3D HUMAN ACTION RECOGNITION AND MOTION ANALYSIS USING SELECTIVE REPRESENTATIONS

D LEIGHTLEY

PhD 2015

MANCHESTER METROPOLITAN UNIVERSITY

3D Human Action Recognition and Motion Analysis using Selective Representations

Daniel Leightley

A thesis submitted in partial fulfillment of the requirements of the Manchester Metropolitan University for the degree of Doctor of Philosophy

Faculty of Science and Engineering
School of Computing, Mathematics and Digital Technology

December 2015

Abstract

With the advent of marker-based motion capture, attempts have been made to recognise and quantify attributes of "type", "content" and "behaviour" from the motion data. Current work exists to obtain quick and easy identification of human motion for use in multiple settings, such as healthcare and gaming by using activity monitors, wearable technology and low-cost accelerometers. Yet, analysing human motion and generating representative features to enable recognition and analysis in an efficient and comprehensive manner has proved elusive thus far. This thesis proposes practical solutions that are based on insights from clinicians, and learning attributes from motion capture data itself. This culminates in an application framework that learns the type, content and behaviour of human motion for recognition, quantitative clinical analysis and outcome measures.

While marker-based motion capture has many uses, it also has major limitations that are explored in this thesis, not least in terms of hardware costs and practical utilisation. These drawbacks have led to the creation of depth sensors capable of providing robust, accurate and low-cost solution to detecting and tracking anatomical landmarks on the human body, without physical markers. This advancement has led researchers to develop low-cost solutions to important healthcare tasks, such as human motion analysis as a clinical aid in prevention care. In this thesis a variety of obstacles in handling marker-less motion capture are identified and overcome by employing parameterisation of Axis-Angles, applying Euler Angles transformations to Exponential Maps, and appropriate distance measures between postures.

While developing an efficient, usable and deployable application framework for clinicians, this thesis introduces techniques to recognise, analyse and quantify human motion in the context of identifying age-related change and mobility. The central theme of this thesis is the creation of discriminative representations of the human body using novel encoding and extraction approaches usable for both marker-based and marker-less motion capture data. The encoding of the human pose is modelled based on the spatial-temporal characteristics to generate a compact, efficient parameterisation. This combination allows for the detection of multiple known and unknown motions in real-time. However, in the context of benchmarking a major drawback exists, the lack of a clinically valid and relevant dataset to enable benchmarking. Without a dataset of this type, it is difficult to validated algorithms aimed at healthcare application. To this end, this thesis introduces a dataset that will enable the computer science community to benchmark healthcare-related algorithms.

Acknowledgements

I would like to express my thanks and gratitude to my supervisors, Jamie S. McPhee and Nicholas Costen for their patience, support and guidance during this work. In particular, I am indebted to the kindness, dedication and support that Moi Hoon Yap has afforded me as Director of Studies. I appreciate the guidance and support Baihua Li provided in the early stages of this work. I am also thankful for the School of Computing, Mathematics and Digital Technology for funding my studentship.

This work would not have been possible without the help of students and staff at Manchester Metropolitan University. I appreciate those have given up their time to take part in the numerous data collection sessions I have held. In addition, I appreciate the assistance of Jessica Coulson who provided her time, knowledge and connections for obtaining data from the elderly.

I have been very fortunate to undertake this work within the Biological and Sensory Computational Group. I owe many thanks to my fellow students. I enjoyed the many conversations, discussions and scrutiny that have impacted this work. Adrian K. Davison, Brett Hewitt, Anna Mölder, Choon-Ching Ng and Ezak Shaubari, thank you all very much.

I especially appreciate and acknowledge John Darby for being a friend, office-mate, advisor and mentor. He took the time to introduce me to several concepts and techniques, most notably within the Matlab environment. Our inspiring discussions laid the foundation for this work. Further, Ryan Cunningham has been courteous and generous in sharing his knowledge and experiences whilst under taking his own Ph.D. To both John, and Ryan, thank you for your support these last few years. Finally, I'd like to express my deepest thanks to my family and friends for their continuous support and encouragement during my studies.

Publications

This thesis is based on contributions of the following publications:

1. **D. Leightley**, J. Darby, B. Li, J.S. McPhee and M.H. Yap. Human Activity Recognition for Physical Rehabilitation. In IEEE International Conference on Systems, Man and Cybernetics, pp. 261-266, 2013.

2. **D. Leightley**, B. Li, J.S. McPhee, M.H. Yap and J. Darby. Exemplar-based Human Action Recognition with Template Matching from a Stream of Motion Capture. In AIMI International Conference on Image Analysis and Recognition, pp. 12 - 20, 2014.

3. **D. Leightley**, M.H. Yap, J. Coulson, Y. Barnouin, J.S. McPhee. Benchmarking Human Motion Analysis Using Kinect One: an open source dataset. In IEEE International Conference on Signal and Information Processing, 2015.

Contents

List of Figures

List of Tables

Abbreviations

AP	**A**nterior-**P**osterior
ANN	**A**rtificial **N**eural **N**etwork
CoM	**C**entre **o**f **M**ass
DIS	**D**elegate **I**dentification and **S**election
DKPI	**D**iscriminative **K**ey and **P**ose **I**dentification
DMM	**D**epth **M**otion **M**ap
DTW	**D**ynamic **T**ime **W**arping
DP	**D**ynamic **P**rogramming
DOF	**D**egree **o**f **F**reedom
FPS	**F**rames **P**er **S**econd
GRBM	**G**aussian **R**estricted **B**oltzmann **M**achine
HCI	**H**uman **C**omputer **I**nteraction
MCC	**M**atthews **C**orrelation **C**oefficent
MoCap	**Mo**tion **Cap**ture
MEI	**M**otion **E**nergy **I**mage
MHI	**M**otion **H**istory **I**mage
ML	**M**edial-**L**ateral
MHV	**M**otion **H**istory **V**olumes
NN	**N**eural **N**eighbour
PCA	**P**rincipal **C**omponents **A**nalysis
RF	**R**andom **F**orests
ROI	**R**egion **O**f **I**nterest
ROC	**R**eceiver **O**perating **C**haracteristic
SD	**S**tandard **D**evation
SPPB	**S**hort **P**hysical **P**erformance **B**attsery

SVM	**S**upport **V**ector **M**achines
TUG	**T**imed-**U**p-and-**G**o
WCSS	**W**ithin-**C**luster-**S**um-of-**S**quares

Chapter 1

Introduction

In this chapter a brief introduction to the field of human motion analysis, representation and recognition is given (expanded further in Chapter 2). A number of important terms are defined and the thesis objectives are stated.

1.1 Introduction

Human action recognition has been actively researched since the early 1980s, this is due to its promise in many application domains, including surveillance, entertainment, healthcare and human-computer interaction [6]. The goal of human action recognition is to recognise human motion, both in real-time and after-the-fact from an unknown sequence of motion capture (MoCap) data. It is possible to identify and distinguish different actions because the brain is capable of both learning new actions and recognising them. However, in computer vision, this same problem has proven to be one of the most difficult and lasting challenges in the field. Given the current state-of-the-art [6–9], a successful algorithm for human action recognition requires one to define the problem with a more specific focus, notably on representation, interpretation and analysis.

Human action is inherently complex, however it is possible to decompose body motion into three different types of action groups. The typical objective of human action recognition is to classify actions that belong to specific "action" groups. Following the taxonomy of Aggarwal and Ryoo [7], actions can be categorised into three groups: *gestures*, *actions* and *activities*. Broadly these can be defined as follows:

- *Gestures* are fundamental atomic components describing the meaningful motion of a participant. For example, "extending a leg" and "raising an arm".
- *Actions* are single activities that may be composed of multiple gestures arranged in temporal order. For example, "walking", "waving" and "drinking".
- *Activities* are complex action sequences comprised of a combination of actions. For example, "Timed-Up-and-Go" is an activity consisting of multiple actions, such as chair rise, walking and turning.

Anthropometric and performance variations differ with participant groups, notably within the elderly population. Understanding the inter-/intra-class variation between these groups can aid in developing tailored solutions. However, the variations are so subtle that it is very difficult to track these changes with the human eye alone. While many people remain healthy, active and engaged into later life, studies have indicated that a minority, approximately 9% of the elderly population, defined as above the age of 60 years, suffer from age-related illnesses such as frailty [10]. Frailty relates to the general decline in multiple body systems, which leaves participants vulnerable to illness or trips and falls. Frailty is an indicator of general health and well-being, it is usually assessed by asking the participant to perform several standardised test (*e.g.* walk back and forth, sit to stand) during which a clinician observes the activity for stability, duration, coordination and posture control resulting in clinical outcome measure [11–13]. However, only a handful of works have sought to unite Computer Vision approaches with healthcare methodology to provides a more informative decision of the participants performance when understanding a range of clinically relevant motions, such as Wang et al. [14], Prochnow et al. [15] and Galnaa et al. [16]. In the majority of cases, these approaches provide a single indicator instead of a detailed analysis, which would provide clearer detailed supportive of clinical measures.

In this thesis, three problems are considered. Firstly, human action recognition; in which the focus is placed on recognising MoCap solely on the characteristics within the patterns of motion they exhibit. This approach differs from other techniques of human action recognition, such that the proposed methods seek to address viewpoint difference, style diversity, anthropometric variations and execution-rate variations that are ever present in human motion by use of efficient representations. Secondly, motion analysis and

evaluation; where identification of different participant groups mobility based solely on decomposing the spatio-temporal relationship of human skeleton. Finally, uniting these two solutions a fully realised system for human action recognition and analysis is developed and deployed in the field. These include benefits such as a novel clinical assessment tool to provide a novel low-cost for stability and mobility assessment. Its main features include real-time performance, ability to classify a diverse range of clinically relevant actions/activities and detect subtle motion variations between participant groups.

1.2 Motivation

The ability to detect, track, recognise and analyse human motion is advantageous for a wide range of high-level applications that rely on a representations extracted from visual input. Interacting with humans and understanding their activities are at the core of many problems in human-computer interaction (HCI). During the past 30 years, many approaches have been proposed to address these problems [6, 8, 9]. However, these proposals are far from suitable for practical application. Due in part to the challenging task of tracking, detecting and analysing human motion, which in part is due to a reliance on noisy low-level indiscriminate features.

Some examples of applications that could benefit from reliable efficient human action recognition are:

- Automated surveillance systems that monitor the general public, used in places such as airports, government buildings and banks. Applications to monitor and detect suspicious activity without human interference have yet to be realised. Having an automated solution vastly improves the detection of suspicious activity as it reduces the possibility of human misinterpretation and misunderstanding.
- Safety systems for detecting vulnerable users, most notably the very young or the elderly. Systems to monitor users in and around busy train stations or on cars to warn others of possible danger or for security monitoring.
- Health monitoring and preventative care for patients. Applications to seamlessly detect and track humans within their own environment. Coupled with analysis

of typical every day activities. For example, a system to monitor the elderly and alert the local hospital if they have a fall or trip.

Currently, to the author's knowledge no framework exists that can perform human action recognition reliably for the above applications. Although the problem as a whole remains unresolved, specific focus should be made to address the associated challenges to classification and recognition.

One of the main challenges is how to efficiently extract and represent features from raw data. These features should be distinctive and uniquely represent each action class (where a class represents a specific action, for example *walking* or *jumping*), and be similar in nature to those actions from the same class. Nevertheless, this is a difficult task due to viewpoint difference, style diversity, anthropometric variations and execution-rate variations exhibited by participants that can lead to large intra-/inter-class variations. In addition, occlusion and overlapping limbs will further introduce noise to the process. In the literature, feature detection and representation are two main components in representing MoCap for classification and recognition. In particular, several approaches focused on encoding the articulated skeletal from each frame of MoCap to combine or link into discriminative representative elements [17–20]. These approaches have demonstrated the ability to handle variability in data volumes, however are computationally expensive when the time to process is considered.

More recent efforts have focused on modelling MoCap with a limited number of representative elements, otherwise referred as the *exemplar* paradigm [21]. A small number of "exemplars" are generated by generalising the spatio-temporal variation in the state space (*e.g.* [21], [22], [23]). If we consider that by representing the most descriptive and representative elements of each action with the exemplar paradigm, the number of samples required for training will be greatly reduced. Further the reduction for the rate of false-positives by reducing the intra-/inter-class variations would aid in the recognition process. However, the performance of various descriptors depends on the selection of appropriate action sequences, for example selecting the most optimum performance of a "walk". Obviously, this is an unassailable task as we inherently all walk differently and a "walk" between two persons can be different in gait and speed.

The advancement of imaging technologies in the past decade, such as the depth sensor, has presented a number of new solutions to marker-less-/marker-based tracking when

compared to traditional RGB-based approaches [6]. One such approach has revolutionised the field of marker-less tracking, Shotton et al. [24] introduced the underlying Microsoft Kinect^{TM}360/One (referred as "Kinect 360" or "Kinect One") sensor algorithms which are quick and efficient in predicting the 3D positions of body joints from a static depth map image provided by a depth camera, without using any temporal information. The approach uses a single depth map and applies randomised decision forest algorithm to automatically detect, segment and locate pre-defined anatomical joints on the body of the user in front of the device. But, of further interest is the capability of the algorithm to provide 3D orthogonal coordinate locations - otherwise referred to as MoCap (similar to that extracted from marker-based approaches).

In a clinical setting the most prevalent methodology suggests using one or a combination of intrusive sensors, such as body-based accelerometer or markers. Still, in recent years the computer vision community has proposed an array of solutions to aid in the decision-making and provide simple kinematic measures within the healthcare sector. These works have predominantly focused on depth sensor technology, such as the Microsoft Kinect, which has demonstrated a capability for tracking in home and healthcare settings (*e.g.* [16, 25–27]).

1.3 Problem Statement

The problem statement addressed by this thesis is concisely stated as follows:

Human motion is inherently complex, while techniques exist to extract anatomical tracking information, understanding and analysis of the data is a challenging and difficult task which remains unresolved. The process should be optimised and refined for use in real-world settings. Firstly, a method that represents high-level human motion effectively and efficiently as well as handling intra-/inter-class variations would improve the recognition process. Secondly, a framework that incorporates recognition to analyse and quantify human mobility for use in a clinical setting would provide greater understanding of the motions themselves.

1.4 Contributions of this Thesis

The main contributions of this thesis as follows: (i) Feature extraction and representation (ii) Recognition of MoCap (iii) Motion analysis of human mobility. This thesis focuses on human action itself and does not explicitly consider the context such as background, interactions between participants or objects. The motivation and context of this work is carefully introduced over the course of the following two chapters, but a concise list of the resulting contributions with relevant sections forward referenced is given below.

- Chapter 4 introduces two novel approaches for human action recognition using the exemplar paradigm, feature selection and pose ranking. Overall both approaches identify and extract the key poses of marker-based MoCap to provide a more compact and efficient representation.
- Chapter 5 introduces a novel framework for identifying and selecting key poses from marker-based MoCap, and a framework for recognising human action in *real time*.
- Chapter 6 introduces a detailed analysis of the Microsoft Kinect 360 suitability for detecting typical daily movements utilised in a clinical setting using state-of-the-art machine learning techniques and marker-less tracking techniques.
- Chapter 7 presents a detailed analysis and evaluation of the Microsoft Kinect One depth sensor to identify age-related mobility changes between age groups using marker-less tracking techniques.
- Chapter 8 presents an application framework utilising marker-less depth sensor technology to detect mobility concerns using a digitalised assessment framework. This framework provides clinical feedback to aid in prevention of mobility-related disease in later life.

Chapter 2

Literature Review

This chapter explores the state-of-the-art work in the field of feature extraction, feature representation, human motion analysis and recognition. Supplemented with a concise discussion on the field and highlighted areas of possible research interest is presented.

2.1 Introduction

In this chapter an overview of the field of articulated human feature representation, motion analysis and recognition is presented. Due to the large volume of work in this area, the overview is not intended to be exhaustive but rather define the areas for which this thesis attempts to contribute. Comprehensive reviews of the literature can be found in a number of review papers (*e.g.* [6–9]). This literature review starts with an analysis of feature extraction and representation methods, which is the first step in any recognition and/or analysis framework. Instead of working with raw video sequences (*e.g.* depth map images and RGB) which contain many pixels and an abundant source of high-dimensional complex data, it is necessary to extract a set of features which are considered as more compact representation of the input data. This process is referred to as *feature extraction*. Then existing techniques for feature extraction and representation of MoCap data extracted from depth map images are discussed. Since human motion is inherently complex, and will generate different vectors and features we need to generate a vector representation for MoCap sequences which is consistent for use with different motion capture (MoCap) types and action types. The process to generate a

unique-feature based representations based on extracted features is *feature representation*. Finally, while human motion can be represented in the form of vector-features, analysis and understanding of the temporal/spatial relationship to allow for robust and efficient understanding is referred to as *motion analysis*.

2.2 Feature Extraction and Representation

The last 20 years has seen a large shift in the types of image modalities utilised, early works focused on extracting two-dimensional features from RGB video sequences. Most recently, the community has shifted to extracting three-dimensional features from depth map images (obtained via depth sensors). Each aspect, and the major works of each domain, is discussed hereafter.

2.2.1 Feature Extraction and Representation from RGB Videos

Broadly following the taxonomy defined by Poppe [6], feature extraction techniques for RGB video sequences can be divided into two categories; *global* and *local* representations. Global representations encode extracted features as a whole, and are obtained in a top-down structure. Problematically, they are very sensitive to environmental variables such as noise and occlusion. Local representations describe the extraction and collection of local features specifically in the spatial-temporal domain, and are obtained in a bottom-up structure. Local representations are less sensitive to noise and partial occlusion yet rely heavily on the accuracy of local interest detectors. These categories are described briefly to provide context to the domain and evolution of the state-of-the-art into depth imagery.

2.2.1.1 Global Representations

Global representations usually involve two steps; a person is localised in the image using background subtraction and/or tracking. Then, the region of interest is encoded as a whole, which results in an image descriptor being formed. Common global representation approaches are obtained using silhouettes, edges or optical flow.

One of the early works to use silhouettes is by Bobick and Davis [28]. The silhouette is extracted from single-view two-dimensional images and aggregated differences between subsequent frames of the sequence is computed. The differences are combined to generate a binary motion energy image (MEI), which identifies where motion occurs. Also, a motion history image (MHI) is constructed to indicate motion intensity. Ziaeefard and Ebrahimnezhad [29] extended the use of silhouettes, by modelling the spatio-temporal dynamic of human action captured encoded in normalised polar-histograms. The authors describe human action within normalised polar-histograms in two ways; firstly, the spatial element, which is the body at each time step. Secondly, temporal element which is the evolution of body poses over time. These elements build a top-level histogram representation of the sequence. Image skeletonisation of silhouettes was introduced by Fujiyoshi and Lipton [20], in which a star representation is extracted based on motion points in relational to a central point. Chen et al. [30] extended this technique further. The authors extracted contours to form a star skeleton, which describes the angles before a reference line, and the lines from the centre (star point) and outlying points of the contour.

