. Springer, Cham. DOI: 10.1007/978-3-319-97982-3_25.

David Ada Adama, Ahmad Lotfi, Caroline Langensiepen and Kevin Lee. (2018). Human Activities Transfer Learning for Assistive Robotics. In: Advances in Computational Intelligence Systems. UKCI 2017. Advances in Intelligent Systems and Computing, vol 650. Springer, Cham. DOI: 10.1007/978-3-319-66939-7_22.

David Ada Adama, Ahmad Lotfi, Caroline Langensiepen, Kevin Lee, and Pedro Trindade. (2017). Learning Human Activities for Assisted Living Robotics. In Proceedings of 10th Conference on Pervasive Technology Related to Assistive Environments (PETRA'17), June 21-23, 2017. DOI: 10.1145/3056540.3076197.

Contents

Nomenclature

Acronyms

AAL Ambient Assisted Living

ADL Activities of Daily Living

ANN Artificial Neural Network

ASSL Adaptive Segmentation and Sequence Learning

AT Assistive Technology

CI Computational Intelligence

CNN Convolutional Neural Network

DL Deep Learning

DNN Deep Neural Network

FL Fuzzy Logic

GMM Gaussian Mixture Models

GMR Gaussian Mixture Regression

HAL Human Activity Learning

HAR Human Activity Recognition

HMM Hidden Markov Model

KNN K - Nearest Neighbour

LOOCV Leave-One-Out Cross Validation

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MASE Mean Absolute Scaled Error

MCAR Missing Completely At Random

ML Machine Learning

RBM Restricted Boltzmann Machine

RF Random Forest

RGB-D Red, Green and Blue - Depth

RMSE Root Mean Square Error

RNN Recurrent Neural Network

SVM Support Vector Machines

TL Transfer Learning

DoF Degree of Freedom

MA Moving Average

SMA Simple Moving Average

Greek Symbols

α Moving average period

θ Joint angle

μ Gaussian membership function of fuzzy partition

Roman Symbols

D	refers to a domain

Other Symbols

E_l	Motion Energy
Q	Activity Segments
s_o	RGB-D sensor origin coordinates
a	an observed human activity
F	feature space
f	feature vector
J	A sample, frame, action, pose or observation within an activity
T	a task within a domain
x	is an observed instance of data input
Y	is the set of activity labels
y	is a corresponding class label for input x

Subscripts

N	Number of frames/ observations of an activity
s	a source space
t	a target space

List of Figures

List of Tables

Chapter 1

Introduction

An individual who learns how to ride a bicycle, when faced with the task of riding a motorcycle is able to learn faster than another person who has no knowledge or experience of riding a bicycle. Correspondingly, think of a person who has never used *chop-sticks* to have a meal and is faced with a task of learning how to use it for the first time. Just by observing a second person who has experience of using *chop-sticks*, the initial person is able to learn and acquire the necessary skills to subsequently use *chop-sticks* to have a meal. Imagine the process involved from the initial stage of *zero* knowledge to the stage of using the *chop-sticks* conveniently or the process of reusing knowledge gained from riding a bicycle to riding a motorcycle. In all these cases, the ability to transfer knowledge through underlying processes involved is crucial to the successful completion of the tasks. Therefore, in the context of assistive robots, when a robot is used for assisting a human, the robot is required to learn tasks from a human. A Transfer Learning (TL) process is necessary to endow the robot with abilities to exploit information generated during the execution of tasks. This thesis investigates solutions for tasks TL and proposes computational frameworks appropriate for an assistive robot to learn tasks from observation of a human. All the work presented in this thesis is in the context of human Activities of Daily Living (ADL).

This chapter presents an introduction to this thesis. In Section 1.1, the background and motivation for the research conducted is presented. Section 1.2 moves on to discuss the overview of the research and describes the schematic of the work proposed. The research questions identified are outlined in Section 1.3.

Section 1.4 outlines the aim and objectives of this research and the major contributions are highlighted in Section 1.5. The structure of the thesis with a summary of the contents of each chapter is given in Section 1.6.

1.1 Background and Motivation

Understanding the process of learning in humans has been an area of interest for decades. This has attracted interest from different areas of study which use different approaches such as; Computational Intelligence (CI), biology, psychology, amongst many other approaches. One key aspect of the learning process that has been challenging to researchers in the artificial intelligence community is designing systems which leverage knowledge gained from solving a task into improved performance in solving similar or dissimilar problems. This is where the concept of TL focuses on. The importance of TL cannot be over emphasised; time spent learning new tasks is reduced, more situations can be handled effectively and the information required of human experts is also reduced.

With an increase in ageing population, performing ADLs by the ageing population becomes challenging and this increases the cost of having to support with caregivers and other measures. According to a survey conducted by the Department of Health and Human Services [122], 423 respondents reported on the use of assistive technology to provide care for their ageing relatives. Figure 1.1 shows the statistics of the respondents who used assistive technology in different ADLs. This indicated a greater percentage of respondents incorporated a form of assistive technology as care support for their ageing relatives. Furthermore, these statistics are forecast to rise in the coming years. This demand for assistive technology motivates researches related to Ambient Assisted Living (AAL) to develop solutions to promote quality of life and independent living. One such solution is the use of assistive robots to support elderly people while carrying out ADLs. These robots are trained to perform ADL. However, there are constraints that exist in performing pre-programmed functions and the robots are not capable of utilising learned experiences in solving new unseen problems. In the case of learning human activities,

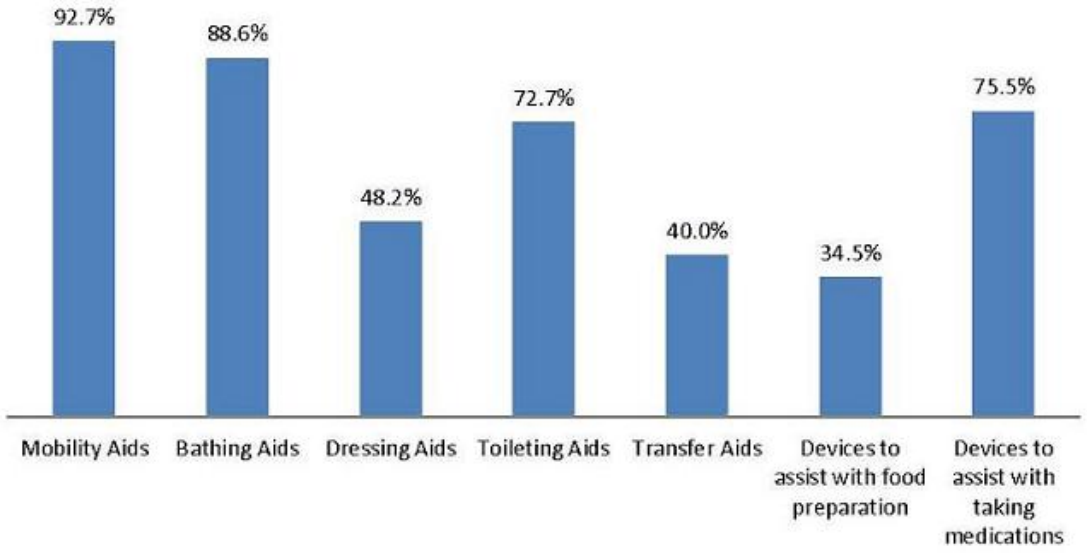


Figure 1.1: Assistive devices used in ADL by 423 respondents according to the Department of Health and Human Services [122].

classifying them correctly from some given set of data is key to understanding how these activities are learnt and how one experience relates to another. The use of Machine Learning (ML) approaches to tackle these constraints is limited due to the characteristics of the training and test data having to come from the same feature space and data distribution. This limits its ability in situations where there are differences in data distribution between the training and test data, which can result in the predictive learner being degraded [106]. Obtaining training data to match the feature space and predicted data distribution of test data is often times expensive and difficult [130]. This prompts for creating high-performance learners for target task(s) from related source task(s) i.e. assistive robots that are able to autonomously learn skill options for target task from related prior knowledge.

The initial process of learning a task(s) is required prior to the transfer of the acquired knowledge to the target. Different approaches have been used to address learning ADL: the proposed approaches include programming by demonstration [87], an approach in which the robot imitates a task demonstrated either by a human operator observed with a motion capture system or by manually moving the robot itself. Statistical approaches like Hidden Markov Models (HMM) ,

Gaussian Mixture Models (GMM) and Gaussian Mixture Regression (GMR) [18] are used to recognise and reproduce different thought tasks. Dynamic Movement Primitives (DMP) is another approach used [109]. The authors in [109] proposed a Simultaneous On-line Discovery and Improvement of Robotic Skills (SODIRS) algorithm that is able to autonomously learn skill options for task variations.

The CI techniques have been applied to TL, amongst which deep learning approach has been widely researched [80]. This is one of the most popular CI techniques that have been significantly applied to the domain of TL [70, 79, 117, 118]. However, it requires a large amount of training data and it also works as a *black box* (learning only relationship between input and output without providing knowledge of the relationship which is key in making decisions) due to its computational framework.

Fuzzy TL (FTL) on the other hand has recently gained interest in the research community and different researchers have used it in various applications. Authors in [149, 150] proposed methods for FTL by incorporating GMM for active learning while trying to address the problem of domain adaptation occurring across heterogeneous spaces. In [104, 105], the author's proposed a framework of FTL to function as a model for prediction in intelligent environments. These works demonstrate the advantages of incorporating Fuzzy Logic (FL) framework over other CI techniques. The FL framework is found to reduce computational complexities, addresses uncertainties associated with data and is easily adapted. Therefore, considering the diverse nature of human activities and how an activity executed by one person can differ in process from another person executing the same activity. It is expedient to take into consideration the associated imprecisions and uncertainties. Thus the FL framework poses to be a tool which will greatly improve on the constraints associated with computational complexities and also providing a generalisable platform for TL across different activities.

From recent surveys [80, 130], the key challenge in TL has been defining the evaluation metrics related to *what to transfer*, *how to transfer* and *when to transfer*. This is mainly because there are various possible measurement options and/or algorithms. The algorithms used so far focus on three main steps namely; First, given a target task, select an appropriate source task or sets of

tasks from which to transfer. Second, learn the relationship between the target task and source task(s). Third, transfer knowledge effectively from source task(s) to target task. The work in [126] focus on learning inter-task relations which are modelled using a three-way Restricted Boltzmann Machine (RBM). This model captures the similarity between samples from source task and target task. The method, however, is computationally complex since it requires large amount of training data and it also does not capture the uncertainties associated in the task constraint. In [82], a TL technique is employed to speed up learning robot models using Local Procrustes Analysis but this method requires correspondence between data sets to be provided and requires large amount of training data.

This research leverages the benefits of TL to promote human-assistive robot transfer of ADL knowledge to aide the development of better solutions for assistive technology. It is with much expectation that this will reduce the learning curve associated with equipping assistive robots with the knowledge required in executing tasks.

1.2 Overview of the Research

The prime motivation of the research presented in this thesis can be summarised by a simple example as demonstrated in Figure 1.2 when an assistive robot observes and learns a task from a human.

Assistive robots deployed in living environments for applications such as elderly care and support for independent living should learn tasks by observing human carers performing routine duties. To achieve this goal, the assistive robots must be equipped with abilities to learn activities. This requires extracting descriptive information of the activities and classify them while they are performed by a human.

Learning human activities by an assistive robot can be classified under two methods [130]: 1) *Independent Learning* which is concerned with learning an activity from scratch and 2) learning by making use of transferred knowledge and/or information which is referred to as *Transfer Learning*. Independent learning is a method whereby an assistive robot learns to perform an activity

Figure 1.2: An illustration transfer learning of a human activity from a human to an assistive robot learning.

independently without any prior knowledge of the activity. For example, an assistive robot learning an activity as illustrated in Figure 1.2 without prior information of how a person would perform the activity - that is, the person performing the task would not be present. This requires more time in learning and more cost incurred which are limitations of the method. On the other hand, TL methodology allows information acquired from prior experience to assist in learning an activity [80].

In the context of this research, an assistive robot should be capable of learning to perform an activity from knowledge acquired as it observes a person perform similar activity. This enables faster learning of activities and allows collaboration and adaptation of robots within living environments. Regardless of the method applied to learning an activity, the availability of descriptive information affects the understanding of an activity. Variations in information and understanding about an activity performed by a person and a robot performing similar activity can be defined as contained within a *knowledge gap* and TL helps to bridge this gap.

Human activities are diverse in nature with imprecision, vagueness, ambiguity and uncertainty in information about the way activities are performed. Thus, variabilities are encountered when an assistive robot tries to

learn activities. This can affect the correct classification of human activities which is relevant in improving the amount of knowledge that can be used by a robot in learning. To capture imprecisions and uncertainties, fuzzy logic has proven to be a suitable method which allows incorporation of imprecisions and uncertainty expressiveness within information [80, 105] and thus can be applied to classify human activities. Combining this method with TL would improve assistive robots learning human activities by observing while activities are performed. Other learning techniques applied to learning or classifying human activities are limited in their ability to handle vagueness, imprecision and uncertainties in activities when considering acquiring knowledge that can be transferred across different learners.

Prior to assistive robots performing a human activity through TL, the information extracted by such robots is a vital component of the system. Observing activities as they are performed through the use of visual or non-visual sensors makes it a lot easier to obtain information of human activities in an environment [38, 112, 113]. It would be extremely hard to understand and interpret activities using a normal visual sensor such as RGB cameras which provide 2D visual data [48]. These sensors provide limited information for an activity performed in a real world environment. However, recent development in RGB-Depth (RGB-D) sensors show that they are better devices for observing human activities [48]. These sensors provide a means of better observing the world to detect human poses used to build activity recognition systems [38, 113]. They also provide a platform for exploiting depth maps, body shape and skeleton joint detection of humans in 3D space which are used in developing sophisticated recognition algorithms.

This research proposes a novel framework for TL in Human Activity Recognition (HAR) through the use of depth information from an RGB-D sensor. This is with the motivation of incorporating the framework in an assistive robot. The schematic representation of the proposed framework is given in Figure 1.3. The framework comprises five major steps: 1) observing human activity with an RGB-D sensor, 2) extract sequence of joint motions, 3) learning activities by recognition, 4) activity representation and 5) the transfer of modelled activity to an assistive robot.

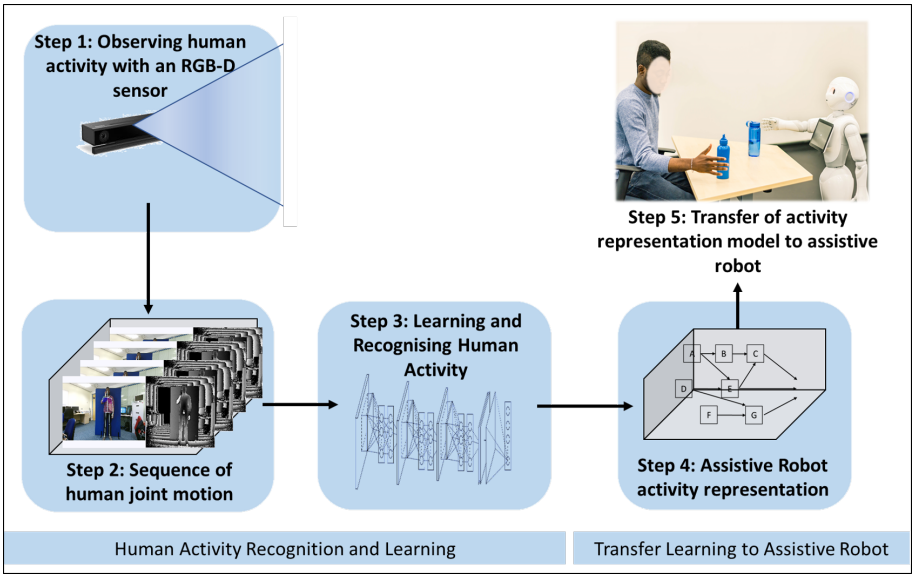


Figure 1.3: Schematic representation of the proposed transfer learning framework for human activities.

The first and second steps have to do with how human activity information is obtained. This entails using the RGB-D sensor to extract point cloud information of different joints of a human body. This gives information of the position of joints throughout an activity. The third step involves activity recognition and learning. This step is key since in TL an initial knowledge is required for transfer to be achieved in a target. Therefore, the proposed framework is capable of recognising activities and learns to predict the constituent tasks. Furthermore, the fourth and fifth steps are concerned with activity representation and knowledge transfer to a robot feature space.

1.3 Research Questions

Following the research overview, the main questions identified as the basis for this thesis are as follows:

- How to learn human activities using a mode of information which is computationally efficient? Most existing methods for recognising human activities using visual information usually rely on the combination of

multiple information modality (i.e. RGB, depth, infrared, etc.) to achieve impressive performances. This often leads to increased computational resources.

- Can activity sequences be modelled from unlabelled data? The differing nature of human activities create challenges when trying to define the true sequence of occurrence of constituent actions. Where unlabelled sequences exist, the challenge is even greater, and thus a reliable method to obtain true sequences required for the transfer knowledge base is needed.

- How can transferred human activity be adapted in a target domain? The bottleneck many TL methods encounter is the adaptation of transferred knowledge in the target domain such that it does not have a negative effect on the primary goal of performance improvement.

To address these questions, the following section outlines the aim and objectives of this research.

1.4 Research Aim and Objectives

The aim of this research is to investigate TL of human activities in an AAL environment. This involves the combination of three concepts; Transfer Learning, Fuzzy Logic and Human Activity Recognition to address the problem of learning human activities and transferring the knowledge acquired to be used in performing activities that have little or no direct contextual knowledge. A suitable method / algorithm for learning of tasks that have no prior direct contextual knowledge will be developed, modelled through the data collected from visual observations of humans executing tasks and a physical robot platform. This creates a refined understanding of human ability to retain and use previously acquired knowledge to solve tasks with no direct prior knowledge. To achieve this aim, the following research objectives have been identified:

1. Investigate existing learning methods and models of human activities and propose a model for recognising and learning activities from visual sensor information.

2. Propose a technique for modelling human activity sequences for better understanding of constituents of human activities required in knowledge transfer.

3. Investigate the current TL approaches employed in skill transfer across different but contextually related activities.

4. Incorporate a rule-based approach with the proposed TL model to capture uncertainties which are evident in performing tasks and also to simplify the TL process. This will reduce the complexities associated with most commonly used CI approaches which rely on large amount of numerical data.

5. Propose a model for transfer of learned human activities to an assistive agent which can be incorporated across different platforms.

6. Implement the improved learning model using an assistive robot simulator environment.

1.5 Major Contributions of the Thesis

The major contributions of the work presented in this thesis are summarised as follows:

- An extensive literature review of the state-of-the-art on TL which encompasses algorithms proposed and validated results from experiments. This also features its applications in human activities, specifically, in human-robot interactions.

- A philosophical investigation and discussion into TL and its applications in ambient assisted living applications.

- A novel framework for human activity learning using an ensemble method from a combination of handcrafted and statistical features.

- A novel adaptive sequence learning methodology for human activities from 3D skeleton joint coordinates information.

- A methodology for predictive modelling of human actions from limited datasets.
- A novel proposal for TL of human actions to assistive robots through heterogeneous feature space learning.
- Application of a TL framework on human activity datasets to achieve learning of ADLs in assistive robots.

The outlined contributions of the thesis are addressed in different chapters of this thesis. A summary of these chapters is presented in the following section.

1.6 Thesis Outline

This thesis consists of seven chapters. Figure 1.4 shows the structure of the thesis with an indication of how the chapters are linked. This gives readers an overview of the organisation of the thesis and a direction on how the chapters are grouped. The summary of contents of this thesis are presented as follows:

Chapter 2 presents a comprehensive literature review of TL, its definition and applications in pervasive computing. This covers approaches employed in TL as related to human activities. The chapter also discusses literature in HAR in the context of assisted living, comprising information obtained from visual information. Details including, approaches, sensor information, preprocessing of information, activity segmentation, feature computation and classification from related literature are reviewed. Specifically, the technical and practical applications of HAR in assisted living environments incorporating assistive agents such as robots are discussed.

Chapter 3 presents a description of TL and how it is applied across domains or across tasks. The chapter discusses TL challenges related to *what to transfer*, *how to transfer* and *when to transfer*, and the limitations in realisation of this concept in day to day applications. To address the challenges, this chapter presents an overview of the concept of TL and how it can be applied in human-robot interaction for assistive robots requiring to learn human tasks in

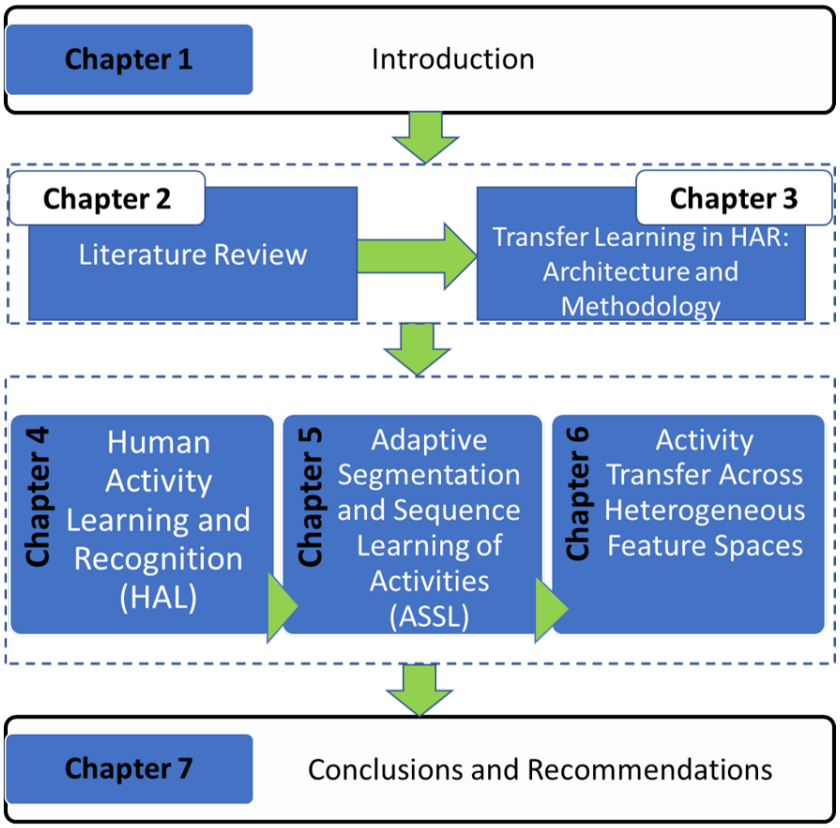


Figure 1.4: Thesis structure showing the organisation of the chapters and their respective dependencies.

AAL environments. The methodology proposed for TL in this thesis is also introduced in this chapter.

Chapter 4 presents a novel Human Activity Learning (HAL) system proposed for recognition of activities that can be incorporated in an assistive robotics as the initial stage in the process of TL. An RGB-D sensor is used to acquire information of human activities and a set of statistical, spatial and temporal features for encoding key aspects of human activities are extracted from the acquired information of human activities. The features are then fed as input to a classifier for the learning and recognition of activities. The experimental results show the overall performance achieved by the proposed system is comparable to the state-of-the-art and has the potential to benefit applications in assistive robots for reducing the time spent in learning activities.

Chapter 5 presents a novel Adaptive Segmentation and Sequence Learning method for the prediction of activities. Following from the recognition of activities, to understand the composition of actions in an activity, it is important to understand the actions that constitute an activity. This is key in predicting future actions for robots learning an activity from observed movements. This chapter aims at segmenting unlabelled observations of recognised human activities and sequence learning of obtained segments to provide assistive robots with intelligence for solving human activities. Results of the process is evaluated experimentally on human activity dataset and compared with existing models for sequence learning model based on probabilistic inferences and regression.

Chapter 6 is directed towards addressing the challenge associated with differing feature spaces when considering TL from human domain onto an assistive robot domain. The chapter presents a novel method of effective TL across heterogeneous feature spaces for the purpose of TL for an assistive robot. A fuzzy latent space exploration is used to obtain mappings of feature spaces. Then, representations of both feature spaces are obtained by applying Labanotation for describing body joints movement. Afterwards, the knowledge transfer is established. This approach is used in simplifying the learning of primitive actions from predicted sequences of activities for assistive robots seeking to execute human actions.

Chapter 7 presents a summary of the findings of this research. The major findings obtained in this thesis are discussed with reflection to the research questions identified in Chapter 1. Following the summary of the findings, the chapter also presents recommendations for applications of the work in this thesis and possible areas of future work.

Chapter 2

Literature Review

2.1 Introduction

Transfer learning (TL) and Human Activity Recognition (HAR) are two broad areas widely studied in Computational Intelligence (CI) applications with so much effort put into developing more suited solutions to advance current performance of existing systems. In this regard, many works have been published in these areas. Therefore, it is important to review the current state-of-the-art related to both areas to justify the intent of the work in this thesis. This chapter is focused on the review of literature related to the work presented in this thesis.

This chapter is structured as follows: Section 2.2 gives an overview of HAR based on RGB-D information of human activity as applied in this work. The features extracted and CI methods so far applied in HAR are discussed. To give a general understanding of TL as related to HAR, Section 2.3 reviews the current research on TL using CI techniques including its applications in human activities. In Section 2.4, assistive technologies related to human activities specifically in AAL environments are discussed. This reviews different technological solutions used, with a primary focus on assistive robots. Section 2.5 follows from the review of previous research to identify the research gaps and highlights how this work differs from previous research works. Section 2.6 summarises the chapter.

2.2 Human Activity Recognition (HAR) with RGB-D Sensors

Learning and classification of human activities using some CI techniques is often referred to as HAR [55, 58]. Over the last few decades, the study of HAR has been carried out to detect, recognise and/or classify activities of humans. The advantages of HAR has seen many applications in several domains such as security, health care, manufacturing, gaming, amongst many others. Owing to this, several approaches have been investigated. An integral component of HAR is how information of activities are obtained or observed. Based on the published literature, HAR approaches are divided in two main categories: visual sensor based and non-visual sensor based HAR. Observing activities through the use of visual [38, 48, 112, 113] or non-visual sensors [19] makes it a lot easier to obtain information of human activities in an environment. Non-visual sensor based approaches utilise information such as environmental conditions - like temperature, motion detection or ambient light, location and information from wearable devices. A comprehensive review of HAR using non-visual sensors can be found in [71] and more recently in [128]. Although these information have some advantages, they are sometimes invasive and burdensome. On the other hand, HAR using visual sensory information mainly rely on the interpretation of images to predict activities [46, 48, 88].

One of the main objectives of HAR is to extract descriptive information (i.e. features) from human activities to be able to distinctly characterise and classify one activity from another. Visual sensor-based approaches are mainly based on 2D or 3D information obtained from the sensor devices. However, it would be extremely difficult to understand and interpret activities using regular visual sensors such as RGB cameras which provide 2D visual information [48]. These sensors provide limited information for an activity performed in a real world environment. Recently, most researches in HAR based on visual sensors have employed RGB-D sensors which prove to be better devices for observing human activities [38, 48, 83, 89]. These RGB-D sensors provide a means of better observing the world to detect human pose used to build HAR systems [113]. They provide a platform for exploiting depth maps, body shape and detecting

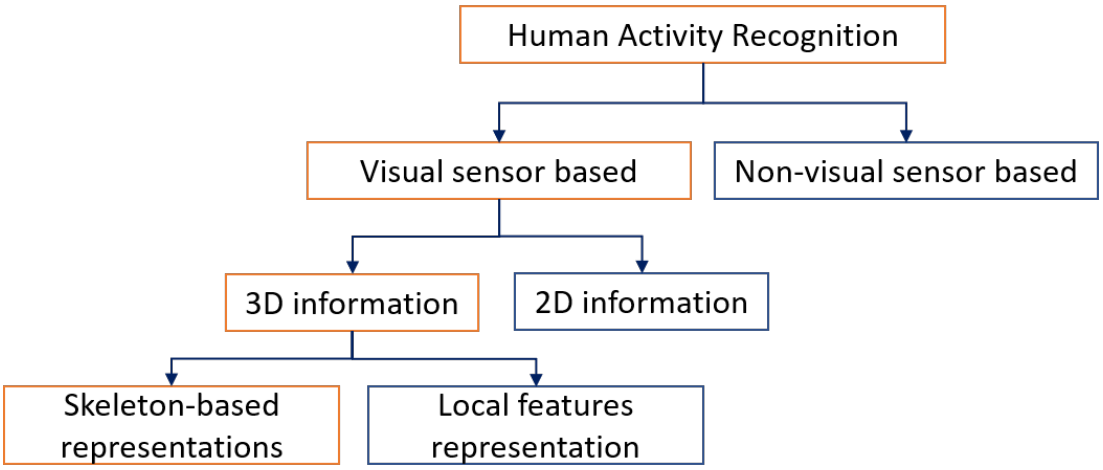


Figure 2.1: Classification of approaches widely used in human activity recognition based on the source of information.

skeletal joints of humans in 3D space which are used in developing sophisticated recognition algorithms. Furthermore, among the many approaches to human representation based on 3D information [1, 10, 16, 48], the majority of the existing methods can be generally grouped into local feature-based representation [140] and skeleton-based representations [49, 111, 113]. Figure 2.1 summarises the categorisation of HAR based on the grouping of activity information employed. Representations based on local features identify relevant points in space-time dimensions, interpret patches at the points as features and encode them into representations which can locate notable regions. However, local feature-based representation methods do not take into consideration the spatial relationships between features. As a result, they are unable to represent multiple humans in the same scene. The local features-based methods can also be computationally expensive due to the complexity involved in the extraction process.

However, skeleton-based representations have shown promising performance in real-world applications including gaming and assisted living [48]. These methods consider the spatial relationships between features which enable the modelling of human joints relationship for encoding the whole body structure. Also, skeleton-based representations are robust to variations in illumination, scale, view and

motion speed. Due to these advantages, such representations are used in real-time applications and many researchers [1, 38] have introduced techniques to facilitate different applications.

2.2.1 Background and Challenges of Vision-Based HAR

Over the past decades, research on HAR has seen much improvement with technological advances in the field leading to the availability of low cost, small and low power consumption sensors. Sensory devices used to obtain human activity information have become less intrusive as they are able to be incorporated in an AAL environment without being noticed. The sensor networks are not left out of the advancements as well. Wireless technologies [127] used in sensor networks have enabled unobtrusive recognition of activities with information accessible from any location. The benefits of these advancements cannot be over-emphasised; remote monitoring, individual profiling, intrusion detection, abnormality detection and so much more.

In the field of computer vision, HAR with vision-based methods is one of the most studied areas. The goal is usually to automatically detect and analyse human activities from a sequence of images captured using camera sensors or other vision sensing modalities. These activities take on different forms which range from elementary actions to complex activities depending on the environment. Aggarwal and Xia [1] categorised such activities into four groups: atomic actions, activities containing sequences of distinct actions, activities including person-object and person-person interactions, and lastly, group activities. The most difficult of all the categories mentioned is group activities. Research in this area has encountered several limitations which could be as a result of the difficulty in collecting the data required or the limitation of existing vision-based sensors.

Here, the challenges of vision-based HAR systems are discussed. From the review of past researches on vision-based HAR, four main challenges are identified. First, the low-level challenges encountered from occlusions, shadows, varying illuminations and cluttered backgrounds [21, 88]. This type of challenges are encountered in most cases when using visual sensors. They create

difficulties in motion segmentation which alter the form in which actions are observed. Zhou and Zhang [143] proposed a technique used in filtering background clutter, occlusions and unstable camera motions for recognising human activities. The technique used a combination of multiple-instance formulation and Markov model in a framework to select elementary actions for encoding movements of local parts. This technique allowed for long-range temporal information of actions in video sequences to be encoded. Chen et al. [22] also attempted to address the challenge of identifying human actions using Conditional Random Fields (CRFs) to differentiate between unknown movements and intentional actions which may occur in a scene through the ordering of video regions and identifying the actors for actions. Also, 3D sensor information [1] has been introduced as a solution to mitigate the low-level difficulties due to their ability to provide structure information from a scene.

The second challenge has to do with changes in view of an activity [1, 10, 11, 133]. Information of the same human action can generate different representations depending on the perspective such information is obtained. This poses a challenge when using stand-alone cameras in acquiring activity information. To tackle this challenge with a single camera is an extremely challenging task. However, solutions proposed to address this challenge have adopted multiple synchronised cameras. Although, implementing such cameras in applications can be a daunting task. One of such solutions is the introduction of 3D Motion Capture systems (MoCap) [1] which have enabled recognition algorithms to alleviate this challenge. The use of depth information from such MoCap systems to obtain skeletal joint information of a human can be used in constructing view-invariant information for algorithms used in HAR [56].

The third challenge identified with vision-based HAR is scale variance [1, 133] which occurs when a subject or different subjects appear to be different sizes when viewed from differing distances to the camera. A solution to this when using 2D information is by extracting features at multiple scales. Also, using 3D information solves this challenge since the depth information of a subject is easily known and can be adjusted through the activity sequence.

Finally, there is the challenge of inter-class similarity and intra-class variability of actions [97]. This occurs as a result of the uncertainties in the way actions

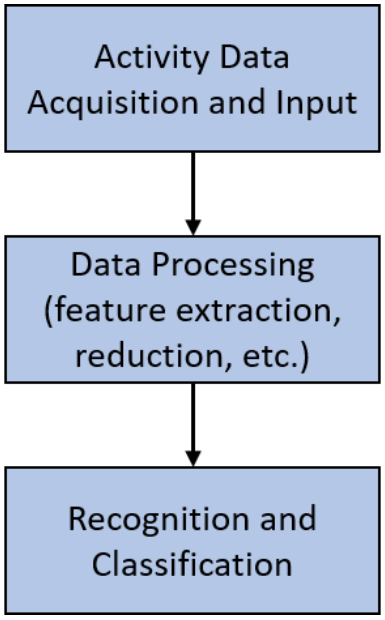


Figure 2.2: Typical steps involved in Human Activity Recognition.

are performed by humans. A single action can be carried out by individuals in different directions with varying characteristics of body movements and similarly, two actions may only be differentiated by subtle spatio-temporal information [1]. This poses a challenge for real-world applications of vision-based HAR and to date, it remains a difficult problem for recognition algorithms using the different modalities of visual data.

To achieve recognition of human activities, three main steps are involved. Figure 2.2 identifies these steps which correspond to data input, processing and classification. Data input step is the acquisition of human activity data with the means of a sensory device and the data is then processed, which entails stages of feature extraction, feature reduction, standardisation, etc. The processing step prepares the data for fitting in the model which will be used in identifying activities. The following sections discuss different methods proposed by researchers that have been applied in the HAR steps presented in Figure 2.2.

In Table 2.1, a general taxonomy of vision-based HAR based on the categorisation in Figure 2.1 is provided. The main features of both the 2D and 3D vision-based approaches are highlighted, example of sensors used, main advantages and disadvantages are given.

Table 2.1: Taxonomy of vision-based HAR based on the grouping of information used.

Grouping	Summary	Benefits	Short-comings	Example sensors
2D information	Infer human activities from 2D points extracted from images.	Processing does not require as much computational resources as 3D information	- Information obtained is limited. - Not robust to variations in scale of subjects.	RGB cameras. E.g. Webcams.
3D information	Identifies human activities from point clouds of changes in human movement.	- Overcomes the scale variance problem. - Provides more information of human activities. - Are robust to view changes in activities.	- Usually require more computational resources. - MoCap systems require installation of multiple sensors.	Motion Capture systems, RGB-D sensors. E.g. Microsoft Kinect [86].