A single, two-dimensional view of a person may not always be suitable for detection and representation. Several works have sought to utilise multiple cameras, with silhouettes extracted from each. Using this system, Zhu et al. [31] extracted silhouettes from each camera to represent each time period. Cherla et al. [32] used two orthogonally placed cameras and combined the features of both. This resulted in a somewhat view-invariant approach, however fails when certain limbs are occluded (such as arms or legs). Weinland et al. [33] combined silhouettes from multiple camera points into a three-dimensional voxel model. These models are very accurate, so long as they are calibrated accordingly. The authors proposed the use of motion history volumes (MHV), which is an extension of Bobick and Davis [28] motion intensity images into a three-dimensional domain. To overcome view invariance, each volume is aligned using Fourier transforms on the cylindrical coordinate system around the medial axis.

Several works have sought to utilise motion information instead of silhouettes. Selecting a region of interest (ROI), it can be described using optical flow techniques, the pixel-wise oriented difference between subsequent frames. Liu et al. [34] calculated the optical flow using spatial-temporal patches. Due to large volumes of data, they extract a Pyramidal Motion Feature and select the most discriminative frames based on a ranking scheme. Ali

and Shah [35] utilise optical flow to obtain a range of kinematic features. These include divergence, vorticity and gradient tensor features. Principle Component Analysis (PCA) is applied to determine dominant kinematic modes.

Global representations are troubled by noise, partial occlusion and viewpoint differences. By dividing the ROI into a fixed spatial or temporal grid, some of these limitations can partly be overcome. Kellokumpu et al. [36] obtained a local binary pattern along the temporal dimension of a sequence and store a histogram of non-background differences in a spatial grid. This allows each cell in the grid to provide an observation of the image locally. Optical flow in a grid-based representation is used by Vrigkas et al. [37]. They propose a learning-based framework to describe an image sequence by using a time-series of optical flow of motion features.

2.2.1.2 Local Representations

Local representations focus on detecting local spatial-temporal interest points first, then, local patches are calculated around these points [38–40]. Features that describe the observed video sequence are a collection of these detected local patches. Compared to global representations described previously, local representations are less sensitive to noise and full/partial occlusion, and do not require background subtraction or point tracking [41, 42].

FIGURE 2.1: Space-time volume of stacked silhouettes captured by a single camera (©IEEE 2007. Reprinted, with permission, from Gorelick et al. [1]).

Laptev and Lindeberg [43] extended the Harris corner detector [44] to include a space-time element (three-dimensional). The extension detects local structures of the image that have significant local variations in both space and time. A drawback to these types of approaches is the stability in the interest points. Dollar et al. [45] applied Gabor filtering on the spatial and temporal dimensions individually. The number of

interest points is adaptive based on the neighbourhood in which local minima is set. This provides more stable interest points. Blank et al. [1, 46] proposed the stacking of silhouettes over a given time sequence to form a spatio-temporal volume (as seen in Figure 2.1), with a Poisson equation used to derive local space-time saliency and orientation features. By using global features that are computed by calculating the weighted moments over the local features robust localisation, alignment and background subtraction are required to work efficiently.

Schüldt et al. [47] calculated patches of normalised derivatives in space and time to provide local descriptors. This provided a more robust feature representation. Laptev et al. [48] utilised local grid-based descriptors to summarise local observations within grid cells, allowing them to ignore small spatial and temporal variations. They use Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow descriptors. Dou and Li [49] extended the use of Scale Invariant Feature Transform (SIFT) and motion temporal templates with spatio-temporal interest points based appearance descriptor. The authors unite MHI and MEI with spatio-temporal points detector. With these representations they are transformed into a three-dimensional SIFT.

Similar to global representations, local representations use grid-based techniques to bin the patches into spatial or temporal elements. Spatially, Zhao and Elgammal [50] bin a spatial grid of local descriptors in histograms, with different levels of granularity. Each patch is weighted according to their current temporal distance to the current frame. Laptev and Pérez [51] use a temporal grid instead of a spatial grid. The authors use HOG and optical-flow, obtained from the interest points to form a spatial-temporal grid-based representation.

Spatial-temporal grid-based representations model the relationship between local descriptors. Yet, they often contain irrelevant and erroneous information. Scovanner et al. [52] developed a word co-occurrence matrix, and iteratively compared all the different pairs of words with similar co-occurrences until a threshold is reached. This ultimately leads to a reduced codebook size, as similar actions are likely to generate similar word distributions. Patron-Perez and Reid [53] represents features as binary variables, which indicate the presence of a code word. The framework approximates the joint distribution of features using first-order dependencies. To enable analysis of these features a graph between all pairs of features is formed.

2.2.2 Feature Extraction and Representation from Depth Sensors

The introduction of low-cost RGB-D sensors (*e.g.* Microsoft Kinect 360/One, ASUS Xtion PRO) has had a positive effect in the research domain of action classification and human motion analysis by providing both depth map images and RGB images simultaneously. These devices, otherwise referred to as range sensors, output a two-dimensional array of distances (typically millimetres) which correspond to each pixel point in the image. Research focusing on feature extraction and representation has been extensively explored for RGB images, however research on depth maps has been limited in scope.

Early works focused on extracting two-dimensional silhouettes of simple human body shapes, then model the evolution of silhouettes in the temporal domain. In a two-dimensional image (as discussed previously) this is a difficult task. However, in a depth map image, the silhouette of a person can be extracted more easily and with greater accuracy [9]. This is because the depth image provides a greater number of descriptive features to enable a more robust detection of the silhouette. Depth-map based methods rely mainly on features, either local, or global, extracted from the space time volume.

Several works have sought to utilise MHI and space-time volume approaches of the depth map image silhouettes to provide a more robust representation. Li et al. [54] extended the bag-of-words technique for use in the three-dimensional domain. They sample a bag of three-dimensional points on the planar projection of the three-dimensional depth map image to characterise a set of salient postures, which corresponds to the nodes in the action graph. Importantly, the number of planar projections used controls the number of points. However, due to occlusion and noise within the depth map images, loss of spatial context information between interest points is observed. This makes it very difficult to accurately sample interest points given the geometry and motion variance across different persons and environments. To address this, Vieira et al. [55] proposed the Space-Time Occupancy Patterns feature descriptor. The depth map sequence is represented as a 4D space-time grid with a saturation framework employed to enhance the roles of the sparse cells. These cells typically consist of interest points and contours of the silhouettes. Wang et al. [56] proposed the Random Occupancy Pattern to address the issue of noise and occlusion by treating the three-dimensional action sequence as a 4D shape. These were constructed by randomly sampling three-dimensional sub-volumes

at different locations with different sizes. The framework was further complimented by a weighted random sampling scheme enabling effective identifying of the participants' contours by identifying a player mask.

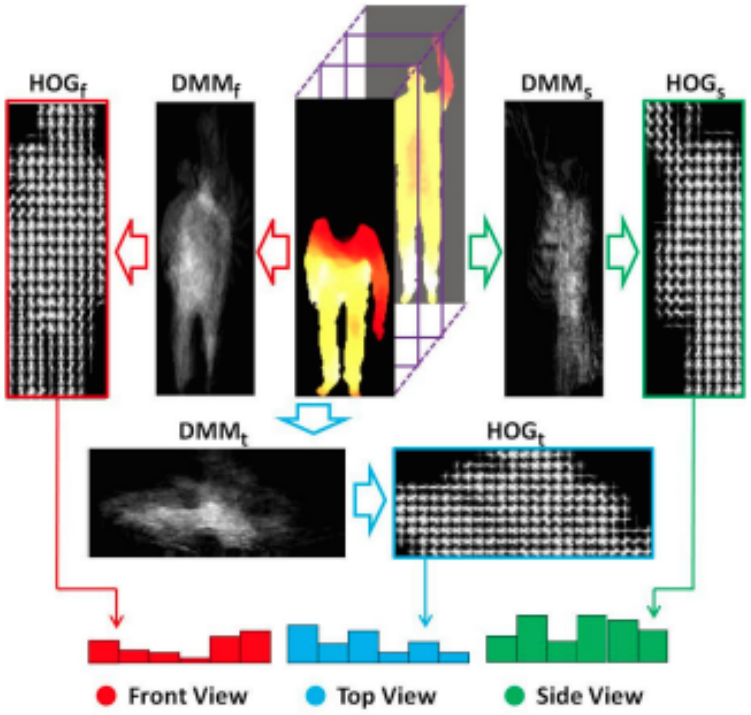

FIGURE 2.2: Depth Motion Map framework and visual examples (©ACM 2012. Reprinted, with permission, from Yang et al. [2]).

Yang et al. [2] introduced Depth Motion Maps (DMM), which are formed from projecting onto three orthogonal planes (as seen in Figure 2.2). The DMM stacks the aggregated MEI, computed in the temporal domain of a video sequence for each plane. HOG is employed to describe the DMM. Ni et al. [57] proposed a Three-Dimensional Motion History Image (3D-MHI), which extends the original technique [28] with two additional depth change induced MHI. A forward and backward motion image which encode the history of a sequence of images is computed and combined. Wu et al. [58] proposed extending MHI by combing MHI with Gait Energy Information (GEI) and inversed recording for use with depth map images. The GEI compensates for non-moving regions and multiple-motion-instance regions. This is complimented by inversed recording which assigns a larger value at initial motion frames instead of the last motion frames. The extended technique has been demonstrated to outperform the original MHI [28] on action-oriented datasets.

Most recently, Oreifej and Liu [59] introduced a 4D histogram descriptor which encodes the distribution of the surface orientation of the 4D volume of time, depth and spatial coordinates. As a depth map image (or sequence) is a depth function of time and space, histograms are constructed. While depth map images can provide descriptive and useful information, Zhang and Parker [60] proposed a 4D local spatio-temporal descriptor to encode human activities. The 4D feature is a weighted linear combination, combing a visual and a geometric component. The method concatenates per-pixel responses and their corresponding gradients within a spatial-temporal window into a high-dimensional feature vector (which contains 10^5 elements). A clustering technique, k-means [61] is employed as a dimensionality reduction technique to separate and group actions into vocabularies.

Other approaches have focused on extracting silhouettes and generating other forms of representations. Jalal et al. [62] utilise Random transform ($\mathcal{R}$) to compute a two-dimensional projection of a silhouette which has been extracted from a depth map image along a specified view direction. $\mathcal{R}$ transform-to-transform are employed to convert the two-dimensional projection into a 1D profile for each frame. A representation is then encoded for each frame of the sequence. Wang et al. [18] introduced the *actionlet* ensemble model of depth map images. The authors sought to recognise interaction between objects and/or other humans. This was achieved by capturing the inter-/intra-class variations for a number of spatio-temporal features between the subject and the object/human. Fanello et al. [63] extend the classic form of HOG to encode global representation. The Global Histogram of Oriented Gradients describes the appearance of a depth map silhouette without splitting it into cells (or regions). The gradient of the depth map image shows the highest response on the contours of the subject, thus indicating the posture of the subject.

2.2.3 Skeleton-based Feature Extraction and Representation

The human body is an articulated system of rigid segments connected by joints, with human motion being a continuous evolution of the spatial configuration. The study of skeleton-based feature extraction and representation dates back to early work by Johansson [64]. This work demonstrated the ability to recognise a large set of actions solely from the joint positions. The concept of joint positions has been explored extensively

ever since. In contrast to depth map based approaches, the majority of skeleton-based approaches model the temporal dynamics. One fundamental reason is the natural evolution of corresponding skeletons across time which can be used to model motion velocity (chapter 3 introduces the main concepts of MoCap). There are three prominent approaches to obtain skeletons: marker-based MoCap systems, monocular or multi-view colour images and depth maps [38, 54, 65, 66]. Overall, MoCap data is the cleanest (the most noise free of the approaches) compared to other approaches. Yet, when using multi-view frameworks (such as colour or monocular) they provide a more stable estimated skeleton. Early works focused on testing algorithms against MoCap data, however there has been a shift in recent works to test on noisy skeleton data obtained via depth map images. This is due to a number of factors such as the simplicity in setting up the system, usability and ability to be used in a large number of environments.

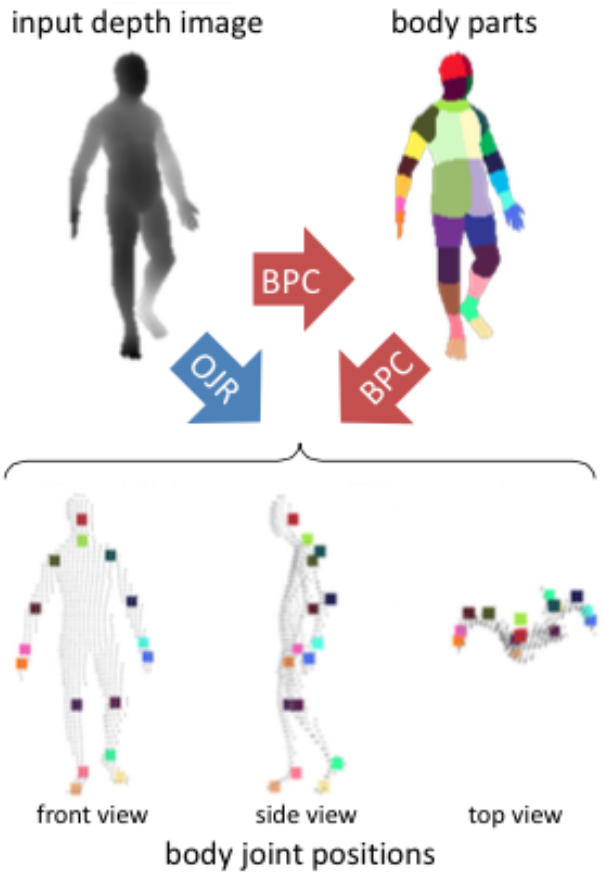

FIGURE 2.3: Three-dimensional joint estimation and extraction from single depth images (©IEEE 2012. Reprinted, with permission, from Shotton et al. [3]).

Of major importance, in 2011 Shotton et al. [24] (further discussed in [3]) proposed a pose estimation and detection framework for extracting three-dimensional body joint

locations from single depth map images. The authors take a depth map image of the area in front of the range sensor and apply Offset Joint Regression (OJR) algorithms to automatically detect, segment using Beyond Pairwise Clustering (BPC) and locate anatomical joints on the body of the user and predict the three-dimensional coordinate location. A total of 20 three-dimensional joint positions are extracted, *e.g. hip (left and right)*, *shoulder (left and right)* and *spine*, an example of which can be observed in Figure 2.3. The success of the tracking algorithms has resulted in the development of the Microsoft Kinect sensor range, which offers the ability to access skeletal data with ease.

Lv and Nevatia [67] designed a set of spatially local features based on single joints, or a combination of joints (that are linked in the hierarchical structure). The authors introduce a framework that is capable of encoding features in real-time. Normalisation is employed to avoid dependency on the environment, body orientation, pose size and view point. Their work suggests that using just pose vectors may, in some cases, cause a loss of some relevant information and reduce the discriminative power of the encoded features. The authors consider three motion types that require motion from different primary body parts (*legs and torso*, *arm (left and right)*, *head*). This results in a high dimensional 141 vector, including the full pose and multiple feature types. To efficiently represent each feature and action class a Hidden Markov Model (HMM) [68] is constructed to model the temporal dynamics. An ensemble of HMM are combined into a weak classifier, namely AdaBoost [69] to improve the feature discriminative power. To support more complex action sequences Xia et al. [70] proposed the Histogram of 3D Joint Locations (HOJ3D). The HOJ3D encodes the spatial occupancy information relative to a pre-defined joint (ideally hip centre or skeletal root). A modified spherical coordinate system (on the pre-defined joint) partitions the three-dimensional space into n bins. Interestingly, to handle scale invariance, the radial distance is not considered. However, this approaches struggles when the participant is not directly facing the range sensor, or the tracking algorithm is unable to robustly track a skeleton.

Briefly, the above methods are limited to single (or simple) gestures; this is due to a failure to model the hierarchical motion evolution. To address this, Koppula et al. [71] evaluated the interaction between a human and an object. The authors encode a Markov Random Fields using a spatio-temporal sequences. They encode two types of nodes, namely object nodes and sub-activity-nodes, and edges representing the relationship

between an object and the human. The modelling of human motion in a hierarchical structure allows for complex activities that include interaction with single and/or multiple objects. The basis of the work is to utilise a skeletal tracking algorithm to assess a joint-by-joint interaction with objects in the scene. Other works such as [72] extend the human-object interaction further by explicitly modelling the motion interaction between the human and object using joint kinematics.

Recently, works have sought to combine depth map images and RGB data in the form of point clouds. This approach is very new area of research, therefore only a few approaches exist. Yang and Tian [73] introduced the EigenJoints features which are extracted from RGB-D sequences. Posture (f_{cc}), motion (f_{cp}) and offset (f_{ci}) features are extracted. The posture and motion features encode the spatial and temporal configuration with pairwise joint differences within a single frame, and intra frame were computed. The offset feature initial pose is neutral (and 'clean'). Due to the high dimensionality of the feature set, PCA is applied to remove irrelevant data and reduce noise to obtain the EigenJoints descriptor.

2.3 Motion Capture: Datasets and Benchmarking

In the computer vision community there exist multiple datasets composed of different modalities such as MoCap, RGB images and depth images. For human MoCap data these are composed entirely of two modalities, namely marker-based systems such as Vicon and marker-less systems such as Microsoft Kinect sensor. Marker-based systems typically place retro-reflective markers on key anatomical joints (typically 40/50 markers) on a participants body, with infrared lights mounted around the room to capture the motion. Orientation and angles are then extracted to represent human motion, section 3.1 discusses these theories in detail. It is important to acknowledge that there are other forms of Vicon, such as instead of using markers colour stickers, however these have not been utilised extensively by the Computer Science community. Marker-less systems rely solely on pose estimation and skeletal extraction techniques. These systems use either RGB or depth-map images to identify the body of interest and segment the body into key anatomical areas. Skeletal tracking algorithms are applied to the points-of-interest to extract MoCap data.

Existing datasets (*e.g.* [74–76]) are vast in number and seek to address a specific area of research (*e.g.* daily living, first person, and gesture). They are captured using marker-based, marker-less or RGB systems, which can be defined into three main categories. Firstly, those that are action recognition datasets such as marker-less G3D dataset [5] which contain simple action sequences obtained in a controlled environment. Secondly, surveillance datasets such as RGB-based i-Lids dataset [77] which are obtained in realistic environments such as airports and ports. The third type of movie datasets are obtained from movie scenes such as RGB-based Hollywood2 [78]. This thesis is exclusively interested in the first category, which are composed of marker-based and marker-less systems. Selecting the most suitable dataset for benchmarking healthcare-based algorithms is a challenge, as can be observed in the limited types of datasets available, a discussion is presented hereafter.