2.2.2 Data Collection of Human Activities in 3D Skeletal Data Space

Data obtained from RGB-D sensors gives information relevant for a robot to understand an activity. By exploring human pose detection using RGB-D sensors, activity recognition has advanced recently [38, 112]. Using RGB-D sensors extracts 3D skeleton data from depth images and body silhouette for feature generation. In [38], the RGB-D sensor is used to generate a human 3D skeleton model with matching of body parts linked by its joints. They extract positions of individual joints from the skeleton in a 3D form x, y, z. Jalal and Kamal [58] use similar RGB-D sensor to obtain depth silhouettes of human activities from which body points information are extracted for the activity recognition system. Zhou et al. [142] also used an RGB-D sensor to capture human skeleton information as part of a system for controlling a mobile robot

using human gestures which is also a similar application proposed by [20]. Another approach is shown in the work in [42] where the RGB-D sensor is used to obtain orientation-based human representation of each joint to the human centroid in 3D space. These researchers [20, 42, 142] use different devices for the acquisition of data. In the following section, methods of acquisition of 3D human skeletal data are discussed.

2.2.2.1 3D Human Skeletal Data Direct Acquisition from Sensors

Direct methods of acquisition of 3D skeletal data of human activities is carried out using different devices commercially available which include, MoCap systems [1], structured-light cameras and time-of-flight sensors. These devices detect the kinematics of human body models in order to identify the relevant joints in the body. Figure 2.3 shows an example representation of tracked skeletal joints obtained from a Microsoft Kinect v2 RGB-D sensor [86].

MoCap systems obtain 3D skeletal information by tracking markers placed on a human in its scene [48]. These systems are based on either visual cameras which utilise multiple cameras at different positions around a subject to track reflective markers that are attached to a subject's body or 3-axis inertial sensors that estimates body part rotations with reference to a fixed point. It should be noted that the inertial sensor-based MoCap systems can obtain the skeletal data without any visual cameras involved. The existing MoCap systems have the software to enable collection of the 3D skeletal data with a high degree of accuracy. However, most of the systems can only be used in controlled environments and are typically expensive.

Structured-light cameras which are types of camera devices that utilise infrared light to capture depth information is also used in the direct acquisition of 3D skeleton data [86]. Light is projected through the infrared sensor in a known pattern and the distortion observed in the pattern when it meets a subject allows the device to decide the depth. The RGB image of the scene observed can also be acquired. Most of the RGB-D sensors are inexpensive which makes them available for use in most applications. This source has been popularly used in research for HAR [38, 41, 89, 94].

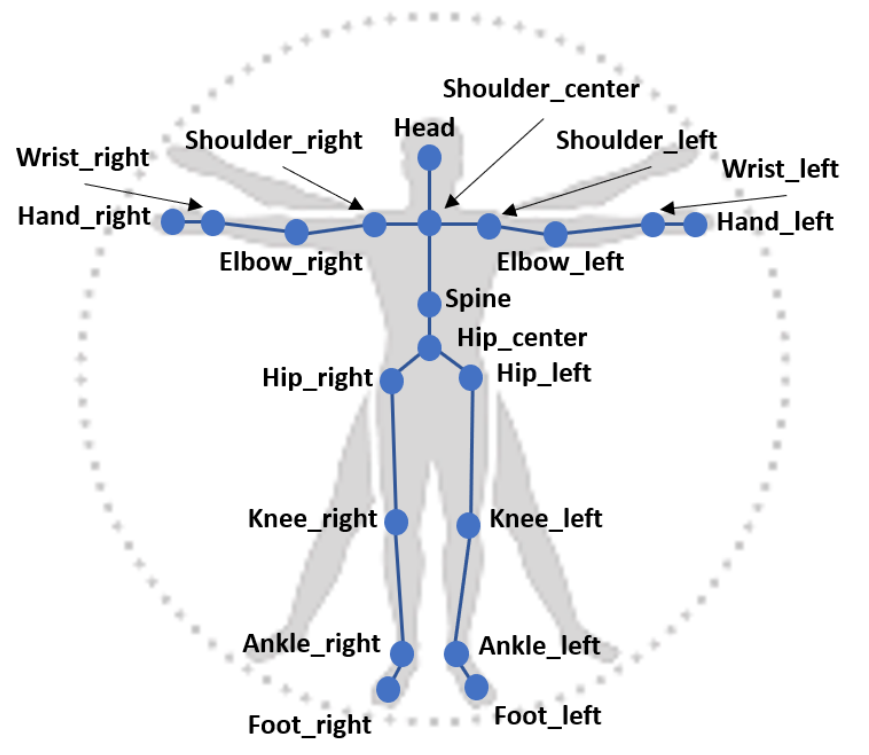


Figure 2.3: Example of skeletal body model obtained from a Microsoft Kinect device. This shows the 20 tracked joints.

Time-of-flight sensors [48] acquire 3D information by emitting light and measuring the time it takes for the light to be returned. Some examples of such sensing technologies are radar and Light Detection and Ranging (LiDAR) [13]. These sensors acquire very accurate 3D information at high frame rates. Comparing all three methods of direct acquisition of 3D skeletal information, the RGB-D sensors are the most affordable and can be installed in an environment. Also, they provide additional RGB data which can be accessed and processed with the depth information.

2.2.2.2 3D Skeleton Construction from Pose Estimation

3D skeletal information can also be acquired through human pose estimation and construction of skeleton [58, 88, 113, 146]. A number of approaches have been proposed to estimate human joints and pose recognition from the knowledge

of available data. Such approaches take advantage of depth images or extra information accessible from the visual sensing device. A majority of the methods are based on the identification of body parts which are fitted to models which extract specific locations of the identified parts. This section provides a review of such methods of human skeleton construction based on visual data.

The first approach considered is the construction of 3D human activity information from depth images. Human skeleton can be constructed from a single observed depth image or from acquired sequences of depth images. This approach is widely used in acquisition of activity information due to the additional geometric information depth images provide. Jalal and Kamal [58] introduced a vision-based life logging system using depth images to track human body points and location. Their work identifies 15 joints from a depth silhouette and an additional 8 centre points of limbs joints are constructed using Gaussian contours mechanism. The work was further extended in [59] using temporal depth motion identification to obtain depth human silhouettes from other objects within the scene. Recently, in [60] another model for human body parts estimation and detection is proposed using depth imagery. A colour space transformation based on heuristic thresholding segmentation technique [5] is used to obtain salient regions and then skin tone detection through foreground segmentation of silhouettes. Afterwards, the body parts are estimated using a proposed body parts model through pixel-wise searching and computation of the distance from the top to the bottom of the silhouette. A novel approach for pose estimation from a single depth image called Model-based Recursive Matching (MRM) was introduced in [132]. This approach combined a depth image and 3D point cloud of the input to create a human skeleton model with customised parameters based on T-pose to fit different body types. The results reported from the work in [132] show the proposed method is able to give accurate estimations in cases where there are occlusions in human pose. The method used a MoCap for depth image acquisition which is able to handle occlusions better than a single RGB-D sensor device. The downside to the use of depth images for pose estimation is that most of the systems are computationally complex to setup.

Another approach widely employed in human skeleton construction is from

traditional RGB images. Typically, most of the methods using RGB images extract visual features using Deep Learning (DL) architectures and other methods to match poses of segmented silhouettes for identifying body parts. Deep Neural Networks (DNN) have demonstrated their ability in construction of human skeleton from RGB images [37, 56, 121]. Toshev and Szegedy [121] applied DNNs in an approach to estimate human poses called 'DeepPose'. They formulate the pose estimation as a regression problem by proposing a cascade of DNN regressors for high precision estimates. Fan et al. [37] adopt a Dual-Source Deep Convolutional Neural Networks (DS-CNN) approach for both joint detection and localisation from a single RGB image. The approach takes image patches as input and learns the appearance of each body part by considering the integrated views in the full body.

Apart from DNNs, other methods have also been used for human body parts estimation from RGB images. For example, Li et al. [74] in a recent work proposed an algorithm for estimating sequences of upper-body parts in unconstrained videos. They use a two-step approach in which a spatial model is constructed to capture relationships between adjacent parts and then a method to select the best out of different pose configurations. Also, a general parametrisation of body pose method to estimate 3D human poses from 2D joint locations is seen in [2]. The method uses priors that are learned from joint limits in poses. The use of multiple images acquired using multiple cameras in different views can be used in observing human and then image processing techniques employed in estimating human depth maps from the combined images. After obtaining depth maps human skeleton models can be composed using some of the methods already described. Although, there are solutions using the construction of depth maps from multiple images to construct human skeletons, such solutions are usually slow and encounter problems relating to noisy depth data and correspondence search failures.

2.2.3 Feature Extraction in HAR from 3D Skeletal Human Activities Data

Feature extraction is a vital component of any HAR system. The goal of feature extraction is to find recognisable characteristics of human activity data that can be used in accurately differentiating between activities, one from another. Due to the importance in the process of feature extraction and the role features play in a HAR system, the performance of any HAR system is largely attributed to the quality of features obtained from the available data.

Following the acquisition of human activity data using methods as reviewed in Sections 2.2.2.1 and 2.2.2.2, the raw data obtained from these sensors have to be preprocessed prior to feature extraction. This process is carried out to reduce redundancy in data for better representation of features of an activity. Most of the works [89, 102] employing 3D joint coordinates data of skeleton use a preprocessing step to offset the data centroids (usually obtained with reference to the sensor origin) to the human centroid as the origin. This makes the data scale-invariant and easier for recognition algorithms to attain improved performances.

According to Subetha and Chitrakala [110], approaches to HAR using RGB-D information fall into two categories: *feature-based* and *model-based*. Feature-based techniques such as Histogram of Oriented Gradients (HOG) and subspace clustering based approach (SCAR) are used to extract features for recognising human activity from data acquired using sensors. Hussein et al. [54] applied statistical covariance of 3D joints (Cov3DJ) as features to encode the skeleton data of joint positions which are then used as input to an SVM model for activity recognition. Another approach applied by [129] used a sequence of joint trajectories and applied wavelets to encode each temporal sequence of joints into features used in activity classification. Model-based techniques have to do with the construction of a human model for recognition either as a 2D, 3D or skeletal model. Vemulapalli et al. [124] construct models using kinematic approach that extract features from frame sequences for human structure representations. Du et al. [32] used a neural network technique to propose an end-to-end hierarchical Recurrent Neural Network (RNN) for representing skeleton based construction. They make use of the raw positions of human

joints as input to the RNN. A combination of both feature-based and model-based approaches for classification of activities is seen in [113]. The authors used a Maximum Entropy Markov Model (MEMM) for classification of activities using features from skeleton tracking combined HOG.

2.2.4 Recognition and Classification of 3D Skeletal Human Activity

Following the extraction of features from 3D skeletal human activity data, the processed features are used in a classification step for learning/recognition of human activities. A number of approaches have applied different techniques which range from statistical to CI methods in the recognition process of vision-based human activities. The classification process involves grouping activities from observed sequences based on the similarities identified from features.

2.2.4.1 Classification with Statistical and Machine Learning Algorithms

Statistical and ML techniques such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Naive Bayesian, or Latent Dirichlet Allocation (LDA) are some of the commonest methods applied in HAR from 3D human skeleton data [120]. Classification of human activities is carried out by extracting relevant features from data obtained using RGB-D sensors. The work in [24] proposed a method for activity recognition using RGB-D data. The 3D joint position information extracted from the sensor are transformed into feature vectors by applying selected soft computing techniques to group key postures of an activity. The posture features are used as input to a learning algorithm for classification of human activities. SVM and KNN algorithms were used separately in classifying activities and the results compared. The SVM algorithm used in classifying 3D human activity skeleton data [24, 41, 88] works by finding the optimal hyperplane which allows separation between distinct classes in an observed feature space. It uses a kernel function ϕ that allows the transformation of activity feature spaces to a higher dimensional space where the data is separable. Nunes et al. [89] applied Random Forest (RF) in a

framework using max-min features from human activity skeleton data. They proposed an extension to the traditional RF which combines a DE meta-heuristic algorithm with RF to optimise recognition performance.

In the work presented in [38], the authors propose using a probabilistic classification in a framework that combines multiple classifiers to form a Dynamic Bayesian Mixture Model (DBMM) for characterising activities from features obtained from distances between different parts of the body. The use of the Bayesian Mixture Model is integrated into a dynamic process that takes into consideration the temporal information of activities. The use of non-parametric approaches which are capable of dealing with large number of classes and the problem of overfitting has been proposed as a solution for HAR from 3D skeleton data. For example, Yang and Tian [135] proposed a Naive Bayes Nearest Neighbour (NBNN) approach to recognise human actions from the accumulated motion energy computed from 3D human skeleton joints. Such methods require no learning process. Other techniques have been applied for sequence-based classification of human activities using 3D skeleton information, among which are Dynamic Time Warping (DTW) and Markov Models [96, 102]. Markov Models like HMMs are very useful in modelling activity sequences and thus they are very resourceful in recognition of activities. By defining the elements of a HMM which are given to be the prior distribution for initial states, the emission matrix and the the transition matrix, a HMM can be used to calculate the probability of an action for a given activity sequence consisting of observed human key poses.

2.2.4.2 Recognition of Human Activities using Computational Intelligence Techniques

Apart from the use of statistical ML techniques for 3D skeleton data HAR, CI methods have also been extensively studied by researchers. CI is a collection of nature-inspired computational models that are used to solve complex real-world problems which traditional statistical or ML techniques might be incompetent due to the reasons of - uncertainties inherent in the problems, such problems might be too complex for mathematical inference or may be stochastic in structure.

Human activity actions when observed through 3D visual information can be complicated with a lot of uncertainties in distinguishing one activity from a set of related activities. Computational methods such as Fuzzy logic [55, 75, 136], neural networks [37, 56, 117] and evolutionary computation [78, 101] are suited for such recognition applications.

Yao et al. [136] have used a fuzzy logic model for human behaviour recognition. Silhouette slices and movement speed from human silhouettes are used as input to the fuzzy system. A fuzzy c-means clustering algorithm is used to learn the fuzzy membership functions and the human behaviour is then identified via selecting the behaviour category with the highest membership degree. Similarly, the work in [75] employed fuzzy logic in proposing a view invariant HAR system using a single camera. They have used a fuzzy qualitative Poisson human model to extract fuzzy qualitative human contour descriptor for human viewpoint analysis. Clustering algorithms are then applied to classify the viewpoints. These methods achieved reasonable performance in HAR. Other variations of fuzzy systems such as evolving fuzzy systems in [55] have also been use. Fuzzy models are good at handling uncertainties in human activity data which makes them a good tool in HAR.

Traditional Artificial Neural Networks (ANNs) have been applied extensively in 3D human skeleton based activity recognition. Parisi et al. [94] employed an ANN model in their work on HAR. They extract pose and motion features from video sequences of activities and apply a clustering technique for grouping actions in prototypical pose-motion trajectories. The classification model consisted of Self-Organising Growing When Required (SOGWR) networks to obtain continuous representations of inputs and determine the latent spatio-temporal dependencies. Other works using neural networks [117, 121] take advantage of its ability to model complex and non-linear relationships which occur in human actions to attain high accuracies. Also, ANNs when compared to other ML techniques do not impose restrictions on input data due to their ability to learn hidden relationships in data, which makes them good in predicting scenarios.

With the recent evolution in technology, DL models [56] have also more recently been applied in activity recognition problems with results showing

robustness of the method in activity recognition. Du et al. [32] proposed an end-to-end hierarchical RNN human skeleton recognition model that models long-term contextual information of temporal activity sequences. DL models are good at automatically learning the features from any dataset and this makes them suitable for large and complex applications. Ijjina and Chalavadi [56] applied extreme learning machines for classification of features obtained using a Convolutional Neural Network (CNN). The method was tested on 5 human activity datasets and achieved high performances. In [145], a sequence-to-sequence model based on DL is used to recognise ADLs taking advantage of activity state representations. Many other applications using DL architectures in HAR can be seen in [37, 117, 121]. However, DL models require large amount of data to achieve for concise predictions of activities and in most cases more resources such as time and reliable processing architectures. Also, using DL limits the flexibility of defining the features to be used in the classification stage. To implement such DL architectures require high processing power with a huge amount of computational resources to train the networks as some architectures take several days or weeks to train.

2.2.5 Discussion

From the review presented, it is evident that HAR is a well-studied area with applications seen in many disciplines, thus the need to further research into solutions to improve current HAR systems. Although there have been many successes recorded in vision-based HAR, the complexities associated with occlusions, varying illuminations, changes in view, scale variance and activity similarity, remain challenging in many applications. These have effects on the computational requirements of many systems. The conclusions from the review on HAR presented are outlined as follows:

- Suitable data for HAR systems must be obtained as this has a defining impact on the system. In addition, the algorithms used for recognition should be investigated and selected based on the performance obtained with the information modality and other relevant factors.

- Most research works focus on activity classification from single-persons, however, action detection and activity pattern discovery require more investigation to provide better understanding of the nature of activities.

2.3 Transfer Learning in Computational Intelligence

TL methods usually employ various computational techniques as training models such as neural networks [79], support vector machines [119], and rule-based models [105, 150]. This section discusses TL methods which apply such CI techniques as solutions to learning problems.

2.3.1 Neural Network Transfer Learning Methods

Neural network architectures have been used in TL applications over the years with results demonstrating superior performance compared to statistical models. However, most applications of neural network in TL apply deep ANN architectures to propose solutions often referred to as *Deep Transfer Learning* (DTL) solutions. In a recent survey by Tan et al. [117], DTL is defined as a case of learning a target task where the objective predictive function, $f(\cdot)$, is a non-linear function that reflects a deep neural network. The effectiveness of deep neural networks in TL is the flexibility of its architectures in extracting high level features which are transferable. This is possible due to the multiple hidden layers which can capture sophisticated non-linear representations in a dataset. In [123], a TL approach using deep neural networks is proposed for vehicle classification. The authors investigated the possibility of TL of a pre-trained CNN model parameters for classifying truck images generated from 3D point cloud data from LiDAR. Also, in [63] four strategies of TL based on different configurations of CNN models are proposed for plant classification applications. The success of the many applications DTL have been applied to can mostly be attributed to the accessibility to DL architectures such as AlexNet [70], GoogleNet [115], VGG [107] and other architectures which can be

pre-trained and configured to suite a variety of applications. Other methods of TL using neural networks for various applications can be found in [80].

2.3.2 Genetic Algorithms Transfer Learning Methods

Genetic Algorithms (GA) are evolutionary computation methods inspired by natural selection to handle optimisation and global search problems. The algorithms are based on biological evolution operators such as selection, mutation and crossover. Initially, GA's were used to solve complex non-linear optimisation problems and later, they were used in hybrid techniques with other CI methods (like fuzzy logic and neural networks) to solve classification and clustering problems. The authors in [64] proposed a genetic TL model which used two similar fitness functions to predict solutions for source and target tasks. The model aimed at maximising both functions by choosing the best samples and label variables. The results showed that the transfer of inter-task mappings was able to reduce the time required to learn a more complex task. However, there are not many researches focusing on the application of GA's to TL.

2.3.3 Fuzzy Logic Transfer Learning Methods

Attempts to learn activities when there are little information available are often plagued with concerns of imprecision, vagueness, approximation and ambiguity of information. Therefore, it can be drawn that the level of certainty in any activity learning system and the availability of information are co-dependent. This is the reason many researches have incorporated fuzzy logic techniques into TL [7, 103]. Incorporating fuzzy logic allows for approximation and expression of uncertainty encountered in the transfer of knowledge as earlier mentioned in Chapter 1.

The concept of fuzzy logic was introduced in [138] as fuzzy set theory and further expanded to include other aspects such as fuzzy rules [12]. The major elements of fuzzy logic are the if-then rules and the linguistic variable which captures imprecisions in a way similar to humans abilities, thus this makes it relevant in TL. A fuzzy-based transductive TL model for predicting long-term bank failure was developed in [7, 8]. The model applied a fuzzy similarity

measure to refine predicted labels for samples in a target domain. Afterwards the authors improved on the model by proposing a fuzzy refinement domain adaptation method which considers the similarity and dissimilarity in the refinement stage [9]. Shell [103] proposed a framework for Fuzzy Transfer Learning (FTL) for prediction in intelligent environments. The framework introduced the use of a transferable fuzzy inference system from a source domain that is adapted to a target domain. The method was applied in two simulated intelligent environments and the experimental results indicated the proposed FTL framework outperformed classical prediction models, although the model was not compared with other TL models.

2.3.4 Human Activities and Transfer Learning

Developing solutions to aid assisted living is an ever growing field of interest in the research community. This involves the incorporation of a range of technological solutions in assisted living environments to enhance the quality of life and well-being. The rapid evolution of artificial intelligence techniques which are used to learn and model real world behaviours has left the classical ML methods behind in terms of the performance obtainable. The classical learning models usually rely on situations where similar distributions of data are used in training and testing the model [105]. When there are changes in data distribution, such models fail. The models will need to be retrained from scratch which is a slow process and learning a new model will require much data which is always not readily available.

The differences in data distributions can be observed in many applications which involve AAL, for example, in assistive care for monitoring a person living independently [91], detecting changes/abnormality in an AAL environment [35] or learning daily routine activities of a person by an assistive agent. These and many more applications are increasingly encountered in pervasive technologies developed for assisted living. A solution to learning the difference in (or lack of sufficient) data distribution is TL. TL applies the knowledge acquired from one domain in a different but related domain to reduce the time needed for training the models from scratch and performance improvement [92]. This method has seen many applications in assisted living [26, 105].

2.3.5 Discussion

The relationship between the feature spaces in which TL is targeted influences the approach applied to achieve transfer of knowledge. This relationship can be either homogeneous or heterogeneous [92]. In the case of homogeneous TL, the feature spaces of the data in both source and target domains are equal. Situations involving homogeneous TL are much simpler to accomplish when compared to heterogeneous transfer. The work proposed in [93] attempts TL by proposing a method of Transfer Component Analysis (TCA) for domain adaptation. This work entails a dimensionality reduction framework for reducing the distance between domains in a latent space with similar features. The authors in [105] proposed a method of FTL for knowledge transfer. The approach considered the case of applying fuzzy logic to learn and transfer knowledge in intelligent environments. The authors showed that the performance achieved using the proposed FTL framework was comparable to other conventional methods of TL. Although the method in [105] performed well, it considered a situation in which labelled data is only present in the source domain and did not focus on the case of differing feature spaces.

Heterogeneous TL on the other hand is more challenging due to the fact that the feature spaces in both domains are drawn from different distributions of data [92]. The work in [150] proposed a method for a fuzzy rule-based approach to TL in both homogeneous and heterogeneous spaces. Also, a heterogeneous TL method is seen in [77]. An incorporation of fuzzy systems computational technique as seen in [105, 150] show its advantage when applied in transfer of knowledge to a target domain where critical information is inadequate. The benefits of heterogeneous TL enables it to be applied in many real world applications [26, 29].

The works reviewed in this section have used different approaches to TL. Although these works achieve impressive performances when used in their respective applications, not much attention is given to applications in activities of daily living. Especially, when dealing with human activities in assisted living environments which this thesis attempts to address, TL would be of great use in driving technological advancements.

2.4 Assistive Technologies Related to Human Activities

Assistive Technology (AT) refers to the use of adaptive, rehabilitative and assistive devices for either the aged population, people with disabilities or any individual, as means to simplify activities. Such devices are used to improve the functional capabilities of individuals. A categorisation of AT as proposed by [44] is based on the devices and services used for AT. Due to the broad spectrum of AT, the categories identified range from, aids for daily living, mobility, communication, telecare/ telehealth and environmental controls among others. These have applications in different areas of assisted living.

Focusing on the aids for daily living category of AT, devices in this category promote independence in Activities of Daily Living (ADLs) which include activities such as cooking, eating, dressing, moving objects and other daily activities in and around a living environment.

Recently, assistive robots are widely used as aids for daily living. Such robots are equipped with capabilities to carry out functions as required for assisted living. However, the challenges of getting assistive robots to act similarly to human abilities remains a bottleneck. A number of robots exist which are able to perform some basic ADLs but are limited in functionality since they are incorporated with preset information [141]. Koppula and Saxena [67][66] proposed a method of robotic reactive response for anticipating human activities by using object affordances. Human activity information were obtained from videos collected while activities were performed. The system was proposed to aid better incorporation of assistive robots in day-to-day human activities such that the robots are able to anticipate human actions and respond accordingly. To achieve the aim of the work, the authors used an anticipatory temporal conditional random field to model rich spatio-temporal relations through objects. Duckworth et al. [33] recently proposed an unsupervised human activity analysis for intelligent mobile robots framework with the aim of providing assistive robots with a means to understand human activities performed from long-term observations in real-world environments. In the work [33], the authors propose a method of learning human activities from visual

information obtained from a mobile while a person performs an activity. The approach used unsupervised ML techniques to learn activities from extracted features. This approach was intended for assistive robots to be able to learn activities by just observing while activities are performed. Through their approach a mobile robot can be able to infer activities from visual observations which are used to capture different aspects of relations between a human subject and their environment. However, the proposed method focused on analysis of the parameters of the unsupervised techniques used in spatio-temporal representations of observed activities. Furthermore, the method was not extended to practical implementations on a real robot.

To conclude this section, the key points considered in developing robust assistive technological solutions for human activities are highlighted as follows:

- The need to capture the rich context for modelling human activities [67]. This would provide adequate information needed to acquire sufficient knowledge of an assisted living environment. Taking advantage of visual information in 3D space is one of the solutions to providing rich activity information.

- The devices used should be capable of working independently in providing assistance. For the case of assistive robots, the development of intelligent robots which are able to sense, observe and act without human intervention should be investigated more. This also links to the ability to improve the knowledge base as new situations are encountered. Therefore, there is the need for TL incorporated within assistive robots.

2.5 Research Gap

The gaps identified from current research as discussed in the review are highlighted in this section. Also, the section discusses how this work differs from previous research.

A different approach to HAR have used non-visual sensory information due to the advantage that some of the sensors such as, Passive Infrared (PIR), temperature and pressure sensors are non-intrusive. However, other non-visual

sensors like wearable sensors can be intrusive, and as such may not be a best fit for HAR. Also, people often find them uncomfortable and may forget to wear them while carrying out activities. Furthermore, as human activities differ in nature and sequence of occurrences, non-visual sensors are limited in the information they provide. It is often difficult to understand the nature of human actions such as the position/ orientation of different parts of the human body during an activity using the information from non-visual sensors. This results in limitations in effectively creating models for human activity. On the other hand, vision-based approaches to HAR offer rich information (for example, depth, heat map, coloured images and many others) from which a range of features can be extracted for high performance activity modelling and recognition algorithms.

Previous approaches to vision-based in HAR mostly focused on the technical aspects (a systems ability to accurately recognise activities) of the proposed systems [38, 89]. These researches have been directed towards evaluating an algorithm/model's ability to attain good performances on AR. However, not much has been directed towards the practical applications of HAR.

TL has been studied in many context and applications. Most successful applications have been in object recognition from images [29, 115]. Other applications in activity recognition [26, 39] and robotics [52] have not achieved much success due to the complexities of TL. The work in [52] considered a multi-robot TL system. The work addressed TL from a control systems perspective by evaluating the performance of controllers. Feuz and Cook [39] proposed TL through feature space remapping with tests on activity recognition datasets. However, they only considered the case of a feature-rich dataset but did not address situations with sparse data. A similar strategy is considered in this work for human-robot TL which would be a novel approach by combining HAR and TL for human-robot interaction.

To address the gaps identified, this research uses a vision-based sensor to obtain information for developing a framework capable of TL human activities for assistive robots. In the following chapter, the methodology applied in developing the framework is described.

2.6 Summary

This chapter presented the state-of-the-art research related to HAR, TL and assistive technologies. The review presented HAR research works based on visual sensory information as related to the work in this thesis. Different techniques to recognise activities have been investigated. In assisted living, HAR plays a major role in the development of technological solutions to meet the needs of independent living. Although, there are still gaps in practical implementations of such systems, its importance cannot be overemphasised.

TL as an alternative to traditional learning methods, exist to aid the transfer of knowledge across different but related situations of learning, so as to reuse knowledge and avoid having to train models from scratch which is the case with traditional learning methods. By incorporating this concept in HAR, systems such as assistive robots, can adapt to situations which require learning of activities by knowledge transfer from a human to robot space. From the literature review, it is seen that the use of simple, low-cost RGB-D sensors can be used to obtain rich information (which is relevant to any computational system) of activities. This is investigated in this research. To reiterate the focus of this research, an RGB-D sensor is used to obtain information of human activities for the purpose of TL of activities for assisted living applications, such as robots used for assisted living.

Chapter 3

Transfer Learning in Human Activity Recognition: Architecture and Methodology

3.1 Introduction

A motivation for Transfer Learning (TL) is to learn information from a source reference which is transferred to improve on the performance achievable in a target reference. This thesis draws on this motivation to accomplish TL in the context of Human Activity Recognition (HAR). The idea is to develop a framework for TL human activities from visual information which can be adapted in a different setting, such as into a robot, to accomplish the task with the acquired knowledge. In Chapter 2, a broad review of previous studies on HAR with a focus of vision-based approaches, TL and assistive technologies were discussed. This chapter presents the architecture and methodology of the TL in HAR framework developed in this thesis. The main components of the framework are outlined.

This chapter is structured as follows: Section 3.2 gives an insight into the concept of TL and definitions of TL. Section 3.3 gives an ontology of TL as applied in HAR. In this section, an in-depth discussion on how TL is carried out is presented. Section 3.4 follows by presenting the approach employed in

the thesis and the architectural framework detailing the key stages in proposed methodology is presented in Section 3.5. Lastly, Section 3.6 draws conclusions to summarise the chapter.

3.2 Background and Definitions of Transfer Learning

TL is an area that has been well studied across different fields ranging from psychology, education, biology, Computational Intelligence (CI) and many other areas [105]. In psychology, TL which is often referred to as transfer of learning is described as:

> "the process and the effective extent to which past experiences (also referred to as the *transfer source*) affect learning and performance in a new situation (the *transfer target*). It should be conceptualised and explained in the context of its prevalence and its relation to learning in general" [36].

In CI, TL involves developing computational models which are capable of mimicking humans ability to learn and reuse knowledge in different but related tasks. For example, the knowledge acquired while learning to eat with a spoon can be applied in learning to use *chopsticks*. This knowledge is transferred across related tasks. Traditional ML techniques work under the assumption that both source and target data are drawn from a similar distribution of information or similar data domains. This assumption holds in situations where the ML model is applied in classification of data which occur in both source and target information. However, in situations when source and target data are drawn from different information distribution, the traditional ML techniques struggle to correctly identify the target data [105]. This poses a limitation to ML techniques being used in such situations [105]. To address the limitation of traditional ML techniques, TL models seek to apply knowledge learned from a previous/source information to a new, but related target information to improve the performance achieved and to reduce the time needed in training the model from scratch [39].

3.2.1 Notations and Definitions

The notations and definitions relating to TL are introduced in this section. From the review of literature [80, 92, 149], the main elements of TL are *domain* and *task*. Therefore, the relevant definitions of TL and its elements adopted from [92] are given as follows:

Definition 3.2.1. Domain: A domain denoted by D and consists of a feature space F and marginal probability distribution of instances of $P(X)$, $\{x_1, ..., x_N\}$, where $X = \{x_1, ..., x_N\} \in F$ [92]. For example, if the learning task is a HAR problem and the 3D skeleton joint positions are the features, F is the space of all joints vectors and X is a particular observation of an activity action. Therefore, if two domains are different, they may have differing distributions and feature spaces.

Definition 3.2.2. Task: A task T is defined as having a label space Y and an objective predictive function $f(\cdot)$ which is not observed but is used to learn from the available data [92]. The objective function is used to predict analogous labels for new occurrences of X. For the HAR problem, Y is the set of activity labels contained in the dataset.

Definition 3.2.3. Transfer Learning: Given a source domain D_s with a task T_s and a target domain D_t with task T_t, TL aims to improve the learning of target task T_t using the knowledge acquired in D_s and T_s by learning a predictive function in D_t. It assumes that either $D_s \neq D_t$ or $T_s \neq T_t$ [92]. In other words, when both the source and target domain and task are equal, the learning problem is reduced to a traditional ML problem.

Practical implementations of TL aim to transfer as much knowledge from a source task or domain over to the target task or domain. The knowledge transferred varies depending on the application and data from the source available. According to Pan and Yang [92], the key challenge in TL is defining the metrics related to *what to transfer*, *how to transfer* and *when to transfer*. This is mainly due to the fact that there are various algorithms that can be applied in TL. In trying to solve this challenge, TL algorithms used so far have focused on three main steps namely:

- Given a target task T_t, select an appropriate source task T_s or sets of tasks from which to transfer knowledge
- Learn the relationship $f(\cdot)$ between a target task T_t and source task T_s, and
- Transfer knowledge effectively from source task(s) T_s to target task T_t.

These steps have been used by many authors including [26, 92] to propose TL models that can handle the challenges encountered in TL.

Effective implementations of TL aim to improve learning in a target task with the advantage of knowledge acquired from the source task. To measure the effectiveness of TL, Torrey and Shavlik [120] identified three measures by which the transfer of knowledge might improve learning in the target task.

1. The initial performance achievable in the target task T_t using the transferred knowledge before any further learning is carried out, compared to the initial performance of an ignorant agent.
2. The cost in terms of time to fully learn the target task T_t given the transferred knowledge compared to the time to learn it (i.e. target task) from scratch.
3. The final performance attainable in the target task T_t compared to the final performance without any transfer.

Adopting these measures in TL implementations guide in evaluating the improvement of learning of target tasks. Thus, the same measures are employed in the experimental chapters of this thesis.

3.2.2 Variations of Transfer Learning

Following the definitions and notations of TL given, the variations of TL common in most survey papers [26, 92, 105, 130] are categorised under three settings, *inductive TL*, *unsupervised TL* and *transductive TL*. The categorisation is based on the differences in relationship between both source and target domains and tasks. However, within each setting TL can be further

grouped in relation to the type of knowledge transferred. These groups are identified as instance transfer, parameter transfer, feature representation transfer and relational knowledge transfer. Table 3.1 shows the classification of TL based on the type of knowledge transferred in different settings . Descriptions of the different settings are presented as follows:

Inductive TL: This is derived from traditional inductive learning. It defines situations which the target learning task is different from the source task, i.e. $T_s \neq T_t$. The aim of inductive TL is to improve learning of the target predictive function with induced training data [92]. These are few labelled data contained in the target domain. It should be noted that both D_s and D_t are known in induced TL.

Unsupervised TL: unsupervised TL as with other forms of TL aims to improve learning the predictive function in the target domain using information from the source in the target [103]. Similar to inductive TL, $T_s \neq T_t$. The difference is

Table 3.1: Summary of classification of transfer learning based on the type of knowledge transferred

TL approach	Description	TL setting
Instance transfer	Reuses information in the source domain to train a target learning model, usually by re-weighting the source information using a defined metric [26].	Inductive and transductive TL.
Parameter transfer	Explores the shared parameters between the source and target domains/tasks which are useful in transfer [26, 92].	Inductive TL.
Feature representation transfer	Discovers relevant features to reduce the differences between source and target spaces, usually by mapping of feature spaces [92].	Inductive, unsupervised and transductive TL
Relational knowledge transfer	Assumes data is not independent and identically distributed (i.i.d), although both domains are relational. Therefore, seeks to obtain mapping of relational knowledge between both domains [26].	Inductive TL

that the data contained in both D_s and D_t are not labelled.