2.3.1 Marker-based Datasets

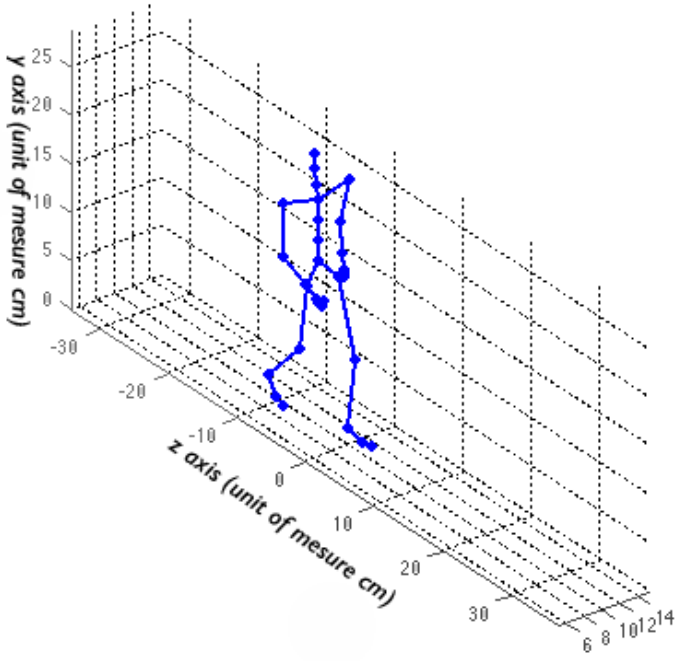

FIGURE 2.4: An example of a participant *walking*. Extracted from the marker-based Carnegie Mellon University Motion Capture Database [4].

Marker-based systems require markers to be placed on the participant at anatomically significant locations. Using multiple cameras, these markers are tracked resulting in Euler angles, rotations and orientations (otherwise referred to as MoCap data) that are

relative to the camera coordinate system. There are numerous marker-based datasets [79–82], which are typically aimed at benchmarking pose estimation, gesture, gaming and security systems. Marker-based datasets are very challenging to construct and require a vast amount of hardware and expertise.

One of the earliest publicly available MoCap datasets was developed by Sigal et al. [83], who introduced the HumanEva dataset. It contained synchronised RGB video and MoCap to support the development of new articulated motion and pose estimation algorithms. Over 40,000 frames of data were collected at 60Hz, encompassing typical every-day tasks including *walking* and *drinking from a cup*. This dataset has been popular in comparing pose estimation techniques with a ground-truth (the MoCap data). Van Der Aa et al. [84] further complimented this dataset with the introduction of the UMPM benchmark dataset, a multi-person dataset which contains synchronised RGB images and MoCap. The dataset focuses on human interaction, with the authors paying particular attention to recording scenarios that involve human-to-human interaction, and human-to-object interaction.

CMU Motion Capture dataset [4] is the most popular marker-based dataset in use for benchmarking action recognition frameworks. It consists of a large amount of game-orientated trials recorded at 120Hz in a lab-based setting. It includes 2600 trials across 23 action categories captured using a marker-based Vicon system. While the number of trials and action categories are diverse, the set-up includes a rigid recording protocol and participants are sourced from a young student population. The dataset has been developed for gestures, actions and interactions in game-based scenarios. An example of a visualised sequences can be observed in Figure 2.4.

To address the lack of diversity in action execution and variable environment, Müller et al. [85] introduced the HDM05 dataset. The dataset contained a limited number of realistic fitness workout trials (1,500) captured by five participants, captured using a strict recording protocol. The dataset has been a popular for benchmarking activity recognition frameworks.

To be more realistic to home-based environments, Tenorth et al. [86] introduced the TUM Kitchen dataset which consists of multi-modality dataset including video and MoCap. Participants were captured in multiple daily living scenarios performing specific

tasks, with participants asked to perform motions as they would in the home with large amounts of freedom given.

2.3.2 Marker-less Datasets

FIGURE 2.5: An example of a participant *fighting*. Extracted from the marker-less G3D dataset [5]. Left: Standard definition RGB image. Middle: Depth map image, with participant mask identified. Right: Extracted skeleton sequence consisting of 20 joints.

Recent technological advancements have led to the availability of low-cost and easy to use image sensor technology (*e.g.* Microsoft Kinect 360/One, ASUS Xtion etc). These systems are marker-less, where the human skeleton is extracted from depth map or video sequences to provide orthogonal coordinates for specified anatomically significant landmarks on the human body. Please refer to Section 2.2.3 where skeletal extraction algorithms are introduced. Marker-less datasets are able to be construed with ease, and require very little hardware and/or software expertise.

The first publicly available dataset constructed using the Microsoft Kinect 360 sensor was by Wang et al. [18]. The MSRDailyActivity3D [18] consists of 10 participants performing daily living activities such as *eating* or *reading a book*. The dataset includes MoCap and synchronised depth map images. As is common in dataset construction, the MSRDailyActivity dataset was captured using a rigid protocol to ensure uniformity in action trials.

Building on the success of the MSRDailyActivity [18] dataset, Sung et al. [87] introduced the CAD-60 dataset, which comprise of RGB-D image sequences of humans performing activities which are recording using the Microsoft Kinect 360 sensor. The participants were given great freedom in undertaking activities as they would in the home, with difference recording environments used, such as *office space*, *bedrooms* and *kitchen*. The

dataset features 60 short RGB-D image sequences, in five recording environments, including 12 activities (*e.g.* talking on couch, relating, drinking water) by four participants. To improve the diversity in actions obtained, Koppula et al. introduced the CAD-120 dataset [88]. As in [87], participants were given the freedom to perform actions as they would in the home. This resulted in 120 RGB-D image sequences of four participants undertaking complex long daily activities such as making cereal and removing plates from a stack in a specific order.

While daily living activities are important, Bloom et al. [5] introduced the G3D [5] dataset which provides image, depth and skeleton data captured using a Microsoft Kinect 360 sensor. The dataset contains a range of typical action sequences you would expect to find in a game-based environment. A total of 200 trials across 20 categories (*e.g.* jump, run and walk) using 10 participants were recorded in a controlled lab-based setting. In addition, Fothergill et al. [76] introduced one of the largest dataset of its type, MSRC-12 Kinect gesture dataset, which comprises of image, depth and skeleton data captured using a Microsoft Kinect 360 sensor. The dataset consists of sequences of human movements, and associate gestures used in gaming environments. The data set includes 594 sequences and 719,359 frames that equates to approximately six hours and 40 minutes-collected from 30 people performing 12 gestures. The dataset is diverse in nature, however it was record in a lab-based setting.

To handle other research domains, other marker-less datasets have been introduced. To handle complex human activities and multi-human interaction, the authors of [5] introduced the G3Di [89] dataset which captured 12 participants split into 6 pairs in a multiplayer game setting. The dataset contains collaborative interactions, such as volleyball, in which one player serves and the other has to hit the ball to return and offensive interactions in which one player has to punch while another has to defend. Being able to re-identify a participant that has been observed previously is an interesting research topic, to that end, Barbosa et al. [75] introduced the RGB-D person re-identification dataset, which consists of *walking* movements captured using the Microsoft Kinect 360 sensor. Kwolek and Kepski [90] introduced the UR Fall Detection dataset which comprises of participants falling and daily living activities. Two Microsoft Kinect 360 sensors are employed with synchronised RGB, depth and MoCap extracted.

2.4 Understanding and Analysis of Human Motion

The understanding of human motion, in terms of recognition, quantification and analysis is an important area of computer vision [6–8]. The goal of understanding human motion is to automatically analyse ongoing actions (and/or activities) from unknown video sequences. In a simple case of one action execution, the objective is to correctly classify it to an action category either by using a propriety framework or machine learning (*e.g.* [17, 91]). Secondly, a system could seek to analyse the way in which the action was undertaken, such as gait speed and body sway, enabling a greater understanding of the motion kinematics.

This section follows broadly the taxonomy of Aggarwal and Ryoo [7] and will explicitly focus on three-dimensional joint features represented as MoCap . It is structured as follows: Space-time and Sequential Approaches (Single-layered) are those approaches that represent human actions and activities directly based on sequential data. *Gestures* and Semantic Approaches (Hierarchical) are those that represent high-level human activities by describing them in terms of other simpler representations, such as basic actions that build a gesture.

2.4.1 Space-time and Sequential Approaches (Single-layered)

Space-time and Sequential Approaches represent human motion and activities directly based on the sequential MoCap data itself. Such approaches consider the sequence as a class of skeletal sequences, and perform recognition from an unknown number of MoCap sequences. Various representation and matching methods have been proposed to enable accurate decision-making as to whether a sequence belongs to a specific class, and/or analysing the motion to enable better understanding [92–94].

A skeleton sequence is seen as a time sequential set of three-dimensional joints, therefore an input sequence containing an execution of an action and/or activity can be represented as a three-dimensional space-time volume constructed by analysing the inter-frame differences. Masood et al. [95] made the assumption that the first frame of any sequence is a neutral pose, then the authors' take the difference between the first frame and each subsequent frame to generate an offset feature. A generalised framework has been proposed by Yao et al. [94], where the skeleton motion is encoded by relational

pose features [96]. The features encode the geometric relations between specific joints in a single pose or a short time-ordered sequence of poses. The framework couples the task of action classification and pose estimation uniformly and performs very well compared against the benchmarks. Several works have sought to identify and group sets of joints to enable joint-to-plane distance and motion evolution. Yun et al. [97] encoded a feature that captures the geometric relationship between a joint and a specific plane which must be spanned by three joints. The feature is designed to encode specific information such as how far the right foot lies with respect to a frontal plane such as the *left knee*, *hip* and *upper torso*. This approach is capable of identifying subtle differences between actions.

Early work by Sheikh et al. [98] used trajectories to represent and classify actions. The authors represented each action as a set of 13 trajectories in a 4-D XYZT space. The authors used affine projection to obtain a normalised representation in order to measure the view-invariant similarity between two sets of trajectories. This was further complimented by Sung et al. [99] who computed each joint's rotation matrix with respect to a defined location on the torso, hand position with respect to the torso and joint rotation motion as features and used a Maximum-Entropy Markov model, which is a graphical model for sequence labelling that combines features of an HMM and maximum entropy models to learn each action class.

However, others have sought to use simple skeleton features provided by skeleton tracking algorithms. Bhattacharya et al. [100] introduced a simple vocabulary approach for gesture classification of MoCap using an Support Vector Machines (SVM). The authors zero mean the skeletal sequence to normalise the sequence and train an SVM on MoCap features. Classification is performed for each frame individually and an accumulative voting strategy determines the winning class. Further, Piyathilaka and Kodagoda [101] proposed a Gaussian mixture-based HMM to detect activities using MoCap data. To infer human activities, the authors utilise the Gaussian mixture output of an HMM to capture the relationship between joints in a sequence. Testing is undertaken offline by constructing a Gaussian mixture model for comparison against the trained model. The authors note that the skeleton data itself is descriptive to be able to robustly detect human action.

Using skeleton data as the core, Vieira et al. [102] used distance matrices as an invariant

feature for classifying MoCap data. The authors rely on representing each MoCap sequence as distance matrices. Firstly, given two postures (skeletal frames) with the same semantic structure may have different MoCap coordinates depending the body position, orientation and viewpoint. An equivalence exists if there is a rigid transformation between two poses. As such, rigid transformations are grouped and used to identify each action group . Then, an Action Graph is built based on a set of silent postures that are represented as a distance matrix. Each group of silent posture is a node on the Graph. A test sequence is compared to the Action Graph to determine the corresponding class. Barnachon et al. [103] extended the concept of Action Graphs. The authors propose an exemplar-based framework for encoded motion graphs from MoCap. MoCap is transformed into a skeleton-centric coordinate system, with spatio-temporal characteristics of each action modelled. Classification is undertaken by comparing Action Graphs of the training with testing data.

Other recognition approaches have focused on analysing a set of features based on deducing if an action has occurred if a specific characteristic has been observed. Utilising three-dimensional joint features, these approaches analyse the MoCap to measure how likely it is that the feature vector was produced by a specific action. If the likelihood between the MoCap feature and action class is above a defined confidence value, a decision is made. Utilising skeletal locations extracted using skeletal tracking algorithms, joint orientation can be computed. Importantly, joint orientation is a useful feature as it is invariant to human body size, and point of view. Sempena and Maulidevi [93] built a feature vector from joint orientation along a sequential time series and applied Dynamic Time Warping (DTW) on the features to enable human action recognition. Bloom et al. [5] concatenated five different features: pairwise joint position difference (otherwise referred to as offset), joint velocity, velocity magnitude, joint angle velocity with respect to the x-y plane and x-z plane, and a three-dimensional joint angle between selected joints. In total, a feature vector of 170 features was generated to enable recognition of gaming actions.

There only exist a few approaches that have sought to identify the user based on the pose. Sinha and Chakravarty [104] proposed a system that analysis the gait and balance of the user for person identification. Unlike in other approaches, Sinha and Chakravarty [104] extract a set of spatio-temporal dynamic and static features to describe each frame. These features are grouped to form a *template* for each user, and an SVM is trained

on each template. More recently, Kapsouras and Nikolaidis [105] introduced a k-means-based framework which identify key action segments, with each segment encoded with a frequency of occurrence histogram. The framework relies on the correct selection of sequences to be modelled, with person identification performed by a bag-of-words approach. However, the framework struggles when the complexity of the codebook increases. These types of approaches have been utilised in industry in areas such as security and gait analysis, for example Kastaniotis et al. [106] sought to recognise the gender of the participant based on their gait style. The proposal struggles when a participant intentionally alters their own gait, or copies another participants gait.

Thus far, this review has focused on action recognition frameworks that utilise marker-based MoCap as an input. These frameworks have been shown to be useful, however several works have sought to assess and evaluate the motion itself. There are many alternatives to marker-based mocap, such as accelerometers, wearable technology and force platforms [107–109]. However, due to its popularity, several works have sought to assess the ability of the Microsoft Kinect 360/One for use in healthcare applications [27].. Kargar et al. [110] utilised a depth sensor to automatically measure the physical mobility of participants. The authors analyse and classify human gait in relation to the "Get-up-and-Go-Test". Two types of features are extracted from the MoCap data provided by the sensor. The first type of feature is related to the human gait (*e.g.* number of steps, duration of each step, and turning duration); whereas the second type describes the anatomical configuration. The authors state that using these features provides a descriptor for charactering physical mobility. To enable classification on the severity of the gait imbalance, the authors implement an SVM.

Recently, several frameworks have been proposed to enable greater understanding of human motion. These frameworks follow a similar structure. They first seek to identify the human action, using action recognition frameworks, and then undertake quantitative analysis on the motion to provide greater understanding [27, 111, 112]. Typically these methods are utilised in the healthcare domain for clinical outcome measures. Dolatabadi et al. [113] proposed a home-based system for assessing changes in gait and balance. The authors utilise a Microsoft Kinect 360 sensor to observe gait recovery in a participant that had undergone surgery. They found that they were able to track the gait and changes of the participant over a number of weeks. This enabled the authors to make clinical judgements based on the information extracted. Gonzalez et al. [114] proposed

a solution for *real-time* balance estimation by deriving Centre-of-Mass (CoM) feature from a Microsoft Kinect 360 sensor and a Wii Balance Board. The authors unite the CoM and angular momentum to quality the stability of user. While this work presents a novel solution to balance measurement, it has been tested on a limited population (two users). Other approaches such as [15, 115, 116] have utilised virtual reality and gaming systems to aid in motion analysis and understanding.

2.4.2 *Gestures* and Semantic Approaches (Hierarchical)

Gestures and Semantic Approaches seek to decompose maker-based and marker-less action sequences into basic "building blocks". These *gestures* are modelled to enable a better understanding of each action sequences for use either in recognition or motion analysis. Such approaches firstly decompose complex action sequence into individual gestures (or key segments) and then seek to represent each gesture to construct a hierarchical framework [22, 70, 117].

Early work by Liu et al. [118] introduced the content of a Motion Index Tree (MIT). With the development of MoCap systems, the authors identified that there is a need for a three-dimensional motion retrieval algorithm. The authors propose a method of partitioning MoCap based on a hierarchical motion descriptor. The MIT serves as a classifier to determine the tree index that contains similar motions based on the test case. To achieve the partition, the Nearest Neighbour (NN) rule-based dynamic clustering algorithm is adopted to detect the similarity between all samples and partition based on a threshold criterion value.

Instead of direct concatenation of three-dimensional joint features, Xia et al. [70] cast the joint positions into three-dimensional cone bins and built a histogram of three-dimensional joint positions to represent each action class. Each bin represents a unique aspect of an action sequence, enabling machine learning techniques to be more discriminative in detecting the correct action sequence. The benefit of using a hierarchical structure is that human motion can be considered as a combination of a set of sub-actions over time. Reyes et al. [119] represent a human model based on a feature vector which composed of 15 joints extracted from a three-dimensional human skeletal model. In addition, the authors implement DTW and combine automatic feature weighting on each joint to achieve real-time action recognition. Wang et al. [18] further introduced the

local occupancy pattern, which models the relationship between the human body parts and the environment surrounding the human. The authors define an *actionlet* as the human action with linear combination of "basic block" gestures. Uniting the *actionlet* and machine learning enables a robust method for recognition.

Raptis et al. [120] utilised a similar framework to that proposed by Liu et al. [118]. In [120] the authors propose a framework to classify gestures in real-time using marker-less technology. By transforming the three-dimensional skeleton into angular representations, view invariance and noisy data is handled more robustly. The authors generate a gesture model, which is trained by a cascaded correlation-based classifier, and extend it to include DTW to evaluate differences in motion classes. The core aspect of the framework is a pairwise comparison between "trained frames" and "test frames". Pazhoumand-Dar et al. [121] introduced a low-level joint feature identification framework. In [121], a discriminative approach to identifying and extracting joint movement similarities is presented. The framework is capable of detecting single instance actions, however, more complex action sequences, interactions with the environment and other users fail to be recognised robustly.

Du et al. [122] is one of the first to propose a Deep Learning solution to the task of recognising MoCap. Du et al. [122] decompose skeleton information provided by the Kinect One into joint groups, based on key anatomical joint regions (*left arm, right arm, torso, left leg and right leg*), to enable a more representative feature vector. The authors advocate that decomposing the sequence into joint groups provides an in-depth analysis of the motion. A hierarchical feed forward neural network is trained 8 layers deep, with each layer representing a unique joint group composition. The framework is capable of handling complex action sequences. However, due to the complexity of the Deep Learning network, training and recognition latency are significantly higher compared to other state-of-the-art approaches.

Han et al. [123] predicts the motion pattern between action classes within the manifold subspace. This enables discrete modelling of the intra-/inter class difference to facilitate efficient action recognition. Taylor et al. [19] proposed a modified conditional Restricted Boltzmann Machine evaluated using a non-linear generative recognition approach. Of great interest to this thesis, the authors use MoCap sequences that are parameterised in Exponential Map form. The authors learn the local constraints and the global dynamics

of each sequence to provide a descriptive element to allow for recognition. The approach is able to capture complex non-linearities in the data and distinguish between different motion styles, as well as the transition between action classes.