A deviation from this explanation was proposed in [26]. The authors distinguished learning based on labelled data in the source and target. The terms *informed* and *uninformed* were used. Applying this to the standard learning terms of *supervised* and *unsupervised* learning, Informed Supervised (IS) TL describes when labelled data is available in both domains. Informed Unsupervised (IU) TL implies labelled data is present only in the source domain. By comparison, Uninformed Supervised (US) TL implies labelled data is available in only the target domain and Uninformed Unsupervised (UU) TL implies there is no availability of labelled data in either source and target domains.

Transductive TL: Situations described as transductive TL situations require $T_s = T_t$ but $D_s \neq D_t$ [6]. Following from Cook et al. [26] definition, TL techniques fall under uninformed supervised methods.

3.3 Ontology of Transfer Learning of Human Activities

Assisted living environments are incorporated with different technological solutions to improve the quality of life and well-being. In recent years, there has been a growing interest in the research community on how to develop evolving solutions to aid assisted living, especially in areas of human activity recognition and learning. Different techniques have been studied, as discussed in Chapter 2, to address the need for technological systems which are intelligent enough to evolve their knowledge to solve task which have not been previously encountered. One such approach is TL, for example, getting assistive robots to learn human activities through TL.

TL has recently attracted interest in recent years due to the potential benefits it offers in artificial intelligence applications including assisted living [105], computer vision [84] and robotics [52]. It has not recorded as much success as the long existing traditional Machine Learning (ML) methods partly

due to the challenges which yet remain unresolved in the research community [39], although, it has potential to become a fundamental driver for the success of ML in the coming years. As stated earlier in the literature review in Chapter 2, the challenges facing TL implementations depend on defining the metrics associated with following aspects *what to transfer*, *how to transfer* and *when to transfer*. Providing solutions to address these three aspects has been the focus of many researches, thus, motivating the proposal of different TL algorithms.

In relation to assisted living, different applications of TL have been studied. Shell and Coupland [105] proposed a model called Fuzzy TL which was applied in an intelligent environment. Data from the source domain was learned by constructing a fuzzy inference system from generated fuzzy rules. The constructed fuzzy inference system is then applied to a new domain referred to as the target domain through stages of adaptation of the generated fuzzy rules with the target data. Results from the model tested on real datasets from two intelligent environments (source and target environments) which were different but related showed the model achieves better performance in the target with transfer of knowledge when compared to performance attained without transfer.

Bócsi et al. [15] proposed a method for improving robot learning manipulation tasks from data obtained from the robot performing other tasks or from similar robot architectures. Their method has made an attempt to address the challenge of *how to transfer* by considering two steps which include, dimensionality reduction of data obtained from the robot to a low dimensional space and manifold alignment of source and target robot dimensions through a transformation function. The work in [52] also follows a similar approach of finding *how to transfer* between multi-robots. Even though these works achieve impressive performances, the challenges of *what to transfer* and *when to transfer* prove to be difficult in TL applications. Addressing these challenges require consideration of properties related to spatial and temporal occurrences of both source/target domains.

This research considers the case of TL from human to robot domains in trying to address some of the challenges of TL. Figure 3.1 shows an illustration of TL of a human activity between a human and an assistive robot in which the robot learns to perform a similar task by extracting relevant properties from the activity source

Figure 3.1: An illustration of Transfer Learning of a human activity with an Assistive Robot.

(a human). This thesis aims to follow a similar approach to TL in the context of transfer of human activity between a human and a robot by: 1) identifying requirements for TL in applications using human/robot as source/target domains respectively, 2) propose a method to address the differences between both domains through a remapping of feature spaces. From the review of related works [15, 105], it is evident that once an optimal mapping between source and target domains is known, what/when to transfer would be achievable.

3.4 Methodology for Transfer Learning of Human Activities

To proceed with the description of the novel framework proposed, a number of elements need to be predefined. Reiterating the definitions given earlier in this chapter, a source domain D_s is defined as:

$$D_s = \{F_s, P(X)\} \tag{3.1}$$

where F_s is the feature space and $P(X)$ is a marginal probability distribution within the source domain, given that,

$$X = \{x_1, x_2, \ldots, x_n, \ldots, x_N\} \in F \tag{3.2}$$

The source domain usually consist of a task T_s to be learnt and this is represented as:

$$T_s = \{Y, f(\cdot)\} \tag{3.3}$$

where $Y = \{y_1, y_2, \ldots, y_n, \ldots, y_N\}$ is a label space with an objective predictive function $f(\cdot)$ to be learned by the pairs $\{x_n, y_n\}$ within the source domain. Therefore, for any given scenario, the source domain can be redefined more specifically as:

$$D_s = \{(x_{s_1}, y_{s_1}), (x_{s_2}, y_{s_2}) \ldots, (x_{s_n}, y_{s_n}), \ldots, (x_{s_N}, y_{s_N})\} \tag{3.4}$$

where x_{s_n} is an observed instance of data input and y_{s_n} is a corresponding class label for prediction in the given scenario. Similarly, for a target, the domain D_t, feature space F_t and task T_t can be defined the same way.

Consider a source domain D_s with a feature space F_s and a target domain D_t with a feature space F_t such that $D_s \neq D_t$, implying $F_s \neq F_t$. TL aims to learn a task in D_s and the knowledge acquired is used in solving a different but related task in D_t. An overview of the method proposed to address the challenges of TL discussed in this work by a remapping of feature spaces between source and target domains is presented in Figure 3.2. Information from both domains is required as inputs from which the feature spaces are constructed. For a model applied in a domain to be effectively transferred to a different domain, the features related to both domains need to be studied. The proposed approach assumes transfer is achieved when an effective mapping of F_s is obtained in D_t.

A human performing an activity is assumed to be the source domain D_s with feature space F_s and an assistive robot needed in learning to perform a similar activity is assumed to be the target domain D_t with feature space F_t. The goal is to be able to learn an activity from a human and transfer the knowledge acquired

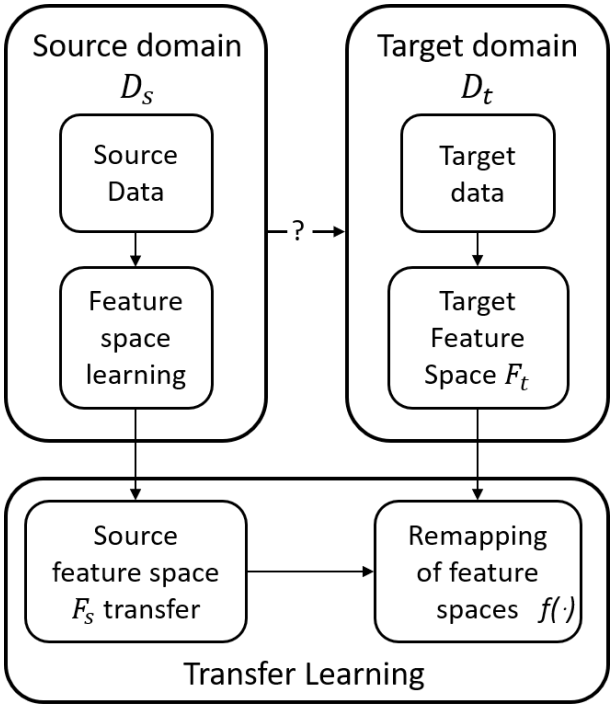


Figure 3.2: Transfer Learning overview by a remapping of features in both source and target domains.

to an assistive robot which would be capable of learning similar activity within an assisted living environment such as the example presented in the illustration in Figure 3.1.

Obtaining sufficient data from a robot to train a model for performing activities is a daunting task with a lot of complexities. However, sufficient data for a model to learn an activity can be obtained as humans perform activities and transferred to a robot. This would also enable assistive robots to learn human activities by observing while a human performs activities. As shown in Figure 3.3, human activity data is obtained from visual cues as activities are performed. The position and orientation features of joints of the human body are extracted. In addition, features of temporal occurrence, velocity, space and motion energy are formulated from the visual information of the activity performed. These features from D_s are used in a learning model for identifying the task performed within the activity.

For a robot to be able to learn to replicate a similar activity, it needs to

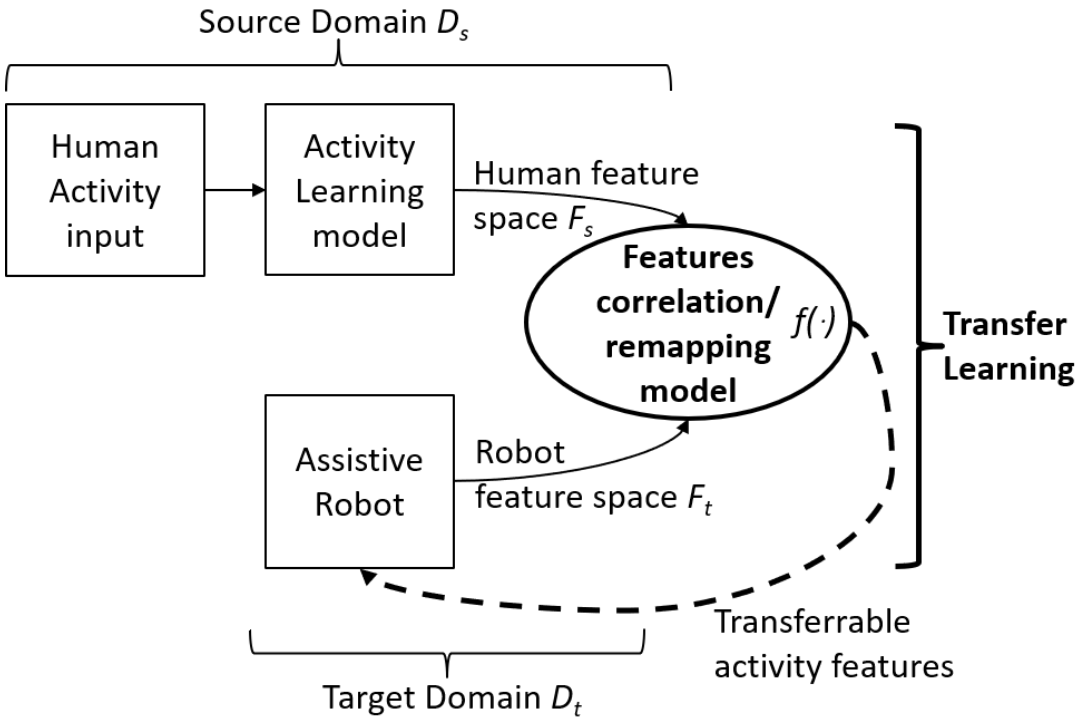


Figure 3.3: Human activity TL from human to robot domains.

understand the feature space of the activity source and how it can be transformed into its own space. The TL model requires the robot feature space F_t as input as well. This feature space can be in the form of joint positions and orientations, forward or inverse kinematics of the robot being used.

3.5 Overview of the System Design for the Proposed Framework

This section describes the proposed framework of TL in HAR. The system design incorporated in this thesis is given in Figure 3.4 and shows the four key stages within the framework. These include:

1. Data acquisition from an RGB-D sensor. More details are presented in the following section.

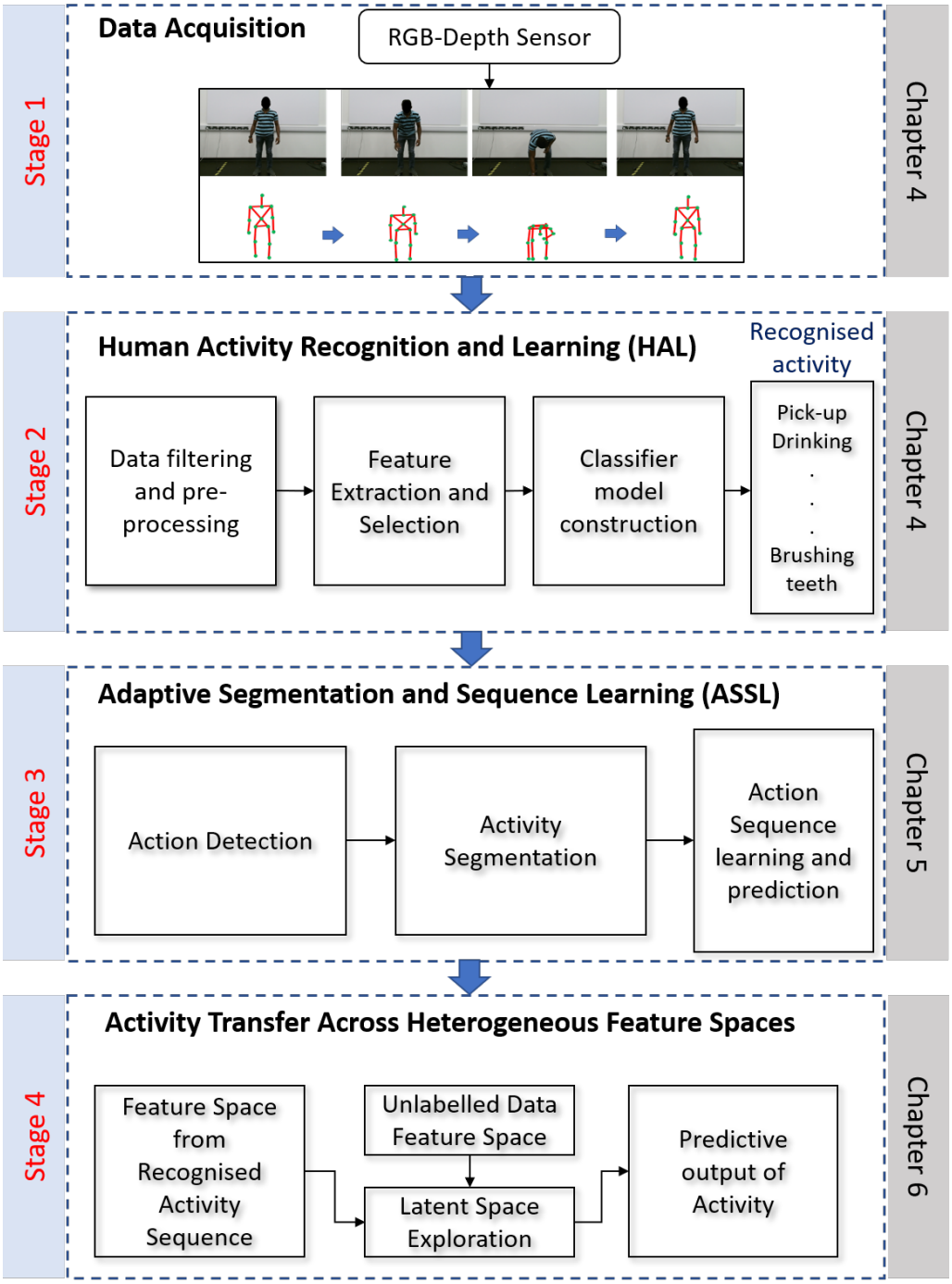


Figure 3.4: System design for the transfer learning in human activity recognition framework.

2. Human activity recognition and learning. A description of the approach used is presented in this chapter and Chapter 4 presents more details of the methodology.

3. Adaptive segmentation and sequence learning of actions. A description of the approach used is presented in subsequent sections of this chapter and more technical details are presented in Chapter 5.

4. Activity transfer across heterogeneous feature spaces. Subsequent sections present a description of the approach used with more details given in Chapter 6.

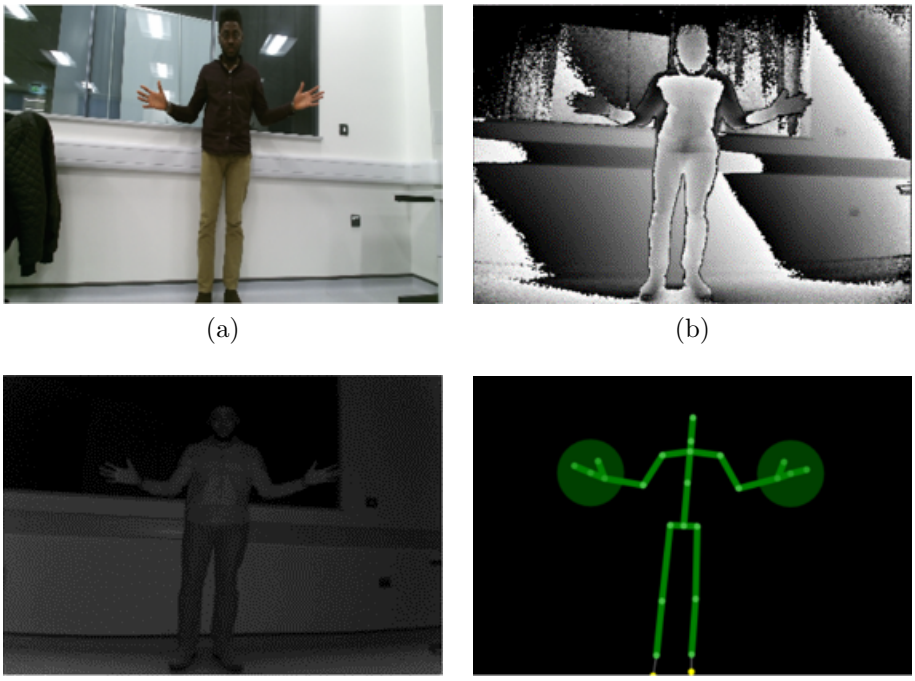
3.5.1 Data Acquisition

In the framework as shown in Figure 3.4, the process starts with obtaining RGB-D sensor information of activities performed by a human. Incoming data are obtained using a single Microsoft Kinect RGB-D sensor [86] which tracks human joint movements and their transitions over time. As mentioned in Chapter 2, RGB-D sensors offer three modes of information which are: RGB (colour), infrared and depth images. These modes of information can be accessed for desired purposes. Figure 3.5 shows samples of different information modes obtained from the sensor used in this work.

The depth information obtained from the sensor is used in this research. From this information, tracking human joints position through each frame is possible. This gives information of each tracked joint location in 3D space of humans during activities. A visual example of tracked skeleton joints obtained from an activity is shown in Figure 3.5(d). This is tracked from the depth information.

Data is obtained from 3D skeleton detection of an actor performing an activity. The skeleton of the actor is tracked using an RGB-D sensor for obtaining positions of joints of the human body. The data representing an activity consist of N number of frames (observations or activity poses). In this work, an activity, a, which is represented by:

$$a = \{J_1, J_2, \ldots, J_n, \ldots, J_N\} \tag{3.5}$$



(a) (b)

(c) (d)

Figure 3.5: Sample frames for different information modalities obtained from an RGB-D sensor; (a) RGB (colour), (b) depth image, (c) infrared image, and (d) tracked skeleton.

where J corresponds to an observation within the activity. Also, each observation consists of input, x_n, and an associated activity label, y_n, represented as:

$$J_n = \{(x_n, y_n)\}, \tag{3.6}$$

and $x_n = [j_1, j_2, \ldots, j_m, \ldots, j_M]$ are 3D human skeleton joint coordinates for j_M joints.

Therefore, by extension of the relation given in Equation 3.4, an activity, a, is represented as:

$$a = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), \ldots, (x_N, y_N)\} \tag{3.7}$$

The Kinect RGB-D sensor considers the skeleton frame of reference from the sensor. Therefore, for better representation of activity features, preprocessing the

data is necessary. The methods used in preprocessing are described in the next section.

3.5.2 Human Activity Recognition and Learning

The second stage in the proposed framework is concerned with the recognition and learning of human activities. This section gives a descriptive overview of the methodology for this stage in the framework. Detailed descriptions of the experimental setup and results evaluation are provided in Chapter 4.

Prior to TL of activities, there is an emphasis on the interpretation of activity information as in the proposed case of activities TL across differing domains. An example is its application in assistive robots. A robot required to perform a human activity would need to be capable of distinguishing one activity from the other which is the process of recognition of activities. This makes it necessary for proper observation of the environment to rightly interpret activities. The information acquired using the RGB-D sensor is used as input in this stage.

3.5.2.1 Data Preprocessing

Data preprocessing is a necessary step in the activity learning process. This is because data obtained during the acquisition process is often noisy (for example too many outliers), may contain missing values, and may be unbalanced in terms of scale (data collected from different experiments could have varying ranges). Therefore, the purpose of preprocessing is to transform the raw data into the right form needed for a model. In this regard, the following steps are taken in preprocessing the data acquired:

Handling Missing Data: Missing data usually occur due to software or hardware faults. In general, there are three types of missing data. These are, Missing At Random (MAR), Missing Completely At Random (MCAR), and Missing Not At Random (MNAR) [62]. In MAR cases, a systematic relationship exists between the inclination of missing data and some other observed data, but not the actual values of the missing data. For instance, if the sensor used in obtaining data is out of action, it is unlikely to be related to the activity performed. Data MCAR occurs when the missing data are not related to either

specific values to be obtained or observed. The missing data points are random subsets of the data obtained. This is a more realistic case in the human activity data used in this work, as the missing data points do not follow any systematic order. For example, when a human carrying out an activity moves out of the sensor's range and the skeleton cannot be tracked or the sensor's speed of recording an activity is not sufficient leading to the loss of some samples. In the case where the characteristics of the missing data do not meet those of MAR and MCAR, they fall in the category of MNAR. The only way to tackle such cases of MNAR is to model the missing data.

There are various techniques for handling missing data. The techniques commonly used include, Pairwise deletion, listwise or case deletion, mean substitution, regression imputation, multiple imputation, maximum likelihood, Expectation-Maximisation and Last Observation Carried Forward (LOCF) [62].

The list-wise deletion technique is often used in many studies which involve acquiring repeated measurements over a time series. It discards those observations with missing data and uses the remaining data for analysis. This work adopts this technique in handling missing data points of observed human activities. Human skeleton data obtained using RGB-D sensors often contain large amounts of samples due to the sensors [86] ability to attain high recording frame rates. Since the data is large enough and the missing data assumptions satisfies the MCAR, the list-wise deletion method is an ideal solution.

Data offset: Human activity data obtained using RGB-D sensors are dependent on the position of the sensors. Therefore, when an activity is performed by many subjects, there is an added variation in information due to the distance of the sensor from the subject. A process of translation is applied to offset the data from the sensor coordinate. This resolves the problem of scale variance encountered in many vision-based human activity learning systems.

In addition to offsetting the data, in primitive human activity learning, there are many variations in the way an activity is performed from one subject to another. For instance, performing an activity of picking up an object, one subject might be left-handed while another subject right-handed. This situation may lead to limitations in a learning models performance. Therefore, in this work, each subjects data obtained is transformed by rotating 180 degrees about the y-axis.

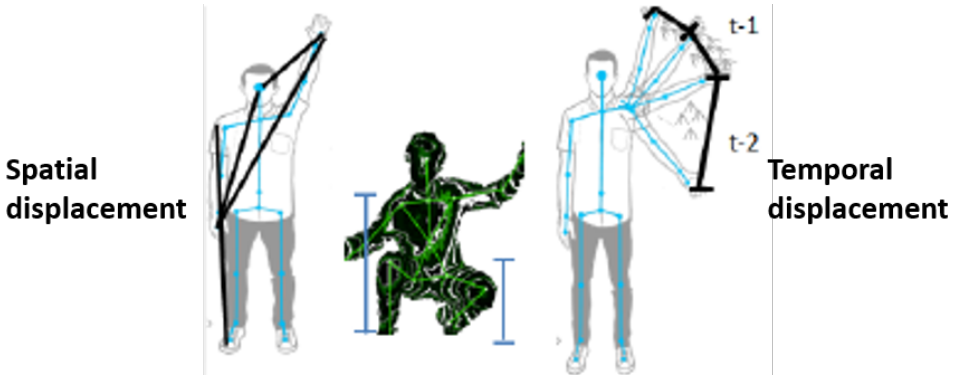


Figure 3.6: A representation of spatial and temporal features from skeleton joint coordinates information.

3.5.2.2 Feature Extraction

Feature extraction is an important aspect of any activity recognition system as raw data obtained from activities do not provide enough information to allow implementing an activity recognition system. Features obtained in HAR systems can be computed using the human skeleton joints coordinates obtained from an RGB-D sensor.

After the preprocessing step, the information obtained, $a = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), \ldots, (x_N, y_N)\}$, are converted into a set of useful feature vectors, F, that model human activities by passing the information to a feature extraction system. The features extracted in this work are based on raw joint positions and displacement-based representations when considering temporal and spatial information, and statistical features in time domain. Figure 3.6 illustrates representations of spatial and temporal information features. Following the feature extraction, the output set of features $F = \{f_1, f_2, \ldots, f_n, \ldots, f_N\}$, where f_n is an identified feature vector within the set. For example, the feature f_1 is computed as the Euclidean distance between two joints through the sequence of activity a. The computation of these features are presented in Chapter 4.

3.5.2.3 Activity Classification

The final stage in learning and recognising human activities is classification of activities using the extracted feature vectors. This step aims to associate feature vectors to the correct activity. From the literature review, it is observed that there is no classification model that is best for HAR and works for all datasets. Therefore, an ensemble of classifiers method is used in this work for classifying instances of the input $F = \{f_1, f_2, \ldots, f_n, \ldots, f_N\}$ to the corresponding labels $Y = \{y_1, y_2, \ldots, y_n, \ldots, y_N\}$. Three classifiers namely, SVM, KNN and RF are investigated and combined in the construction of the ensemble of classifiers. The configuration of these three classifiers are summarised as follows:

- A multi-class SVM implementation similar to [24] is applied for activity recognition. The multi-class SVM is an extension of the SVM from binary classifier. A *one against-one* approach which is based on the construction of several binary SVM classifiers is stated to be the most suitable for practical use. This method is necessary for Y classes dataset, where, $Y > 2$. A training phase is carried out during which the activity features are given as input to the multi-class SVM together with activity labels. In the test phase, activity labels are obtained from the classifier.

- KNN is among one of the simplest ML algorithms and is a method of classifying objects based on closest training points in the feature space. An object is assigned to a class most common among its K nearest neighbours (where K is a positive integer) by a majority of votes of its neighbours. In most cases, the Euclidean distance is used as the metric in finding the nearest neighbours to an object. Applying this method in the proposed approach, in the training phase, the activity feature vectors and activity labels of the training set are stored. During the classification phase, the user defined constant, K, and unlabelled activity feature vectors are classified by assigning a label most frequent among the K training samples.

- RF is an ensemble learning method based on decision trees. A group of decision tree classifiers are trained on different random subsets of the input

information and the output is obtained as the class that gets the most votes from the predictions of individual trees.

The ensemble method takes advantage of the performance achievable with combined classifiers which is most times better than a single classifier model.

Evaluating the performance of a recognition model is important to know how well the model performed in learning and recognising the activities. The experiments carried out in this work to evaluate the performance of the activity learning model employ a cross-validation technique. This technique takes a proportion of the input data which is used for training the model and afterwards, the trained model is tested on the data left - *new data* - out during training. A Leave-One-Out Cross Validation (LOOCV) technique is used [51]. This is a k-fold cross validation method where k is the number of subjects in this case. For example, for an activity performed separately by four subjects, $k = 4$. The model is trained using three subjects leaving one subjects' data for testing. This is done iteratively and the average error is computed and used to evaluate the model.

Several metrics are used to evaluate the performance of HAR models [128], some of which are used in this work. These include the accuracy, precision and recall, and are defined as follows:

1. Accuracy: is a widely used statistical metric in HAR to evaluate how well a model correctly identifies a condition [128]. It is the proportion of true results, that is, both True Positives TP, and True Negatives TN, to the total number of cases considered. It can be represented as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.8}$$

   where FP are the False Positives and FN are the False Negatives.

2. Precision: is the proportion of the True Positives to all positive results. In terms of classification performance, the precision measures the substantial results that are relevant. This is given as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.9}$$

3. Recall: also called the true positive rate is the proportion of correctly identified positive instances to the total number of correctly classified instances. The recall finds the true class accuracy from a given model and is given as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.10}$$

3.5.3 Adaptive Segmentation and Sequence Learning

Following the recognition of human activities in the second stage, the third stage of the proposed framework is concerned with the adaptive segmentation and sequence learning of actions in an activity. This section gives a descriptive overview of the methodology for this stage in the framework. Detailed descriptions of the experimental setup and results evaluation are provided in Chapter 5.

Humans have the ability to learn activities by observing while activities are executed by another human. One important aspect of this process is extracting segments of key aspects of activities and exploiting this information to be able to replicate the constituent actions. This involves generating activity representations required to understand sequential movements of different body parts towards actualising the activity. Therefore, to understand the constituent actions, segmentation is performed and then the sequence of actions are learned from obtained segments.

3.5.3.1 Action Detection

Activity information obtained using the RGB-D sensors contain several actions and not all actions are relevant in determining the sequence of an activity. Therefore, it is necessary to identify key actions. However, detecting key actions can be a tedious task which requires much computational resources to process the entire activity information obtained. To detect key actions, this work investigates the motion energy, E_l feature of human skeleton joints in an activity. The movement of joints through an activity show changes in acceleration and deceleration. Therefore, by exploring this feature key actions can be identified. For example, the motion energy for an observation, $E_l(J)$, is

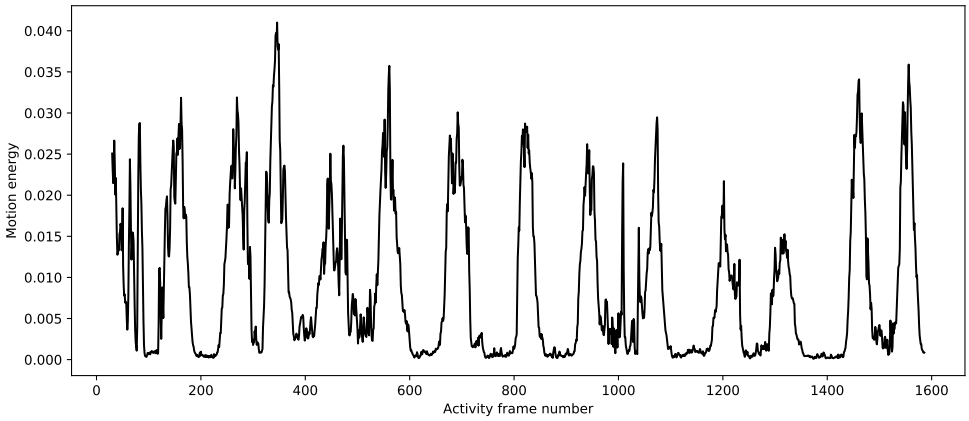


Figure 3.7: An example of motion energy data for an activity sequence obtained from one subject.

the cumulative energy of all joints in that observation given as:

$$E_l(J) = \sum_{m=1}^{M} E_l(j_m) \tag{3.11}$$

A key action, $\overline{J}$, is identified using the functions $max(E_l)$ and $min(E_l)$.

Figure 3.7 shows an example of the motion energy for a sequence an activity performed by a subject. The peaks of both acceleration and deceleration represent key actions of an activity.

3.5.3.2 Activity Segmentation

Prior to sequence learning and prediction, the number of segments that an activity comprises of need to be known. This information is not easily obtained from mere observations of the extracted key actions. Similar key actions are grouped using a clustering technique to obtain activity segments, Q. Clustering techniques differ in terms of the way feature spaces are grouped. A generic grouping of these techniques are based on parametric and non-parametric methods in which clustering is done. Parametric methods rely on some assumptions of certain parameters (for example, the number of clusters expected) prior to analysis of the dataset [25]. However, in situations (like that of this work) where such assumptions cannot be made, non-parametric

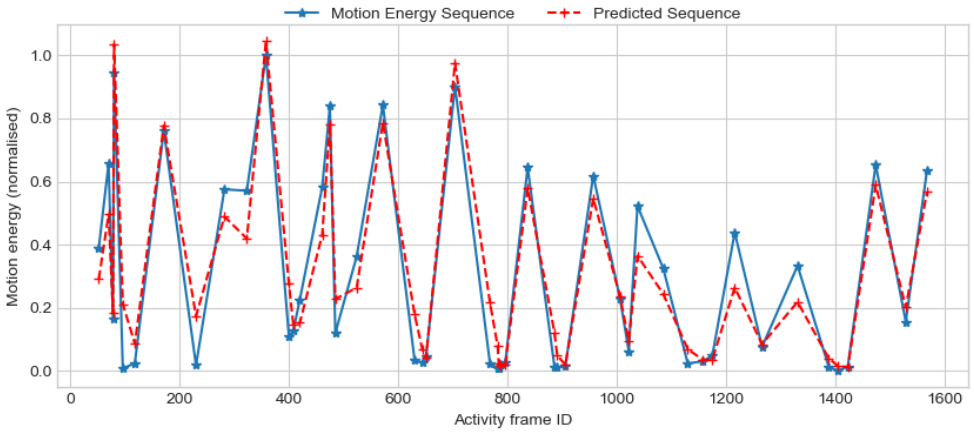


Figure 3.8: An example of the predicted activity sequence for an activity performed by a subject.

clustering methods are best suited. These methods provide a flexibility in the analysis of complex multi-modal feature spaces [25]. Therefore, a non-parametric clustering method is explored in this work for the segmentation of the obtained key actions using the expression:

$$Q_z = C(\overline{J}_b) \tag{3.12}$$

where $z = \{1, 2, \ldots, Z\}$ for Z activity segments and represents each unique segment, $b = \{1, 2, \ldots, B\}$, for B key actions and C is a function assigning each key action to a unique segment.

3.5.3.3 Sequence Learning and Prediction

This step aims to learn the sequence of actions from identified segments. This work employs an RNN method in sequence learning and prediction. A Long Short-Term Memory (LSTM) [53] network is used due to its ability to recall past occurrences over a long period from time series information. The key actions, $\overline{J}_b$ and their respective segments, Q_z are inputs to the network.

The performance of the sequence learning and prediction model is done using the LOOCV method as described earlier in Section 3.5.2.3. Figure 3.8 shows an example of predicted activity sequence for one subject. The figure shows the

actual sequence obtained from motion energy of identified key actions and the sequence predicted after learning. However, the metrics used in evaluating the performance as with most time series prediction models are based on the Mean Absolute Error (MAE), Mean Absolute Scaled Error (MASE), and Root Mean Square Error (RMSE). These metrics are defined as follows:

1. Mean Absolute Error: is a measure of the average magnitude of errors in a set of predictions, without consideration of the direction. It is obtained as follows:
$$MAE = \frac{1}{B}\sum_{b=1}^{B}|\overline{J}_b - f(\overline{J}_b)| \tag{3.13}$$
where $f(\overline{J}_b)$ is the predicted value of $\overline{J}_b$

2. Mean Absolute Scaled Error: is used as a measure of accuracy of predictions. It is computed as the ratio of the MAE of predicted actions, $f(\overline{J}_b)$ to the MAE of in-sample one-step forecast.