Several methods have sought to identify "key poses(s)", which are the best pose(s) to represent an action sequence in the fewest number of frames possible [92]. Early work by González et al. [124] proposed a framework for automatic keyframing of human action for use in computer animation. The authors propose an action model that is comprised of key frames representing each action class. The authors build an eigenspace by employing PCA. Each key frame is selected based on a temporal interpolation scheme. Each action is represented by a temporal sequence of key frames. However, the authors acknowledge that computation expense make the system unfeasible for real-world application. Barnachon et al. [125] proposed utilising histogram of poses, extracted from MoCap which were computed based on a weighted Hausdorf distance. The authors identify delegate poses of an action sequence, which are selected manually. Then, a histogram is computed based on the distance between the delegate poses and each pose in the sequence. To enable real-time functionality the authors implement dynamic programming to enable decomposition of action sequences. The authors were able to detect action sequences with a high recognition rate.

Using the spatio-temporal relationship of skeleton poses Bloom et al. [117] proposed a recognition framework based on clustering spatio-temporal manifolds. The authors extract MoCap data and convert 13 joints into Euler angles for each frame, then each action class is clustered using k-means. By using clustering techniques, the authors are able to decompose a set of sequences into groups, typically corresponding to phases of a motion sequence. A rank scheme is employed to select a *peak pose* for each cluster. These peak poses form a template for each action class. An online framework utilising DTW is employed to perform online recognition. The recognition results obtained are in line with current state-of-the-art. Conversely, Thanh et al. [126] proposed to extract discriminative patterns within action classes for use in classification. Given a MoCap sequence, a Shape Histogram to represent the three-dimensional skeleton is formed to measure the distribution of relative positions of neighbouring points. Using a self-similarity matrix, the matrix is grouped and refined using a Conciseness Cost Matrix function to extract key fames. For classification, the authors implement a confidence scheme to identify (with confidence) to which class a test sequence belongs.

There is a requirement to have a sufficient number of training samples in order to represent each action effectively [7]. Several works have focused on extending key poses to encapsulate the exemplar paradigm to model action/activities using a limited number of samples. Elgammal et al. [21] sought to capture the dynamics of gestures in an exemplar-based recognition system. The approach is based on representing each motion as a sequence of key exemplars. The key exemplars were selected based on a non-parametric estimation framework. A probabilistic framework is employed for matching sequences with the exemplar set. By utilising this type of approach, the authors have been able to model complex gesture sequences using a limited number of samples. In their most recent work, Barnachon et al. [22] extend the framework in Barnachon et al. [125] to propose the integral histogram approach by representing action sequences with either one or three exemplars. They extended the concept of integral histograms in the spatial domain by clustering action sequences to generate pose similarity. This method decomposes each cycle of a repetitive action sequence into histogram bins. For recognition, they decompose a continuous action sequence of poses into integral histograms based on DTW to compute a confidence score based on distance.

There are relatively few approaches that seek to unite classification and replacedquantifyquantification analysis of human motion in a hierarchical context. Cary et al. [127] proposed a system to unite the Microsoft Kinect for xBox 360 and Artificial Neural Networks (ANN) to aide in classification for physiotherapy assessment. The authors design a feature vector based on grouping of joints. The first group is composed of the torso joints (defined as joints of the spine, and neck); with the second group the remaining joints (defined as the outer joints such as hands and feet);. The feature vector is computed by extracting the associated angles between the groupings. The work employs a multi-level ANN that decomposes each limb into a separate model. This allows the authors to recognise complex action sequences and assess their correctness in relation to a predefined model.

There has been a shift towards utilising depth sensor technology for measuring movement in people with disease [16]. Predominately, Wang et al. [14] proposed a system for monitoring muscular-skeletal disorders with a single depth sensor. The authors introduce the Temporal Alignment Spatial Summarization method to decouple the complex spatio-temporal information. The framework detects and extracts *Action Units* which represent the different phases of human action (*e.g.* chair rise, one repetition). Then multiple

measurements are extracted from the skeleton to provide an indicator for well-being and health leading to clinical outcome measures. Amini Maghsoud Bigy et al. [128] introduced a framework to assess gait movement in patients that have been diagnosed with Parkinson's disease. The work proposes a framework to detect tremors and motions that are typical of a falling motion. The features proposed are gait-based features that are capable of providing discriminative representation of the motion being observed.

2.5 Discussion and Conclusions

This chapter has given a brief overview of the state-of-the-art in the field of feature extraction, feature representation and human motion analysis. In keeping with other reviews [6, 7, 9], there are clear advantages to uniting complementary algorithms and techniques for application in human motion analysis. For example, combining human action recognition with motion analysis to provide intervention analysis of human mobility for use in a clinical context. Furthermore, a novel area of contribution has been identified: the combination of feature representation, classification and motion analysis within a single framework for the context of health. In the remainder of this thesis three approaches for achieving this are explored: (i) the representation of MoCap data that is view-invariant, accommodates action dissimilarity and anthropometric variations; (ii) the ability to classify action sequences robustly in a real-time environment; (iii) extraction of relevant features to support clinical outcome measures. This chapter concludes by briefly reviewing each of the taxonomies that have been highlighted and discussing their relevance to this thesis.

2.5.1 Feature Extraction and Representation

Local and global approaches to extraction and representation of RGB images have proven to yield good results in identifying the human pose, and can be extracted with relatively low cost [6]. These types of approaches are typically limited to a single view-point, which have limited their applicability [28, 33, 52]. Yet, increasing the number of view points (cameras) such as in [31–33] goes some way in solving this issue but at the cost of increasing algorithm complexity and computation. The early work of Bobick and Davis [28] in which the silhouette of the human subject is extracted, discussed in Section 2.2.1

is fundamental to these types of approaches. Current algorithms using silhouettes are suitable for single person action recognition and perform best on single-layered motion sequences. There is difficulty in recognising complex activities due to the information lost either when computing projections or when stacking information along the temporal domain. Occlusion and noise can distort the silhouettes drastically. Importantly, many approaches that utilise RGB images assume that the video is readily segmented into sequences that contain at most one instance of an action and can be undertaken offline. Often, the human figure mask can be readily extracted from scenarios favourable to algorithm design. While several works (*e.g.* [40, 49, 54]) have addressed this issue, it remains a challenge to perform direct detection for online applications. This is an important research topic that is receiving much attention, however this thesis will not focus on this area, nor will any contribution be made.

The use of RGB images, while still relevant, has somewhat been overshadowed by the introduction of low-cost RGB-D sensors which enable easy access to three-dimensional data to compliment traditional RGB data. Acquiring three-dimensional data from depth sensors is more convenient than pose estimation from RGB or using motion capture systems. The approach of Bobick and Davis [28], which introduced the intensity image could aid in partially recovering information that is typically lost from three-dimensional to two-dimensional projections. However, designing both effective and efficient depth extraction and representations for human pose and classification is a difficult task. First of all, depth sequences may contain serious occlusions, which makes the global features unstable. In addition, depth maps do not have as much texture as RGB images do, and they are usually too noisy (both spatially and temporal) to apply local differential operators such as gradients on. Yet, the community is actively developing algorithms to combat these issues, as in [56].

The approaches of Shotton et al. [24] (further discussed in [3]) and Koppula et al. [71] discussed in Section 2.2.3 seek to extract three-dimensional motion capture representations for anatomical landmarks on the human body. These types of approaches have allowed the community to focus on the motion being performed, and not focus on extracting meaningful data points. In this thesis, motion capture data extracted from a depth sensor using the skeleton tracking algorithm of [24] is extensively utilised. This algorithm is incorporated into the Kinect 360/One sensor, which is a low-cost portable device ideal for utilisation in the clinical domain.

2.5.2 MoCap Datasets

Existing datasets have been created for benchmarking specific research domains, for example Kwolek and Kepski's [90] objective is to aid in detecting participants who fall; whereas Bloom et al. [5] provides gaming-based actions to aid research in improving action recognition for entertainment. Extending this further, by placing anatomically significant markers (*e.g.* [83]) on the human body could limit the range of motion and impair the result. Existing datasets lack the clinical support and validity required to answer the objectives of this thesis. For example, gait analysis requires an algorithm that can deal with a range of measurements such as angles, stride length and foot positioning. Using a game or gesture oriented dataset may not arrive at a reliable, clinically objective benchmark result.

Datasets have been proposed to solve specific tasks and early work introduced the marker-based systems which require expensive hardware and software expertise. It is not possible to extract sophisticated recording environments due to hardware requirements. As a consequence, rigid recording protocol, which restrict the participant to a defined motion zone, limting their ability to express the motion as they may otherwise would have, as seen in Sigal et al. [83] and Van Der Aa et al. [84] have resulted in similar dataset being obtain. Further, rigid protocols can limit the type of motion the participant is performing, and alter their typical motion pattern - resulting in poor representation. As mentioned previously, hardware constraints limit the environments in which data can be collected. For example, recording in a local General Practitioner Office is not possible to due the time required to place markers on the human body and calibrate them. A key objective of any dataset should be its ability to represent the norm, therefore out in the wild data capture is important.

Datasets such as CMU Motion Capture dataset [4], HDM05 dataset [85] and TUM dataset [86] were never intended for use in benchmarking healthcare-based applications. Without clinical support it is different to reliable benchmark algorithms with a high degree of accuracy; an aspect which is highly sought after by the medical community.

The introduction of the MSRDailyActivity3D [18] dataset started a recent trend in generating and publishing datasets for specific tasks - however they are all similar in nature and scope. The G3D [5] (later complemented with G3Di [89]) contain action

categories which are similar to [4] and leave little scope for benchmarking of healthcare-based algorithms. To further complicate matters, the actions undertaken provide little clinical scope, and are either captured using a rigid protocol [5, 75] or a very loose protocol [18] making it difficult to extract quantitative judgements. It is important that a clinically supported dataset be introduced to the community to aid in clinical assessment, quantification, analysis and clinical validation.

2.5.3 Understanding Human Motion

The ability to understand human motion is applicable in many domains, such as healthcare, gaming and security [14, 104, 117]. Extracting human motion features, such as mocap it is possible to classify the action and provide supportive clinical outcomes (as discussed briefly in Section 2.4). This section has exclusively discussed methodologies and frameworks that utilise three-dimensional motion capture information. The execution and performance of any action by two participants can result in a varying number of frames. To overcome this, techniques have been proposed to use the temporal domain to align these two sequences. The vast majority of existing algorithms solve this problem through temporal modelling, which models the temporal evolution of different action sequences. For example, the HMM has been widely used to model the temporal evolutions [70]. Further, Bloom et al. [117] predicts the motion pattern between action classes within the manifold subspace. Finally, DTW utilised by Müller and Roöder [129] computed the optimal alignment between motion templates composed of three-dimensional joints. However, each of the methods mentioned previously have been obtained utilising marker-based MoCap Vicon systems. The introduction of the skeleton tracker by Shotton et al. [24] (further discussed in [3]) may undermine the performance of these models and result in a modelling of the noise rather than the action sequence.

The main body of work for understanding human motion is for action recognition/classification. Achieving a reliable algorithm that can operate for complex action sequences (such as gymnastic) is an active research topic. Barnachon et al. [22] goes some way in addressing this issue by encoding an integral histogram with "sub-actions" that are key to an action overall. Yet, the algorithm is vulnerable to noise. Latency is an important factor, [18, 22, 119] presents a low-latency online recognition system, which process the

input without considering temporal alignment. This type of system is in high demand for applications in gaming and healthcare.

Human action recognition/classification is only part of the "Understanding Human Motion" aspect, several works have sought to unite classification with clinical frameworks for aid in the decision-making process for healthcare application. With the aim of providing more information to the clinician to allow for an informed decision pertaining to the well-being of the participant. Most notably, Wang et al. [14] developed a framework for assessing those with Parkinson's disease. However, the framework provides only basic clinical outcome measures, making it difficult to draw objective conclusions. A framework to provide objective outcome measures is desirable and missing from the current literature.

Whilst noise and occlusion are problematic in all types of data, they present an ever greater problem in markerless tracking. In the literature, data noise and occlusion for a home-based setting have not been taken into account. Finally, further refinement for analysing human motion to obtain outcome measures is required to allow objective decision-making.

Chapter 3

Theory and Techniques

This chapter describes each of the component techniques that are brought together to define the frameworks proposed in later chapters. These consist of feature descriptors for representing the human body and for ambient projection of high-dimensional data. A number of different classifiers are introduced for use throughout this thesis.

In the remainder of this thesis a number of techniques are introduced to address the problem statement defined in chapter 1. These are all based around a body representation, classification and an analysis framework that provides classification and outcome measures that provide detailed analysis of human motion . Uniting these approaches into a single unified framework leads to a novel streamlined framework that provides three-dimensional marker-less human action classification and motion analysis using descriptive representations. In this chapter the methods used to construct the feature descriptors, body representation (Section 3.1) and classification (Section 3.2) which are used in later chapters are reviewed and discussed.

3.1 Feature Descriptors and Body Representation

This section introduces the key techniques for feature descriptor and body representation that are brought together to define the framework in later chapters. This section will introduce the theory and issues of mathematically modelling spatial rotations and rigid body orientations in the physical world (Section 3.1.3) (three dimensional space of

yaw, *pitch* and *roll*), grouping semantically similar human body poses (Section 3.1.7) and grouping and projection of high-dimensional motions into a "latent" pose space (Section 3.1.8).

Human motion is highly complex, with subtle variations between different participants. There are many ways to represent the motion over time, that are mathematically, computationally and practical. This section provides a background of spatial rotations from first principles by introducing *Euler's theorem* of rotation in relation to the human body. Then four popular rotation representations are presented and discussed. These are: Euler Angles, Coordinate Matrix, Axis-Angle, and Exponential Map. The first three representations are introduced and compared from a mathematical and computational point of view. Each representation approach will be utilised in this thesis to solve a particular problem. Particular attention will focus on distance metrics, computational speed and discriminatory power. Finally, a generalised articulated skeletal model and notation is introduced.

Briefly, there are subtle differences between *rotations* and *orientations*. It is important for the reader to be aware of these.

A *rotation* is the action of transforming one vector into another vector. A rotation preserves the magnitude of a vector and preserves the handedness of the space (it observes the direction of the cross product between base vectors). All rotations obtained in three-dimensional space have three degrees-of-freedom (DOF), therefore we can imply that they need at least three features to define them.

An *orientation* is the viewpoint of a rigid body in any given space. Confusion is observed between these two terms, because orientations are usually (not always) represented as a rotation *with respect* to a fixed, known coordinate or relative point. Figure 3.1 demonstrates a rigid body represented in a local coordinate system (dx, dy and dz), which is measured against a fixed global coordinate system (x, y and z). It is important to be aware that in this case, displacement implies action (otherwise referred to as *angular displacement*). For the purpose of this thesis, we ignore the translational component and focus on rotation component, unless stated otherwise.

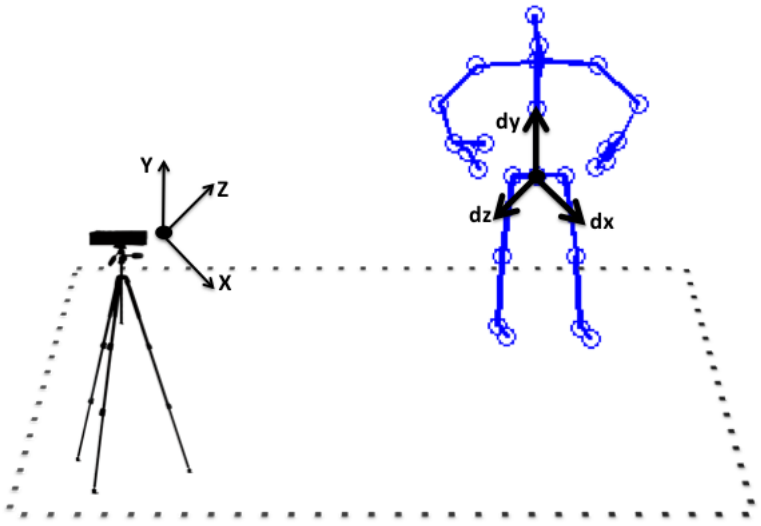

FIGURE 3.1: A global coordinate system which represents the orientation of the body with respect to a known world coordinate system of the sensor device. Example 1: The coordinate system of the body with respect to the Kinect device (x, y, and z). Example 2: Localising the coordinate system to the body (dx, dy, and dz).

3.1.1 Articulated Skeletal Model: Terminology

Rigid skeleton models of the subject are typical in any development framework that utilises MoCap. A *skeletal model* consists of a hierarchy of rigid *bones* which are connected together by *joints* (as demonstrated in Figure 3.2). Bones are rigid bodies as they cannot change or bend - therefore they can be described in length. It is important to be aware that for marker-less technology, bone lengths can sometimes vary due to noise, occlusion and poor skeletal tracking. Conversely, varying bone lengths could indicate potential noisy outlier data which can be efficiently managed [104]. Joints connect bones together and allow them to move with respect to each other (this is denoted as w.r.t), either rotationally or translationally. For marker-based systems, joints have an attachment point on the bone to which they connect between one and three rotational DOFs. For marker-less-based systems, joints are represented by between one and three orthogonal DOFs in respect to a fixed coordinate world system. For clarity, different frameworks, both marker-based and marker-less represent different types of rotational joints. In this work, 1 DOF represents a hinge joint, 2 DOFs is a universal joint and 3 DOFs are ball-and-socket joints.

When manipulating the motion of a skeleton, either that obtained by a marker-based or marker-less system, it may be useful to determine the root motion direction. As

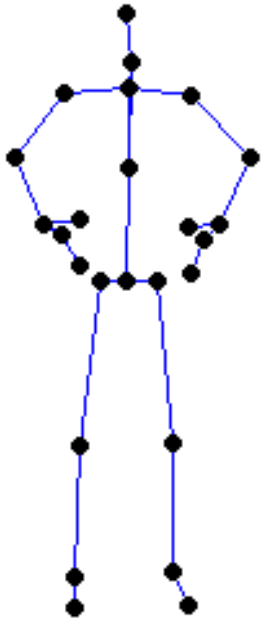

FIGURE 3.2: A figure representing the skeleton structure extracted from the Kinect One. A total of 25 tracked joints using the algorithm presented in [3].

mentioned briefly, a rigid body means that the distance between any two joints should remain the same irrespective of any motion or force placed upon the body. Formally, this means for any two joint positions $x = [p_t^j x, p_t^j y, p_t^j z]$ and $y = [p_t^j x, p_t^j y, p_t^j z]$ are:

$$||x_t - y_t|| = x(0) - y(0) = c \tag{3.1}$$

where c is a constant value.

In dealing with motion capture data, it is important to allow joints and segments of the body to rotate relative to each other but still maintain as a rigid body. This thesis represents the body as a *kinematic chain*. All translations are performed in world coordinate space and applied to the entire rigid body. The subject has specific orientation, rotation and translation in a world space w.r.t to a fixed position. The skeleton of a subject, denoted as P, is represented as a tuple of joint represents (*e.g.* Euler Angle, Axis-Angle), denoted as $p_t^j = \{x, y, z\}_{t=1:T}^{j=1:J}$. This thesis will refer to each pose at any given time as p_t, a set of poses as $\mathcal{P}$ and a specific joint at time t as p_t^j. Note that $\mathcal{P}$ and p_t are functions of time. An important decision for chain representation is how to parameterise the joints. Numerous parameterisations exist and have been discussed previously in this thesis.

3.1.2 Background: Euler's Theorem and Distance Matrices

The basic principle of rigid body orientation is *Euler's Theorem* (Figure 3.3). Euler's Theorem can be succinctly define as follows:

Euler's Theorem: Every angular displacement (or orientation) of a rigid body can be described as some angle θ about three mutually orthogonal coordinate axis fixed in space.

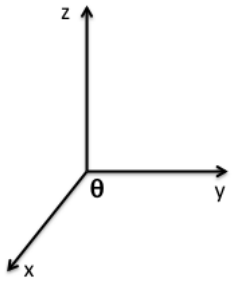

FIGURE 3.3: Euler's Theorem: Any angular displacement of any rigid body can be described as a rotation about a fixed axis (*e.g.* x) by an angle θ.

Providing context to human motion, Euler's Theorem suggests that if you grab a tennis ball, which is at some orientation in space, and rotate the tennis ball to some other orientation, there will *always* exist a *fixed* axis that can be rotated around in order to get a specific orientation, where magnitude is the angle. Therefore, taking this example we can state that case as: the axis will indicate *which way* to rotate an object and the angle indicates *how far*.

Interestingly, Euler's Theorem postulates the angle that directly provides us with an indicator of a distance metric on rotations, and therefore orientations. Thus, we are able to determine the distance between two orientations by using the angle of rotation between two orientations. This angle is easier to compute in some representations, and is motivation for the quaternion approach. It is easy to see why human motion, in terms of MoCap is represented in this way.

To conclude, Euler's Theorem denotes that spatial rotation have three degrees-of-freedom - two to specify the axis and one for the angle. Therefore, the "minimum" number of expected parameters to describe a rotation is three.

3.1.3 Euler Angles

The most common way to represent a rotation in MoCap systems is to represent it into three sequential rotations around principal orthogonal axes (namely x, y and z other known as *yaw*, *pitch* and *roll*) and represent the rotation as triple (θ_3 or θ_x, θ_y and θ_z), with each being around a particular axis (Figure 3.3). This is based on the fact, that in this thesis, the rotation has three DOF, therefore the three angles are capable of describing a rotation. This can be expressed in principal rotation matrices as:

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_x & \sin\theta_x & 0 \\ 0 & \sin\theta_x & \cos\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$Y = \begin{bmatrix} \cos\theta_y & 0 & -\sin\theta_y & 0 \\ 0 & 1 & 0 & 0 \\ \sin\theta_y & 0 & \cos\theta_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

$$Z = \begin{bmatrix} \cos\theta_z & \sin\theta_z & 0 & 0 \\ -\sin\theta_z & \cos\theta_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

A transpose will be present if a row vector basis is used. It is important to note that Euler Angles are different to what was proposed in Euler's Theorem (although in the 2D case they are the same, in three-dimensional this isn't the case). General rotation can be done by composing rotation over axis. For example, a rotation matrix $\mathbf{R}(\theta_x, \theta_y, \theta_z)$ in respect to a joint angle (*e.g.* extracted from a MoCap system) $\theta_x, \theta_y, \theta_z$ is represented as:

$$\mathbf{R}(\theta_x, \theta_y, \theta_z) = \mathbf{R}_x \cdot \mathbf{R}_y \cdot \mathbf{R}_z = \begin{bmatrix} C_y C_z & C_y S_z & -S_y & 0 \\ S_x S_y C_z & S_x S_y S_z + C_x C_z & S_x C_y & 0 \\ C_x S_y C_z + S_x S_z & C_x S_y S_z - S_x C_z & C_x C_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.3}$$

where $S_i = \sin(\theta_i)$ and $C_i = \cos(\theta_i)$. Also note that using matrix multiplication and the order is important (*e.g.* $\mathbf{R}_x \cdot \mathbf{R}_y \cdot \mathbf{R}_z \neq \mathbf{R}_y \cdot \mathbf{R}_x \cdot \mathbf{R}_z$). In a MoCap system setup, the rotations are *assumed* to be relative to a fixed world axis initially and then projected into a locally relevant axis of the human skeleton. For example, any orientation (and therefore rotation) can be specified by *yaw* (*e.g.* around the thumb on the right hand), *pitch* (*e.g.* around the middle finger on the right hand) and *roll* (*e.g.* around the middle finger of the right hand).

Euler Angles are a compact, three vector method of representation, for example a motion curve can be extracted and visualised with relative ease. In addition a benefit of Euler Angles is that angles are used to provide direct meaning, with no normalisation required and the angles are invariant to participants. Conversely, drawbacks exist with gimbal lock, singularities and discontinuity.

Gimbal lock: This is a coordinate singularity when two axes effectively line up, resulting in a temporary loss of a degree-of-freedom. This, albeit temporally, results in two angle vectors, *e.g.* θ_1 and θ_3, become associated with the same degree-of-freedom.

Singularities and Discontinuity: Euler Angles, contain three vectors which can cause a singularity - with a degree-of-freedom being lost during certain rotations. In addition, when an angle passes from 2π to 0, the components are within a circle space and not vector space and cause discontinuity in the data - making it very troublesome to model.

3.1.4 Coordinate Matrix

A group of rotations of Euclidean 3-space ($\mathbb{R}^3$) is usually denoted as $\mathbf{SO}(3)$, which is defined as a group of *special orthogonal* 3 by 3 matrices. In the literature it is also referred to as a Rotation Matrix. An *orthogonal matrix* consists of orthogonal row and column vector which are of unit magnitude. The orthogonal matrices, called $\mathbb{O}(3)$

is represented by two subgroup determinant (value associated with a square matrix - represented as *det*), $det = +1$ and $det = -1$. The subgroup represented by negative determinant ($det = -1$) are *reflections* as they are capable of changing axis of space. The positive subgroup, represented by determinant ($det = +1$) are *special* orthogonal matrices, as each matrices $\mathbf{R} \in \mathbf{SO}(3)$ will project a column vector to a new column vector $\mathbf{x} \in \mathbb{R}^3$ as:

$$y = Rx \tag{3.4}$$

by rotating and retaining its magnitude.

It is possible to extract a coordinate matrix for any given rotated space provided the original system is known. This is given as:

$$\mathbf{R} = \begin{bmatrix} | & | & | \\ \hat{x}_{new} & \hat{y}_{new} & \hat{z}_{new} \\ | & | & | \end{bmatrix} \tag{3.5}$$

where a vector from an unrotated basis space (*e.g.* x, y and z) into a new space defined by the column vectors. For example, Figure 3.1, the orientation is defined by the global coordinate system of the Microsoft Kinect (360/One) sensor, and the *hip center* will serve as our basis. Therefore, a coordinate transformation, columns-wise, can be undertaken.

Taking into account Euler's Theorem, the axis of rotation of a matrix is the eigenvectors. An eigenvector of a transformation is scaled by the transformation. Explicitly in this case, the rotation is a set of points that do not move under rotation, which thus determines a fixed axis of rotation. The remaining two eigenvectors will be complex in nature (thus also have complex eigenvalues). They are important as they describe a plane orthogonal to the axis of rotation. An angle can then be computed by finding the angle between the original and result vectors using a dot product and inverse cosine (arcos). Mathematically, the matrix representation is used to define rotations - $\mathbf{SO}(3)$ is the group we seek to represent. Matrices in this form map 1-to-1 with the angular displacement of rigid bodies (further discussed in Section 3.1.1).

Unfortunately, problems arise when using coordinate matrices. Firstly, it requires 9 parameters to represent the structure of only three DOF. Therefore, six constraints are enforced to remove extra DOF. If computational time is an important factor, this type of representation is inefficient. Secondly, as found with other representations, when rotations are concatenated (numerically) precision and round-off errors occur, causing minor differences from the special orthogonal, which introduces shearing and scaling issues. Finally, these types of representations are very difficult to visualise, in terms of the MoCap sequences they represent, since the axis is an eigenvector.

3.1.5 Axis-Angle

It is possible to represent an $\mathbf{SO}(3)$, which is a rotation in three-dimensional Euclidean space as a pair of vectors (unit vector $\hat{\mathbf{e}}$ indicating the direction of the axis rotation, and an angle θ representing the magnitude of rotation about the axis), as observed in Figure 3.4. This is called an *axis-angle* or sometimes referred to as a *Euler axis*. This form of representation is commonly used in marker-less based MoCap systems due to a lack of a rigid body form.

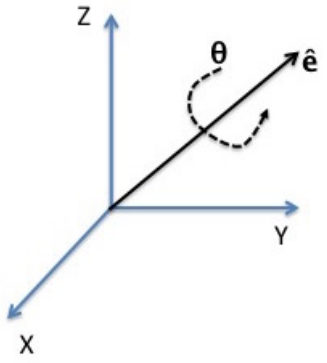

FIGURE 3.4: A visualisation of a rotation represented by a Euler Axis Angle.

To compose a rotation in this form is computationally expensive, as an intermediate step, such as a coordinate matrix is required. However, the representation does lend its self to Euler‘s Theorem, meaning we can use the rotation between two orientations as the metric.

3.1.6 Exponential Map

In computer vision, the Exponential Map is a generalisation of the exponential function of a matrix. It is based on generating a manifold which encodes the geometric features on which to estimate trajectories, for human motion joint-based trajectories [130]. A benefit, as will be demonstrated further, is that exponential maps are relatively easy to compute and stable to gimbal lock.

The Exponential Map maps a vector in $\mathbb{R}^3$ describing the axis and magnitude of a three degrees-of-freedom rotation to a corresponding rotation [130]. It is possible to compute an Exponential Map in different ways, however in this thesis the proposal by Grassia [130] is adopted due to its simplicity and popularity in the community. To formulate an exponential map from $\mathbb{R}^3$, a set of real numbers corresponding to a rotation, to a higher-dimensional analogue glome representation, $\mathbf{S}^3$ which is given as:

$$exp(\mathbf{p}) = \left\{ \begin{array}{lll} [0,0,0,1]^T & & \text{if} \quad \mathbf{p} = 0; \\ \sum_{m=0}^{\inf}(\frac{1}{2}\tilde{\mathbf{p}}^m) & = & [\sin(\frac{1}{2}\theta)\hat{\mathbf{p}}, \quad \cos(\frac{1}{2}\theta)]^T \quad \text{if} \quad \mathbf{p} \neq 0. \end{array} \right\} \tag{3.6}$$

where $\mathbf{p}$ is the pose (in $\mathbb{R}$), $\theta = |\mathbf{p}|$ and $\hat{\mathbf{p}}/|\mathbf{p}|$ to provide the representation of a single three-dimensional Euler angle Exponential Map form.

A set of transformed poses is denoted as $\bar{\mathcal{P}}$. The above equation maps $\mathbf{p}$ to the union quaternion representing a rotation of θ about $\mathbf{p}$ (or $\mathbf{p} = [\theta_x, \theta_y, \theta_z]$). The use of an Exponential Map encodes both the magnitude and axis rotation into a single three-dimensional vector. The formula demonstrates that by parameterising an axis/angle rotation in three Euclidean parameters is acceptable. However, there are still outstanding limitations with this method. It is clear from the equation above that exponential maps must have singularities ($\mathbf{p} \mapsto 0$). These are located on the glome (in $\mathbb{R}^3$) of a radius $2n\pi$. This is because any rotation of 2π about any axis is equivalent to no rotation at all as the rotation maps back on itself.

3.1.7 Human Motion Segmentation and Similarity Grouping

The need to group unlabelled data, otherwise referred to as clustering, arises in many different applications, such as data mining and knowledge discovery [131], pattern recognition [132] and health [133]. The object of any clustering approach is to determine "sensible" groups (clusters) formed by analysing available patterns in the data to extract information relating to similarity and dissimilarity.

A basic definition of *clustering* is as follows. Given a set of data vectors $X = \{x_1, \ldots, x_n\}$, the task is to group them such that "more similar" vectors are in the same cluster and "less similar" vectors are in different clusters. A set, typically denoted as $\mathcal{R}$, containing these clusters is called a *clustering* of X. Consider a motion sequence, such as walking, formed by a group of poses. Employing clustering it is possible to group each walking phase (gait cycle), based on each stride, therefore the strides are identified as unique groups. This section will explore k-means clustering [61], which has been selected due to its popularity and high accurate rate and use in clustering different data types.

3.1.7.1 k-means clustering

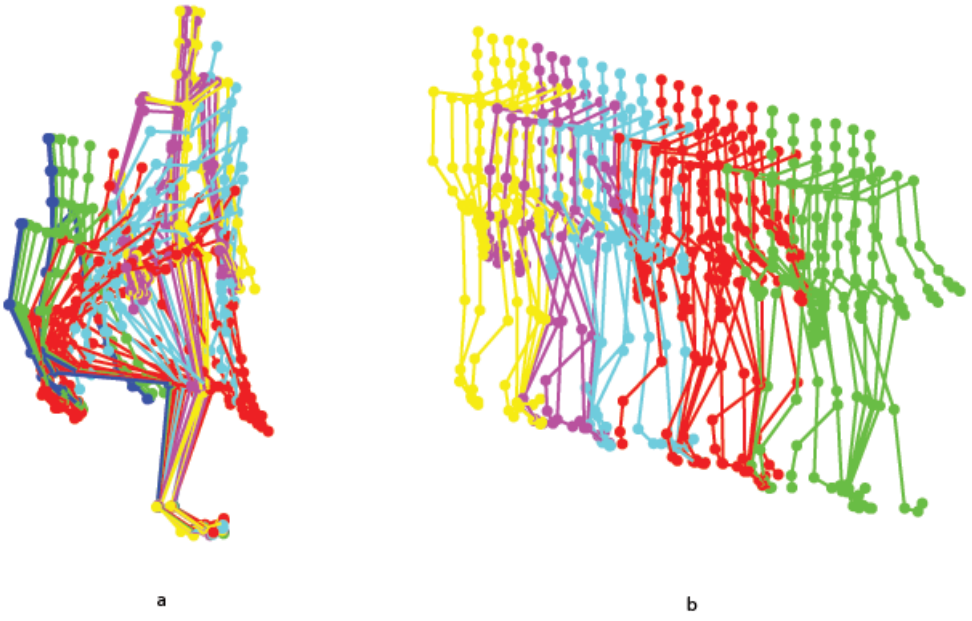

FIGURE 3.5: An example of k-means being utilised on MoCap data. Observe the ability of k-means to identify the typical phases of an action sequence. a) Chair rise clustered into unique phases. b) Walking forward clustered into unique phases.

The k-means clustering [61] rational is relatively simple. The algorithm assumes that the number of clusters, k, is known. Its objective is to move *cluster representative* centroid of each k cluster into regions that are dense points of data X. The k-means [61] algorithm is iterative in nature. In general, k-means is defined as follows. Consider n data points $x_i, i = 1, \ldots, n$ which need to be partitioned into k . The objective is to assign a cluster to each data point, where the aim is to find positions $\mu_i, i = 1, \ldots, k$ of the cluster which minimises the distance from the data points to the cluster centroid location. At each iteration t:

- The centroid initialisation for each cluster: $\mu_i = value, i = 1, \ldots, k$.
- The closest cluster for each data point is assigned: $c_i = \{j : d(x_j, \mu_i) \leq d(x_j, \mu_i), l \neq i, j = 1, \ldots, n\}$.
- Each cluster position is set to the mean of all data points within the cluster: $\mu_i \frac{1}{|c_i|} \sum_{j \in i} x_j, \forall_i$.
- Repeat until convergence.

The algorithm is capable of clustering highly complex data representation. An example of clustering MoCap sequences can be observed in Figure 3.5. k-means is suitable for unravelling compact cluster [132, 134], the algorithm is considered to be fast, as it is based on a iterative processes, which only requires a few passes through the data until it achieves full convergence. This enables k-means to perform robustly when processing complex and large datasets. However, k-means has several drawbacks that have yet to be addressed by the community. k-means cannot guarantee convergence to the global minimum, which, mathematically it can be assumed would represent the best possible clustering distribution. The algorithm returns the cluster corresponding to the local minima. Thus, different initialisation of centroid locations, on different machines may lead to different cluster groupings . Further, the algorithm is very sensitive to outliers and "noisy" data. Because each point is assigned a cluster, outliers will influence the centroid mean location. Finally, accurate estimation of the number of k clusters is vital for the success of the algorithm. However, it is not a trivial task. The following section discusses a method for selecting the optimum k.

3.1.7.2 The k problem - optimum number of clusters

The number of clusters should match the data. An incorrect choice could result in an invalid cluster distribution that does not represent the data itself. More specifically, if a large number of clusters are used, it is likely that at least one cluster will be split into two or more, all containing very similar poses - which generally lies between sparse regions in-between those clusters. k-means is robust for unravelling compact sequences, such as MoCap, where accurate estimation of the number of clusters is crucial. There is no single solution to estimating the optimum k value, with several works selecting k manually (*e.g.* [22, 135]) or using automatic selection methods (*e.g.* [136–138]). In this work, the Elbow method [138] is extended to represent the *within-cluster-sum-of-squares* (WCSS) via the gap statistic proposed by Tibshirani et al. [134] (please refer to [134] for in-depth algorithm presentation). This method enables automatic selection of the k which is free from human bias.

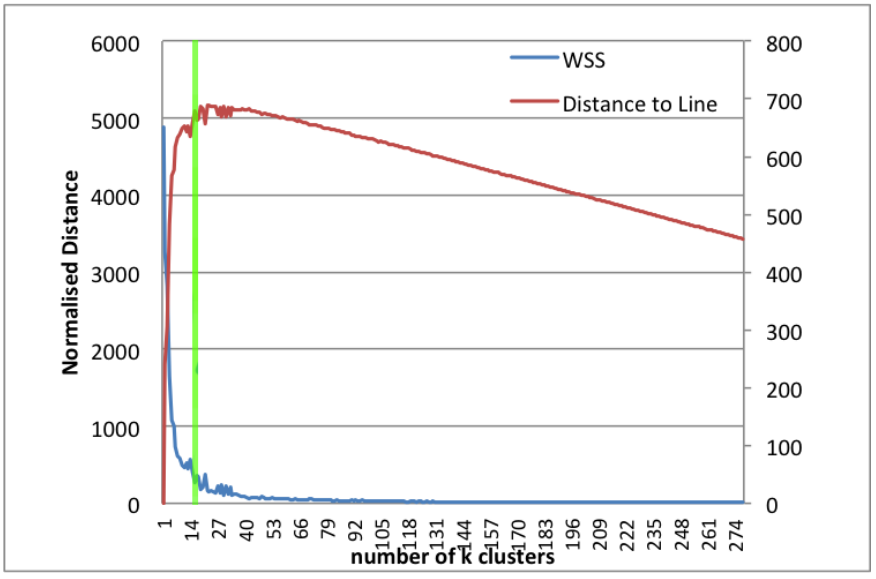

FIGURE 3.6: A visual demonstration of selecting the most suitable k. With the green line indicting the optimum k. With MoCap as input.