3. Root Mean Square Error: is measured as the square root of the average of squared differences between a predicted action, $f(\overline{J}_b)$ and the actual action, $\overline{J}_b$ which is represented as:
$$RMSE = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\overline{J}_b - f(\overline{J}_b))^2} \tag{3.14}$$

3.5.4 Activity Transfer Across Heterogeneous Feature Spaces

The fourth and final stage of the proposed framework is concerned with the transfer of the learned human activity across differing feature spaces. This involves transfer of the activities and actions learned from human domain, D_s, with a feature space, F_z to robot domain, D_t with feature space, F_t. This section gives a descriptive overview of the methodology for this stage in the framework. Detailed descriptions of the experimental setup and results evaluation are provided in Chapter 6.

Algorithm 1 Guided algorithm for TL by feature space remapping from source to target domains.

Input:

Source domain feature space F_s and Target domain feature space F_t.

Output:

Mapping function $f(s)$ from F_s to F_t.

Procedure:

1: Check and remove all duplicate features in F_s and F_t using the preprocessing approach described in Section 3.5.2.1.
2: For every observation in the source domain D_s^i, a weight W_s^i is estimated for each feature *for* $i > 0$.
3: Similarly, weights are constructed for the target features and represented by a matrix W_t.
4: For identical features in F_s and F_t, return corresponding weights W_s and W_t.
5: For the non-identical features in F_t, find correlation between weights W_s and W_t.
6: $f(s)$ is obtained by running a similarity function on weights W_s and W_t obtained, and a transformation of learned model to the target domain.

In situations where $F_s = F_t$, there can be a direct mapping from source to target to achieve transfer. This case is a much simpler case of TL where the challenges of *what/when to transfer* can be addressed with less computational effort. However, in applications involving human-robot interaction where a robot is required to learn an action from a human, the difficulty remains *how* transfer can be achieved. The differences in both feature spaces makes it not feasible for a direct mapping of features across the robot/human domains. This work assumes the robot domain needed for transfer of knowledge differs in feature space from that of a human, that is, $F_s \neq F_t$ and therefore for TL across such domains this thesis proposes a remapping of feature space from source to target domains.

The proposed method for a remapping of feature spaces is summarised in Algorithm 1. It should be noted that this algorithm presented at this stage is a guided algorithm. A more detailed algorithm is given in Chapter 6. The method requires both source and target domain feature spaces as inputs and the output obtained is a mapping function $f(s)$ which is a transformation of source features into relevant target features. Duplicate features within the feature spaces are discarded and weights W_s and W_t are assigned to features through a measure

of feature importance in both domains. Identical features are extracted in a matrix while a method of correlation is applied to the weights of non-identical features to deduce a relationship between the features. A rule-based approach is used to identify the similarities between both feature spaces and as a common ground for representation of complexities in activity sequences. Once this stage is completed, a mapping function is defined which is used in the transformation from F_s to F_t. It is worth noting that the proposed TL by feature remapping method is generalisable to different applications. This is possible if the feature spaces for transfer of knowledge are identified and not specific to an application or information distribution.

3.6 Discussion

This chapter presented a detailed description of the framework - shown in Figure 3.4 - developed in this research. The concept of TL and its applications in an assisted living environment is discussed with a proposed application in assistive robotics - which is increasingly being incorporated in assisted living environments and explored in other applications to provide meaningful services to the end-users. The ontology of TL of human activities is discussed and a description of the methodology adopted in this work which is based on learning the relationship between feature spaces. This builds upon the primary motivation for assistive robotics applications.

Furthermore, to achieve the aim of TL human activities, the chapter presented the architectural framework showing the different stages involved. This comprises of the acquisition of data from an RGB-D sensor, human activity recognition and learning, the adaptive segmentation and sequence learning of actions, and the transfer of activity across differing feature spaces. In the following chapters, detailed descriptions of all the stages including the technical formulation of methodologies, experiments conducted and evaluation of results are presented.

Chapter 4

Human Activity Learning and Recognition for Assistive Robotics

4.1 Introduction

Ambient Assisted Living (AAL) is an active research area that has attracted a lot of interest in recent years through the development of various solutions to enable independent living and promote quality of life and well-being for an ageing human populace [14]. AAL solutions utilise assistive robots and other technologies to aid in daily routine activities. The robots are incorporated in various applications which involve human-computer interaction that traverse humans of all ages. Such applications include care for older adults [61, 134].

Due to the dynamic nature of the environment in real world applications, it is quite challenging to have assistive robots execute functions easily. A specific case is assistive robots that can interact with older adults as carers. These robots learn tasks by observing a human carer execute the tasks. Such robots learn human activities by extracting descriptive information of the activities in order to classify them as they are executed. This process involves a transfer of knowledge/information of the activity performed which is *Transfer Learning* [130].

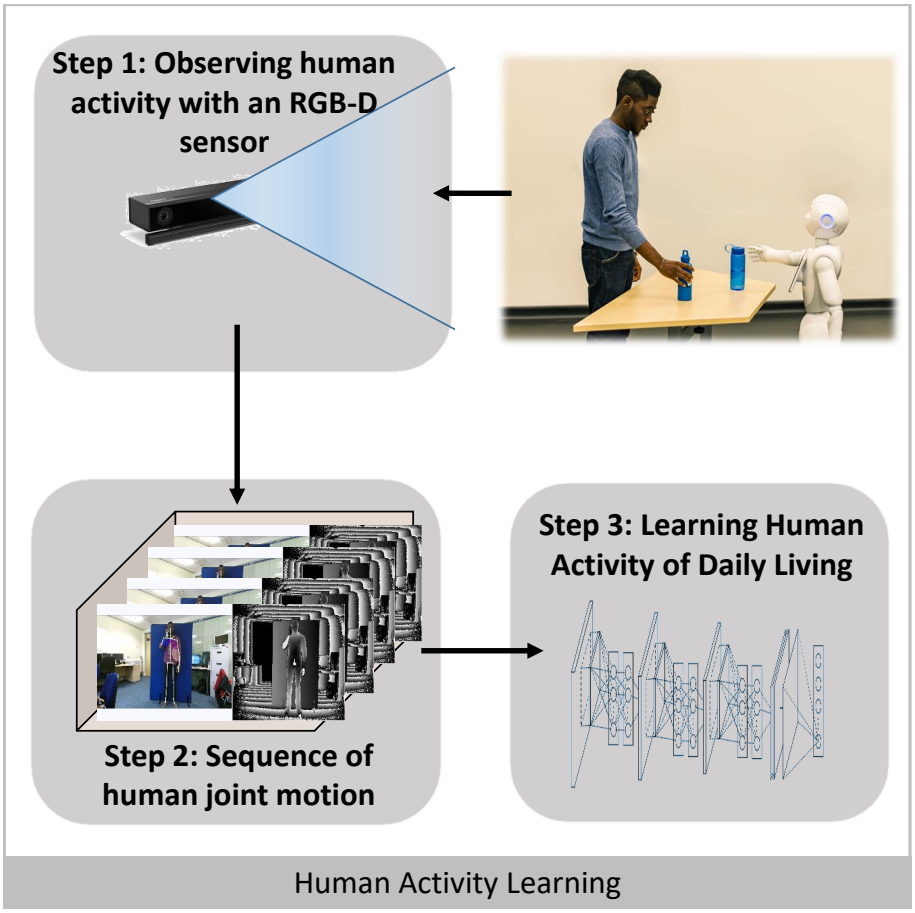


Figure 4.1: A conceptual overview of learning of human activity by an assistive robot using information from an RGB-D sensor.

Regardless of the method applied to learning an activity by a robot, there is a knowledge gap contained in the varied information acquired of a person executing an activity and a robot carrying out a similar activity. Transfer Learning (TL) helps to bridge this gap by providing faster learning of activities and better collaboration of assistive robots in AAL environments [52]. A conceptual overview of the processes involved in learning of human activities for assistive robotics is given in Figure 4.1. It is evident in this context that the ability to correctly recognise a human activity, and correctly learn (as highlighted in steps 1-3 of Figure 4.1) such activity plays a significant role in the amount of knowledge which can be transferred to an assistive robot to be used in learning.

This chapter presents a novel Human Activity Learning (HAL) system for assistive robotics. This will act as part of the process of TL for assistive robots as mentioned in the previous chapter. The focus is on the three steps shown in Figure 4.1. An RGB-D sensor is used to obtain 3D skeleton information of body joints during activities as they are executed by a human. Descriptive features are then extracted from the skeleton information obtained and the most informative features are selected to be used in training a classifier model. These features are extremely valuable in evaluating the performance of the system because redundant and noisy features can have negative effect on the system performance. An ensemble of classifiers model is used in building the learning model for activities. The approach presented here employs three classifiers - Multiclass Support Vector Machines (MSVM), K-Nearest Neighbour (K-NN) and Random Forest (RF) - in creating the ensemble model. These classifiers are classical algorithms used in ML problems. The proposed method is not only focused on using the selected algorithms but a combination of them in an ensemble. The reason for using an ensemble of classifiers is to improve performance compared with a single classifier model [116]. The results discussed in subsequent sections show the improved performance.

The remaining sections in this chapter are structured as follows. In Section 4.2, details of the methods applied in 3D data processing and feature representation are explained. Section 4.3 explains the classifier ensemble model approach for human activity learning. Section 4.4 presents experimental results and their evaluation, Section 4.5 summarises the main results and provides discussion of the future work.

4.2 Methodology for Human Activity Data Processing and Feature Representation

The proposed approach to HAL described in this chapter works by extracting features from 3D skeletal data and applying feature selection techniques for selecting the most informative features used in building a learning model for human activities. The overview of the system architecture shown in Figure 4.2

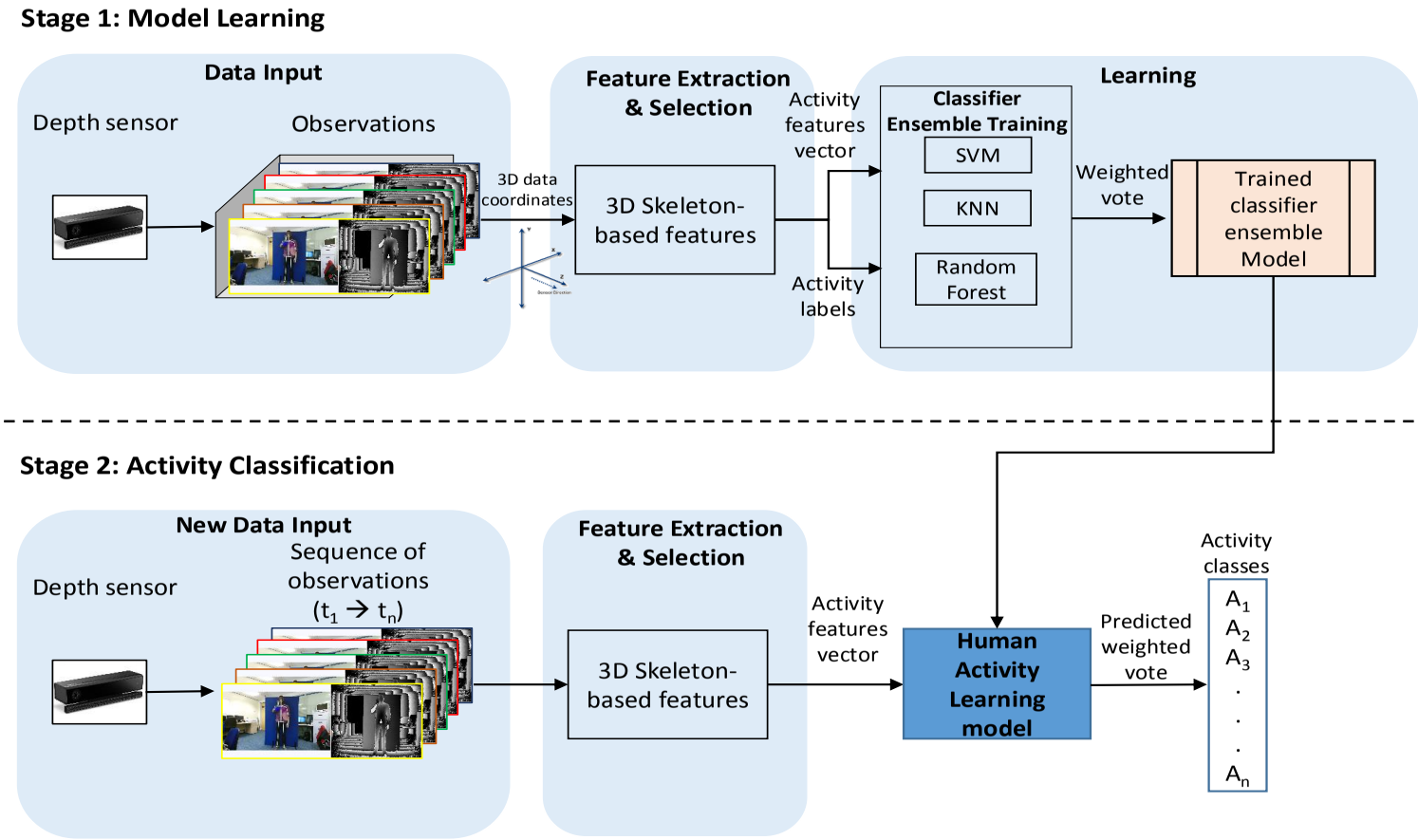


Figure 4.2: Architecture of the human activity learning model. Stage 1: Model Learning (top): learning human activities by training a set of classifiers (SVM, KNN and RF) from 3D skeleton features obtained from activity frames captured using an RGB-D sensor. Stage 2: Activity Classification (bottom): observations from human activity are used to extract/ select relevant features which are fed into the trained classifier models, and activities performed are detected.

illustrates the main stages within the process. This is divided into two stages as follows:

Stage 1: Model learning

- Data Input: Data input into the system from a dataset containing 3D skeleton information of human joints. This data is captured using an RGB-D sensor and pre-processed before it is used in training activity classifier ensemble model.
- Feature Extraction and Selection: Features representing activities are computed from the data. This step also includes the selection of optimal features relevant for learning activities.
- Learning: Training selected classifier models through supervised learning of activities. The output of this step is the learned classifier ensemble model ready to be utilised in activity classification.

Stage 2: Activity classification

- New Data Input: Data input in this stage is similar to that described in the model learning stage. However, this has to be unseen data in order to validate the performance of the learned models. The data can be obtained from a dataset or on-the-fly from an RGB-D sensor.
- Similar features are extracted from the data to be classified. The key difference in this stage is the data used is unlabelled unlike the model learning stage which is based on a supervised approach. The features extracted are passed into the learned classifier ensemble model for identification of activity classes.

4.2.1 3D Activity Data Preprocessing

Human activity is composed of a continuous transformation of a series of human poses. Preprocessing the information is necessary to reduce irregularities in the data obtained from the sensor. RGB-D sensors provide information in three modes namely; 1) RGB image, 2) depth image and 3) skeleton joint coordinates.

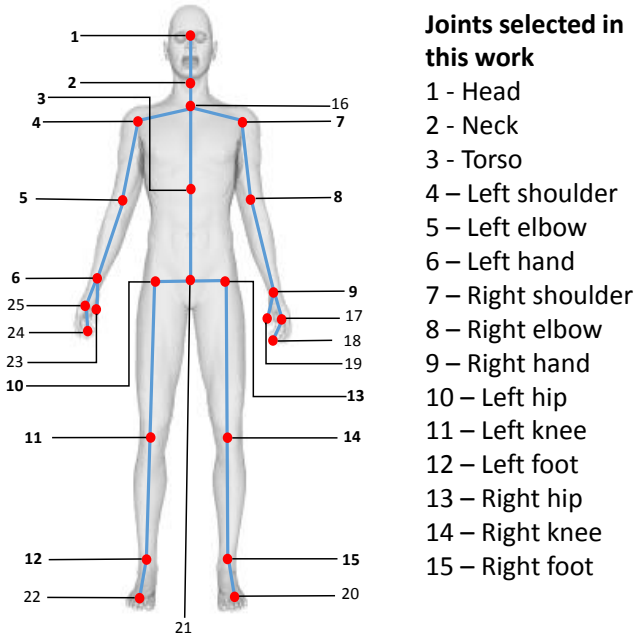


Figure 4.3: Skeleton representation of Microsoft Kinect V2 with 25 joints. 15 key joints (i.e. the highlighted joint labels) are used in this work as shown in the label definition in the figure.

However, this work uses only the skeleton joint coordinates information. A Microsoft Kinect v2 [86] RGB-D sensor which has a skeleton model consisting of 25 joints as shown in Figure 4.3 is used in this work. From the information obtained from the Kinect sensor, 15 key joints (i.e. the highlighted joint labels) as outlined in Figure 4.3 are selected for use. Data is acquired from the sensor as frames containing different poses that make up an activity. 3D skeleton joint coordinates J are obtained from pose approximation in each frame [135] with coordinates relative to the sensor position where,

$$J = [j_1, j_2, \ldots, j_m, \ldots, j_M], \quad \text{for } J \in \mathbb{R}^{3\times M}, M = 15 \tag{4.1}$$

j_m represents a joint in the frame with coordinates x, y, z corresponding to horizontal, vertical and depth positions respectively and M is the number of skeleton joints in a frame.

To make the joint coordinates invariant of the sensor position, the origin of

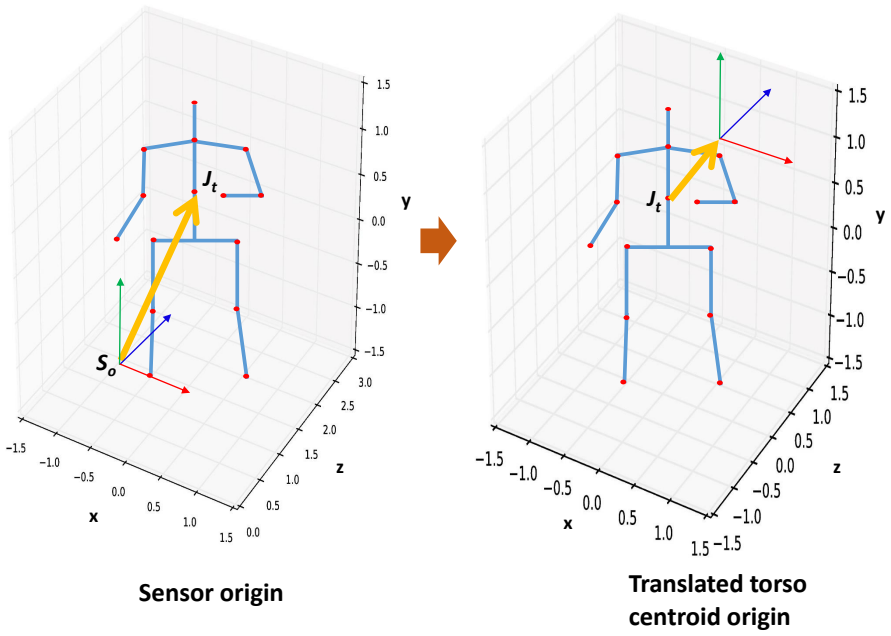


Figure 4.4: Translation of skeleton coordinate system from the sensor origin to the torso centroid origin.

the skeleton is translated along the vector $\overrightarrow{s_o j_t}$, where s_o is the sensor coordinates origin and j_t represents the torso centroid joint of the skeleton. Each joint coordinate position $\overrightarrow{j_m}$ (j_m is a vector representing each joint coordinates of the skeleton) is computed with reference to the new origin of torso centroid $\overrightarrow{j_m} - \overrightarrow{j_t}$. Thus, the skeleton is independent of the sensor position as shown in Figure 4.4. Each sample posture of activity is then reformulated to the torso centroid origin.

Another stage of pre-processing is done to symmetrise the data in order to eliminate ambiguity in gestures performed by left and right-handed people. This ensures each activity is represented in a variation of its original form as shown in Figure 4.5. The symmetry is computed along the $y - axis$ of the origin (torso centroid).

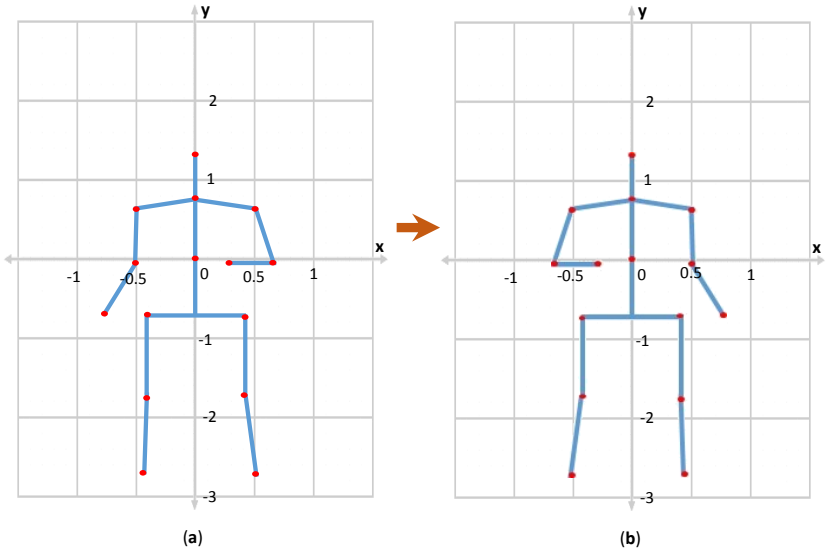


Figure 4.5: Skeleton symmetrisation of an activity posture about the $y - axis$. (a) represents the original activity posture and (b) is the symmetry obtained of same posture.

4.2.2 Extraction and Representation of 3D Features

Extraction of descriptive information from acquired raw sensor information is crucial to any learning system as raw data does not provide adequate information for learning. This is carried out after the data is pre-processed. In this work, the features used are divided into two distinct categories: joint displacement based features and statistical features in the time domain. Joint displacement based features encode information relative to position and motion of body joints [48, 135]. This information considers displacement between joints of an activity pose and 3D position differences of skeleton joints across different time periods of an activity. Similarly, statistical time domain features encode information of variations across a collection of activity poses within a specified time domain. The following sections provide details of the features used in this work.

4.2.2.1 Displacement-based features

Displacement-based features represent the features obtained as a result of a shift in position of a human joint either with reference to a fixed position or change

through time. These features are computed as follows:

1. Spatial displacement between selected skeletal joint coordinates is computed as the Euclidean distance feature, f_1, between any two joints described in Equation 4.2. The joints are selected based on relevance to activities.

$$f_1 = \sqrt{\sum_{x,y,z}(j_m - j_i)^2}, \tag{4.2}$$

   for $1 \leq$ (m,i) $\leq M$ and $m \neq i$. j_m and j_i are any pair of selected joints with coordinates x, y, z.

2. Temporal joint displacement features consider the 3D consecutive motion of joints, f_2, and the overall motion dynamic of joints, f_3, features through an activity. f_2 is computed as the joint coordinates position difference between the current pose c and its preceding pose p in Equation 4.3 and f_3, as the temporal difference between the each joint current pose from the initial pose J_1 in Equation 4.4.

$$f_2 = [j_m^c - j_{m-1}^p]; \quad \text{for } j_m^c \in J_n^c \text{ and } j_{m-1}^p \in J_n^p \tag{4.3}$$

$$f_3 = [j_m^c - j_m]; \quad \text{for } j_m^c \in J_n^c \text{ and } j_m \in J_1 \tag{4.4}$$

4.2.2.2 Statistical features in time domain

This is computed as the projected difference of joint coordinates of the current pose J_n^c (also referred to as the current activity frame) from the mean, variance, standard deviation, skewness and kurtosis of joint coordinates for an activity. These features are computed as follows:

1. Joint coordinate-mean difference;

$$f_4 = j_m - j_{mean} \tag{4.5}$$

where the mean of all positions for a joint coordinate is $j_{mean} = \frac{1}{N}\sum_{m=1}^{N} j_m$ and N is the number of poses in an activity.

2. Joint coordinate-variance difference;

$$f_5 = j_m - \frac{\sum_{m=1}^{N}(j_m - j_{mean})^2}{N} \tag{4.6}$$

3. Joint coordinate-standard deviation difference;

$$f_6 = j_m - \sqrt{\frac{\sum_{m=1}^{N}(j_m - j_{mean})^2}{N}} \tag{4.7}$$

4. Joint coordinate-skewness difference;

$$f_7 = j_m - \frac{\sum_{m=1}^{N}(j_m - j_{mean})^3}{(N-1)\sigma^3} \tag{4.8}$$

where σ refers to the standard deviation of each joint coordinate for all poses in an activity.

5. Joint coordinate-kurtosis difference;

$$f_8 = j_m - \frac{\sum_{m=1}^{N}(j_m - j_{mean})^4}{(N-1)\sigma^4} \tag{4.9}$$

All activity feature vectors computed are concatenated to form a matrix, F, of extracted activity features in which the columns correspond to feature vectors and the rows correspond to features extracted from different frames of activities. F is represented by the following;

$$F = \{f_1, f_2, \ldots, f_8\} \tag{4.10}$$

4.2.3 Features Normalisation

HAL systems can be problematic if the extracted features are not well processed. This is due to the heterogeneity in features. A further pre-processing

of extracted features is needed to deal with the issue of features heterogeneity before classification. This is done through feature normalisation which is often applied in many ML applications [19, 113]. Normalisation of each feature in the activity features matrix obtained in Equation 4.10 is done according to:

$$F_{norm} = \frac{f_n'^{,c} - min(f_N')}{max(f_N') - min(f_N')} \tag{4.11}$$

where $f_n'^{,c}$ is a feature from the column feature vector, f', of the current pose J_n. The obtained feature matrix after normalisation becomes F_{norm}.

4.2.4 Feature Selection

Feature selection is performed on the normalised activity features matrix.This is important to any learning model as it enables faster training, reduces over-fitting, improves accuracy and reduces model complexity (making it easier to interpret [19, 45]. In this work, a filter method for feature selection known as Relief-F [65] is applied. Filter methods are preferred to other methods such as wrapper methods since they do not require a fixed learning mechanism and therefore have more generalisation across different learning models [45].

The Relief-F method uses a statistical approach rather than heuristic to provide relevance weights to rank potential features. The features ranked above a set threshold are selected for the model. In this chapter, the threshold is determined from the number of features that provide the best substitution accuracy with the learning model. The performance achieved using the selected features is presented in the experimental results in Section 4.4.

4.3 Classifier Ensemble Model

The final stage in developing an activity learning system is training a classification model with the selected features to achieve a good learning performance score. In the work presented in [38], a selection of learning models were used separately to identify activities. This thesis employs a combination of different learning models in a framework referred to as a bagging ensemble of classifiers in order to achieve

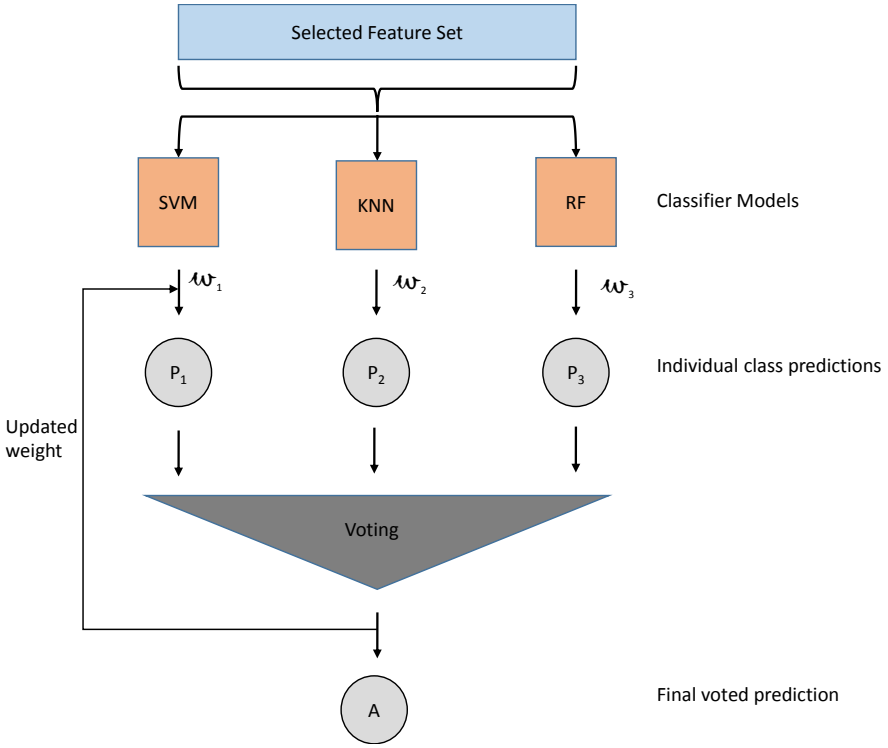


Figure 4.6: Overview of weighted voting architecture of classifier ensemble.

an improved performance of the system. The use of an ensemble of classifiers model generally allows for better predictive performance than the performance achievable with a single model [30, 137]. According to [116] Ensemble Models, are learning models that construct a set of classifiers used in classifying new information based on a weighted vote of individual classifier predictions. Three base classifiers are used in this work to construct a bagging ensemble of classifiers; Multi-class Support Vector Machines (SVM), K-Nearest Neighbour and Random Forest classifiers. The pictorial overview of the bagging ensemble method applied is shown in Figure 4.6.

The weighted votes works by computing the weighted majority vote $\hat{q}$, through allocation of weights ω_r to each classifier C_r.

$$\hat{q} = \arg\max_i \sum_{r=1}^{3} \omega_r \times (C_r(s) = i), \tag{4.12}$$

where $C_r(s)$ is a classifier characteristic function in a set of unique classifier labels.

The weights assigned to individual classifiers in the ensemble are computed during the learning phase by weighted votes. At the initial stage uniform weights are set and updated at each iteration of cross-validation. The updated classifier weights used in succeeding iterations are computed as ratios of the average precision obtained in the preceding iteration of each classifier in the ensemble.

The multi-class SVM model follows the configuration reported in [24] which is an extension of a binary classifier. A *one against-one* approach based on the construction of several binary SVM classifiers suitable for Y classes contained in a dataset ($Y > 2$) is implemented as one of the base classifiers. The K-NN classifier algorithm is one of the simplest ML algorithms used in classifying observations based on the closest training points in the feature space. An instance of observation is assigned to a class most common among its k nearest neighbours by a majority of votes of its neighbours, where $k > 0$. Euclidean distance is used in most cases as a metric in finding nearest neighbours. In the proposed HAL model, a value of $k = 5$ nearest neighbour is used in the configuration. Random Forest classifier consist of an ensemble of decision trees where each decision tree is trained from randomly selected samples of an original training set. In this work, RF is used with 10 decision trees. The configuration used is similar to [89] implementation of RF.

4.4 Experiments and Evaluation

To evaluate the performance of the HAL system, data collected from our experimental setup is used. This is used in order to verify the HAL system via a limited test performed before it is tested on public datasets. Afterwards, the system is also evaluated using publicly available benchmark human activity dataset, Cornell Activity Dataset (CAD-60) [112]. The following sections describe the experiments conducted in this work and discussion of the results obtained.

4.4.1 Experimental Setup

Skeletal data is collected from three actors using a Microsoft Kinect V2 RGB-D sensor as mentioned previously in Section 4.2.1. The data is obtained at a frame rate of 30 frames per second (fps). Four activities are carried out namely; *Brushing teeth*, *Pick up object* (from the ground), *Sit on sofa* and *Stand up*. Each actor performs a single activity for a duration of 45 - 90 seconds. *Sitting on sofa* activity is performed by an actor going through a sequence of sitting and getting up poses with more time spent in the sitting pose. While the *Stand up* activity is performed in a similar way with more time spent staying standing. The summary of the data collected is presented in Table 4.1.

The data acquired is pre-processed following the process earlier mentioned in

Table 4.1: Summary of experimental human activity data collected from 3 actors using Microsoft kinect V2 RGB-D sensor. Activities performed comprise: Brushing teeth, Pick up object, Sit on sofa, Stand up.

Activity	**Number of frames**		
	Actor 1	**Actor 2**	**Actor 3**
Brushing teeth	2202	1876	1781
Pick up object	1804	1663	1355
Sit on sofa	1489	1672	2736
Stand up	2126	2059	2100
Total	**7621**	**7270**	**7972**

Table 4.2: Activity features computed from raw RGB-D sensor information of skeleton with 15 joints used in this work.

Feature description	**Feature label**
Spatial displacement δ between both hands, hands and head, hands and feet, shoulders and feet, hip and feet.	$1-9$
Temporal joint coordinate displacement t_{cp}	$10-54$
Temporal joint coordinate displacement t_{ci}	$55-99$
Joint coordinate-mean difference $j_{(i,mean)}$	$100-144$
Joint coordinate-variance difference $j_{(i,var)}$	$145-189$
Joint coordinate-standard deviation difference $j_{(i,std)}$	$190-234$
Joint coordinate-skewness difference $j_{(i,skw)}$	$235-279$
Joint coordinate-kurtosis difference $j_{(i,kur)}$	$278-324$
Total number of computed features	324

Section 4.2.1. Key features representing activities are extracted from the processed data. Table 4.2 shows the number of activity features computed from the RGB-D sensor skeleton with 15 joints. The number of joints used in computing spatial displacement features are selected based on the importance of the joints while carrying out the selected activities. Nine features are computed which represent the Euclidean distance between both left and right hands, each hand and head, each hand and its corresponding foot, each shoulder and corresponding foot, each hip and corresponding foot. The other features are obtained for each joint coordinate- given that 15 joints are used, each feature description comprises $15 \times 3 = 45$ features extracted.

Features selected from the experimental dataset are fed into the learning model to test the performance of the system. A *K-fold* cross-validation test strategy is applied with $K = 4$. This involves splitting the data into 4-folds in which 3-folds are used as training data for the model and the remaining fold is left out for validation. This process is repeated using each fold for validation and the final result is the average performance of all test validation folds.

4.4.2 CAD-60 Dataset and Experiment

The CAD-60 dataset comprises RGB-D sequence of human activities acquired using an RGB-D sensor at a frame rate of 15 fps. The dataset contains RGB image, depth image and skeleton joint coordinates information of 15 skeletal joints of activities carried out. However, the proposed HAL system utilises only the skeleton joint coordinates information. Four different actors perform 12 activities in five different locations namely; bathroom, bedroom, kitchen, living room and office. The activities performed are; Rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard, working on computer and a random + still activity. The random + still contains random movements sequence and a still pose performed by each actor. The stages described in the proposed HAL system are applied, with the CAD-60 dataset as raw input to the system. The same number of features as shown in Table 4.2 are computed from the dataset.

Table 4.3: Performance of the proposed HAL system on experimental dataset comprising four activities: Brushing teeth, Pick up object, Sit on sofa, Stand up.

Activity	**Performance result**	
	Precision (%)	**Recall (%)**
Brushing teeth	40.38	62.19
Pick up object	100	94.69
Sit on sofa	100	100
Stand up	54.10	35.13
Average	**70.65**	**68.43**

Learning the activities is done as a grouping of activities in the various locations. The grouping shown in Table 4.4 follows the format used by all approaches reported in the state-of-the-art in [27]. For testing the trained model, a method of *leave-one-out* cross-validation is carried out in which the model is trained on three actors and tested on the *unseen* actor. This is also called a *new person* test strategy.

4.4.3 Evaluation and Discussion

The proposed HAL system is evaluated on both datasets mentioned in Sections 4.4.1 and 4.4.2 following the test methods described. The CAD-60 dataset tests are performed following similar test methods described by [112] and other approaches by the state-of-the-art in [27]. Test results and discussions are presented in the following sections.

4.4.3.1 Experimental Dataset Results and Evaluation

Table 4.3 shows the results obtained from the performance of the HAL system on the experimental dataset. These are presented in terms of *Precision* and *Recall*. The system achieves an overall average precision of 70.65% and recall of 68.43% with the dataset. In Figure 4.7, the confusion matrix shows the percentage of correctly classified activities along with the percentage of false classified activities. It can be noted that the performance in activities of Pick up object with recall of 94.69% and Sit on sofa with recall of 100% are quite impressive. However, the model did not perform as impressively in correctly classifying brushing teeth and

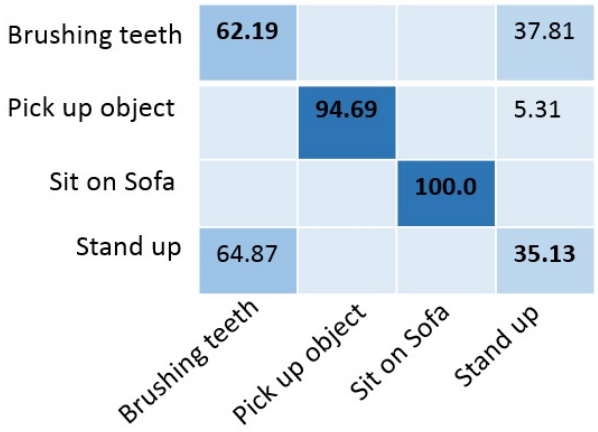


Figure 4.7: Confusion matrix of the proposed HAL system on experimental data.

stand up activities activities. This is due to the fact that the both activities have closely related poses as brushing teeth is performed while in a stand up pose. This gives rise to more stand up data - i.e. 64.87% - characterised as brushing teeth which affects the overall performance achieved. In order to adequately test the robustness of a supervised learning system, the availability of more data samples is required for proper training and validation of learning models. However, the experimental dataset collected contains fewer data samples when compared with other human activity datasets such as the CAD-60 dataset. This can also be a reason for the performance achieved on the experimental dataset. Therefore, the HAL system is also tested with the CAD-60 dataset which contains more samples of human activity.

4.4.3.2 CAD-60 Dataset Results and Evaluation

The results obtained from the performance of the proposed HAL system on the dataset are shown in Table 4.4. This is presented in terms of *Precision* and *Recall* of the HAL system. The proposed system achieved an overall average performance of 92.32% precision and 89.66% recall with features selected using the Relief-F feature selection method described earlier and a performance 90.96% precision and 88.52% recall when all the features extracted are used. In Table 4.5, the result from different locations are shown. When compared with the results

in Table 4.4, the system achieved a better performance with selected features than with all the features as reported in Table 4.5. Table 4.6 shows the proposed system performance compared to the state-of-the-art performances on the same

Table 4.4: Performance of the HAL system with *selected features* on the CAD-60 dataset using a "new person" test in different locations: Bathroom, Bedroom, Kitchen, Living room and Office.

Location	**Activity**	**Proposed HAL system**	
		Prec. (%)	**Rec. (%)**
Bathroom	Rinsing mouth	100	99.97
	Brushing teeth	96.97	75.16
	Wearing contact lens	54.48	92.68
	Random + still	99.98	100
	Average	**95.72**	**93.41**
Bedroom	Talking on phone	98.58	74.55
	Drinking water	91.47	60.99
	Opening pill container	15.39	66.55
	Random + still	100	100
	Average	**94.37**	**84.01**
Kitchen	Drinking water	92.96	74.81
	Cooking (chopping)	31.04	66.67
	Cooking (stirring)	78.43	77.52
	Opening pill container	74.49	75.49
	Random + still	100	100
	Average	**86.85**	**84.76**
Living room	Talking on phone	82.36	88.29
	Drinking water	86.93	74.14
	Talking on couch	94.27	100
	Relaxing on couch	100	100
	Random + still	100	100
	Average	**94.37**	**94.41**
Office	Talking on phone	67.06	93.42
	Writing on board	87.36	73.19
	Drinking water	100	83.84
	Working on computer	100	100
	Random + still	100	100
	Average	**93.28**	**91.71**
Overall average		**92.32**	**89.66**

Table 4.5: Performance of the HAL system with all features extracted from the CAD-60 dataset using a "new person" test. This shows the average performance from different locations

Location	**Performance result**	
	Precision (%)	**Recall (%)**
Bathroom	91.36	90.37
Bedroom	86.72	83.43
Kitchen	86.38	83.54
Living room	95.95	94.36
Office	94.41	90.92
Overall average	**90.96**	**88.52**

Table 4.6: Overall average precision and recall of the HAL system with the state-of-the-art on the CAD-60 dataset in a "new person" setting as reported in [27]. The extended modality column indicates the mode of RGB-D information used by different works i.e. Skeletal joint coordinates only (-) or skeletal joint coordinates with a combination of either RGB image and depth image modes (✓).

Method	**Prec. (%)**	**Rec. (%)**	**Extended modality**
Sung et al. [112, 113]	67.9	55.5	✓
Piyathilaka and Kodagoda [96]	70.0	78.0	-
Yang and Tian [135]	71.9	66.6	✓
Ni et al. [88]	75.9	69.5	✓
Gaglio et al. [41]	77.3	76.7	-
Gupta et al. [46]	78.1	75.4	✓
Koppula et al. [68]	80.8	71.4	✓
Nunes et al. [89]	81.83	80.02	-
Zhang and Tian [139]	86.0	84.0	✓
HAL system (with all features)	**90.96**	**88.52**	-
Faria et al. [38]	91.1	91.9	-
Parisi et al. [94]	91.9	90.2	-
HAL system (with selected features)	**92.32**	**89.66**	-
Zhu et al. [146]	93.2	84.6	✓
Shan and Akella [102]	93.8	94.5	-
Cippitelli et al. [24]	93.9	93.5	-

dataset [27]. The table also shows information of the state-of-the-art works which employ extended modality of RGB-D sensor information which is a combination of skeletal joint coordinates information with either of RGB image and depth image sensor information modes. The proposed HAL system's performance indicates the features extracted in our system sufficiently discriminate the human activities from skeletal joints information.

Comparison of the proposed HAL system's performance with the state-of-the-art based on the CAD-60 dataset is presented in Figure 4.8. The results show that the proposed system is able to attain an impressive performance. While some other proposed systems performance outperforms the HAL systems performance, the proposed HAL system differs from the other better performances in the following ways. The system proposed by [146] reported a performance of 93.2% precision and 84.6% recall. Although their precision exceeds that of the proposed HAL system, our system performs better in terms of recall. Also, the system by [146] uses a fusion of spatio-temporal interest point features obtained from combination of RGB-D sensor information modalities, i.e. depth image, RGB image and skeleton information as indicated in Table 4.6. This process can increase computational cost. The proposed HAL system utilises only the skeleton information offered by the RGB-D sensor to achieve such high performance. This shows that by adding more information for computer vision processing our system has the potential to achieve a higher performance.

Based on the results presented in Table 4.6, the performance attained by [102] slightly out performs our proposed HAL system. This approach performed tests excluding the random + still activity performed by all actors in the dataset which is included in the tests performed using the proposed HAL system. This information is relevant in generalising the robustness of the system across varying human activities.

The proposed system by [24] on the CAD-60 dataset attained a higher performance of both precision and recall of 93.9% and 93.5% respectively. Their system is tested with the dataset in a similar way observed in the system by [102] which excludes test on the random + still activity. Another reason could also be due to the fact that the proposed HAL uses all 15 skeleton joints of the

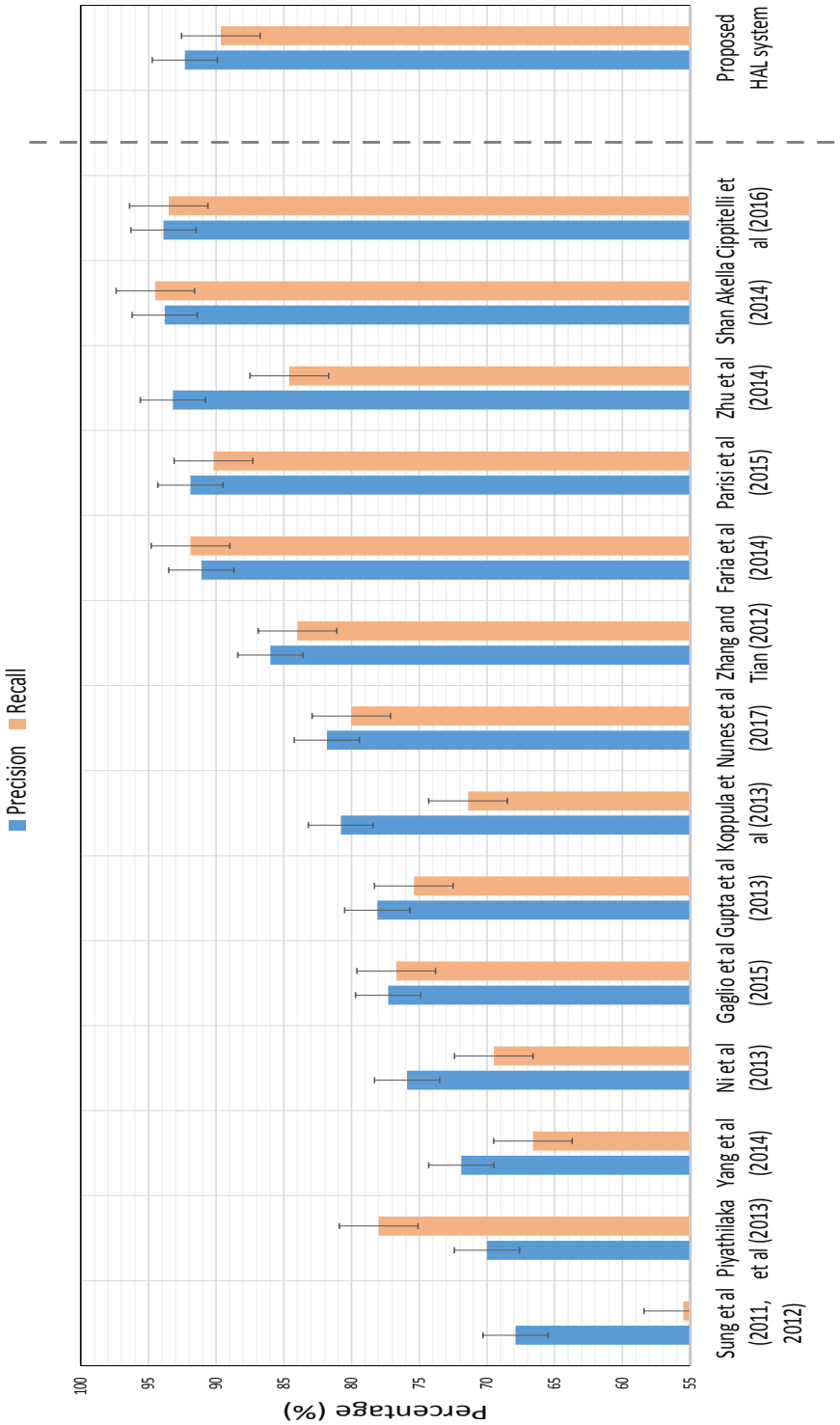


Figure 4.8: Precision and Recall Performance comparison of the HAL system with the state-of-the-art results on the CAD-60 dataset.

CAD-60 dataset whereas [24] used 11 selected skeleton joints to achieve the high performance. The selected joints do not include relevant joints such as the shoulders which are needed for our proposed application in assistive robots effectively executing human activities via TL. However, the HAL system with 15 skeletal joints achieves higher performance when compared with the 87.9% precision and 86.7% recall of [24] using the same number of joints.

With the performance achieved using the proposed HAL system with both experimental and publicly tested CAD-60 datasets, this shows the systems potential in applications of assistive robots learning of human activities.

4.4.3.3 Comparison of Classifier Ensemble with Single Classifier Performance

The method of using a classifier ensemble as proposed in this work shows the increase in activity learning accuracy when compared with other proposed methods using single classifiers. Table 4.7 shows the performance of the proposed classifier ensemble method with other methods which apply single classifiers in learning human activities. Also, it can be noticed that majority of the other approaches apply SVM in recognising human activities which is also used in the proposed classifier ensemble method and results show the classifier ensemble outperforms the other single classifier methods. In addition, the

Table 4.7: Proposed classifier ensemble method performance comparison with single classifier performance on CAD-60 dataset

Proposed by	**Method**	**Prec. (%)**	**Rec. (%)**
Yang and Tian [135]	Naive Bayes Nearest Neighbour	71.9	66.6
Ni et al. [88]	Latent SVM	75.9	69.5
Gaglio et al. [41]	SVM	77.3	76.7
Koppula et al. [68]	Structural SVM	80.8	71.4
Nunes et al. [89]	RF	81.83	80.02
Zhang and Tian [139]	SVM	86.0	84.0
Parisi et al. [94]	Neural Network	91.9	90.2
HAL system	**Classifier Ensemble**	**92.32**	**89.66**

classifier ensemble approach proposed also has the benefit of attaining higher accuracies with a small amount of training samples. This has an advantage over other widely used methods such as deep learning neural networks [56] which require a lot of data and more time in training such networks for concise predictions.

4.5 Discussion

The work presented in this chapter proposes a novel system for human activity learning with the use of skeletal data obtained using an RGB-D sensor. The work has shown explicitly the process of refining the raw sensor data obtained, computing relevant features and training the learning model. The main objective of this work is to have an activity learning system which is able to distinctly recognise activities as they are performed. The system can then be incorporated in an assistive robot to aid learning to perform the activities. The performance attained by the proposed system on the CAD-60 benchmark dataset shows its reliability if used with an assistive robot.

Although a selection of three base classifiers are used in building the ensemble model, this could be extended to include more classifiers which may improve performance and also deep learning neural networks which are increasingly used in human activity recognition systems. The system could also be extended to learning activities on-the-fly as they are carried out by an actor. The direction of research following this chapter is to segment different aspects of each learned activity into representations that any assistive robot platform can adopt in reliably executing human activity. This is presented in the following chapter.

Chapter 5

Adaptive Segmentation and Sequence Learning of Human Activities

5.1 Introduction

There are two main categories of learning algorithms suitable for human activity learning: *Batch learning* and *Sequence learning*. Classical batch learning algorithms predict output for new data when a complete training set of data is used. In this case, the new data samples are presented simultaneously when desired. However, a complete training dataset is often not available in advance for most practical applications. In applications such as human activity prediction [72], healthcare monitoring [91] and industrial functions [114] in which temporal changes within a task are being observed, the classical batch learning algorithms are rather infeasible for learning. Sequence learning is executed in a series of occurrences of samples within a given training dataset. Samples are used in the algorithm one after another and discarded after learning. This implies that the computational time and memory required for learning is reduced, and the learning process can accommodate temporal changes associated with tasks [114]. In most cases of humans executing tasks, the path of actions may vary, however, each path contains approximately a

similar order of true segments. To effectively learn such sequences of tasks, there are two key challenges which are often encountered. Firstly, the *segmentation* of tasks wherein given the observed task path, the start and end positions of constituent actions through the path are identified. Secondly, the *sequential learning* of essential underlying actions [76]. The task segmentation is critical in sequence learning for modelling and interpreting tasks information as it facilitates the adaptation of learning sequences in unseen situations [69].

The remainder of this chapter is organised as follows: Section 5.2 presents an overview of segmentation and sequence learning of activities. Section 5.3 describes the research methodology explaining an overview of the proposed framework. In Section 5.4, the method proposed in this work for unsupervised human activity segmentation is presented and Section 5.5 follows with a description of the sequence learning method used in learning the activity segments constructed. Section 5.6 describes the application of the proposed model to human activity datasets and the results obtained. In Section 5.7, the performance of the proposed ASSL is compared with other sequence learning approach and conclusions of the work are drawn in Section 5.8.

5.2 Overview of Segmentation and Sequential Modelling of Activities

There is a growing interest in research related to learning human activity sequences. This section presents a review of relevant works in two categories; the segmentation of human activities for detecting constituent actions, and activity modelling through sequential learning/prediction.

5.2.1 Action Detection and Segmentation

Most of the proposed activity recognition models [98] can attain impressive performances in their respective areas of application. The majority focus on supervised approaches to activity recognition in which there is a sufficient amount of labelled data available to build training models. However, in real-world situations where obtaining labels for activities is a rather daunting

task, supervised methods for activity recognition may not be feasible [104]. On the other hand, unsupervised learning methods, like clustering [25] are best suited for such applications.

An aspect of activity recognition which tends to be a challenge for many systems is detecting underlying/constituents actions in activities. This information is important in determining the structure of activities which is important when considering trends or sequences in such activities [72]. Therefore, segmentation is performed on data to obtain partitions which represent certain characteristics in activities. This is a vital step in investigating activity sequences. Existing approaches to segmentation of human activity differ in terms of the following categories [3, 4]; the activity types that are modelled, the sensing technology used to acquire information and the Computational Intelligence (CI) methods used in the segmentation process.

With a focus on Human Activity Recognition (HAR) from 3D human skeleton joints information, i.e. the joint positions or angles, different methods have been proposed for detecting actions in an activity. The authors in [72] proposed a method for detecting atomic actions which they call *actionlets* using motion velocity. The method combined the Harris corner detector and Lucas Kanade (LK) optical flow to get velocity magnitudes. Some works using the kinetic energy poses to determine key poses in activities are found in [89, 102]. These methods then apply different Machine Learning (ML) algorithms for classification of actions obtained for activity recognition.

5.2.2 Sequential Modelling of Activities

Sequence learning algorithms are used for the analysis of patterns generated through a series of observed information for recognition or classification of activities [145]. Researchers have studied sequence learning over many decades. This led to the development of statistical models such as Hidden Markov Models (HMM) [40, 99] and Autoregressive Integrated Moving Average (ARIMA) [34] which were introduced for time series and temporal pattern recognition problems [28]. Recurrent Neural Networks (RNNs) have since evolved to solve sequence prediction problems due to their recurrent lateral

structure. Long-Short Term Memory (LSTM), a type of RNN, have a unique ability to selectively pass information across time and are able to model significantly long-term dependencies due to the gating mechanism they possess [53]. LSTMs also can deal with the vanishing gradient problem. This has seen impressive performances in a variety of real-world applications.

Concerning human activities, attempts to model human activity sequences have been studied by various researchers [85, 131] using different temporal models for HAR. Discriminative models for example, Conditional Random Fields (CRF) are employed in modelling human actions. The CRF is used in [50] to estimate motion patterns that correspond to manifold subspace of 3D joint position features for human action recognition. Generative models are also used for modelling human actions. HMM is used over predefined motion features of 3D joint positions to learn the dynamics of human actions [81]. Similar approaches employing generative models to model activities are also proposed in [90, 102]. The 3D joint positions obtained through skeleton tracking tend to be noisy. Therefore, when the change between actions is small, without the accurate selection of features, recognising precise action states becomes difficult. This tends to undermine the performance of generative models. Such models require an adequate amount of data for training as they are prone to over-fitting. Dynamic Time Warping (DTW) [23] is another solution used in modelling actions by defining the distance between two temporal sequences of activity actions. The learning can then be achieved through nearest-neighbour classification. However, the performance of DTW is dependent on a good measure of the samples similarity. It could also suffer from temporal misalignment when handling periodic actions which could lead to degrading its performance [73].

These works demonstrate the effectiveness of segmentation and sequence modelling in exploring the underlying patterns in sequential data. Following from the identification of key actions, the non-parametric segmentation of 3D skeletal data of human activities obtained. This is then used in an LSTM model for the prediction of activity actions. In the following section, the problem statement is described and key definitions used in this work are presented.

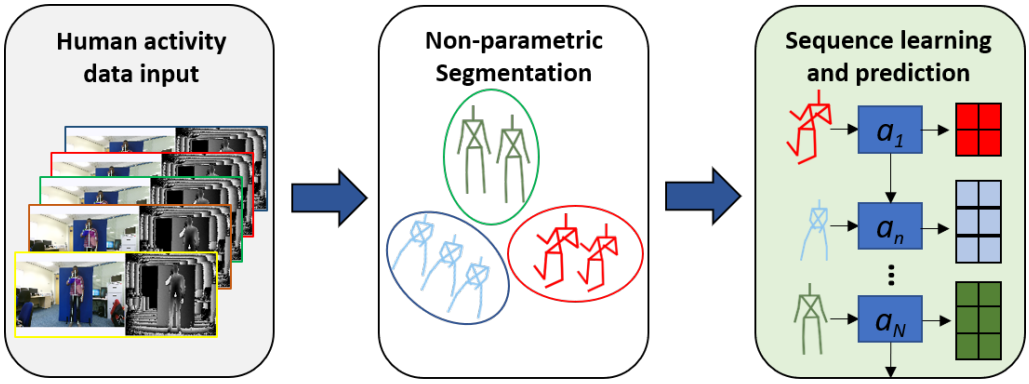


Figure 5.1: Overview of the proposed approach to the Adaptive Segmentation and Sequence Learning (ASSL) of human activity.

5.3 Methodology

To address the challenges of segmentation and sequence learning of human activities, a novel framework for Adaptive Segmentation and Sequence Learning (ASSL) is proposed using visual information of activities. An overview of the ASSL framework is depicted in Figure 5.1. There are three distinct steps in the proposed ASSL framework as described below:

1. Initially, key actions from observed human activity information are obtained. Human activities contain a large number of actions for which only the key aspects are relevant. By exploiting the temporal accumulated motion energy of each action through the sequence, the key actions can be extracted from the points of change in acceleration and deceleration of activity motion.

2. While segments of activities can be inferred from manual annotations, this creates a burden in *supervised* situations where high-dimensional data would require large amounts of annotations to obtain feasible segments which can be learned. A non-parametric technique for feature space analysis is applied for *unsupervised* segmentation of relevant activity actions.

3. From the segments obtained, a Recurrent Neural Network (RNN) method for sequence learning called Long Short-Term Memory (LSTM) is used to learn activity sequences.

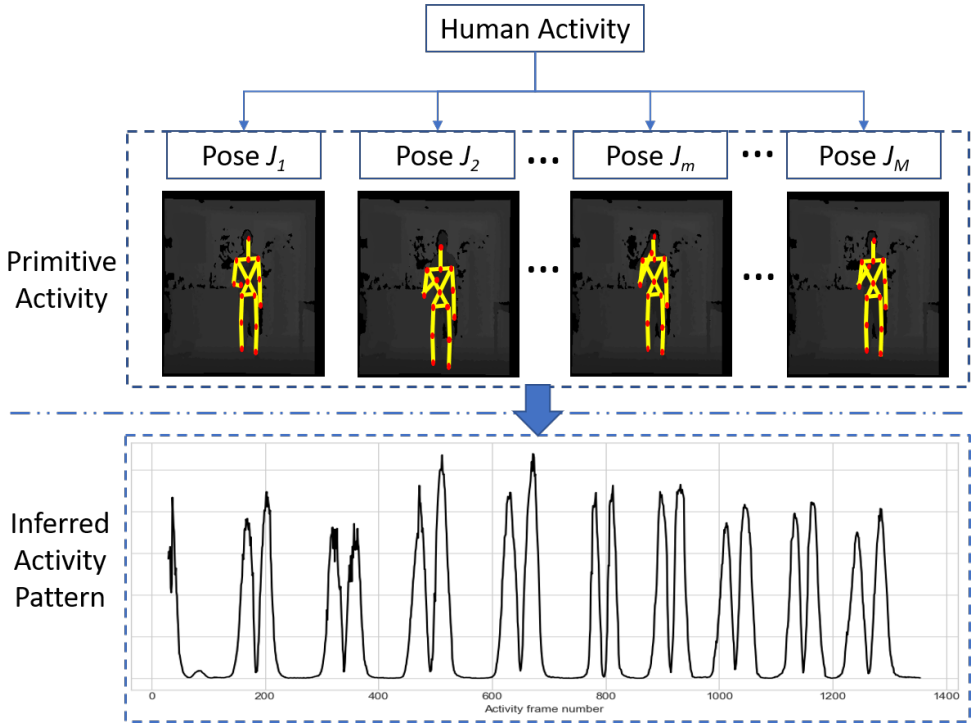


Figure 5.2: An illustration of learning underlying patterns of simple primitive human activity sequences from 3D temporal information.

Figure 5.2 illustrates the underlying concept of how human activity patterns can be inferred and learned from processing extracted visual 3D information. This work will benefit applications which require learning the underlying sequences in human actions through activities.

5.3.1 Definitions

Given a set of observed human activities $A = \{a_1, a_2, \ldots, a_n, \ldots, a_N\}$ performed by actors, the observations are obtained using an RGB-Depth (RGB-D) sensor. Each demonstration of an activity a_n within the observed activities set is a discrete time sequence of activity poses. An activity pose J_n as represented by:

$$J_n = [j_1, j_2, \ldots, j_m, \ldots, j_M], \quad \text{for } J \in \mathbb{R}^{3\times M}, \tag{5.1}$$

is a feature space which represents 3D human skeleton joints with coordinates. M

represents the total number of joints in J_n with each joint, j_m, with coordinates x_m, y_m, z_m corresponding to horizontal, vertical and depth positions respectively.

Definition 5.3.1. Key action, $\overline{J}$ is defined as the important atomic level action performed during an activity. Key actions extracted from an activity represent a subset of poses $\overline{J} \subset a_n$, for $n = 1, 2, ..., N$, which occurs in varying time instants of an executed activity.

Definition 5.3.2. Activity segmentation is defined by a function C in which each key action, $\overline{J}_b$, $b = 1, 2, ..., B$, of an activity a_n is assigned a value, Q_z, $z = 1, 2, ..., Z$, corresponding to a unique activity segment represented as:

$$C : a_n \longmapsto (\overline{J}_b)_{1,2,...,B}, \quad \text{for } \overline{J}_b \in Q_z \tag{5.2}$$

where b is the index of the key action through the activity sequence and B is the number of key actions contained in a_n. Each segment derived comprises similar activity key actions through a temporal sequence.

Definition 5.3.3. Activity action sequence, S, is defined as the temporal ordering of all B key actions obtained from activity a_n. A repetition of similar key actions may be observed in the sequence at points where a_n contains actions which are repeated at different temporal instances. A representation of this definition is presented as:

$$S = (\overline{J}_b)_{b=1}^{B} \tag{5.3}$$

5.3.2 Assumptions

For the research presented in this chapter, certain assumptions are made. They are:

- The observed sequence of a human activity comprises of unlabelled atomic actions which this work aims to identify through a process of adaptive segmentation.

- The number of key poses $\overline{J}_B$ that make up an activity is not given. This is drawn from the fact that each activity can be segmented into key poses

which make up for the relevant aspects that define the activity. However, this number is not pre-defined from activity observations in the proposed model.

5.3.3 Problem Statement 1

Given an observed sequence of a human activity obtained using an RGB-D sensor, the first phase is the segmentation of an unlabelled sequence into meaningful representations of similar actionlets. The segments obtained represent a collection of similar actions which may (or may not) fulfil temporal order relationship constraints.

The task of segmentation from an unlabelled activity sequence is addressed in this work using an adaptive approach to segmentation. The following steps are proposed for use in obtaining the function C for the segmentation of an activity.

Detection of key actions (or poses): Key actions of an activity are relevant in the process of learning an activity sequence. This is mainly because an activity can be executed in different forms whilst certain key aspects through the observation of an activity can uniquely identify it. As mentioned in the Introduction section, the motion energy feature of actions through an activity can be used in obtaining these key actions. The key actions are therefore identified by applying a filtering method of moving average crossovers of the motion energy. The description of how this is implemented is presented in the next Section.

Non-parametric feature space clustering: The key actions obtained from the filtering process of the motion energy feature are clustered using a Mean-Shift feature space analysis method. This method performs the clustering in terms of similarity of the motion energy of key actions.

5.3.4 Problem Statement 2

To learn the sequence S of transition of actions from one activity segment to another, it is important to note that an activity is not executed in only one possible sequence. An activity can be executed with different temporal orders of

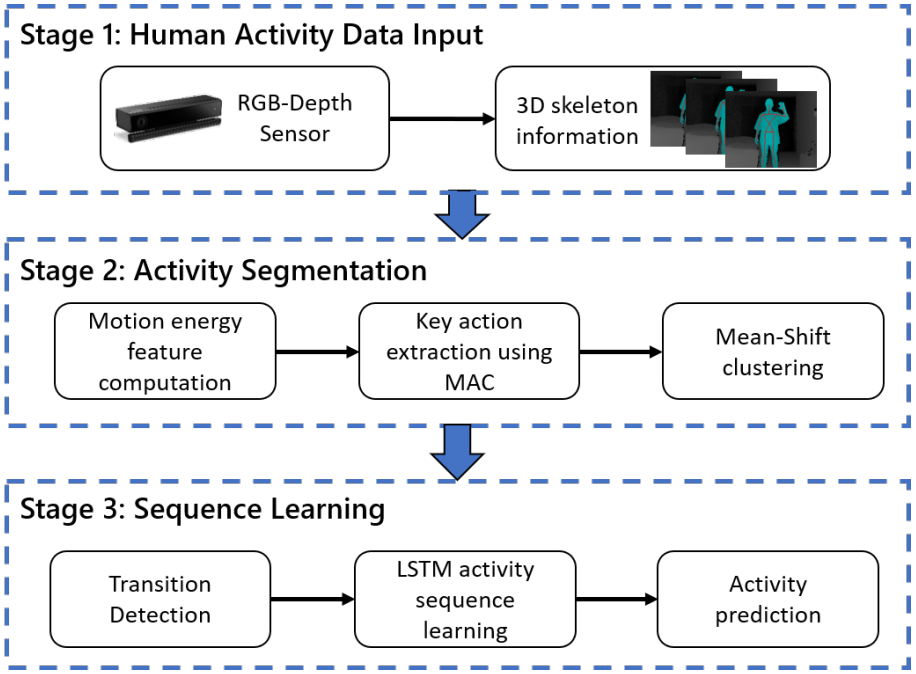


Figure 5.3: Architecture of the proposed ASSL approach for human activities from 3D skeleton information which comprises activity input, segmentation and sequence learning stages respectively.

constituent actions. This results in a challenge of learning a generalised sequence for an activity.

The sequence of actions from one segment to another occur in intervals. The LSTM-RNN algorithm, which is predominantly used in predicting time series, is applied in learning the sequence of distinct actions within the activity segments. This method is used as the algorithm is able to capture infinitely long sequences and predict succeeding occurrences based on the memory gates.

The architecture of the ASSL approach for human activities from 3D skeleton information as proposed in this thesis is depicted in Figure 5.3. This comprises three stages of activity data input from an RGB-D sensor, segmentation of human activity and sequential learning and prediction. Details of these stages are provided in the proceeding sections.

5.4 Activity Segmentation

Segmentation of human activity is relevant in the analysis of trends in transitions from one activity state to another. This section describes the process of activity segmentation using the extracted human activity information.

5.4.1 Key Action Point Detection with Motion Energy

Human activity consists of movement sequences generated by different body parts. It is worth noting that not all aspects of an activity movement sequence are necessary to define an activity. Certain aspects of the sequence can be executed in different forms and still result in a similar activity. To simplify an activity to the relevant action points that constitute the sequence, key poses are selected. This is achieved by leveraging the motion energy obtained from activity sequences.

5.4.1.1 Extraction of Motion Energy

The motion energy of activity poses as first proposed by [102] is based on the fact that joints show changes in acceleration and deceleration through an activity. This information is significant when considering the identification of the key action points of activities. Following from the representation of an activity pose given in Equation 5.1, the motion energy E_l for each pose is computed as the sum of motion energies for each joint in the pose;

$$E_l(J_n) = \sum_{m=1}^{M} E_l(j_m) \tag{5.4}$$

where j_m is a joint in the pose. It is assumed that the mass of all joints to be equally one unit due to the fact that it is impossible to obtain the actual mass of a joint from the information obtained using RGB-D sensors. Computing the joint velocities using the temporal change ΔT in the position d of joints during

an activity, the motion energy can be expressed as:

$$E_l(J_n) = \frac{1}{2}\sum_{m=1}^{M}(v_{j_m})^2 \tag{5.5}$$

where, v_{j_m} represents the velocity of joint j_m and is expressed as $v_{j_m} = \frac{d_m^c - d_m^p}{\Delta T}$, d_m^c is the current joint position and d_m^p is the previous joint position. By substituting v_{j_m} in Equation 5.5, the motion energy of each joint is computed using the following equation:

$$E_l(J_n) = \frac{1}{2}\sum_{m=1}^{M}\left(\frac{d_m^c - d_m^p}{\Delta T}\right)^2 \tag{5.6}$$

5.4.1.2 Moving Average Crossover of Motion Energy

The Moving Average (MA) is a filtering technique often applied to get overall trends in data. This technique is used to highlight long-term cycles in time series data by smoothing out short-term variations [31]. It works by creating series of averages of different time windows from a dataset over a given distribution.

Most of the works employing motion energy for identifying key action points of activities set threshold values of energy from a random exploration of selected points in order to extract the relevant points of interest in an activity [89, 102, 144]. The energy thresholds are selected by repeated experiments of different threshold values and the observations below the threshold value are selected as key poses. The MA of the extracted motion energy of poses are used in filtering the motion energy signal extracted from an activity sequence.

A different approach is proposed to use crossovers of two Simple Moving Averages (SMA) of the extracted motion energy in identifying the relevant key poses of an activity. The SMA is an un-weighted mean of a set of data points. This is taken from equal sets of data to ensure variations in the mean and data points are aligned and not shifted in time. Given the motion energy obtained in Equation 5.4, the SMA for the motion energy signal of an activity can be

computed using the following expression:

$$SMA = \frac{\sum_{r=0}^{\alpha-1} E_l(J_n)_{t-r}}{\alpha} \tag{5.7}$$

where α is the value of the period selected for MA and $t-r$ is the position of the selected observation within α. This is expressed in a simplified form as follows;

$$SMA_{E_l} = \frac{E_l(J_n)_t + E_l(J_n)_{t-1} + ... + E_l(J_n)_{t-(\alpha-1)}}{\alpha} \tag{5.8}$$

Two moving averages are selected in this work - a short-term average (fast moving average) α_f and a long-term moving average (slow moving average) α_s. The MA crossovers are obtained from points where the SMAs for both α_f and α_s intersect. These points indicate significant changes in motion energy of activity poses and are used as reference points for their corresponding key actions in an activity sequence as presented in the following equation.

$$\overline{J_b} = SMA_{\alpha_s} \cap SMA_{\alpha_f} \tag{5.9}$$

Following the acquisition of the key action points, activity segments are obtained by application of a non-parametric feature space analysis technique - In this case, mean-shift clustering for associating key actions to clusters of similar actions.

5.4.2 Non-Parametric Clustering for Segmentation

Prior to learning the sequence of actions in an activity for prediction, it is necessary to know the segments that make up an activity. This information is not easily determined by mere observation of the key actions obtained from exploration of the motion energy feature. Also, the number of segments that can be defined for an activity can vary depending on the sequence observed. Therefore, the use of a non-parametric method of clustering key actions is proposed to determine the number of segments in an activity sequence and assign the obtained key actions to their respective segments before learning can be achieved. A mean-shift clustering approach is adopted here [25]. The mean-shift approach builds upon the concept of Kernel Density Estimation (KDE) [95] which estimates the hidden distribution for a dataset by placing a

Algorithm 2 Segmentation of human activity from joints coordinate skeleton information.