To determine the optimal k, an iterative process is undertaken which is described as follows: k-means is performed n times, where $n = \sqrt{T/2}$, k is increased for each iteration by a value of one *e.g.* $k = 2, 3, \ldots, n$. For each iteration (of increasing k), *within-cluster-sum-of-squares* is minimised. This is defined as the sum of the distance between each pose and its assigned cluster centroid. This allows for a numerical value to represent the compactness of each cluster. After n iterations a set of $\mathbf{W} = \{w_1, \ldots, w_n\}$ scores are computed. The *within-cluster-sum-of-squares* for each iteration are averaged to provide

a single unit value. The suitable k value is the $log(w_i)$ index which falls farthest below a reference curve, defined as:

$$G_t(k) = E_T^*\{\log(w_i)\} - \log(w_i) \tag{3.7}$$

where E_T^* denotes expectation under the sample size of T from the original sequence. The optimum k is the value maximising $G_t(k)$ after taking the original distribution into account, an example of optimum k is shown in Figure 3.6.

3.1.8 Pose State Space

Assuming that a rigid skeleton is fully specified (and not inferred) at time t by a high-dimensional "ambient" joint series $p_t^j = \{x_t^j, y_t^j, z_t^j\}_{t=1:T}^{j=1:J}$, it is assumed that the coordinate system is shared (Figure 3.7 provides a visual example of the motion signals). However, when a feature vector is large, such as with MoCap, an alternative, more informative solution is desired to recover a low-dimensional "ambient" encoding of (or a part of) the original pose state space (Figure 3.8 demonstrates the "ambient" representation). As mentioned in Section 3.1, a number of different options exist for parameterisation of such as Euler Angles, Axis-Angles and Exponential Maps. The data projection to a latent pose space reduces the amount of redundant information to provide a meaningful and manageable representation. A mapping from the latent space to the original pose state space exists and enables the parameterisation of the rigid skeleton for objective function evaluation. In this section notation for the rigid skeleton is introduced and the technique for recovery of the associated latent pose space from MoCap data is presented (please refer to Section 2.3 for an introduction to MoCap datasets and data).

3.1.8.1 Principle Component Analysis

Principle Component Analysis (PCA) is used to decompose the variation in a set of $\mathcal{X}$ pose vectors, $\mathcal{X} = \{x_1, \ldots, x_n\}$. The mean $\bar{x}$ and a covariance matrix $\mathbf{S}_i$ are calculated and Eigen decomposition is used to compute the l eigenvalues and eigenvectors, λ_i of $\mathbf{S}$. PCA technique is defined as follows:

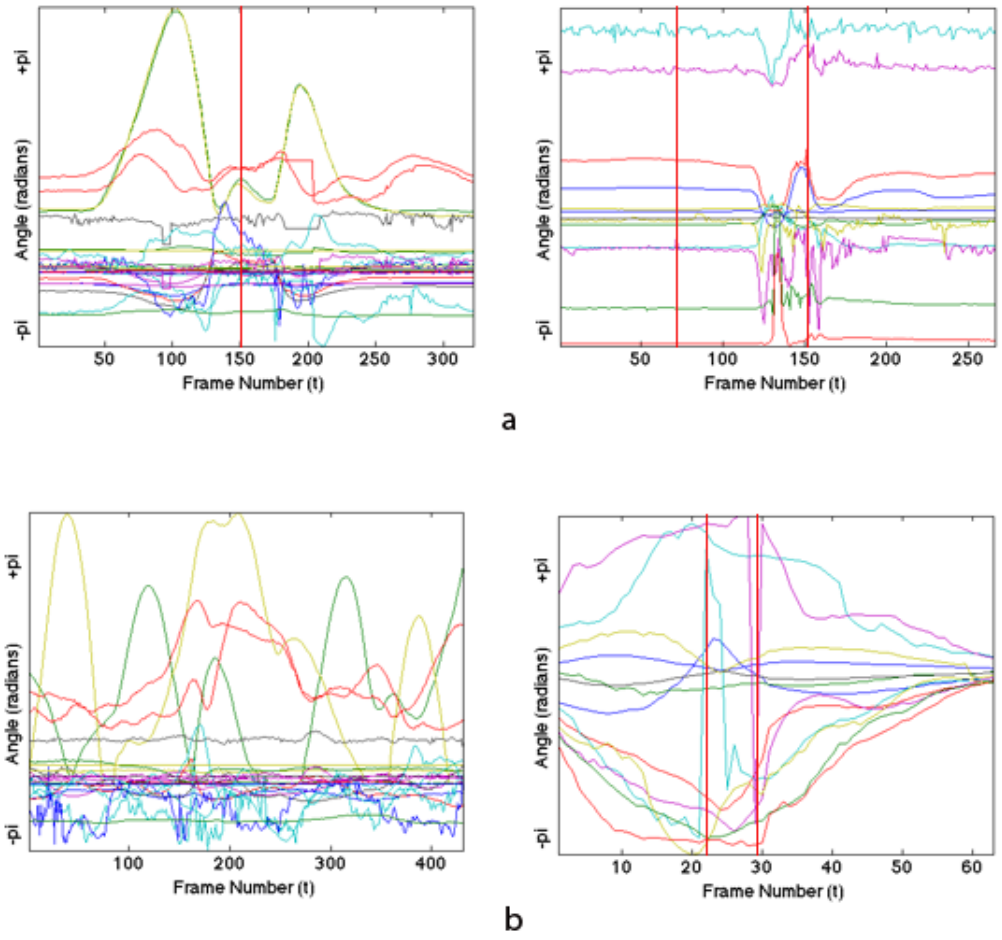

FIGURE 3.7: High dimensional MoCap pose vector visualisation: **a** left) *CMU Walk*; **a** right) *Kinect Walk*; **b** left) *CMU Jump*; **b** right) *Kinect Jump*. Vertical red lines denote singularities of bad MoCap data.

1. Estimation of $\mathcal{X}$ into a covariance matrix $\mathbf{S}$. The mean $\bar{x}$ is subtracted.
2. Eigendecomposition of $\mathbf{S}$ and compute the l eigenvalues and eigenvectors, λ_i (where $i = 0, 2, \ldots, l-1$).
3. Eigenvalues are arranged in descending order, $\lambda_0 \geq \lambda_1 \cdots \lambda_{l-1}$.
4. Selection of the number of eigenvalues to retain. A user defined parameter. These are known as the principal components.
5. Transform each l-dimensional vector x in the original "ambient" pose space to an m-dimensional "latent" vector space sy via the transformation of $y = \mathcal{S}^T$. Such as those demonstrated in Fig. 3.8.

The MoCap data used in this thesis is composed of a set of poses over time. These poses are represented as a set of three-dimensional positional markers (either Euler or Axis-Angle). For MoCap poses, PCA has been introduced for dimensionality reduction, compression and comparing motion sequences [139, 140]. PCA is capable of handling noise data that is inevitably contained within MoCap, Vieira et al. [102] note that in practise only a few (≤ 5) are required to represent postures in a discriminative way. Nevertheless, selecting the number of principal components to retain and avoiding "curse of dimensionality" is not an easy task - and is yet to be solved. However, a disadvantage of this resides within the anthropometric variations in action performance, while subtle variances can be over generalised by using PCA [139]. A discussion and implementation of PCA is presented in chapter 6.

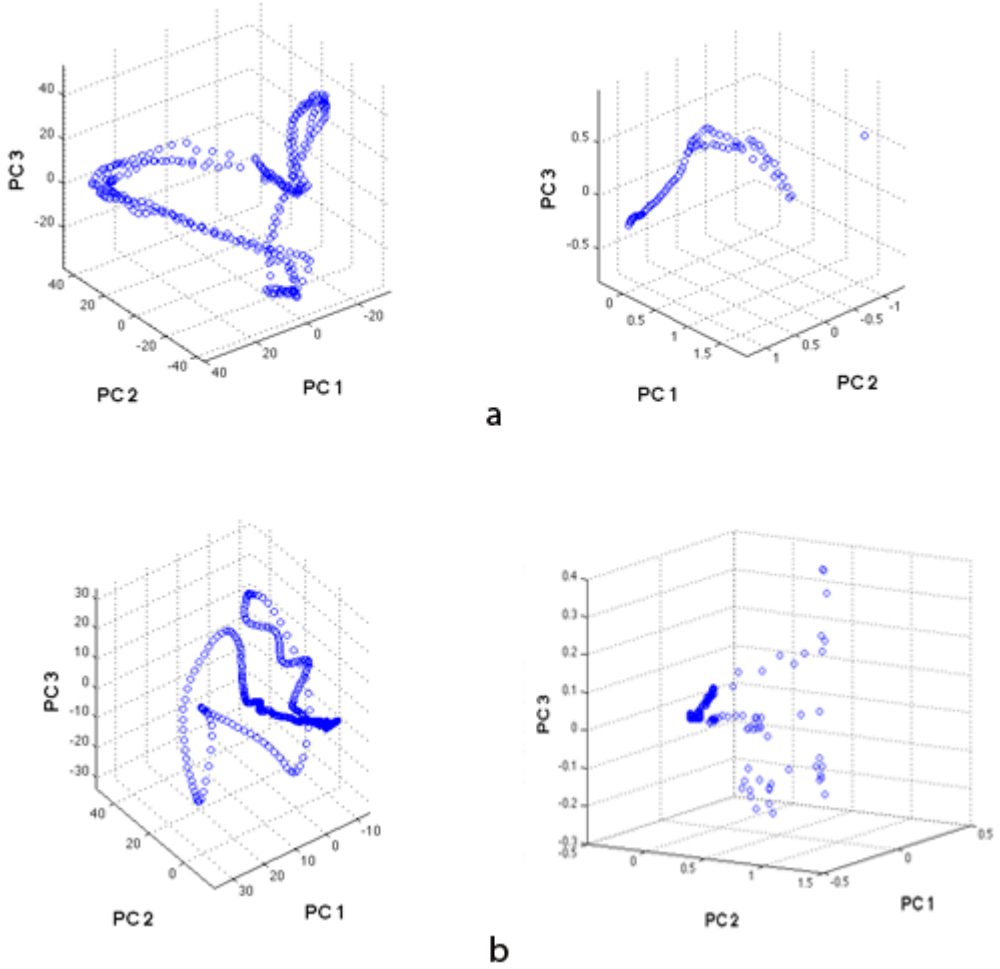

FIGURE 3.8: Low dimensional latent MoCap pose vector visualisation for 3 Principal Components (PC: **a** left) *CMU Walk*; **a** right) *Kinect Walk*; **b** left) *CMU Jump*; **b** right) *Kinect Jump*. Joint angle data is that shown in Fig. 3.7. Where PC represents Principal Component dimension.

3.2 Machine Learning

This section briefly introduces the basic concepts of several well-known machine learning techniques. These techniques have been utilised throughout this thesis to aid in the decision making process and/or to provide validation to the proposed framework.

3.2.1 Support Vector Machines

Based on statistical learning theory, Support Vector Machines (SVM) is a supervised machine learning classifier [141]. In simple terms, the SVM produces a model that represents the training data by learning the optimum separating margin between the linear hyperplanes of each class. The data points on either side of each hyperplane are defined as the *support vectors*. SVMs have been implemented with varying degrees of success within the MoCap research community. Barnachon et al. [125] noted that the ability of SVMs to function efficiently is dependent on the type of features being trained, with histogram-based representations struggling due to their complexity. Conversely, Sinha and Chakravarty [104] noted that MoCap data itself is very difficult to separate, however by extracting meaningful representations with reduced dimensionality, SVMs may operate more efficiently. In cases where the classes of data are not linearly separable, such as MoCap, the points are projected to a higher dimensional space where linear separation may be possible. SVM requires the user to state the parameter C, which controls the trade-off between model complexity and empirical error in SVM.

The radial basis function kernel, which non-linearly maps samples into a higher dimensional space has been found to be efficient when handling MoCap data [47, 142]. The radial basis function is able to handle cases when the relationship between the class labels and data vectors is non-linear [143]. It is defined as:

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2), \gamma > 0 \tag{3.8}$$

where x are the training vectors and i, j are matrices index identifiers. In addition, the parameter γ determines the shape of the separating hyperplane in the radial basis kernel.

3.2.2 Random Forest

Random Forest (RF) [144] has proven to be easy to implement, and very efficient for training and testing in various computer vision problems [24, 139, 145] as a multi-class classification method. It is an ensemble classification that contains n_{tree} randomised decision trees. RF combines bootstrap aggregating (otherwise referred to as bagging [146]) and randomised feature selection [147, 148] to reduce the correlation between trees and to aid in noise reduction.

A RF consists of multiple decision trees constructed by supervised learning of a training set. An RF model is constructed by using the bootstrap aggregating method to randomly generate n_{tree} decision trees which are each provided with randomly selected samples of the training input and then all the decision trees are combined into a decision forest. For this thesis each bootstrap, a random m_{try} (default 3) sample of the training data is used which determines the size of an unpruned tree. For classification (and regression) the model predicts a corresponding class based on the voting of all trees, where the class with the greatest number of votes is determined as the probabilistic choice. RF only requires one user parameter, n_{tree}, which sets the number of decision trees to grow.

3.2.3 Artificial Neural Networks

Artificial Neural Networks (ANN) is a popular technique used for a variety of classification tasks. For MoCap data it has been used extensively to aid in healthcare, recognition and feature detection [107, 127, 149]. In mathematical terms, a *neuron* is an operator that maps $\Re^p \rightarrow \Re$. If we consider a neuron j receives a signal z_j that is the sum of p inputs x_i scaled by associated connector weights w_{ij}:

$$z_j = w_{0j} + \sum_{i=1}^{p} x_i w_{ij} = \sum_{i=0}^{p} x_i w_{ij} = \mathbf{X}^T \mathbf{w}_j \tag{3.9}$$

where $x = [x_0, \ldots, x_p]^T$ is the input vector with T denoting transposition, $\mathbf{w}_j = [w_{0j}, \ldots, w_{pj}]^T$ is the weight vectors of a neuron j, and w_{0j} is the bias parameter, which is treated as an extra connection with a constant unit value of $x_0 = 1$. A neuron outputs a logistic sigmoid activation function.

A *neural network* is a set of interconnected neurons. In this thesis, a *feed forward neural network* is implemented, with the neurons organised in layers so that a neuron in layer l receives inputs only from the neurons in layer $l-1$. The first layer is referred to as the input layer, the last layer is denoted as the output layer and any intermediate layers are called hidden layers.

3.3 Discussion and Conclusions

The techniques described in this chapter provide the basis for the contributions that are presented in later chapters. A framework, for simplicity, is presented as follows:

1. Marker-based and/or marker-less MoCap data is represented in a new feature basis to enable a more efficient representation for use in recognition and healthcare applications.
2. Identify and rank poses based on its discriminative power to reduce the number of training samples required.
3. Attempt to recognise the appropriate action class for unseen MoCap sequences.
4. Identify subtle differences between MoCap sequences to enable clinical outcome measures of human motion. Specifically, identifying motions between the young and old in relation to their mobility.

These steps are implemented in chapter 4, to rank and extract key poses for use in human action recognition. Further utilised in chapter 5 and chapter 8. How can human action be decomposed to reduce training time? Is it possible to detect subtle differences between participant groups? How should a framework be developed to realise a clinical framework? This thesis attempts to answer these questions, a list of specific contributions is presented in Section 1.4 and also at the beginning of each chapter.

Chapter 4

Exemplar Paradigm: Discriminative Key Pose Extraction from Marker-based MoCap

In this chapter, two approaches for identifying descriptive and discriminative key pose for use in human action recognition by using feature ranking are proposed and evaluated. The ability to recognise action sequences robustly is important for a large number of applications, such as health or motion analysis. The first approach, Delegate Identification and Selection (DIS) identifies delegate postures using a statistical ranking and joint discrimination function. The second approach, Discriminative Key Pose Identification (DKPI) determines key poses by assessing the maximum subspace score of the dissimilarity space. The resulting approaches are evaluated for recognising known and unknown human motions using temporal-window-based classification and machine learning techniques.

4.1 Introduction

This chapter combines the competing benefits - flexibility and efficiency to propose two approaches that are capable of detecting human action accurately and robustly across

a number of datasets. The first approach, Delegate Identification and Selection (DIS) identifies delegate postures using a statistical ranking and joint discrimination function. This approach is based on a generative exemplar-based framework for human action classification, with MoCap as an input. The proposed representation and selection approach combines generalised local representations by parameterising joint information to generate global exemplars to describe the different phases of an action sequence. DIS is partly inspired by the use of statistical analysis to determine key poses [117]. The second approach, Discriminative Key Pose Identification (DKPI) determines key poses by assessing the maximum subspace scoring of the dissimilarity space. The proposed approach computes local representations based on joint dissimilarity and mutual joint respect to identify key poses of a MoCap sequence. The number of delegates for this approach is varied based on the estimation of the complexity given a particular action model with their discriminative power and mutual respect to other poses taken into account [18, 22, 29, 150].

The main contributions of this chapter are as follows:

1. Delegate Identification and Selection: Introduction of a novel algorithm for identifying key poses using statistical ranking schema derived from the t-test. Poses are selected based on their inter-/intra class discriminatory power using novel ranking function to form an action model (Section 4.2.1).

2. Discriminative Key Pose Identification: Introduction of a novel framework for identifying key poses based on k-means clustering of human action. The framework encodes a joint dissimilarity matrix to identify those active joints of a segmented sequence which best represent the action sequence (Section 4.2.2).

3. A window-based algorithm for detecting marker-based MoCap sequences is introduced and evaluated. A group of poses are placed within a window-based approach for classification. Using dynamic programming principles the algorithm is capable of providing classification results with little latency (Section 4.3).

4.2 Approach Methodology

This section introduces DIS and DKPI approaches. For simplicity and generalisation, human motion, typically captured by a marker-based MoCap system, is modelled using a *kinematic chain*. A kinematic chain consists of *body segments* that are connected to various body *joints*. A sequence can be seen as a time-sequential of 3D joint coordinates that relate to the fixed kinematic chain. In this thesis, a motion sequence is a series of frames (otherwise denoted as poses), with each frame specifying the 3D coordinates of the joints at a certain time period. Recall, in Section 3.1.1 a sequence is denoted as $\mathcal{P} = \{p_t^j | t = 1, \ldots, T; j = 1, \ldots, J\}$, where t denotes the time and j is the joint index.