Input:

Instances of 3D skeleton joints coordinate of human activities $A = \{a_1, a_2, \ldots, a_n, \ldots, a_N\}$, in which each observation of activity a_n is a pose $J_n = [j_1, j_2, \ldots, j_m, \ldots, j_M]$;
Activity time window t;
Moving average periods α_s, α_f;

Output:

Activity segments obtained as a function C for assigning each key action to a segment;

Procedure:

1: **for** a_n, $n = 1$ to N **do**
2: Find the velocity of each observation J_n within t;
3: Compute the motion energy for J_n: $E_l(J_n) = \sum_{m=1}^{M} E_l(j_m)$;
4: Compute the simple moving average of the motion energy with the periods α_s, α_f: $SMA = \frac{\sum_{r=0}^{\alpha-1} E_l(J_n)_{t-r}}{\alpha}$;
5: Key action points, $\overline{J_b} = SMA_{\alpha_s} \cap SMA_{\alpha_f}$;
6: **end for**
7: Assign $\overline{J}_b$ to a cluster Q_z which is determined by a non-parametric mean-shift clustering technique;
8: **return** $Q_Z = C(\overline{J}_b)$.

kernel on each point contained in the dataset. The description of the mode of application for the proposed segmentation of human activity is provided below.

Given B key action points, $\overline{J_b}$, $b = 1, ..., B$, on a 2-dimensional space computed for an activity. As described in Section 5.4.1, these points correspond to the motion energies of key action positions. The kernel density estimate for the key action points with kernel K with a bandwidth parameter h is:

$$f(\overline{J}) = \frac{1}{Bh^2} \sum_{b=1}^{B} K\left(\frac{\overline{J} - \overline{J_b}}{h}\right) \tag{5.10}$$

with K satisfying the following two conditions:

1. $\int K(\overline{J}) d\overline{J} = 1$, and
2. $K(\overline{J}) = K(|\overline{J}|)$ for all values of $\overline{J}$.

The first condition is required to ensure normalisation of the density estimate while the second condition relates to the symmetry of the data space containing all key action points of an activity. By applying a Gaussian symmetric kernel function for $K(\overline{J})$, the gradient of the density estimator in Equation 5.10 takes the form:

$$\nabla f(\overline{J}) = \frac{2}{Bh^4}\left(\sum_{b=1}^{B} g\left(\left|\frac{\overline{J}-\overline{J_b}}{h}\right|\right)\right)\vec{X}(\overline{J}) \tag{5.11}$$

where $\vec{X}(\overline{J})$ is the mean-shift vector pointing in the direction of increasing density and is represented as:

$$\vec{X}(\overline{J}) = \left(\frac{\sum_{b=1}^{B} \overline{J_b} g\left(\left|\frac{\overline{J}-\overline{J_b}}{h}\right|\right)}{\sum_{b=1}^{B} g\left(\left|\frac{\overline{J}-\overline{J_b}}{h}\right|\right)} - \overline{J}\right) \tag{5.12}$$

and $g(|\overline{J}|)$ is the derivative of the Gaussian kernel.

With the KDE computed, the mean-shift procedure is carried out by successive:-

- Computation of the mean-shift vector $\vec{X}(\overline{J_b})$ at the location of each key action point $\overline{J_b}$,
- Translation of each action point $\overline{J_b} \rightarrow \overline{J_b} + \vec{X}(\overline{J_b})$,
- Repeat until convergence, that is, where the gradient density function is zero.

Afterwards, the key action points identified at the same points are segmented as belonging to the same cluster Q_z. For further details of convergence, readers are referred to [25]. Algorithm 2 list the procedure for activity segmentation proposed in this thesis.

5.5 Sequence Learning and Prediction Model

The sequence learning stage involves the learning of activity sequences from the segmented key actions. An LSTM network [53] is used to learn the long-term

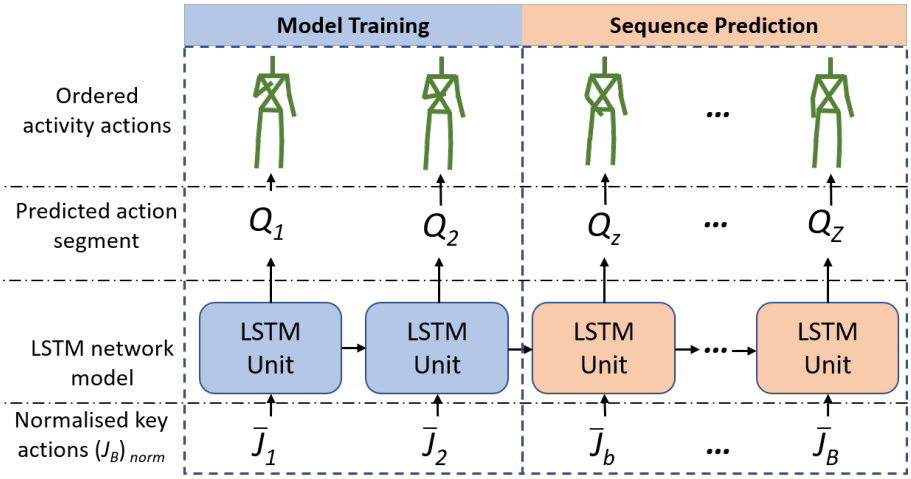


Figure 5.4: LSTM structure for sequential learning and prediction of key action segments of human activity.

contextual dependencies between key actions of an activity. The segmented key actions are used as input to the network for learning the dependencies between the action segments. This is further extended to predicting sequential actions of activities. Figure 5.4 illustrates the structure of an LSTM network as applied in this work. The LSTM comprises of the following components: input gate i_t, forget gate f_t, a cell with a self-recurrent connection and output gate o_t. The key action segments obtained for an activity are normalised for standardisation of the values, thus resulting in $Q_{norm} = \{\overline{J}_{1Q_1}, ..., \overline{J}_{BQ_Z}\}_{norm}$. By taking Q_{norm} as input to the network, the network is updated every t timestep by iterating through all instances of the normalised key actions using the following equations;

$$i_t = \sigma(W^i(\overline{J}_{bQ_z}(t)) + U^i H_{t-1} + V^i) \tag{5.13}$$

$$f_t = \sigma(W^f(\overline{J}_{bQ_z}(t)) + U^f H_{t-1} + V^f) \tag{5.14}$$

$$o_t = \sigma(W^o(\overline{J}_{bQ_z}(t)) + U^o H_{t-1} + V^o) \tag{5.15}$$

$$g_t = tanH(W^g(\overline{J}_{bQ_z}(t)) + U^g H_{t-1} + V^g) \tag{5.16}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{5.17}$$

$$H_t = o_t \odot tanH(c_t) \tag{5.18}$$

where, $\sigma(\cdot)$ and $tanH(\cdot)$ are the sigmoid and hyperbolic functions respectively. W, U, V are parameters of the LSTM model. The operation $\odot$ denotes the element-wise multiplication of two vectors. The use of LSTM is due to its ability to map input activity sequences by recursively transforming current inputs Q_{norm} with the output hidden vector of previous steps H_{t-1}. Also, the vanish gradient problem inherent with RNN's is overcomed by the memory cell c_t which is computed, allowing the error derivatives to flow in a different path.

5.6 Application of the ASSL Framework to 3D Skeleton Data of Daily Human Activity

This section reports the experimental procedure and results of applications of the proposed ASSL framework on 3D skeleton human activity datasets. To illustrate the application of the proposed work of ASSL of human activity sequences, the model proposed was applied to selected human activities. The proposed model is adaptive to different activities and thus gives it the ability to deal with complexities in activities.

To understand the methodology and its ability to solve the problems identified in the earlier Sections 5.3.3 and 5.3.4, the following hypotheses are proposed and evaluated.

Hypothesis 5.6.1. Where an unlabelled sequence of activity data is available, the segmentation technique proposed can be used to identify unique segments of an activity used for label assignments of actions in the sequence.

Hypothesis 5.6.2. Activity segments identified can be used to learn sequences for prediction with a reliable performance.

To address these hypotheses, two activities are selected from two real world human activity datasets; Dataset 1 - An experimental human activity dataset

collected for this work and Dataset 2 - A benchmark public dataset, Cornell Activity Dataset (CAD-60) [112].

5.6.1 Experimental Design and Datasets

The motivation for the proposed ASSL framework is to address the issue of unlabelled sequences of human activities, in such cases where there is no knowledge *a priori* of constituent actions and their order, whilst there is the need to develop a system for identifying the patterns of activities. The experimental design and datasets used in evaluating the proposed framework are presented in this section.

5.6.1.1 Dataset 1 - Experimental Human Activity Dataset

The dataset generated to evaluate the proposed system in this work consists an activity which involves a person picking up an object placed on a surface. A Microsoft Kinect version 2 RGB-D sensor [86] is used to acquire the 3D joint coordinate information of person. This information is obtained at 30 fps. This activity is chosen due to the proposed work being focused on enhancing the ability of assistive robots learning activity sequences for independent prediction of actions. Figure 5.5 shows sample frames of the selected activity carried out by a person.

To obtain adequate amount of data to evaluate the ASSL framework, the activity is performed by three people. Each person is required to pick up an object from a flat surface repeatedly eight to ten times while the joint positions are recorded throughout the sequence. Table 5.1 shows the number of frames acquired from each person while carrying out the activity.

Table 5.1: Experimental dataset acquired from three actors for an activity - pick up object from a flat surface.

Activity	**Number of frames**			**Total**
	Person 1	**Person 2**	**Person 3**	
Pick up object	1804	1663	1355	4822

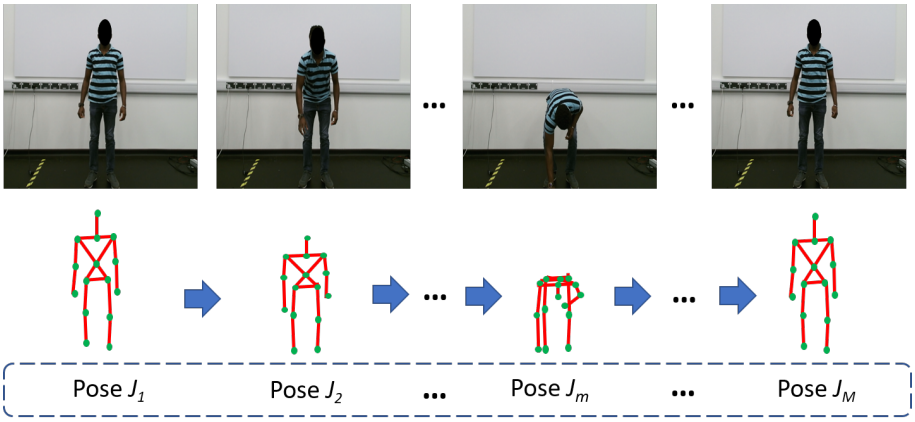


Figure 5.5: Sample frames of *pick up object* activity obtained from the experimental activity dataset using an RGB-D sensor.

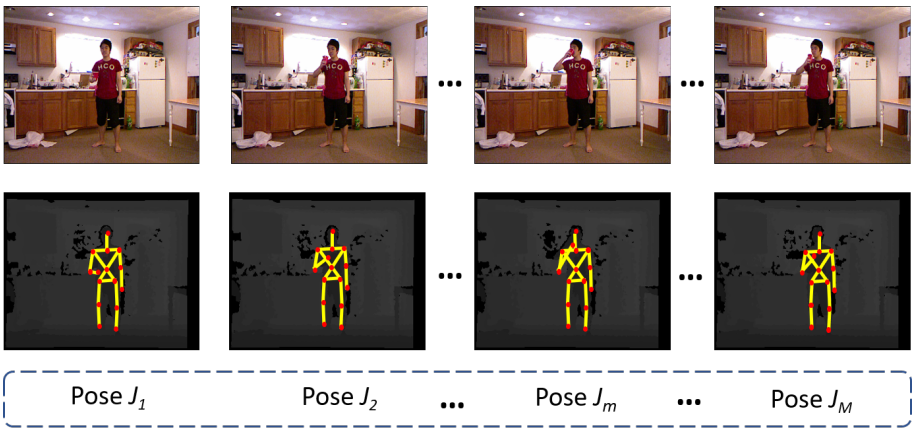


Figure 5.6: Sample frames of *drinking water* activity obtained using an RGB-D sensor contained in the CAD-60 dataset [112]. The sample shows RGB images and the corresponding depth image with the tracked skeleton overlaid.

5.6.1.2 Dataset 2 - Cornell Activity Dataset (CAD-60)

The CAD-60 dataset [112] is based on human activity data obtained using an RGB-D sensor. The dataset comprises three modes of human activities data, RGB images, Depth images and 3D skeleton joint coordinates observed from a person performing an activity. The skeleton joint data consists of joint coordinates information of 15 joints. The dataset is recorded at a frame rate of 15 fps using a Microsoft Kinect sensor and includes recordings for 12 human

activities namely; rinsing mouth, brushing teeth, wearing contact lens, talking on the phone, drinking water, opening pill container, cooking (chopping), cooking (stirring), talking on couch, relaxing on couch, writing on whiteboard, working on computer and a sequence of random plus stationary activities. The data is collected from four participants with each performing each activity.

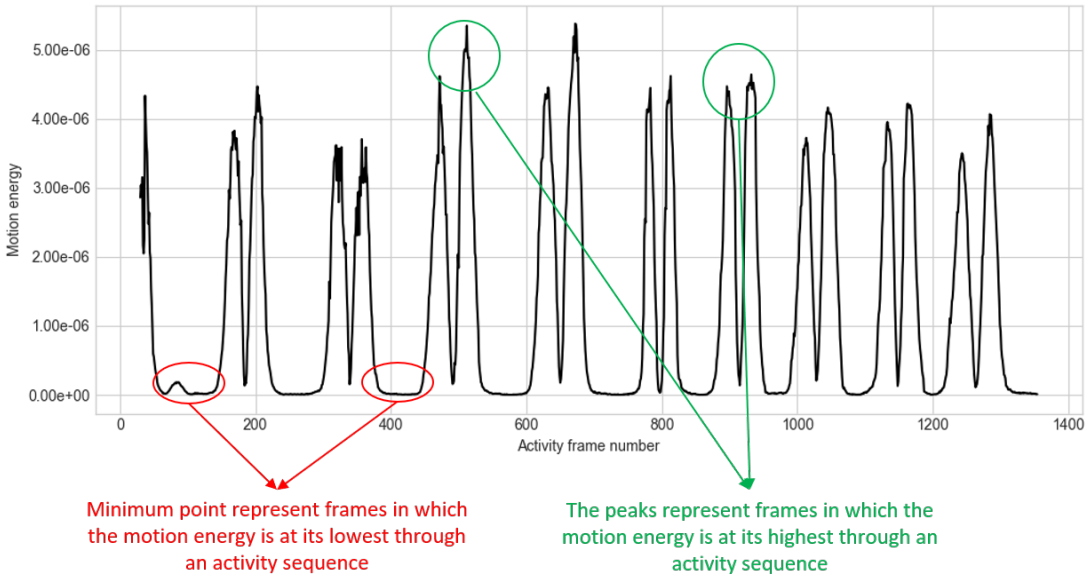
Most applications of this dataset are based on activity classification and therefore involve the use of all activities within the dataset. However, to demonstrate the work proposed in this thesis, a single activity from the dataset is selected and used in our evaluations. The activity chosen is the *drinking water* activity as there are more motions involved in the activity when compared to the remainder activities available in the dataset. This creates a scenario with varying motion patterns to test the robustness of the framework. Sample frames of varying actions occurring throughout the activity sequence are shown in Figure 5.6. The samples show a person drinking water with the tracked skeleton joints overlaid on the depth images. The activity is performed repeatedly $2-3$ times.

5.6.2 Experimental Human Activity Dataset Results and Evaluation

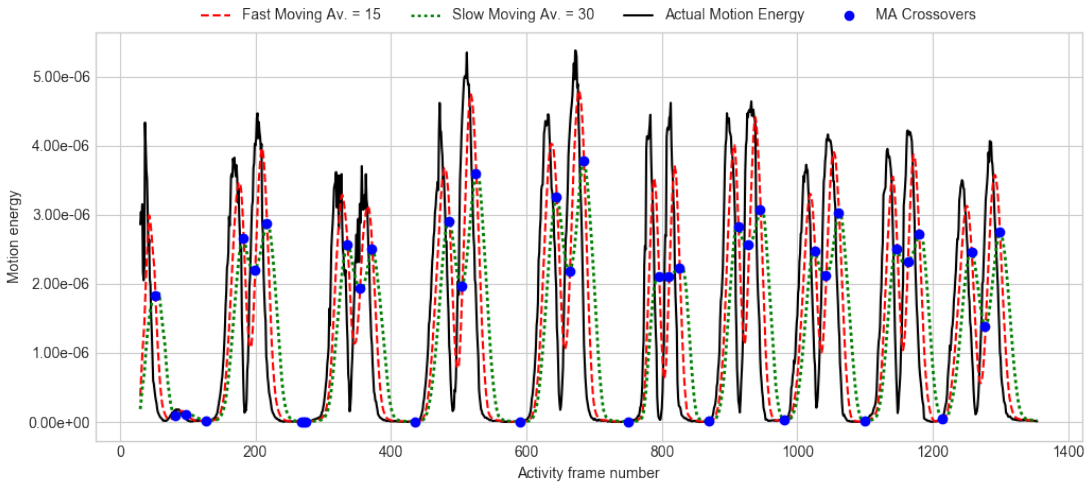
To evaluate the performance of the proposed framework on the experimental dataset, it is implemented in stages, starting with the segmentation process - the computation of motion energy, detection of key action points and the non-parametric clustering for key action segmentation. This is then followed by the sequence learning and prediction of the obtained key actions.

5.6.2.1 Key Action Identification using Motion Energy

Applying the approach to identifying key action points of an activity, the motion energy is computed for 3D joint positions data obtained from each person. A window size, w_s, of one second is used which corresponds to 30 frames of activity to compute the motion energy. Figure 5.7(a) shows the motion energy obtained from person 1 of the experimental dataset. The figure shows the changes in



(a)



(b)

Figure 5.7: Key action identification for *pick up object* activity from person 1 in the experimental dataset; (a) Motion energy plot for person 1 from the experimental dataset. The energy is computed using a 1 second window = 30 frames, (b) Motion energy plot with identified crossover points of two moving averages which represent the identified key action points of the activity. $SMA_{\alpha_f} = 15$ and $SMA_{\alpha_s} = 30$.

the cumulative motion energy which is a result of continuous acceleration and deceleration of body joints through the activity sequence.

In the proposed framework, the key actions are identified at points of minimum and maximum motion energies. Applying the simple moving average technique, after multiple experiments with different values of SMA_{α_s} and SMA_{α_f}, 30 and 15 frames are selected for both moving averages respectively. Figure 5.7(b) depicts the key action points identified from the motion energy computed in Figure 5.7(a). The green plot shows the SMA_{α_s} while the red plot shows the SMA_{α_f}. The crossover points of both moving averages are identified by the blue dots in Figure 5.7(b). These points represent the key actions $\overline{J_B}$ in the activity sequence from the data. Similarly, the key actions are obtained for all participants in the experimental dataset.

5.6.2.2 Non-parametric Clustering of Experimental Dataset

Due to the varying nature of the activities performed from one individual to another, there are variations in motion energy values from person to person. To tackle this difficulty, the motion energy of the key actions identified for each participant's activity are normalised for standardisation across all participants. Figure 5.8 shows the representation of normalised motion energies of identified key actions for all persons in the dataset. A total of 202 key action frames are identified from all three participants which shows a significant reduction when compared to the total number of frames 4822 as shown in Table 5.1. This emphasises the need for the segmentation process to reduce the computational complexities when learning the activity sequence.

The normalised values are then clustered using the non-parametric method described earlier. The results obtained from clustering is also represented in Figure 5.8. It can be observed that for the selected activity three segments corresponding to Q_1, Q_2 and Q_3, are identified and the boundaries of the segments as obtained from the results are represented by the horizontal line plots (green and orange) shown on the figure. Figure 5.9 shows the distribution of the number of key action points identified in each activity segment for all participants.

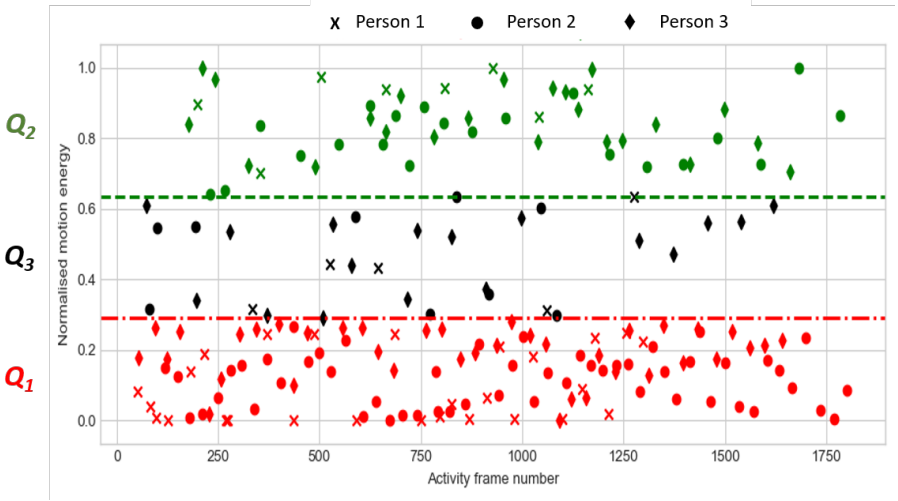


Figure 5.8: Normalised motion energy with action segment identification of key actions for all participants in the experimental human activity dataset corresponding to the *pick up object activity*.

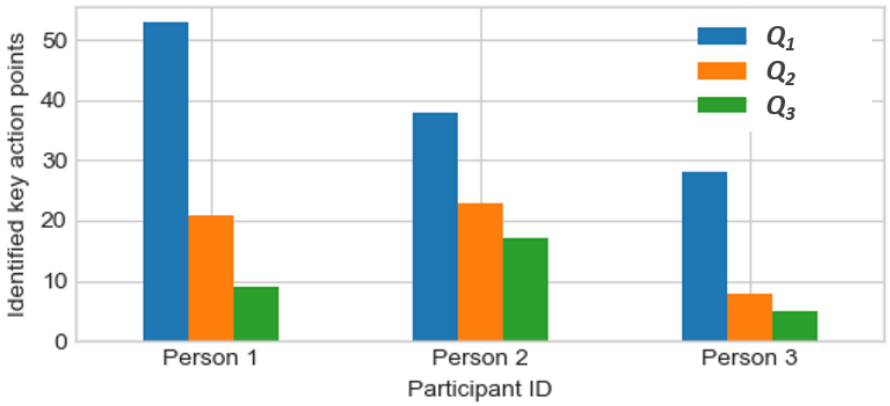
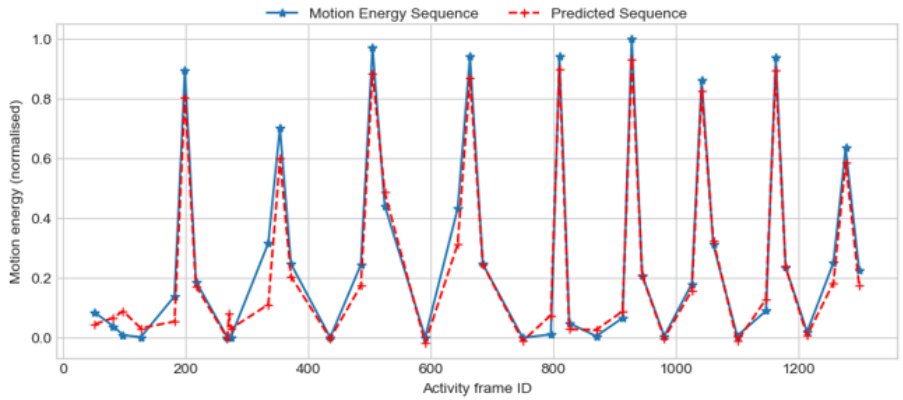


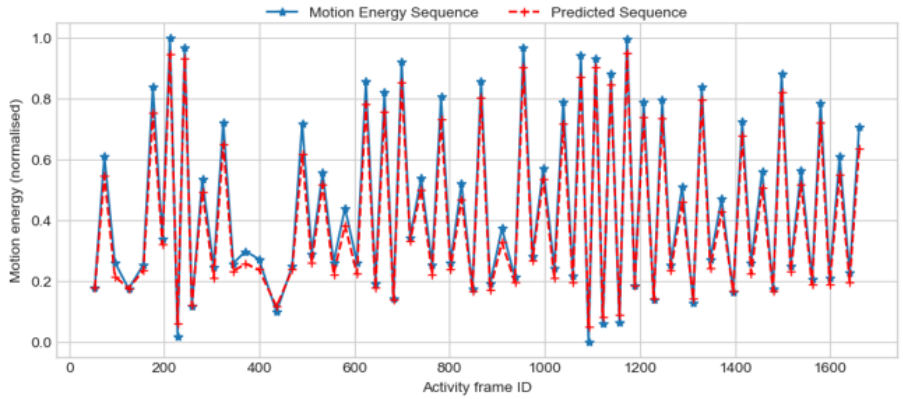
Figure 5.9: Activity segmentation distribution for participants in the experimental human activity dataset.

5.6.2.3 Sequence Learning of Experimental Human Activity Dataset

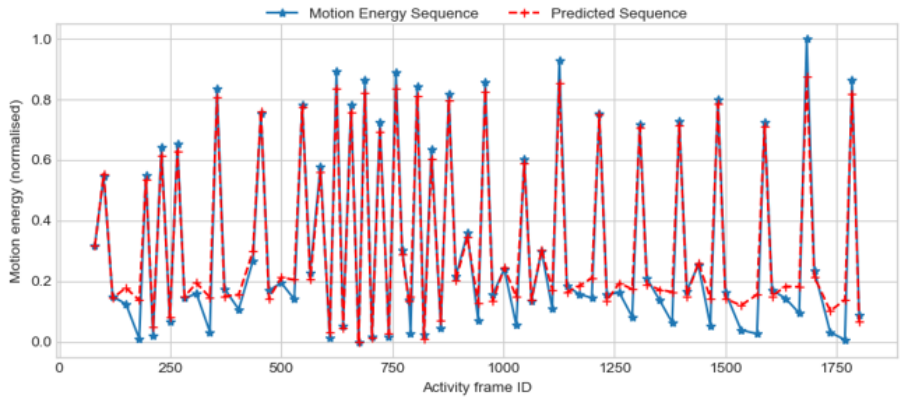
The sequence learning model is grounded on the results obtained from the activity segmentation process. To investigate the performance, the outputs from the segmentation process are fed as input to the learning model and a comparison is made between the results obtained and the actual activity sequence observed. This comparison is done in terms of the MAE, MASE and



(a)



(b)



(c)

Figure 5.10: Performance of sequence learning model on the prediction of experimental dataset activity sequence; (a) Person 1, (b) Person 2 and (c) Person 3.

RMSE for the predictions made as defined in Chapter 3. The performance of the sequence learning model in this work depends on a proper segmentation of the unlabelled activity sequences observed.

The performance of the sequence learning framework is evaluated on the experimental dataset. Considering the dataset consists of 3 participants, a leave-one-out cross validation approach is used in experiments to learn sequences of key action occurrences for an activity. Two participants are used in training the model and the remainder is left out for testing. This is done through consecutive iterations with each participant used in testing the model.

Figure 5.10 shows the result of the sequence learning model on the prediction of the activity sequence contained in the experimental dataset. Table 5.2 shows the result when the experimental dataset is applied to the proposed ASSL model. The results produced RMSE values of 0.055, 0.049 and 0.050 respectively for all three participants in the dataset when each was tested using the leave-one-out cross validation. The lower the RMSE value the better the result in predicting the sequence. The variation in the structure of the sequence between the remainder two person's data used when training the model and the structure of the person 1 used in testing the model produced a higher RMSE value (0.055) in comparison with the RMSE value obtained for other two. This can be attributed to the nature of the activity sequence for person 1, that is, the speed of the activity.

5.6.3 CAD-60 Dataset Results and Evaluation

The segmentation process applied to the CAD-60 dataset using the same values of simple moving averages as in the case of the experimental activity dataset to identify key actions which are segmented resulted in a similar number of activity segments. The distribution of key actions identified in each segment is given Figure 5.11. This shows a similar ratio in the distribution of key actions identified for all actors except for the case of *Actor 1*. This infers that for the activity - *drinking water* - performed by all actors, there are three atomic actions that define the activity. The order in which the actions occur define the activity sequence. It is important to note that the segments identified in the experiments with the CAD-60 experiment are not the same as those of the experimental activity

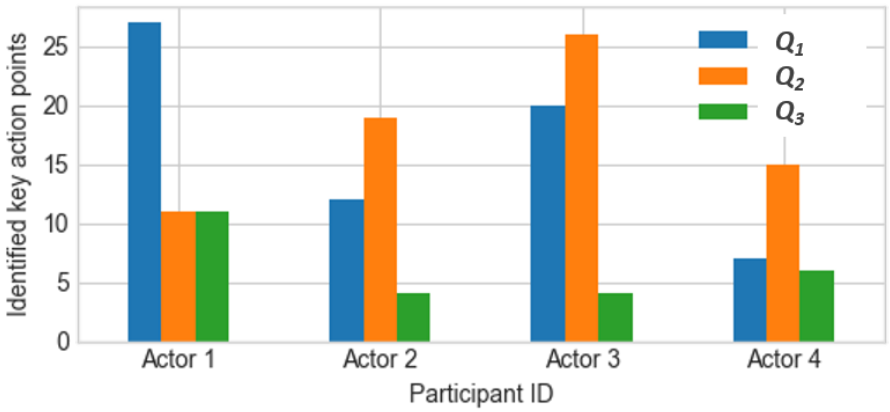


Figure 5.11: Distribution of key action points in identified activity segments for all actors in the CAD-60 dataset.

dataset.

Evaluating the performance of the sequence learning framework with the CAD-60 dataset is implemented in a similar method to the experimental dataset. A leave-one-out cross validation approach is also applied with each participant data used in testing while the remainder three are used in training the model. This is performed in consecutive iterations. In Figure 5.12, the prediction results for all actors are shown. The plots in the figure represent when each actors' activity data is left out from the training process and used to test the trained sequence learning model. Table 5.3 shows the prediction results obtained for the dataset with the ASSL. The RMSE values produced from predicting activity sequences for the data tested correspond to 0.092, 0.053, 0.025 and 0.078 for Actor's $1, 2, 3$ and 4 respectively. The low RMSE values show the model is able to learn with a high degree of reliability the activity sequence.

5.7 Comparison with other Sequence Learning Model

This section presents a comparison of the proposed ASSL framework's performance with another statistical model widely used in learning sequences

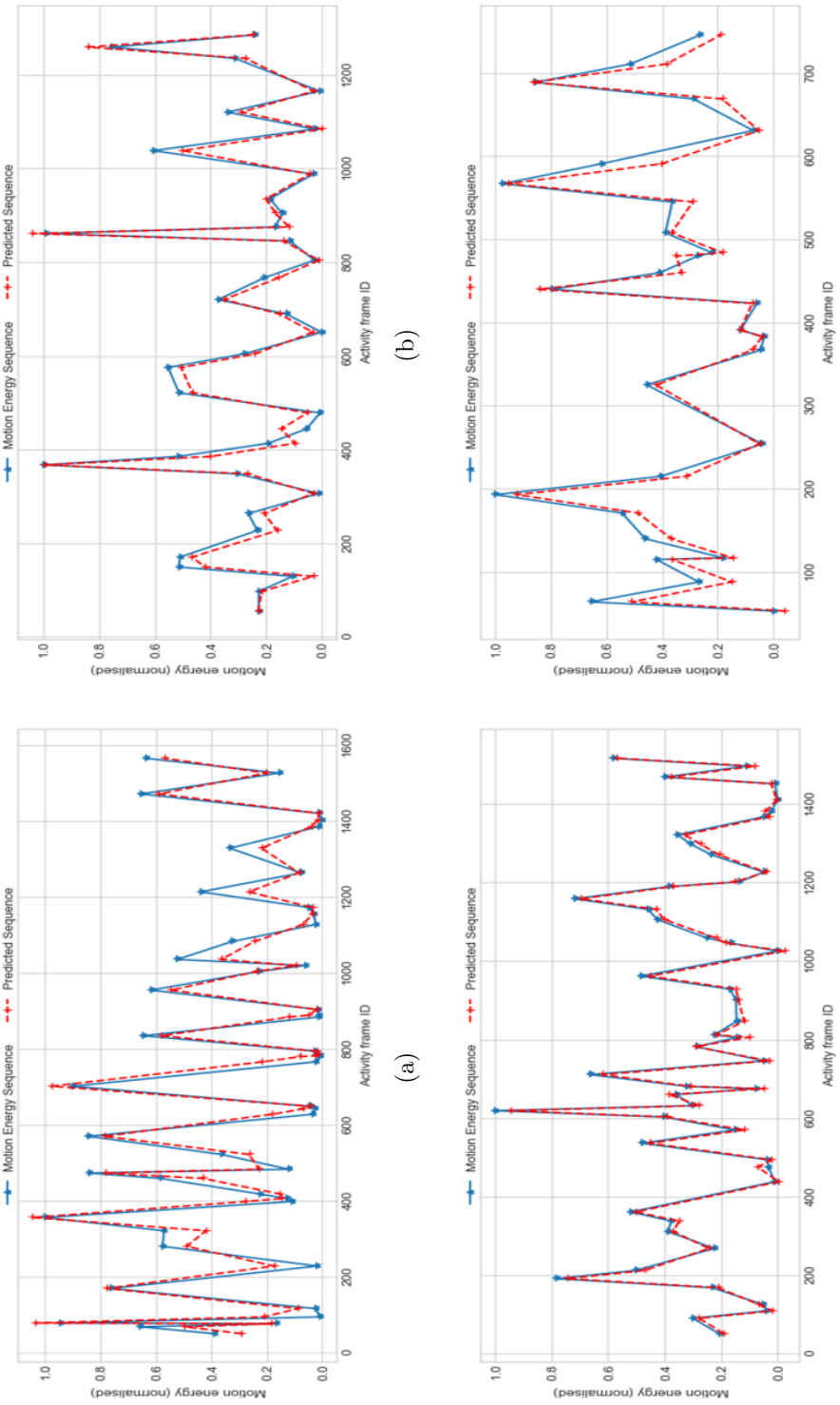


Figure 5.12: Prediction performance of sequence learning model on the CAD-60 dataset; (a) Actor 1, (b) Actor 2, (c) Actor 3 and (d) Actor 4.

from time series data. The adaptive segmentation and sequence learning of 3D skeleton data of human activities framework primarily demonstrates that unlabelled actions and sequences of activities can be modelled for accurate prediction of unseen actions. This is beneficial for applications that require exploiting the underlying patterns to understand human tasks from visual observations while they are executed. This was demonstrated in the previous sections. To further emphasise the ability of the proposed framework to learn activity sequences, a comparison is made with another method of sequence learning used in forecasting applications, an Autoregressive Integrated Moving Average (ARIMA). The basis for selecting this model is because it comes from a well established area of CI. ARIMA models are also widely used in analysis of temporal pattern recognition and time series prediction. The algorithm is applied to both the experimental dataset and CAD-60 dataset described earlier in the experimental design, with the same validation technique already described.