The k-means clustering algorithm underlines the proposed approaches in this section (presented and discussed in Section 3.1.7). k-means is iterative in nature, starting with an initial estimation of the centroid for each cluster which continues until convergence of a motion sequence into an assigned number of k clusters. It has been shown to be efficient in segmenting and clustering MoCap data, as demonstrated by Zhou et al. [136]. As highlighted earlier (see Section 3.1.7), accurate estimation of the number of clusters is crucial. In this thesis, the Elbow method is implemented to identify the optimum number of k clusters.

4.2.1 Delegate Identification and Selection

The DIS framework selects delegate postures using a statistical ranking and joint discrimination power. A generative exemplar-based framework for human action classification, with MoCap as an input, is introduced and discussed in the subsequent sections.

4.2.1.1 MoCap Representation

The data represents coordinates for a human subject performing activities in a predefined action space. That is, within the confide of a fixed global coordinate system. For a HCI classification, it is important to place the human skeleton data at the centre of the coordinate system to become view-invariant. The root orientation and translation is handled in a unique manner because it encodes the transformation over time. As in [19], the representation of each pose is by an incremental “forward” and “sideways”

vector relative to the forward-facing direction of the participant. Height remains non-incremental relative to the distance from the ground plane. Orientation is represented by an incremental change about the gravitational vertices. Finally, the remaining rotations are represented by absolute pitch and roll, once again relative to the forward-facing direction of the skeleton. As has been highlighted in Section 3.1.6, there is no single solution to parameterisation of rotation that is applicable for all application domains. The proposed approach requires the parameterisation of the 3D Euler Angles in to Exponential Map form. This was done to avoid gimbal lock, discontinuities and ball-and-socket joints complications that are typically found in MoCap. In addition, it provides a more usable feature vector that is less noisy and mathematically stable. Recall, that the exponential map maps a vector $\mathbb{R}^3 \mapsto \mathbf{S}^3$ and is formulated from the corresponding quaternion representation of rotation. As part of the mapping process, several joints consisted of constant zero values, therefore they are removed.

4.2.1.2 Delegate Selection

Given a motion sequence, $\mathcal{P}$ that has been clustered into k clusters, a delegate pose for each k cluster is identified. The delegate should contain enough information to describe the motion sequence, and the specific cluster it represents. Therefore, only the most representative pose of a cluster should be identified as a delegate, denoted as d_t^j. The Receiver Operating Characteristic (ROC) Curve framework is modified to measure the within-cluster representation capabilities of each pose at a joint level.

Let Q_k be a recursive relationship between $p_{t,k}^j$ with regards to $p_{t,k}^j$ of a given cluster k and is given as:

$$Q_k = \underset{k \in [1,\ldots,K]}{\operatorname{argmin}} \Big(\sum_{k=1}^{K} \frac{p_{t,k}^j + p_{t,k}^j}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \Big) \tag{4.1}$$

where Q_k is the lowest scoring delegate, $p_{t,k}^j$ and $p_{t,k}^j$ are two poses within $k \in K$ cluster and SE_1, SE_2 refer to the estimated standard error of the ROC for each pose and its associated joints; and r represents the *estimated* correlation between poses when the cluster Euclidean distribution is taken into account. For each cluster, a delegate is selected based on the lowest average ROC error, similar to the Area Under the Curve. The pose which has the lowest standard error, compared to other poses of the same

cluster is selected as a delegate representation for the action class. This results in a set of delegate poses $\mathbf{D} = \{d^j_{1,k}, 2, k^j, \ldots, d^j_{t,K}\}, d \in \mathcal{P}$.

4.2.1.3 Delegate Ranking

It is now possible to represent $\mathcal{P}$ into a set of $\mathbf{D}$ delegate poses, where each delegate represents each phase of the action sequence. While this may be sufficient in recognising human motion, inter-/intra-class variation exists. To overcome this variation and provide a more compact but efficient representation each delegate pose is ranked against all other delegates to determine which are the most representative and discriminative. To rank each delegate pose, a statistical analysis based function is proposed employing the standard significant test referred to as the t-test. The significance between two poses at a joint level, $d^j_{1,k}$ and $d^j_{1,k}$ is computed as:

$$C_{score}(d^j_{1,k}, d^j_{1,k}) = \sqrt{\frac{(j-1)s_x^2 + (j+1)s_y^2}{n_1 + n_2 + 2}} \tag{4.2}$$

where j is the index location for the j-th joint for $d^j_{1,k}$ and $d^j_{1,k}$; s_x and s_y are the standard deviation (SD) between j-th joint and all other joints and n_1, n_2 are the sample sizes of d_1 and d_2. Using the cost function to identify pose level significance, a set of delegates is ranked, given as:

$$R_{score} = \sum_{i \in \mathbf{D}} \left(\sum_{i^* \in \mathbf{D}} C_{score}(d^j_{i,k}, d^j_{i^*,k}) \right) \tag{4.3}$$

where R_{score} is the representative power of each delegate pose compared to all other delegates.

For clarity, an exemplar model can be constructed by selecting the number of exemplars to retain for each action class, denoted as e. For this approach, e is defined to determine the number of delegates to retain for each action class. This is given as:

$$model^l = \sum_{l=1}^{\mathcal{L}} \max_{e} \left(R_{score} \right) \tag{4.4}$$

where $model^l$ is the exemplar model and $l = (1, 2, \ldots, L)$ denotes the action class. An example of which is demonstrated in Figure 4.1.

4.2.2 Discriminative Key Pose Identification

DKPI determines key poses by assessing the maximum subspace scoring of the dissimilarity space of the star skeleton representation. The approach computes local representations based on joint dissimilarity and mutual joint respect to identify key poses. Unlike the approach proposed in Section 4.2.1, the number of delegates are dynamic, based on the action complexity with regards to the discriminative power and mutual respect to other poses. Two user-defined "retain" parameters are necessary; the number of active joints to retain and a percentage value for the number of poses to retain.

4.2.2.1 MoCap Representation

For the DKPI approach, as in [121], only major anatomical joint landmarks associated with human motion are retained, such as *hands*, *head*, *feet* and *torso*, with other joints discounted. Marker-based datasets are discussed at length in Section 2.3, the following joint locations have been retained (where possible), *left hand*, *left elbow*, *right hand*, *right elbow*, *left shoulder*, *right shoulder*, *head*, *upper-torso*, *lower-torso*, *left knee*, *left foot*, *right knee* and *right foot*. In addition, the joints retained must contain three DOF to provide an orthogonal representation of the motion.

After the 3D positions of interest are extracted, as proposed in Section 4.2.1, the representation of each pose by a incremental "forward" and "sideways" vector relative to the forward-facing direction of the participant. Height remains non-incremental relative to the distance from the ground plane. Orientation was represented by an incremental change about the gravitational vertices. Finally, the remaining rotations are represented by absolute pitch and roll, once again relative to the forward-facing direction of the skeleton.

In the next step of pre-processing, Z-score normalisation [151] is undertaken for each action class to make our approach robust to different body sizes and subtle inter-class variations. The normalisation of each joint is undertaken at each j-th joint, at time t

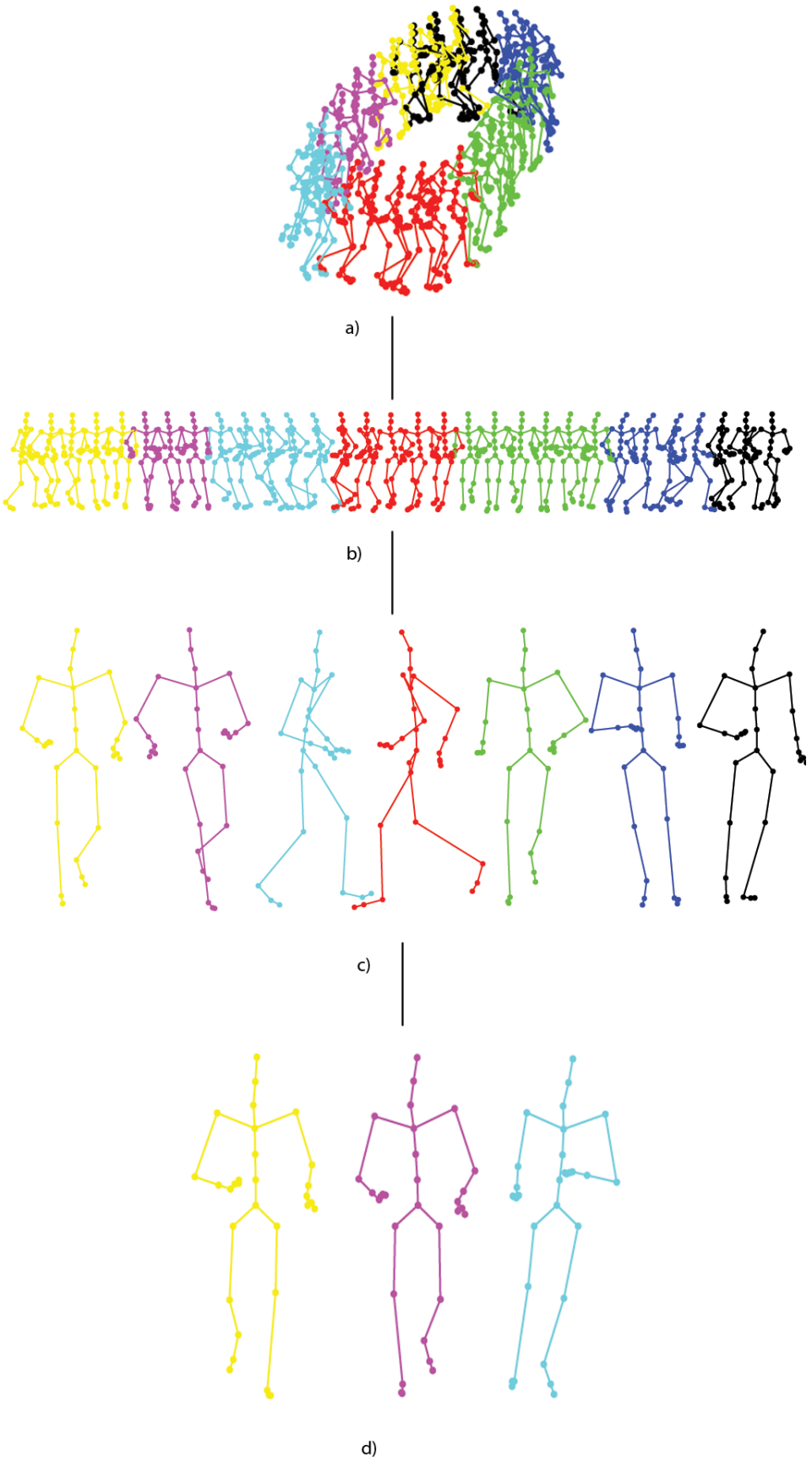

FIGURE 4.1: Delegate Identification and Selection Approach: Decomposition of a human participant running in a circle (extracted from MoCap). (a) original motion sequence of a human running in a circle. (b) k-means clustering, where each colour denotes the cluster. (c) delegate pose for each cluster. (d) the selected delegate exemplars for the motion sequence. where $e = 3$.

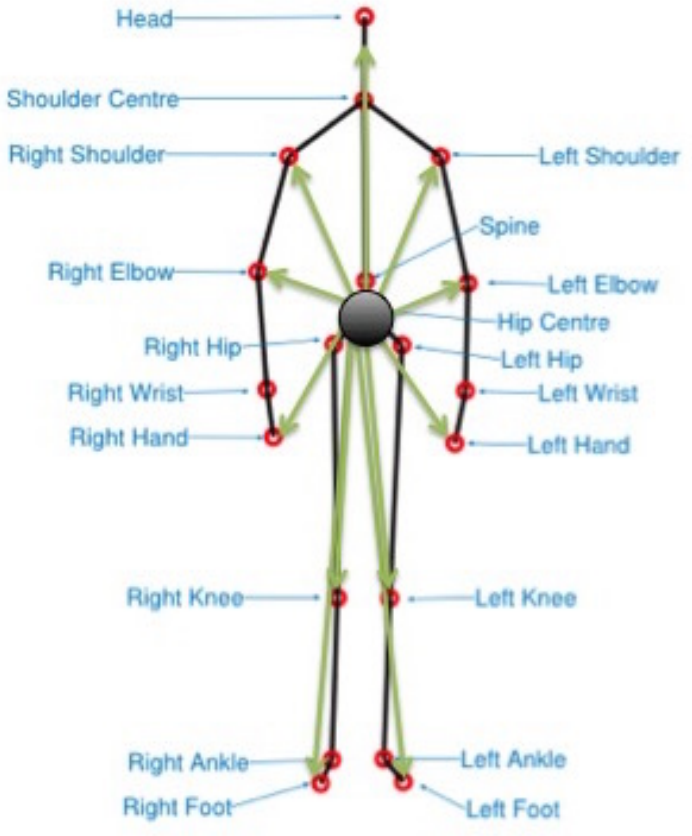

FIGURE 4.2: A visual representation of the skeletal joints of interest when forming a Star Skeleton representation.

using a joint mean vector $\vec{f_j}$, and the standard deviation of the joint group, σ_j calculated across all action sequences which form part of the action class. This is given as:

$$\vec{p_t^j} = \frac{p_t^j - \vec{f_j}}{\sigma_j} \tag{4.5}$$

4.2.2.2 Star Skeleton Dissimilarity Space (Euclidean)

The classical approach to human action classification focuses on three-dimensional representations concatenated to form a feature vector. However, these types of HCI representations struggle when scaled, or to factor in body size and action performance. Following the approach introduced in Wang et al. [18] and Vieira et al. [102], it is possible to encode a "star skeleton" representation to completely represent the global position and orientation (Figure 4.2 provides a visual representation of the skeleton).

Recall, any number of joints can be tracked by a marker-based system, for each joint j, a pairwise relative position is computed by taking the difference between the position of j^* and any other joint j-th for time t, given as:

$$\vec{p}_t^{\,j} = \vec{p}_t^{\,j^*} - \vec{p}_t^{\,j} \tag{4.6}$$

The feature representing the j-th joint is given as:

$$\vec{p}_t^{\,j} = \{\vec{p}_t^{\,j} | j^* \neq j\} \tag{4.7}$$

The relative position, provide by the star skeleton representation allows for complex actions to be represented in pairwise form. The spatial variation contained within $\vec{p}_t^{\,j}$ can be efficiently handled using a clustering strategy, Figure 4.3 demonstrates the skeletal posture and its associated distance matrix. In this approach, as with DIS, a set of poses, $\mathcal{P}$, comprising of $\vec{p}_t^{\,j}$ is clustered into k clusters.

4.2.2.3 Most Active Joints in Clusters

While it may be possible to train a machine learning classifier on the star skeleton alone, it is unlikely to yield accurate, or repeatable results. In this section, the objective is to identify the most active joints of each cluster, which in-turn will be used to describe the "key" phases of the action sequence. As in [70, 121, 152, 153], the selection of the most active joints can aid in reducing inter-/intra-class variation and provide a more discriminate representation.

In order to identify human poses, consider the star skeleton representation as a dissimilarity space representation [154], where each sample is a pairwise dissimilarity to a set of other joints. Using the dissimilarity representation for each time period t, a mutual respect between two poses, $\vec{p}_1$ and $\vec{p}_2$ and their dissimilarity lengths of i, q, Mutual Information can be expressed in entropy, given as:

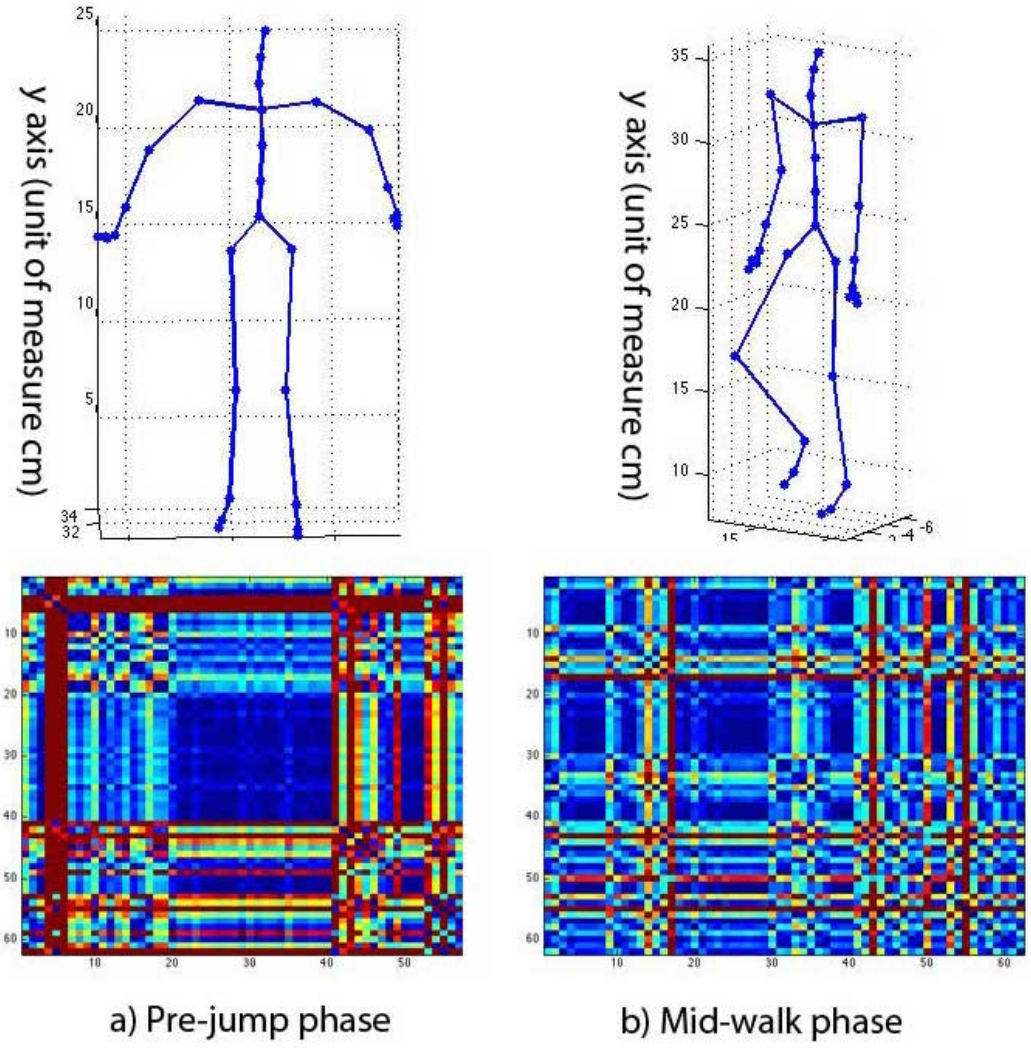

FIGURE 4.3: Example of skeleton postures and their respective distance matrices: a) A skeleton figure of a subject pre-jump phase. b) A skeleton figure of a subject walking.

$$
\begin{aligned}
M(\vec{p}_1;\vec{p}_2) &= H(\Psi) + H(\Omega) - H(\Omega,\Psi) \\
&= -\sum_{i\in\Psi} p(i)\log p(i) - \sum_{q\in\Omega} p(q)\log p(q) + \sum_{q\in\Omega}\sum_{i\in\Psi} p(i,q)\log p(i,q) \\
&= \sum_{i\in\Psi}\sum_{q\in\Omega} p(i,q)\log\left(\frac{p(i,q)}{p(i)p(q)}\right)
\end{aligned}
\tag{4.8}
$$

where $H(\Psi)$ and $H(\Omega)$ represent the marginal entropies, $H(\Omega,\Psi)$ joint entropy based star skeleton joint grouping, and $p()$ is a probability density function to map joint distributions.