Autoregressive Moving Average (ARMA) models are amongst the most widely used statistical algorithms for modelling and predicting time series information [108]. A generalisation of this model is the Autoregressive Integrated Moving Average (ARIMA) which is applied in situations where there is evidence of non-stationarity in data. In such cases, a differencing step, d, corresponding to the *Integrated* part of the model is applied to remove non-stationarity points [17]. Afterwards, the ARMA model is applied on the stationary data. The implementation of ARIMA in this work follows the method described in [17]. The Auto-Regressive, AR, component uses weighted linear combinations of previous values of the data sequence and performs a regression of the sequence against itself. Similarly, the Moving Average, MA, component attempts predicting a target using regression based on past forecast errors. The parameters of the ARIMA model corresponding to coefficients of the orders of the model are d, p and q. p represents the number of time lags to consider. When $p = 0$, the mode is reduced to a MA model of q order. Similarly, if $q = 0$, the model becomes AR of p order. Details of the selection of the optimal parameters for the ARIMA model used are beyond the scope of this work. Readers are referred to [17] for more insight into ARIMA.

5.7.1 Result of ARIMA Model on Experimental Dataset

The normalised key action points of the motion energy extracted from the experimental human activity are used as input to the ARIMA model. The results shown in Table 5.2 present the performance of the ARIMA model on the experimental dataset. As observed from the table, the proposed ASSL model had a better performance in terms of the MAE and RMSE than the ARIMA model across all participants in the dataset. There is a significant difference in the MAE and RMSE performance obtained with the ASSL method outperforming the ARIMA model. However, the ARIMA model did better in terms of the MASE performance. As with most unsupervised learning structures, the ARIMA is able to predict data sequences with only the targeted data.

Table 5.2: Comparison of the proposed ASSL model performance with an Autoregressive Integrated Moving Average (ARIMA) model on the experimental human activity dataset.

Method	**Metric**	**Person 1**	**Person 2**	**Person 3**
ASSL	MAE	**0.044**	**0.025**	**0.032**
	MASE	0.867	0.630	0.690
	RMSE	**0.055**	**0.049**	**0.050**
ARIMA	MAE	0.228	0.135	0.132
	MASE	**0.586**	**0.272**	**0.291**
	RMSE	0.298	0.198	0.175

Table 5.3: Comparison of the proposed ASSL model performance with an Autoregressive Integrated Moving Average (ARIMA) model on the CAD-60 dataset.

Method	**Metric**	**Actor 1**	**Actor 2**	**Actor 3**	**Actor 4**
ASSL	MAE	**0.072**	**0.044**	**0.023**	**0.062**
	MASE	0.914	0.921	1.074	0.968
	RMSE	**0.092**	**0.053**	**0.025**	**0.078**
ARIMA	MAE	0.307	0.202	0.220	0.255
	MASE	**0.865**	**0.690**	**0.983**	**0.802**
	RMSE	0.339	0.267	0.264	0.326

5.7.2 Result of ARIMA Model on CAD-60 Dataset

Table 5.3 shows the results obtained for the comparison of the ASSL framework with the ARIMA model on the CAD-60 dataset. The performance attained using the ARIMA model showed higher RMSE and MAE values for all actors when compared to that of the ASSL. The only exception is in terms MASE, the ARIMA did better than the ASSL by attaining lower MASE values across all four actors. Based on these results, it can be concluded that the proposed ASSL approach outperformed the ARIMA model.

The ARIMA model works as a regression model and therefore does not require labelled samples. However, the proposed approach is able to obtain labels through a non-parametric approach which is used in the later stage of sequence learning. This gives the ASSL method an edge over the ARIMA.

5.8 Discussion

In this chapter, a novel adaptive technique for the segmentation and sequential learning of human activities is presented. The goal is to enable the discovery unknown activity patterns for prediction of future actions in an activity sequence, especially, for use in assistive robotics. Due to the dynamic nature of human behaviour, there are uncertainties associated with modelling actions performed in an activity. This work focused on proposing a model capable of adapting to variations that exist in actions through activity sequences. The use of 3D skeleton joint data obtained with RGB-D sensors makes it possible to acquire representations of actions for learning such activities.

The motion energy of skeleton joints is used as a feature in the segmentation process. This is due to changes in acceleration and deceleration observed in skeleton joints through a continuous sequence of activities. This feature is used in identifying key actions in an activity sequence from the moving average crossovers of the computed motion energy. This step acts as a filter stage as not all actions of an activity are relevant in predicting the activity sequence. This work leverages the ability of LSTM model in learning activity sequences for predicting future actions of activities based on previous instances. The results

show the performance of the LSTM sequence learning model is better than the unsupervised sequence learning approaches. Furthermore, learning sequences of activity from unlabelled activity structures are addressed. The segmentation approach used to identify labels from the structures made it possible to solve the unsupervised learning problem with a supervised technique of learning sequences.

Chapter 6

Activity Transfer Across Heterogeneous Feature Spaces

6.1 Introduction

As stated in the framework design in Chapter 3, Transfer Learning (TL) aims to improve performance on a target using the knowledge learned from a source. The transferal of knowledge between domains involves considerations of the nature of the data contained in each domain and the relationship between feature spaces.

Traditional machine learning approaches work with the assumption that the data for training (within the source domain) and test (within the target domain) are drawn from the same probability distributions and have the same feature spaces [39, 147]. However, in practical situations, it is often not the case. If the data distribution or feature space of the target changes, the trained models become unfit and new models would need to be rebuilt. For example, in the case of human-robot TL of activities. Based on the activities recognised through the HAL model, the robot is required to perform the activities. Due to the different feature space distributions between a human and robot, if the activity model is not well adapted to the robot's feature space, the accuracy in actualising movements will be significantly affected. TL techniques are applied to handle such situations.

Most methods that have been proposed for TL focus on the differences in task labels or the differences in the probability distribution of data between

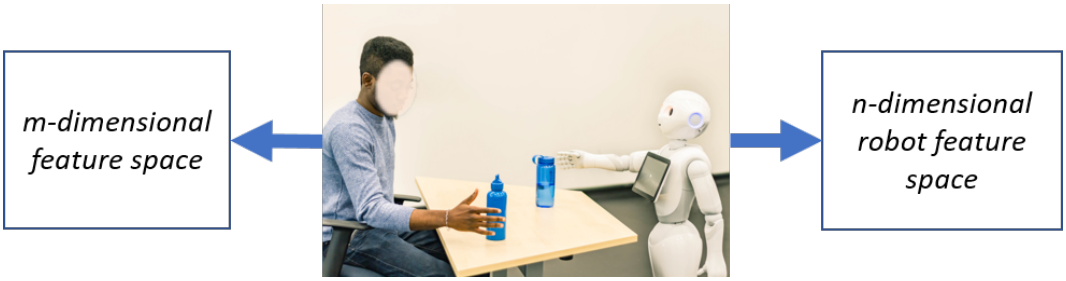


Figure 6.1: Illustration of an activity executed by a human which is intended to be learnt by an assistive robot with a different feature space distribution.

both source and target domains [105, 149]. This chapter presents a novel activity transfer across heterogeneous feature spaces model which is contained in the framework. The method incorporates fuzzy sets concept in a computational approach to acquiring membership states of activity instances from a source domain (observed human carrying out the activity) which are used to identify task states from unlabelled data of human activities. These states are mapped to the target feature space (assistive robot) for effective learning and prediction of activities. The use of a fuzzy computational approach is due to the fact that directly mapping the source and target features would not be feasible. Differences in data distribution in both domains creates a limitation for a direct mapping and thus, the fuzzy approach is important in transformation across the feature spaces. For better understanding of the challenge, consider the illustration in Figure 6.1, the difference in feature spaces prompts for a suitable method of knowledge transfer to enable the robot to learn the activity from the human. This motivates the work in this chapter.

The remainder of the chapter is organised as follows; Section 6.2 presents an overview of TL in heterogeneous feature spaces. In Section 6.3, the method for activity transfer across heterogeneous feature spaces is presented in detail. Section 6.4 describes the application of the proposed method for activity transfer and Section 6.5 follows with discussions and a summary of the chapter.

6.2 Overview of Transfer Learning in Heterogeneous Feature Spaces

Much work has been done relating to TL and this section discusses some of the works related to the methodology in this chapter. TL of a human activity usually involves a process of learning the activity in the source domain to acquire relevant knowledge of the activity which is transferred to the target domain. Exploring the feature spaces in both domains helps in understanding the approach employed in knowledge transfer. Most existing solutions to TL consider cases of homogeneous spaces [147, 148]. This is the case in which the feature space and probability distribution of information in the source and target domains are similar. In homogeneous TL, the methods for the adaptation of the transferred knowledge to the target domain have employed representative models that include information theoretical learning, Transfer Component Analysis (TCA), transfer deep network, feature level domain adaptation, scatter component analysis and a host of other models.

In heterogeneous spaces, the typical methods often applied in the target domain adaptation are alignment-based models, semi-supervised kernel matching for domain adaptation, heterogeneous spectral mapping and kernel Canonical Correlation Analysis (CCA) [150]. These methods have had success in handling the issues in heterogeneous domain adaptation, however, they do not consider the uncertainty inherent in most cases of knowledge transfer problems. The amount of information available in the target domain determines the degree of uncertainty in transfer. Problems with few data in the target domain have a high degree of uncertainty due to the limited amount of information which can be extracted. However, the development of fuzzy systems have had some success in addressing this problem.

6.3 Methodology

This section describes the methodology for human activity transfer when considering situations of heterogeneous feature space relationships. This works

considers the case of human activities involving human-robot interaction where a robot is used in an assisted living environment as an assistive agent. This entails having to carry out activities as a human would. The challenge of learning human activities by an assistive robot requires an adaptation of the source feature space which are the features observed from a human in the target (or robot) feature space. Before going into details of the methodology used in this work, it is important to define some preliminary terms used in this work.

Definition 6.3.1. Heterogeneous feature space transfer: Heterogeneous feature space transfer is a branch of TL in which the feature spaces of both source and target domains differ, $F_s \neq F_t$. Given the difference between human and robot features, transfer of activities in both domains involves exploring heterogeneity in both feature spaces.

In the next section a description of the framework proposed in this work is given.

6.3.1 Description of Activity Transfer Model

In Figure 6.2, an overview of the method proposed in this work for fuzzy TL of human activities in heterogeneous feature spaces is given. The source feature space consists of unlabelled human activity data in the form of angles, θ, between selected joints of the body. These data is extracted as human skeleton coordinates data obtainable using an RGB-D sensor. In order to get labels to identify the states of the joints at any point, Labanotation [43] is used to determine the states of joints. This information is also important to obtain the number of fuzzy membership functions used in building partitions for the fuzzy inference system in the source domain. The model in the source domain is trained and the membership degrees for T inputs are determined. The trained model is applied in the target domain. Since the feature spaces differ, an adaptation of the trained model in the target domain is necessary. To this end the limits of the membership functions are adapted to the target feature space intervals. Knowledge transfer is executed by a transfer of the combined membership degrees from source domain to target domain.

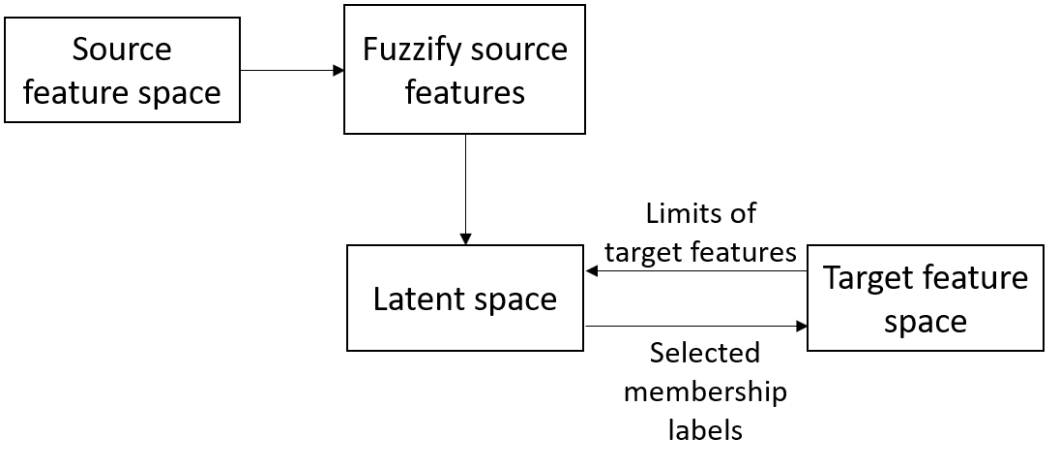


Figure 6.2: Overview of activity transfer across heterogeneous feature spaces methodology.

6.3.2 Extraction of Joint States

An activity performed by a human consists of various movements of human joints. Such movements can be described using Labanotation. This method was introduced by Rudolf Laban as a means of movement notation [43]. It comprises of four components namely; body, time, space and dynamics. The *body* component represents the moving body part, *time* represents the movement duration, *space* stands for the description of motion in terms of distance, directions, degree, or level, while *dynamics* represent the emotional components of motions. This work makes use of the body, time and space components while excluding the dynamics.

The computation of a Labanotation score is drawn in two dimensions, time rows and body columns. Figure 6.3 shows an illustration of these dimensions. The columns correspond to body parts and these contain Labanotation symbols representing the movement of each body part through time. The Labanotation symbols are normalised to fit the starting and ending times which flows from the bottom to the top. The gaps between symbols in a column show a lack of motion in that period of time or the continued previous pose. Also, in Figure 6.3 the columns are divided into the left and the right which correspond to the left and right sides of the body. In Labanotation, the shapes of the symbols

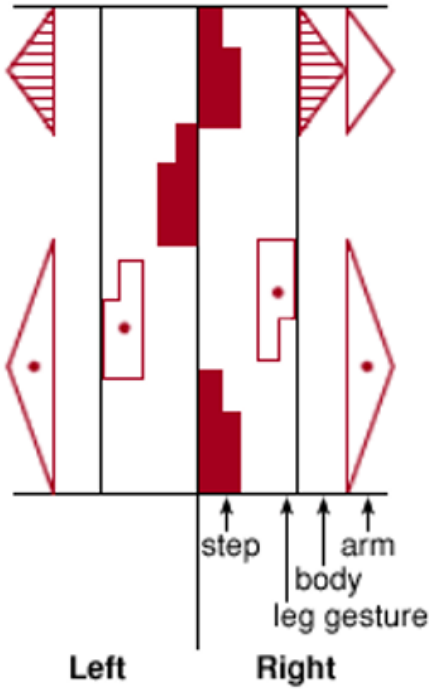


Figure 6.3: Illustration of Labanotation dimensions and score.

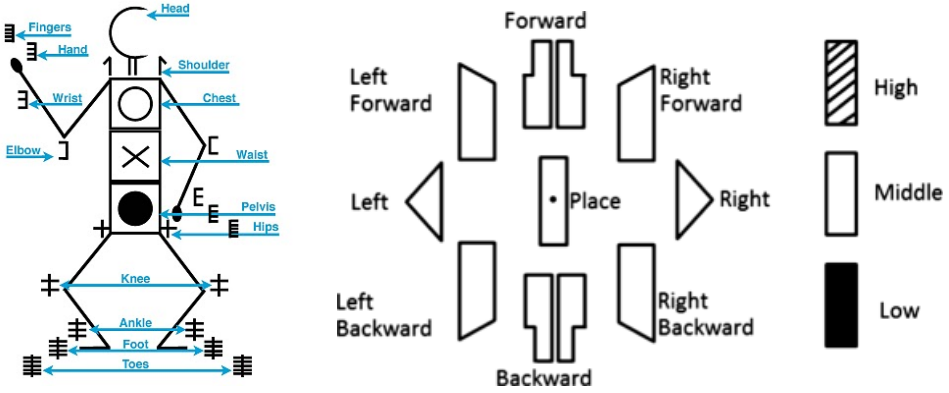
represent the direction of motion of different body parts. This is specified in the x, y and z coordinates system with the centre of the body as reference and the shape symbols are presented in Figure 6.4.

Each body part identified in the Labanotation a local coordinate system which is parallel to the part is defined at the joint near that body part. Depending on the local coordinate system, the Labanotation defines 11 shapes for azimuth directions and 3 types of shadings for levels of movements which is also known as zenith directions [57] as represented in Figure 6.4(b).

When constructing fuzzy models, the optimal number of membership

Table 6.1: Degree of contraction and extension of joints.

Degree of change		**Arm and Leg Position**
Identifier	**Membership Label**	
1	HM	High Movement
2	NM	Normal Movement
3	LM	Low Movement



(a) (b)

Figure 6.4: Labanotation illustration for describing the coordinates of human body movements. (a) shows the representation of joints of a human and (b) shows direction symbols for joints with 3 levels; *high*, *middle* and *low*.

functions required is often not known. Most applications involving fuzzy logic select the number of membership functions at random by using a value of $2N + 1$, where N is a positive integer. Also, the human joint states are not known through an activity from either joint angles (or positions) information. It is difficult to identify labels of these joints as to whether it is in a maximum, minimum, high or low state. Labanotation helps to simplify these difficulties.

In this work, Labanotation is employed in generating the joint states of human movement which is used to obtain labels of joint states and the number of fuzzy membership functions needed in the fuzzification process of the source features of an activity. The angles between human joints are considered as the input data. Table 6.1 categorises the degrees of movement of vital body parts (arms and legs) used in most activities. These categories determine the selection of membership labels and number of membership functions used to express the movement of human joints. The membership labels as shown in Table 6.1 correspond to the labanotation levels which in this work are High Movement (HM), Normal Movement (NM) and Low Movement (LM). The high and low movements correspond to joint angle movements towards the upper and lower limit values of joints respectively and the normal movement representing joint movements at mid-position.

6.3.3 Fuzzy Activity Model

The fuzzy activity model describes the formulation of fuzzy partitions used in creating the inference system for learning activities. Joint states extracted from the source data are fuzzified in an attempt to get the membership degrees representing tasks within an activity. The initial parameters used in defining the fuzzification process are determined from the extraction of joint states described in the previous section. A Takagi-Sugeno fuzzy model [47] is applied in this work in the initial process of determining the antecedent membership functions of the source activity features. The model is composed of s rules with the representation as follows:

$$\text{if } f \text{ is } A_i(f, v_i), \text{ then } p \text{ is } B_i(f, a_i) \quad i = 1, ..., s. \tag{6.1}$$

where $f = (f_1, \ldots, f_n, \ldots, f_N)$ is the set of input features of an activity. A_i is the set of membership functions obtained with each rule, v_i are centre parameters of the fuzzy partitions and a_i are coefficients of linear functions for the input activity features of the fuzzy rules.

The conditions of the rules used to obtain the sets $A_1, ..., A_s$ are constructed using a fuzzy space grid partitioning method [125]. This method is applied to divide the input activity features into the specified number of partitions determined by the membership functions. Each partition defines a fuzzy set A_i associated with a Gaussian membership function, which is normalised and represented by $\mu_{A_i}(f)$ as shown in Equation 6.2.

$$\mu_{A_i}(f_i) = \exp\left(\frac{-(f_i - v_i)^2}{2\delta_i}\right) \tag{6.2}$$

where δ_i is the width of the fuzzy space partition.

6.3.4 Knowledge Transfer Across Domains

The model proposed in this work is aimed at achieving transfer of human activities from a human domain to an assistive robot domain as mentioned in Section 6.1 by considering both domains as source and target domains respectively. The source domain consists of m-dimensional input variables of human joint angles denoted

by $D_s = (f_1^s, ..., f_m^s)$. The target domain also denoted by $D_t = (f_1^t, ..., f_m^t)$.

In this case, the model is built for obtaining fuzzy membership degrees in the human domain. This model cannot be directly applied to an assistive robot domain to perform similar activities due to the reason that the rules need to be modified to fit the robot feature space. In order to achieve this, the following steps are outlined to modify the fuzzy activity model obtained from the human domain for use in the target robot domain:

Step 1 - Applying labanotation to extract states of a robots feature space: This process is used to determine the similarity between $F_s \sim F_t$ in order to obtain the relevant joints that correlate in both domains. Since labanotation is used in the human domain to determine the joint state which gives an indication of the number of membership functions used. Similar representations of joints information is used to describe a robot feature space. Therefore the selected membership functions in both domains are similar.

Step 2 - Fuzzify the target feature space: The target feature space is represented as a fuzzy system. A method of generating fuzzy rules from numerical data as proposed by [125] is applied. Using the number of membership functions obtained in Step 1 represents the number of regions and the limits of each feature in the target as the interval for the feature.

Step 3 - Adaptation of the target feature space to transferred fuzzy activity model: The fuzzy activity model trained in the source domain is adapted in the target domain. This involves the mapping of a robots features space to fit the model obtained from source activity feature space.

Step 4 - Transfer of fuzzy membership degrees from source to target space: Due to the fact that joint movements cannot be assumed to be crisp values for a particular activity, the membership degrees obtained for each task T_s of an activity is taken and mapped to a corresponding label in D_t.

The transferred knowledge is intended to be used by the robot in acquiring the information needed to drive the joints in performing the activity similar to a human. The algorithm for the proposed method of fuzzy TL of human activities in heterogeneous feature spaces is given in Algorithm 3.

Algorithm 3 Transfer Learning of human activities in heterogeneous feature spaces.

Input:

Source domain D_s, Target domain D_t, Activity input a_n, Fuzzy partitions δ_i.

Output:

μ_s for target domain D_t.

Procedure:

Step 1:

1: Extract the joint information of human activity using the data processing method described in Chapter 4.

2: Determine the membership functions using Labanotation. See Section 6.3.2.

Step 2:

3: Fuzzify the input feature space. See Section 6.3.3.

4: Obtain the fuzzy membership degrees using the highest degree of membership of input.

Step 3:

5: Determine the transition sequence of source feature space. See Chapter 5.

6: Determine the intervals of the target feature space and obtain membership functions using Equation 6.2.

Step 4:

7: Use the fuzzy membership degrees obtained from source feature space in target domain to obtain the target features.

6.4 Application of Methodology and Results of the Activity Transfer Framework

To demonstrate the application of the proposed method for activity transfer, the framework is evaluated on human activities obtained from the human (source domain) feature space which is transferred to an assistive robots (target domain) feature space. The activity performed involves a sequence of the left and right arm gestures starting with both hands held down and raised up repeatedly. An RGB-D sensor is used to obtain data of human joint angles while carrying out the activity. Figure 6.5 shows selected key frames of the activity collected for the experiment. Each frame represents a task T_s within the activity.

The activity performed is mostly concerned with arm movements, therefore

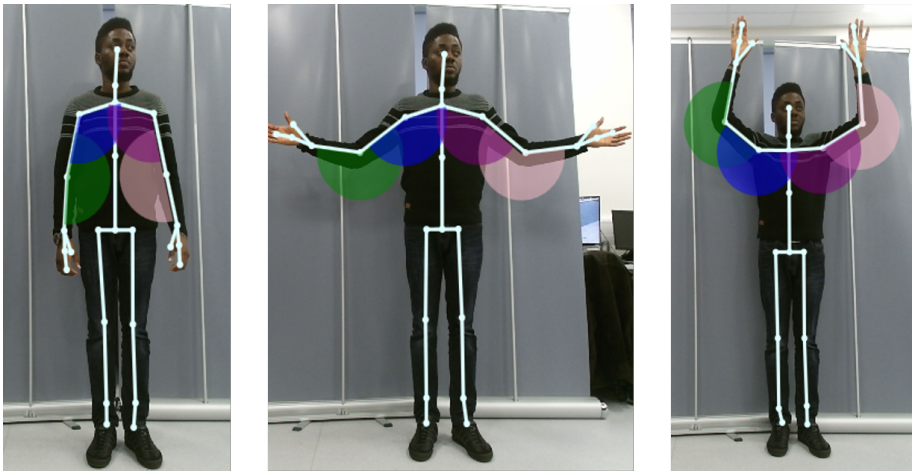
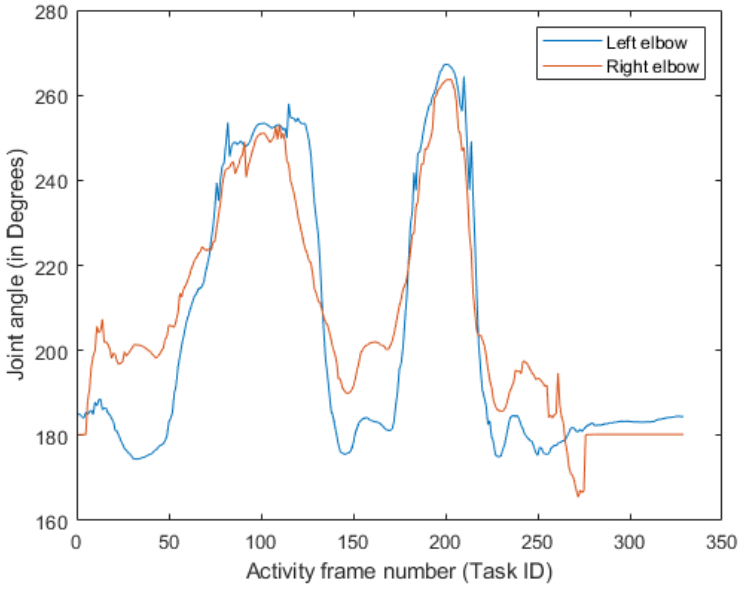


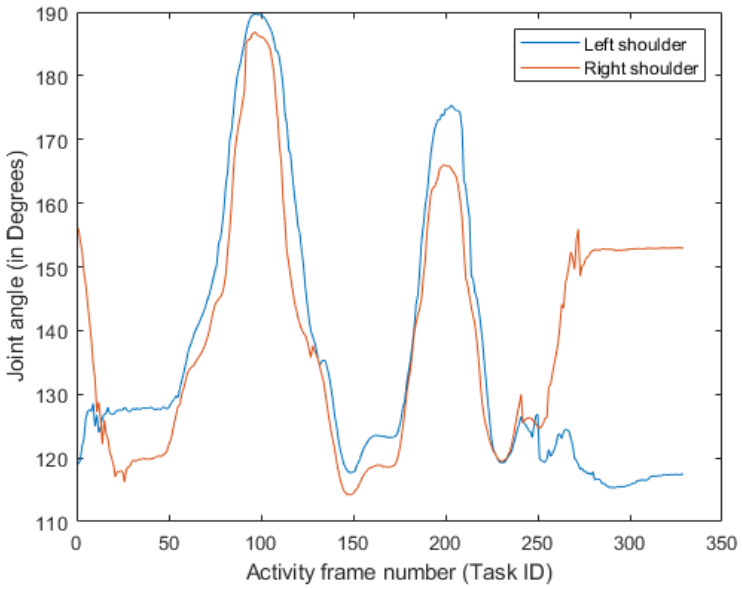
Figure 6.5: Examples of activity frames from a sequence of arm movements from down to up activity positions. The highlighted areas show the joint angles extracted.

four joint angles considered to be the most used in the activity are selected. These are considered as the source feature space and correspond to the angles at the left elbow, θ_{LE}, right elbow, θ_{RE}, left shoulder, θ_{LS} and right shoulder, θ_{RS}. The movement trajectory of these angles in the activity depicted by the sample frames in Figure 6.5 are represented in Figure 6.6. These trajectories show the movement of the selected joint angles through the observed activity. Considering the nature of human movements which are not smooth through the trajectory, a filtering process is applied as a preprocessing step to smoothen the raw data extracted. The fuzzy model for the activity is obtained using the approach earlier described. Three membership functions are used in the fuzzy partitions. These are determined with respect to the degree of joint movement obtained through the label definitions in Table 6.1.

In the experiment, a two-arm Baxter robot [100] as shown in Figure 6.7(a) is considered as the target domain for assistive applications. The robot consists of seven Degrees of Freedom (DoF) on each arm that are identified in Figure 6.7(b). These include two DoFs around the shoulder (roll and pitch), $S0$ and $S1$, two DoFs around the elbow (roll and pitch), $E0$ and $E1$, and three DoFs around the wrist, $W0$, $W1$, and $W2$.



(a)



(b)

Figure 6.6: Joint angles trajectory for source (human) activity with up and down sequential movement of arms. (a) represents elbow movements for both arms and (b) shoulder movements for both arms.

(a) (b)

Figure 6.7: Research robot used in this work. (a) two-arm Baxter robot and (b) Baxter robot joints identification.

6.4.1 Result of Joint States Extraction

The activity used in the experiments describe the movement of both the left and right arms of a human subject. This involved the roll directional movement of the joints. The work in this thesis is limited to only the roll movement. However, it can be extended to more complex cases that involve the three axes of rotation (i.e roll, pitch and yaw). It can be observed that the poses in the activity are also limited to frontal poses. Naturally human activities can have more complicated motions, such as arm movement in a backwards direction. Such motions are ignored in this work because this work focuses on assistive robots observing human activities that are intended to be viewed by a robot in front of a human subject performing the activities.

Following the description of Labanotation presented earlier in this chapter, the joint movements are digitised into the three levels identified: *high*, *normal* and *low*. Figure 6.8 shows the representation of joint states for the activity conducted. The figure shows four columns that correspond to the Labanotations of the selected joints; the shoulder and elbow joints of the left and right arms.

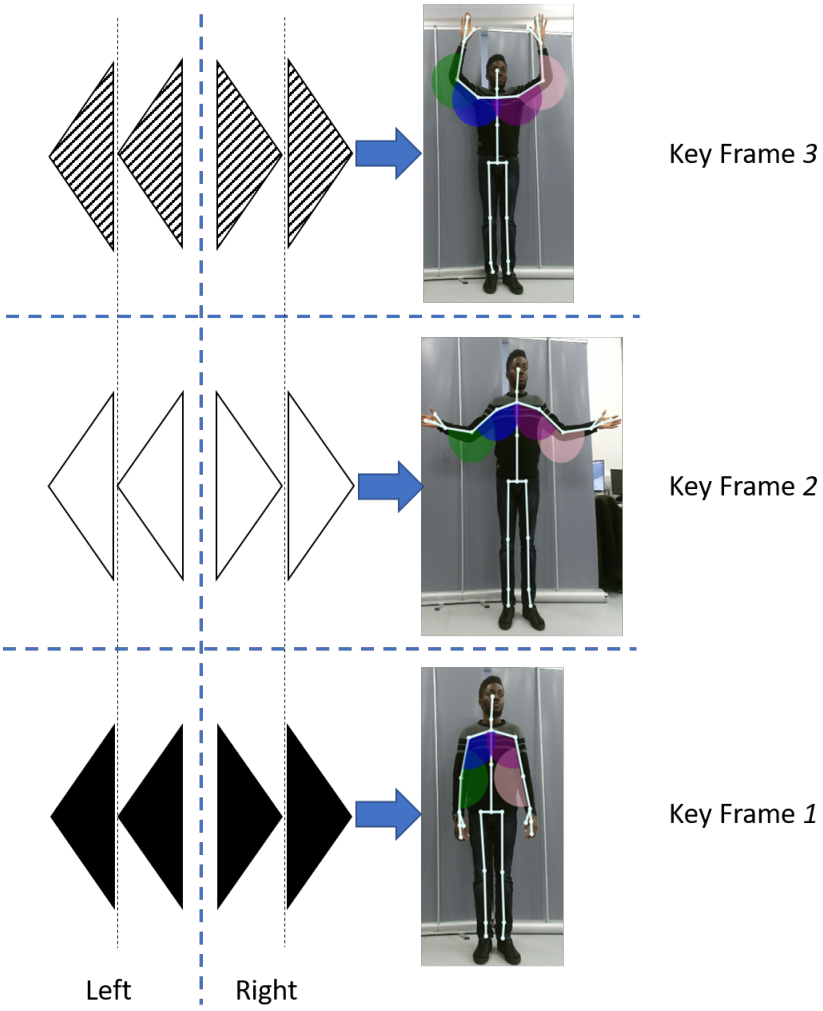


Figure 6.8: Extracting joint states of the elbow and shoulder joints of an activity using Labanotation.

6.4.2 Knowledge Transfer Through Fuzzification

The representation obtained from the Labanotation is used in generating the fuzzy partitions for each joints space. The use of Gaussian membership functions are employed in the fuzzification for each partition. The joint angle variables [θ_{LE} θ_{RE} θ_{LS} θ_{RS}] are partitioned in the universe of discourse [0° 180°] which represent the limit of the joint's movement. Three Gaussian membership functions representing the labanotation joint states; High Movement (HM), Normal Movement (NM) and Low Movement (LM) are defined as shown in Figure 6.9.

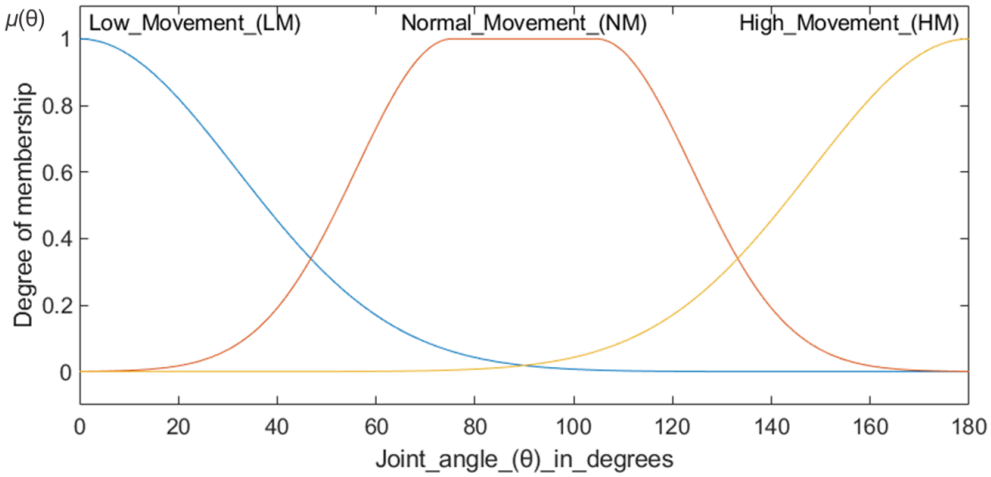


Figure 6.9: Fuzzy partitions using Gaussian membership functions of human elbow and shoulder joint angles.

The fuzzy membership degrees corresponding to the identified key frames in the human feature space are obtained. This creates a latent space for the mapping of joint movements from the human (source) to robot (target) spaces.

In the experiments, the joints of the robot, $E1$ and $W1$, corresponding to the roll directional movements of human shoulder and elbow joints respectively are selected. Table 6.2 shows the universe of discourse of the robot's joint's with the limits, θ_{min} and θ_{max}, of joint angles. The joints used in this work are highlighted in the table. Fuzzy partitions of the selected joint's are created using Gaussian membership functions as applied in the human domain. This makes it possible to transfer the fuzzy membership degrees obtained from the human feature space to

Table 6.2: Baxter left and right arm joint's angle limit.

Joint Name	Joint Variable	θ_{min}	θ_{max}	θ_{range}
$S0$	θ_1	$+51°$	$-141°$	$192°$
$S1$	θ_2	$+60°$	$-123°$	$183°$
$E0$	θ_3	$+173°$	$-173°$	$346°$
$\boldsymbol{E1}$	$\boldsymbol{\theta_4}$	$\mathbf{+150°}$	$\mathbf{-3°}$	$\mathbf{153°}$
$W0$	θ_5	$+175°$	$-175°$	$350°$
$\boldsymbol{W1}$	$\boldsymbol{\theta_6}$	$\mathbf{+120°}$	$\mathbf{-90°}$	$\mathbf{210°}$
$W2$	θ_7	$+175°$	$-175°$	$350°$

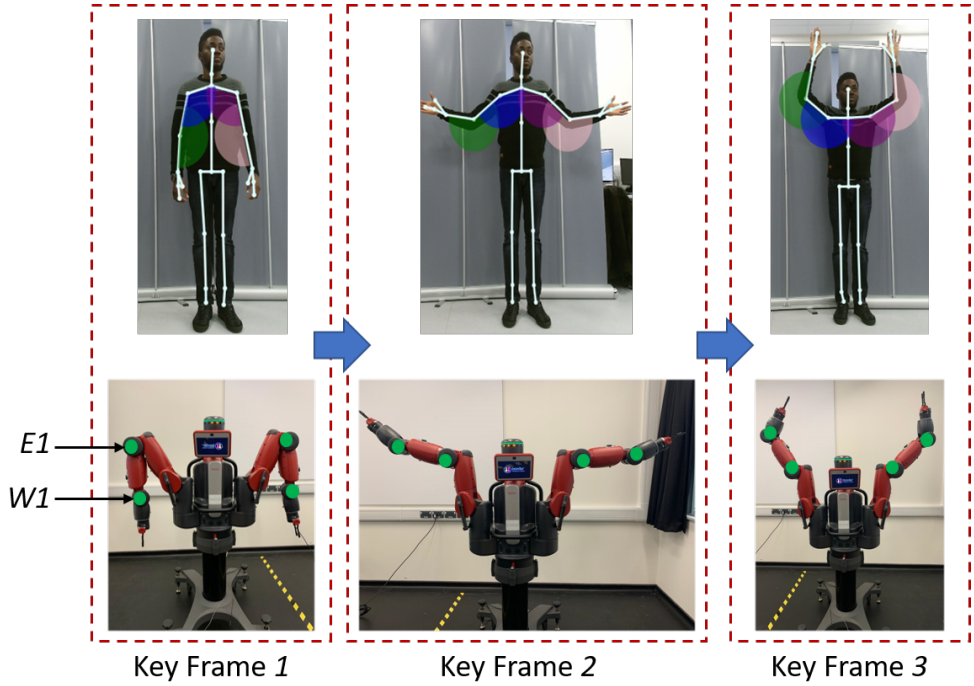


Figure 6.10: Final motions of Baxter assistive robot from the transferred human activity information.

the robot's feature space. The fuzzy membership degrees mapped into the robot's feature space are defuzzified to obtain crisp values for the joint angles that are used in determining the directional movements of the robot's arm joints.

Figure 6.10 shows the final motions performed by the robot. The poses shown are based on the key frames identified for the activity. Based on the visual inspections, it can be concluded that the system can reproduce the original motion to a high degree of certainty. It can be observed from the system implementation that the universe of discourse for the human and robot joints have different range. This can be the case with any assistive robot, thus, the need for applying a fuzzy inference system to handle such uncertainty. Also, the intermediate poses captured between two key frames are neglected. This is because while observing an activity, humans tend to neglect small changes in motion and focus on the key aspects of the activity.

6.5 Discussion

In this chapter, an approach to fuzzy TL of human activities in heterogeneous feature spaces is presented. The proposed method facilitates faster learning of activities by assistive agents which are used in assisted living environments. The method uses a combination of Labanotation and fuzzy logic in representing the observed joint states from a subject while performing an activity. Labanotation is used to determine the number of fuzzy partitions to be created and provides a high level feature space for both source and target feature representation. This approach is experimented on a simple human activity which is transferred to an assistive robot platform. The intervals of the feature space in the target domain are obtained to adapt the membership functions of the trained source model. The outcome from the experiment proves that the proposed methodology for human activity transfer across heterogeneous feature spaces is a useful tool in equipping an assistive robot with skills necessary to perform human activities in an assisted living environment.

Although attempts to address the challenge of differences feature spaces across the domains are currently under much study, the proposed fuzzy approach for knowledge transfer proves to be efficient in achieving the goal of learning and predicting human actions from visual information of observations.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The work in this thesis presented a novel framework for fuzzy Transfer Learning (TL) in human activity recognition with the purpose of enabling assistive agents in AAL environments to acquire the knowledge of human activities as are performed by humans. The motivation for the work is from the perspective of humans ability to excel in dealing with everyday activities through the process of learning and adapting to different activities. This comprises the application of different complex techniques that enable a lifelong learning process from observations.

The thesis attempted to answer the research questions identified when viewed from the theoretical and practical perspectives. Three research questions were identified in Chapter 1 which are summarised as follows:

- How to learn human activities using a computationally efficient information modality?
- Can activity sequences be modelled from unlabelled data?
- Lastly, how can transferred human activities be adapted from a source domain to a target domain?

The chapters in the thesis focused on addressing the questions identified. Hence, the following sections in this chapter summarise the findings and

conclusions drawn from the thesis. The major contributions are discussed along with considerations for future work. The following sub-sections outline the findings of the thesis.

7.1.1 Source Information Can be Considered as 3D Human Activity Data

The use of visual human activity information obtained as 3D skeleton joint coordinates of the human body is used in this thesis for recognising daily human activities. The framework developed for TL in human activity recognition uses an RGB-D sensor capable of extracting information in a 3D space to obtain information while humans perform activities. These RGB-D sensors can be obtained at low cost and deliver reliable information of objects tracked. Therefore, not much processing is required to obtain the skeleton joints' information. It can be observed from the results presented in Table 4.6 of Chapter 4 that the accuracy of the system modelled for human activity learning using only the joint coordinates information obtained higher accuracies in comparison to other methods proposed which use a combination of information modalities.

7.1.2 Human Actions Can be Identified from Unlabelled Data

The results presented in Chapter 5 shows how human actions can be identified from unlabelled human activity information. The method shows that the use of a non-parametric clustering approach described by Equations 5.10, 5.11 and 5.12, simplifies the process of identifying the number of key actions in an activity sequence. Furthermore, the number of key actions in each cluster can be used to infer the difference in activities performed by different people. For example, identifying the difference in speed.

7.1.3 Transfer Learning is Effective When Activities are Well Interpreted

The strategy used in this work in achieving TL of human activities is based on the interpretation of activity information in a manner capable of being adapted in a target domain. Hence, the stages of recognition of activities, segmentation and sequence learning of actions within activities. Afterwards, an effective transfer is performed by employing fuzzy logic for interpreting each movement made during an activity. The realisation of this is obtained using the procedure presented in Algorithm 3. This ensures the differences between both source and target domains are handled so as to enable generalisation across different target platforms.

7.2 Summary of Major Contributions

The approach employed to achieve the aim set out in this thesis resulted in significant contributions. These contributions are discussed as follows:

7.2.1 A Novel Framework for Human Activity Learning

This thesis presented a novel framework for the learning and recognition of human activities from data obtained using RGB-D sensors. The fundamental part of the TL of human activities for application in assisted agents such as robotics is first the ability to recognise activities. This process involved the development of a model for activity recognition from observed information. Human activity information is obtained as coordinates of key joints in a human extracted using an RGB-D sensor. Experiments are conducted on selected activities to acquire enough information for building the model. From the information acquired, relevant features (traditional and hand-crafted) used in identifying activities are detected and are used in a novel classifier ensemble model to recognise different activities. The results obtained in Chapter 4 show the ability of the framework to identify activities with the feature set over state-of-the-art models on experimental and benchmark datasets.

7.2.2 A Novel Framework for Action Segmentation and Sequence Learning from Unlabelled Sequences

The fuzzy TL in human activity recognition framework encompasses the ability to understand constituent actions within each activity identified. Therefore, a novel framework is developed for the adaptive segmentation and sequence learning of the actions of activities. The framework developed consists of three stages with each stage defined as:

1. Extraction of key actions from observed unlabelled human activity information which is described using the process described by Equation 5.4 - Equation 5.9. Key actions through activities are identified as not all actions in an activity are relevant in defining a sequence for representing the activity.

2. Activity segmentation of key actions via clustering presented in Algorithm 2. Similar key actions are grouped and assigned labels used in identifying the sequence order.

3. Sequence learning of the segmented key actions using Equation 5.13 - Equation 5.18. This enables the ordered representation of actions in the identified segments.

7.2.3 A Novel Framework for Human Activity Transfer using Fuzzy Generated Rules from Human to Robot Spaces

The transfer of the learned activity from the source domain to the target domain is based on the exploration of heterogeneous features spaces of both domains. A method of remapping feature spaces is developed using the steps in Algorithm 3 to enable effective mapping of the source features to the target. The framework uses an approach of fuzzy latent space exploration in Equation 6.1 to obtain mappings of the features. The case study used is the transfer of human activity features to an assistive robot.

The thesis developed a novel fuzzy activity model that describes the formulation of fuzzy partitions from human joints states for creating an inference system that derives feature maps. Each joint is represented by a set of membership functions that determine the rules for its movement in activities. These rules are then used to obtain joint movements in a robot's feature space. In summary, the fuzzy activity model comprises the following steps:

1. Extraction of joint states movements: the joint states are obtained by applying labanotation to determine a relation describing the degree of movement across each joint.

2. Fuzzification of the feature spaces: both source and target feature spaces are fuzzified through defined fuzzy membership functions. The fuzzy rules are then generated for movement sequences.

3. Adaptation of the activity model to the target feature space: The model trained in the source domain is adapted in the target domain.

4. Transfer of fuzzy membership identities from source to target feature spaces: The final stage involves mapping fuzzy membership identities for the modelled activity in the source to obtain membership identities in the target for movement actualisation.

7.3 Future Work and Recommendations

Similar to any research, the need for future work for the improvement of the framework is evident. This section identifies the directions for future research and recommendations for improvement of the framework for fuzzy TL in human activity recognition.

- Extension of the HAL model.

  Although the classifier ensemble model applied three base classifiers, this could be extended to include more classifiers which may improve performance and also deep learning neural networks which are increasingly used in HAR systems. Additionally, the ensemble model presented can be

used in other applications such as fall detection systems. The system could also be extended to learning activities on-the-fly as they are carried out by an actor. This is important as the performance of current technology systems are now directed towards real-time applications.

- The incorporation of cloud-based applications for processing activity information.

  The framework developed in this thesis is proposed for assistive robotics applications. To achieve fast processing of the recognition and learning of activity sequences, the incorporation of a cloud-based system which benefit from the low cost of physical hardware resources for processing would improve the efficiency of the system. As such, the framework would support the connection of multiple robots which can be easily integrated into the system.

- Extension to more activities.

  The fuzzy TL framework presented in this research focused on a set of 12 activities as described in the datasets used. Future work should consider more activities that have not been considered in this research. Such activities should involve the active participation of the limbs (hands and legs). Another suggestion would be the fusion of other sensors such as wearable sensors with RGB-D sensors to detect salient movements during activities. This could also be used to provide information such as the orientation of joints for robots.

- Application of TL to more problems.

  TL is yet to be used extensively in day-to-day applications. This concept with its many benefits is yet to be explored in-depth. If well explored through the more incorporation in daily applications to promote independent assisted living, it would be the driver for the next revolution in technology.

References

[1] J. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70 – 80, 2014. 16, 17, 18, 19, 21

[2] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3d human pose reconstruction. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1446–1455. IEEE, 2015. 24

[3] S. Aminikhanghahi and D. J. Cook. Using change point detection to automate daily activity segmentation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 262–267, 2017. 88

[4] S. Aminikhanghahi and D. J. Cook. Enhancing activity recognition using cpd-based activity segmentation. *Pervasive and Mobile Computing*, 53:75 – 89, 2019. 88

[5] D. N. Anh. Detection of lesion region in skin images by moment of patch. In *2016 IEEE RIVF International Conference on Computing Communication Technologies, Research, Innovation, and Vision for the Future (RIVF)*, pages 217–222. IEEE, 2016. 23

[6] A. Arnold, R. Nallapati, and W. W. Cohen. A comparative study of methods for transductive transfer learning. In *Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 77–82. IEEE, 2007. 43

[7] V. Behbood. *Fuzzy Transfer Learning for Financial Early Warning System*. PhD thesis, University of Technology, Sydney, 2013. 31

[8] V. Behbood, J. Lu, and G. Zhang. Fuzzy bridged refinement domain adaptation: Long-term bank failure prediction. *International Journal of Computational Intelligence and Applications*, 12(01), 2013. 31

[9] V. Behbood, J. Lu, and G. Zhang. Fuzzy refinement domain adaptation for long term prediction in banking ecosystem. *IEEE Transactions on Industrial Informatics*, 10(2):1637–1646, 2014. 32

[10] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 16, 18

[11] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1929–1942, 2016. 18

[12] R. E. Bellman and L. A. Zadeh. Decision-making in a fuzzy environment. *Management Science*, 17(4):B141–B164, 1970. 31

[13] C. Benedek, B. Glai, B. Nagy, and Z. Jank. Lidar-based gait analysis and activity recognition in a 4d surveillance system. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):101–113, 2018. 22

[14] S. Blackman, C. Matlo, C. Bobrovitskiy, A. Waldoch, M. L. Fang, P. Jackson, A. Mihailidis, L. Nygård, A. Astell, and A. Sixsmith. Ambient assisted living technologies for aging well: a scoping review. *Journal of Intelligent Systems*, 25(1):55–69, 2016. 63

[15] B. Bócsi, L. Csató, and J. Peters. Alignment-based transfer learning for robot models. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2013. 44, 45

[16] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 16

[17] Ü. Ç. Büyükşahin and Ş. Ertekin. Improving forecasting accuracy of time series data using a new arima-ann hybrid method and empirical mode decomposition. *Neurocomputing*, 361:151 – 163, 2019. 112

[18] S. Calinon, F. Guenter, and A. Billard. On learning, representing, and generalising a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(2):286–298, 2007. 4

[19] N. A. Capela, E. D. Lemaire, and N. Baddour. Feature selection for wearable smartphone-based human activity recognition with able bodied, elderly, and stroke patients. *PLOS ONE*, 10(4):1–18, 04 2015. 15, 73

[20] F. Chao, Y. Huang, X. Zhang, C. Shang, L. Yang, C. Zhou, H. Hu, and C.-M. Lin. A robot calligraphy system: From simple to complex writing by human gestures. *Engineering Applications of Artificial Intelligence*, 59:1 – 14, 2017. 21

[21] L. Chen, J. Hsieh, C. Chuang, C. Huang, and D. Y. Chen. Occluded human action analysis using dynamic manifold model. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1245–1248, 2012. 17

[22] W. Chen, C. Xiong, R. Xu, and J. J. Corso. Actionness ranking with lattice conditional ordinal random fields. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 748–755, 2014. 18

[23] H.-R. Choi and T. Kim. Modified dynamic time warping based on direction similarity for fast gesture recognition. *Mathematical Problems in Engineering*, 2018(2404089), 2018. 89

[24] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante. A human activity recognition system using skeleton data from rgbd sensors. *Computational Intelligence and Neuroscience*, 2016, June 2016. 26, 55, 75, 81, 82, 84

[25] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002. 58, 59, 88, 97, 99

[26] D. Cook, K. D. Feuz, and N. C. Krishnan. Transfer learning for activity recognition: a survey. *Knowledge and Information Systems*, 36(3):537–556, 2013. 32, 33, 36, 41, 42, 43

[27] Cornell University. Cornell Activity Dataset: state of the art results. Available at: http://pr.cs.cornell.edu/humanactivities/results.php, 2009. [Online; accessed 15-February-2018]. xxi, 78, 81, 82

[28] Y. Cui, S. Ahmad, and J. Hawkins. Continuous online sequence learning with an unsupervised neural network model. *Neural Computation*, 28(11), Nov. 2016. 88

[29] O. Day and T. M. Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4(1):29, 2017. 33, 36

[30] R. Diao, F. Chao, T. Peng, N. Snooke, and Q. Shen. Feature selection inspired classifier ensemble reduction. *IEEE Transactions on Cybernetics*, 44(8):1259–1268, Aug 2014. 74

[31] C. Droke. *Moving Averages Simplified*. Marketplace Books, 2001. 96

[32] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. 25, 29

[33] P. Duckworth, D. C. Hogg, and A. G. Cohn. Unsupervised human activity analysis for intelligent mobile robots. *Artificial Intelligence*, 270:67 – 92, 2019. 34

[34] J. Durbin and S. Koopman. *Time series analysis by state space methods*. Oxford University Press, New York, 2012. 88

[35] A. Elbayoudi, A. Lotfi, and C. Langensiepen. The human behaviour indicator: A measure of behavioural evolution. *Expert Systems with Applications*, 118:493 – 505, 2019. 32

[36] H. C. Ellis. *The Transfer of Learning*. Macmillan, Oxford, England, 1965. 39

[37] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1347–1355. IEEE, 2015. 24, 28, 29

[38] D. R. Faria, C. Premebida, and U. Nunes. A Probabilistic Approach for Human Everyday Activities Recognition using Body Motion from RGB-D Images. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, pages 732–737. IEEE, 2014. 7, 15, 17, 20, 21, 27, 36, 73, 81

[39] K. D. Feuz and D. J. Cook. Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (fsr). *ACM Trans. Intell. Syst. Technol.*, 6(1):3:1–3:27, 2015. 36, 39, 44, 116

[40] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine Learning*, 32(1):41–62, 1998. 88

[41] S. Gaglio, G. L. Re, and M. Morana. Human activity recognition process using 3-d posture data. *IEEE Transactions on Human-Machine Systems*, 45(5):586–597, Oct 2015. 21, 26, 81, 84

[42] Y. Gu, H. Do, Y. Ou, and W. Sheng. Human gesture recognition through a kinect sensor. In *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1379–1384. IEEE, 2012. 21

[43] A. H. Guest. *Labanotation the system of analyzing and recording movement (4th ed.)*. Routledge, New York, 2005. 119, 120

[44] A. T. Guide. Definitions and categories of at. http://www.assistivetechnologyguide.co.uk/guides/definitions-and-categories-of-at/?LMCL=UNtoCv, 2019. Accessed: 23-10-2019. 34

[45] P. Gupta and T. Dallas. Feature selection and activity recognition system using a single triaxial accelerometer. *IEEE Transactions on Biomedical Engineering*, 61(6):1780–1786, June 2014. doi: 10.1109/TBME.2014.2307069. 73

[46] R. Gupta, A. Y.-S. Chia, and D. Rajan. Human activities recognition using depth images. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 283–292, 2013. 15, 81

[47] M. L. Hadjili and V. Wertz. Takagi-sugeno fuzzy modeling incorporating input variables selection. *IEEE Transactions on Fuzzy Systems*, 10(6):728–742, 2002. 123

[48] F. Han, B. Reily, W. Hoff, and H. Zhang. Space-time representation of people based on 3d skeletal data: A review. *Computer Vision and Image Understanding*, 158(Supplement C):85 – 105, 2017. 7, 15, 16, 21, 22, 70

[49] F. Han, X. Yang, C. Reardon, Y. Zhang, and H. Zhang. Simultaneous feature and body-part learning for real-time robot awareness of human behaviors. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2621–2628, 2017. 16

[50] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 28(5):836 – 849, 2010. 89

[51] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009. 56

[52] M. K. Helwa and A. P. Schoellig. Multi-robot transfer learning: A dynamical system perspective. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 4702–4708. IEEE/RSJ, 2017. 36, 43, 44, 64

[53] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 59, 89, 99

[54] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2466–2472, Beijing, China, 2013. AAAI Press. 25

[55] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis. Human activity recognition based on Evolving Fuzzy Systems. *International journal of neural systems*, 20(5):355–364, 2010. 15, 28

[56] E. P. Ijjina and K. M. Chalavadi. Human action recognition in rgb-d videos using motion sequence information and deep learning. *Pattern Recognition*, 72:504 – 516, 2017. 18, 24, 28, 29, 85

[57] K. Ikeuchi, Z. Ma, Z. Yan, S. Kudoh, and M. Nakamura. Describing upper-body motions based on labanotation for learning-from-observation robots. *International Journal of Computer Vision*, 126(12):1415–1429, 2018. 121

[58] A. Jalal and S. Kamal. Real-time life logging via a depth silhouette-based human activity recognition system for smart home services. In *11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2014*, pages 74–80. IEEE, 2014. 15, 20, 22, 23

[59] A. Jalal, S. Kamal, and D. Kim. Depth silhouettes context: A new robust feature for human tracking and activity recognition based on embedded hmms. In *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 294–299. IEEE, 2015. 23

[60] A. Jalal, A. Nadeem, and S. Bobasu. Human body parts estimation and detection for physical sports movements. In *2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE)*, pages 104–109. IEEE, 2019. 23

[61] C. Jayawardena, I. H. Kuo, E. Broadbent, and B. A. MacDonald. Socially assistive robot healthbot: Design, implementation, and field trials. *IEEE Systems Journal*, 10(3):1056–1067, Sept 2016. 63

[62] H. Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402, 2013. 52, 53

[63] A. Kaya, A. S. Keceli, C. Catal, H. Y. Yalic, H. Temucin, and B. Tekinerdogan. Analysis of transfer learning for deep neural network based plant classification models. *Computers and Electronics in Agriculture*, 158:20 – 29, 2019. 30

[64] B. Koçer and A. Arslan. Genetic transfer learning. *Expert Syst. Appl.*, 37 (10):6997–7002, 2010. 31

[65] I. Kononenko. Estimating attributes: Analysis and extensions of relief. In F. Bergadano and L. De Raedt, editors, *Machine Learning: ECML-94*, pages 171–182, Berlin, Heidelberg, 1994. Springer Berlin Heidelberg. 73

[66] H. S. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages 792–800, 2013. 34

[67] H. S. Koppula and A. Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, 2016. 34, 35

[68] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013. 81, 84

[69] S. Krishnan, A. Garg, S. Patil, C. Lea, G. Hager, P. Abbeel, and K. Goldberg. Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning. *The International Journal of Robotics Research*, 36(13-14):1595–1618, 2017. 87

[70] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, 2012. 4, 30

[71] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys Tutorials*, 15(3): 1192–1209, 2013. 15

[72] K. Li and Y. Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1644–1657, 2014. 86, 88

[73] L. Li and B. A. Prakash. Time series clustering: Complex is simpler! In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 185–192, 2011. 89

[74] Q. Li, F. He, T. Wang, L. Zhou, and S. Xi. Human pose estimation by exploiting spatial and temporal constraints in body-part configurations. *IEEE Access*, 5:443–454, 2017. 24

[75] C. H. Lim and C. S. Chan. Fuzzy qualitative human model for viewpoint identification. *Neural Computing and Applications*, 27(4):845–856, 2016. 28

[76] R. Lioutikov, G. Neumann, G. Maeda, and J. Peters. Learning movement primitive libraries through probabilistic segmentation. *The International Journal of Robotics Research*, 36(8):879–894, 2017. 87

[77] F. Liu, G. Zhang, H. Lu, and J. Lu. Heterogeneous unsupervised cross-domain transfer learning. *CoRR*, abs/1701.02511, 2017. URL http://arxiv.org/abs/1701.02511. 33

[78] L. Liu, L. Shao, X. Li, and K. Lu. Learning spatio-temporal representations for action recognition: A genetic programming approach. *IEEE Transactions on Cybernetics*, 46(1):158–170, 2016. 28

[79] M. Long, J. Wang, Y. Cao, J. Sun, and P. S. Yu. Deep learning of transferable representation for scalable domain adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2027–2040, 2016. 4, 30

[80] J. Lu, V. Behbood, P. Hao, H. Zuo, S. Xue, and G. Zhang. Transfer learning using computational intelligence: a survey. *Knowledge-Based Systems*, 80: 14–23, 2015. 4, 6, 7, 31, 40

[81] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Computer Vision – ECCV 2006*, pages 359–372, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 89

[82] N. Makondo, B. Rosman, and O. Hasegawa. Knowledge transfer for learning robot models via local procrustes analysis. In *2015 IEEE-RAS*

15th International Conference on Humanoid Robots (Humanoids), pages 1075–1082. IEEE, 2015. 5

[83] A. Manzi, L. Fiorini, R. Limosani, P. Dario, and F. Cavallo. Two-person activity recognition using skeleton data. *IET Computer Vision*, September 2017. URL http://digital-library.theiet.org/content/journals/10.1049/iet-cvi.2017.0118. 15

[84] MathWorks Inc. Transfer Learning Using AlexNet. Available at: https://www.mathworks.com/help/deeplearning/examples/transfer-learning-using-alexnet.html, 2018. [Online; accessed 30-December-2018]. 43

[85] J. Medina-Quero, S. Zhang, C. Nugent, and M. Espinilla. Ensemble classifier of long short-term memory with fuzzy temporal windows on binary sensors for activity recognition. *Expert Systems with Applications*, 114:441 – 453, 2018. 89

[86] Microsoft. Developing with kinect for windows. https://developer.microsoft.com/en-us/windows/kinect/develop, 2017. Accessed: 2017-02-28. 20, 21, 50, 53, 68, 102

[87] Y. Mollard, T. Munzer, A. Baisero, M. Toussaint, and M. Lopes. Robot programming from demonstration, feedback and transfer. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1825–1831. IEEE, 2015. 3

[88] B. Ni, Y. Pei, P. Moulin, and S. Yan. Multilevel depth and image fusion for human activity detection. *IEEE Transactions on Cybernetics*, 43(5): 1383–1394, Oct 2013. 15, 17, 22, 26, 81, 84

[89] U. M. Nunes, D. R. Faria, and P. Peixoto. A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier. *Pattern Recognition Letters*, 99:21 – 31, 2017. 15, 21, 25, 26, 36, 75, 81, 84, 88, 96

[90] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human

skeletal action recognition. *Journal of Visual Communication and Image Representation*, 25(1):24 – 38, 2014. 89

[91] D. Ortega-Anderez, A. Lotfi, C. Langensiepen, and K. Appiah. A multi-level refinement approach towards the classification of quotidian activities using accelerometer data. *Journal of Ambient Intelligence and Humanized Computing*, Oct 2018. 32, 86

[92] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. 32, 33, 40, 41, 42

[93] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22 (2):199–210, 2011. 33

[94] G. Parisi, C. Weber, and W. S. Self-organizing neural integration of pose-motion features for human action recognition. *Frontier in Neurobotics*, 9, June 2015. 21, 28, 81, 84

[95] E. Parzen. On estimation of a probability density function and mode. *The Annuals of Mathematical Statistics*, 33(3):1065–1076, 1962. 97

[96] L. Piyathilaka and S. Kodagoda. Gaussian mixture based hmm for human daily activity recognition using 3d skeleton features. In *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, pages 567–572, June 2013. 27, 81

[97] R. Poppe. A survey on vision-based human action recognition. *Image Vision Comput.*, 28(6):976–990, 2010. 18

[98] L. L. Presti and M. L. Cascia. 3d skeleton-based human action classification: A survey. *Pattern Recognition*, 53:130 – 147, 2016. 87

[99] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986. 88

[100] R. Robotics. Baxter robot. https://www.rethinkrobotics.com/baxter/tech-specs/, 2016. Accessed: 28-02-2019. 126

[101] A. Sargano, P. Angelov, and Z. Habib. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Applied Sciences*, 7(1):110, 2017. 28

[102] J. Shan and S. Akella. 3d human action segmentation and recognition using pose kinetic energy. In *2014 IEEE International Workshop on Advanced Robotics and its Social Impacts*, pages 69–75, Sept 2014. 25, 27, 81, 82, 88, 89, 95, 96

[103] J. Shell. *Fuzzy Transfer Learning*. PhD thesis, De Montfort University, 2013. 31, 32, 42

[104] J. Shell and S. Coupland. Towards fuzzy transfer learning for intelligent environments. In *Ambient Intelligence*, pages 145–160. Springer Berlin Heidelberg, 2012. 4, 88

[105] J. Shell and S. Coupland. Fuzzy transfer learning: Methodology and application. *Information Sciences*, 293:59 – 79, 2015. 4, 7, 30, 32, 33, 39, 41, 43, 44, 45, 117

[106] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244, 2000. ISSN 0378-3758. 3

[107] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 30

[108] E. M. Smith, J. Smith, P. Legg, and S. Francis. Predicting the occurrence of world news events using recurrent neural networks and auto-regressive moving average models. In *Advances in Computational Intelligence Systems*, pages 191–202, 2018. 112

[109] F. Stulp, L. Herlant, A. Hoarau, and G. Raiola. Simultaneous on-line discovery and improvement of robotic skill options. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1408–1413. IEEE, 2014. 4

[110] T. Subetha and S. Chitrakala. A survey on human activity recognition from videos. In *2016 International Conference on Information Communication and Embedded Systems (ICICES)*, pages 1–7, 2016. 25

[111] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 16

[112] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. In *Proceedings of the 16th AAAI Conference on Plan, Activity, and Intent Recognition*, AAAIWS'11-16, pages 47–55. AAAI Press, 2011. xviii, 7, 15, 20, 75, 78, 81, 102, 103

[113] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *2012 IEEE International Conference on Robotics and Automation*, pages 842–849. IEEE, 2012. 7, 15, 16, 22, 26, 73, 81

[114] S. Suresh, K. Dong, and H. Kim. A sequential learning algorithm for self-adaptive resource allocation network classifier. *Neurocomputing*, 73(16): 3012 – 3019, 2010. 86

[115] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 30, 36

[116] M. A. Tahir, J. Kittler, and A. Bouridane. Multilabel classification using heterogeneous ensemble of multi-label classifiers. *Pattern Recognition Letters*, 33(5):513 – 523, 2012. 65, 74

[117] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 270–279. Springer International Publishing, 2018. 4, 28, 29, 30

[118] S. Tan, K. C. Sim, and M. Gales. Improving the interpretability of deep neural networks with stimulated learning. In *Workshop on Automatic*

Speech Recognition and Understanding (ASRU), pages 617–623. IEEE, 2015. 4

[119] T. Tommasi, F. Orabona, and B. Caputo. Learning categories from few examples with multi model knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):928–941, 2014. 30

[120] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2009. 26, 41

[121] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660. IEEE, 2014. 24, 28, 29

[122] U.S. Department of Health and Human Services. Accelerating adoption of assistive technology to reduce physical strain among family caregivers of the chronically disabled elderly living at home. https://aspe.hhs.gov/report/accelerating-adoption-assistive-technology-reduce-physical-strain-among-family-caregivers-chronically-disabled-elderly-living-home, 2012. Accessed: 26-07-2019. xvi, 2, 3

[123] R. Vatani Nezafat, O. Sahin, and M. Cetin. Transfer learning using deep neural networks for classification of truck body types based on side-fire lidar data. *Journal of Big Data Analytics in Transportation*, 1(1):71–82, 2019. 30

[124] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2014. 25

[125] L. . Wang and J. M. Mendel. Generating fuzzy rules by learning from examples. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(6): 1414–1427, 1992. 123, 124

[126] Y. Wang, X. Han, Z. Liu, D. Luo, and X. Wu. Modelling inter-task relations to transfer robot skills with three-way rbms. In *2015 IEEE International*

Conference on Mechatronics and Automation (ICMA), pages 1276–1282. IEEE, 2015. 5

[127] Y. Wang, X. Jiang, R. Cao, and X. Wang. Robust indoor human activity recognition using wireless signals. *Sensors*, 15(7):17195–17208, 2015. 17

[128] Y. Wang, S. Cang, and H. Yu. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, 137:167 – 190, 2019. 15, 56

[129] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu. Concurrent action detection with structural prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3136–3143. IEEE, 2013. 25

[130] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), May 2016. 3, 4, 5, 41, 63

[131] J. Wen and Z. Wang. Learning general model for activity recognition with limited labelled data. *Expert Systems with Applications*, 74:19 – 28, 2017. 89

[132] Q. Wu, G. Xu, M. Li, L. Chen, X. Zhang, and J. Xie. Human pose estimation method based on single depth image. *IET Computer Vision*, 12(6):919–924, 2018. 23

[133] L. Xia, C. C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27, June 2012. 18

[134] Y. Xiao, Z. Zhang, A. Beck, J. Yuan, and D. Thalmann. Human-robot interaction by understanding upper body gestures. *Presence*, 23(2):133–154, Aug 2014. 63

[135] X. Yang and Y. Tian. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation*, 25(1):2 – 11, 2014. 27, 68, 70, 81, 84

[136] B. Yao, H. Hagras, M. J. Alhaddad, and D. Alghazzawi. A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments. *Soft Computing*, 19(2):499–506, 2015. 28

[137] G. Yao, H. Zeng, F. Chao, C. Su, C.-M. Lin, and C. Zhou. Integration of classifier diversity measures for feature selection-based classifier ensemble reduction. *Soft Computing*, 20(8):2995–3005, Aug 2016. 74

[138] L. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338 – 353, 1965. 31

[139] C. Zhang and Y. Tian. Rgb-d camera-based daily living activity recognition. *Journal of Computer Vision and Image Processing*, 2(4), Dec 2012. 81, 84

[140] H. Zhang and L. E. Parker. 4-dimensional local spatio-temporal features for human activity recognition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2044–2049, 2011. 16

[141] J. Zhong, T. Han, A. Lotfi, A. Cangelosi, and X. Liu. Bridging the gap between robotic applications and computational intelligence - an overview on domestic robotics. In *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8, 2019. 34

[142] D. Zhou, M. Shi, F. Chao, C.-M. Lin, L. Yang, C. Shang, and C. Zhou. Use of human gestures for controlling a mobile robot via adaptive cmac network and fuzzy logic controller. *Neurocomputing*, 282:218 – 231, 2018. 20, 21

[143] W. Zhou and Z. Zhang. Human action recognition with multiple-instance markov model. *IEEE Transactions on Information Forensics and Security*, 9(10):1581–1591, 2014. 18

[144] G. Zhu, L. Zhang, P. Shen, J. Song, L. Zhi, and K. Yi. Human action recognition using key poses and atomic motions. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1209–1214, 2015. 96

[145] H. Zhu, H. Chen, and R. Brown. A sequence-to-sequence model-based deep learning approach for recognizing activity of daily living for senior care. *Journal of Biomedical Informatics*, 84:148 – 158, 2018. 29, 88

[146] Y. Zhu, W. Chen, and G. Guo. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image and Vision Computing*, 32(8):453 – 464, 2014. 22, 81, 82

[147] H. Zuo, G. Zhang, W. Pedrycz, V. Behbood, and J. Lu. Granular fuzzy regression domain adaptation in takagisugeno fuzzy models. *IEEE Transactions on Fuzzy Systems*, 26(2):847–858, 2018. 116, 118

[148] H. Zuo, G. Zhang, W. Pedrycz, V. Behbood, and J. Lu. Granular fuzzy regression domain adaptation in takagisugeno fuzzy models. *IEEE Transactions on Fuzzy Systems*, 26(2):847–858, 2018. 118

[149] H. Zuo, J. Lu, G. Zhang, and F. Liu. Fuzzy transfer learning using an infinite gaussian mixture model and active learning. *IEEE Transactions on Fuzzy Systems*, 27(2):291–303, 2019. 4, 40, 117

[150] H. Zuo, J. Lu, G. Zhang, and W. Pedrycz. Fuzzy rule-based domain adaptation in homogeneous and heterogeneous spaces. *IEEE Transactions on Fuzzy Systems*, 27(2):348–361, 2019. 4, 30, 33, 118