The Mutual Information (Eq. 4.8) is capable of identifying complex relationships between elements. In the DKPI approach, each action sequence in dissimilarity space is clustered using k-means. For each cluster, the dissimilarity space for each pose is decomposed into joint-based dissimilarity representations and compared in a recursive manner to all other joint dissimilarity representations of the cluster [155]. To determine which joints represent the most "active" a cost function is defined as:

$$cost(\vec{p}_1, \vec{p}_2) = \min_{\vec{p} \in k} \{M(\vec{p}_1; \vec{p}_2)\} \tag{4.9}$$

where k is a cluster of dissimilarity representations, and those joints that contain the lowest *cost* are identified as "key joints". To identify and retain the key joints, a user-defined parameter is required; this outlines the number of joints to retain. Retaining the same matrix format, for each cluster all joints that have not been selected as "key" are assigned zero values and only those joints which have been identified as "key" contain joint information. This is undertaken to reduce confusion and provide a more representative element for training and classification.

4.2.2.4 Sequence Reduction

While it is possible to train a machine learning classifier solely on the dissimilarity representations defined earlier, data similarity can cause classifier confusion, but also result in higher training times. Therefore, to improve classification accuracy and reduce model complexity (thus reduce training time) a sequence reduction algorithm is required.

Taking the "key" joint representation encoded earlier, k-means clustering is applied to the feature set to obtain a set of k clusters. For each cluster, the average in-cluster distance is computed to determine how close each pose is to each other pose of that cluster. Recall that $\mathcal{P}$ has J multiple star skeleton representations, K clusters and N poses for each cluster. For each cluster, sum of Euclidean distances between each other pose is computed as follows:

$$score = \frac{\sum_{i=0, i \neq m}^{N} \sum_{k=0}^{J} (\vec{p}_{i,k}, \vec{p}_{m,k})^2}{N^2} \tag{4.10}$$

With a user-defined parameter (percentage) used to determine the number of poses to retain within each cluster. Simply, retain X amount of poses that have the lowest *score*.

4.3 Recognition

As has been mentioned previously, human action classification is a relatively difficult task, more so when there is a requirement to perform classification quickly, or otherwise to provide a result within a reasonable time to enable the human user to make a judgement. Given a test stream of MoCap, $A = (p_1^J, p_2^J, \ldots, p_T^J)$, where T denotes the total number of pose frames and J is the number of joints. A fair assumption is that any human action is an evolutionary process over time, and forms a time sequence. Inspired by window-based approaches [22, 125], a set of A poses is decomposed into a variable number of windows groups, denoted as w, that contains s number of poses. This can be represented as:

$$A = \{\underbrace{[p_a^J, \ldots, p_b^J]}_{w_1}^{\propto s}\}, \{\underbrace{[p_c^J, \ldots, p_t^J]}_{w_2}^{\propto s}\} \tag{4.11}$$

where $a \leq b \leq c \leq t$. The decomposition of A into w windows enables classification to be performed at an undetermined instance in time. More importantly, the window-based approach generalises over small variations in MoCap.

To measure the similarity between two poses, p_1 and p_2, the City Block metric, denoted as C_p is employed and given as:

$$C_p(p_1, p_2) = \sum_{j=1}^{J} |p_1^j - p_2^j| \tag{4.12}$$

where j is the joint index and C_p is the similarity between p_1 and p_2 respectively. This method has multiple benefits, not least the ability to undertake a recursive similarity measure between $model^l$ and each pose contained within a specific window group, $w \in W$. Employing Dynamic Programming (DP) ensures efficiency and speed to enable suitable processing power to make decision in a real time environment. Within each w window, a majority voting technique is employed, where the class with the most votes

in a window is selected as the window vote - very similar to the First-Past-The-Post principle used in General Election Voting. A prediction of the action class is determined by aggregating the winning vote from each window, V_w. This is defined as:

$$\mathcal{Y} = \max f\frac{l}{w}(V_w) \tag{4.13}$$

where V_w is the vote for each window, f is a frequency function and $\mathcal{Y}$ is the class prediction based on group majority voting.

4.4 Experiments

This section contains action classification results for detecting typical gaming motions using two proposed approaches defined in Section 4.2.1. To simulate a real world application, all experiments presented hereafter were conducted on *unknown participant actions*, meaning that no data from the participant being tested was included in the training set. This is a very important aspect, as it is reasonable to assume that training and testing sets will be different for real world usage. Protocols that are common within the community do not draw this distinction and perform recognition on known participants.

4.4.1 Protocol and Machine Learning

Marker-based MoCap data extracted from the CMU [4], HDM05 [85] and TUM Kitchen [86] datasets described in detail in Section 2.3. These datasets consist of a number of marker-based MoCap sequences providing a range of human action sequences performed by a number of different participants in a variety of recording environments. Processing these datasets offer two very important advantages over other methods of evaluation, (i) a recorded ground truth of the MoCap permits the quantitative evaluation of the proposed framework and (ii) the same participant performing the same action multiple times and that is intended for training. Each dataset was captured at 120Hz.

To assess the performance of both approaches, the following experiments were performed. Firstly, machine learning techniques SVM [141] and RF [144] and ANN were trained for

each approach with a two-folder cross validation undertaken to fine-tune the algorithms. Secondly, a comparison using the proposed window-based technique is presented and tested.

4.4.2 Selection Criteria

4.4.2.1 Approach 1: Delegate Identification and Selection

The DIS framework selects delegate postures using a statistical ranking and joint discrimination function. A generative exemplar-based framework for human action classification, with MoCap as an input, is introduced and discussed in the subsequent sections.

TABLE 4.1: Delegate Identification and Selection: Obtained k selection based on automatic WCSS for k-means clustering with FPS rate for classification.

Dataset	Avg k	FPS
CMU	23 ($\pm$19)	23
HDM05	12 ($\pm$7)	24
TUM	24 ($\pm$3)	24

The approach has been developed for practical HCI-based autonomy, with an attempt at removing as much human interaction as possible. Hence, there are only two user defined parameters; for exemplar construction e denotes the number of exemplars to retain for each action class, as stated in Eq. 4.4. For classification, s denotes the window size, as seen in Eq. 4.11. However, the alteration of either of these parameters can affect the result, as demonstrated in Figure 4.4, when the window size is altered. A minimum accuracy is reached when approximately 50 frames form the window size, this may suggest that the high frame rate provided by MoCap can result in intra-class confusion when motions are changing from one to the other. However, it is clear that the accuracy fluctuates across the window sizes.

Based on fine-tuning and experimentation, e was set at 3 and the window was set at 10. While higher classification results were possible for each dataset, a default value was set to provide comparability. This was also to enable generalisation for potential real world application where fine tuning is not possible. The window size parameter is not unreasonable considering the data capture rate is 120Hz. It is important to note, emphasis is placed on selecting suitable action sequences for exemplar creation. As these actions are modelled without any artificial representation, if they are not appropriately

selected, classification rate can be affected. The approach relies on k-means to group similar poses, if a poor quality action is provided it would result in a poor representation. Unlike other works, where k is set manually and/or without a protocol, this framework automatically selects the k value, which is presented in Table. 4.1. The k calculation reflects the freedom in which the datasets were captured, for example CMU and TUM allowed the subjects to perform movement relatively freely. The HDM05 was captured with a rigid protocol in place, thus the resulting k was lower.

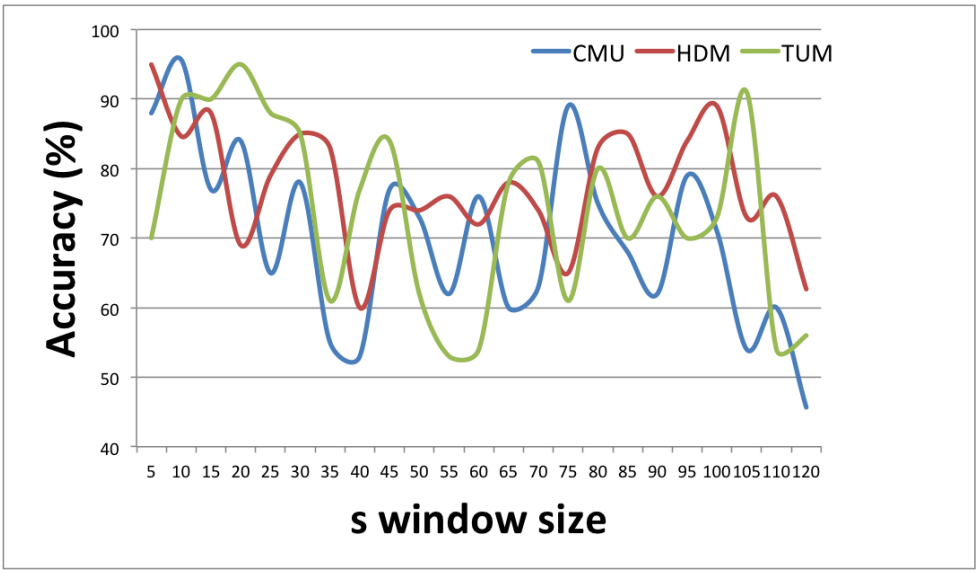

FIGURE 4.4: Delegate Identification and Selection: Accuracy results for each dataset as the window size s is increased.

4.4.2.2 Approach 2: Discriminative Key Pose Identification

DKPI determines key poses by assessing the maximum subspace scoring of the dissimilarity space of the star skeleton representation. The approach computes local representations based on joint dissimilarity and mutual joint respect to identify key poses.

TABLE 4.2: Discriminative Key Pose Identification: Obtained k selection based on automatic WCSS for k-means clustering with FPS rate for classification.

Dataset	Avg k	FPS
CMU	11 ($\pm$4)	38
HDM05	14 ($\pm$4)	39
TUM	17 ($\pm$6)	36

The approach seeks to reduce an action sequence, while maintaining key representations. Unlike the approach proposed in Section 4.2.1, the number of delegates are dynamic,

based on the action complexity with regards to the discriminative power and mutual respect to other poses. Two user-defined "retain" parameters are necessary; number of active joints to retain and a percentage value for the number of poses to retain. As with the previous approach, s denotes the window size, as seen in Eq. 4.11. Altering the size of the window can alter the recognition result, Figure 4.5 demonstrates the change in accuracy as the window size increases. The accuracy fluctuates depending on the type of action sequence and dataset. Introducing more frames within the window tends to improve and maintain classification accuracy. However, as discovered in DIS, accuracy is partly dependent on window size.

Based on fine-tuning and experimentation, the window size was set at 40 and the number of poses to retain was set at 0.40 - meaning we'd keep 40% of all possible poses. While it would be possible to achieve higher accuracy results by fine-tuning the parameters for each datasets, a default value was set to provide comparability. The approach relies on k-means to group similar poses, if a poor quality action is provided it would result in a poor representation. Unlike other works, where k is set manually and/or without a protocol, this framework automatically selects the k value, which is presented in Table 4.2. The k calculation reflects the freedom in which the datasets were captured, however, unlike in DIS approach, clustering is performed on the star skeleton dissimilarity space representation.

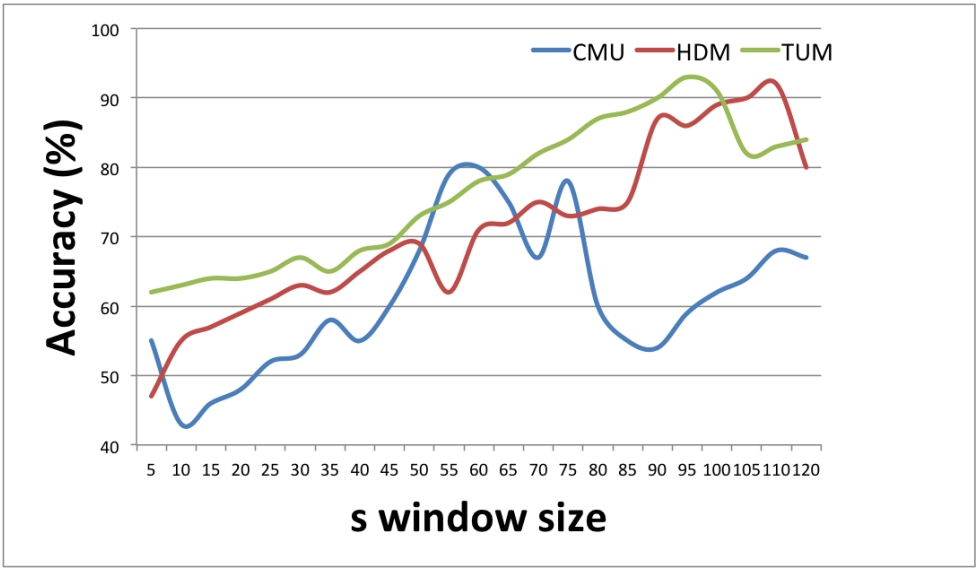

FIGURE 4.5: Discriminative Key Pose Identification: Accuracy results for each dataset as the window size s is increased.

TABLE 4.3: Result Comparison: Classification accuracy results for Delegate Identification and Selection and Discriminative Key Pose Identification approaches compared against previous best.

Dataset	DIS (%)	DKPI (%)	Previous best
CMU	89.01	92.96	90.92 [22]
HDM05	93.19	94.45	97.27 [153]
TUM	86.29	91.30	92.56 [22]

4.4.3 Result Comparison

Both approaches have competing benefits. Approach 1 (DIS) is very quick in developing a model and performing classification, whereas approach 2 (DKPI) takes considerably longer to construct a model and perform classification. Table 4.3 presents a comparison of the classification results obtained through both approaches. It is clear that approach 2 has achieved higher classification results, whereas approach 1 has still performed very well considering the simplicity of the approach.

Evaluating DIS against the CMU dataset against machine learning techniques resulted in a lower-than-expected classification result, with SVM obtaining 58.19%, RF obtaining 62.71% and ANN obtaining 68.75%. The low results indicate the possibility that machine learning struggles without ample training samples to compute a model. In comparison with other state-of-the-art, experimental protocol in [22], the same protocol was utilised to allow for direct comparison. An exemplar model with 9 action classes was constructed and tested on 49 test sequences (as in [22]). The approach presented achieved a classification rate of 87.78% when compared to [22] classification accuracy of 90.92%. This is reasonably good considering the limited number of training samples used and the simplicity of the approach. However, to assess the possible potential of the framework for real world deployment, the number of test sequences was increased. A total of 180 test sequences from the CMU dataset were selected. In this experiment a classification rate of 89.01% was obtained. While this result is slightly lower than the previously reported accuracy, it reflects the capability of the approach to handle a significant number of test sequences with varied body compositions.

The DKPI resulted confident classification results, with SVM obtaining 80.19%, RF obtaining 83.76% and ANN obtaining 84.87%. These results are significantly higher than those obtained using DIS, this may be down to more samples being used for training.

In comparison with other state-of-the-art, the approach presented achieved a classification rate of 94.04% when compared to [22] classification accuracy of 90.92%. This is an improvement, however this may be due to a higher number of samples used for training. To assess the possible potential of the framework for real world deployment, the number of test sequences was increased. Using the same experimental protocol previously mentioned with a total of 180 test sequences a classification rate of 92.96% was obtained.

Benchmarking DIS against the HDM05 dataset, subject b was randomly selected from the dataset, then 9 action classes were selected to construction the exemplar model. For classification, 160 test sequences from the remaining 4 subjects were randomly selected. Using machine learning techniques, an overall low classification rate was achieved, with SVM obtaining 48.62%, RF obtaining 70.10% and ANN obtaining 58.10%. A conclusion can be drawn for this the varied machine learning results; that the low classification results highlights the need for a robust classification framework to take into account spatio-temporal changes. Current state-of-the-art approaches, [156] reported an accuracy of 80% when using their manual *key-frame* approach. While this was a good classification rate, the approach is fully automated and able to operate online without extensive training and manual subjective input. Further, Gowayyed et al. [153] utilised Histogram of Oriented Displacement, where they achieved an accuracy of 97.27%. A classification rate of 93.19% was obtained using the approach presented.

For the DKPI framework, SVM obtained 73.22%, RF obtained 79.12% and ANN obtained 87.29%. Accuracy obtained through machine learning is near that achieved by using the window-based approach, however model complexity and model training are key shortcomings. In addition, machine learning is not able to factor in temporal aspects of human motion. A finally classification rate of 94.45% was obtained. This demonstrates that DKPI framework is robust and efficient at representing human motion.

To provide comparison with other works that have utilised the TUM dataset the DIS framework was compared using a set of sequences $(0{-}2, 0{-}4, 0{-}6, 0{-}8, 0{-}10, 0{-}11, 1{-}6)$ for testing and the remaining 13 sequences to construct the exemplar model. To maintain consistency with others who have used this dataset ([22, 91]) a 10 class test protocol where "jogging" and "walking" were separated was selected. For machine learning, the approach achieved an overall low classification rate, with SVM obtaining 59.52%,

RF obtaining 64.87% and ANN obtaining 48.98%, albeit slightly better than the CMU dataset. Using the window-based approach, a classification rate of 86.29% was achieved. The classification rate achieved by using the approach is comparable to current state-of-the-art. In [22], they were able to achieve an accuracy of 92.56% by using integral histograms. This is considerably better than early work in [54], where an accuracy of 62.77% was reported.

Evaluating the DKPI framework, SVM obtained 78.46%, RF obtained 72.75% and ANN obtained 69.71%. Using the window-based approach, a classification rate of 91.30% was achieved. The dataset contains a number of complex and varied sequences in which the model was required to model intrinsic details; the classification results highlight the versatility of this approach.

Two approaches have been evaluated against benchmark datasets. The DIS framework is capable of efficiently recognising action. While several major works have been capable of achieving high accuracy rates, they are fine tuned to specific actions and/or datasets. This framework is simple, yet elegant and is capable providing suitable classification results. The DKPI framework is capable of efficiently recognise human action. When presented with complex human action, the approach is capable of modelling key phases while reducing the model size.

4.4.4 Phase Detection

The exemplar paradigm requires the selection of descriptive poses for construction of the exemplar model. This chapter has demonstrated the ability to detect key poses, as demonstrated in Figure 4.6. Key exemplars have been important to allow for human action classification, as highlighted with both approaches. Extending the concept of both approaches, it is possible to identify and represent key phases of an action sequence, as demonstrated in Figure 4.6 the second level represents the exemplar sequence for each phase. The exemplars, by visual inspection are representative of the phases for the action. Future chapters of this thesis will explore human action segmentation for the task of classification and human analysis.

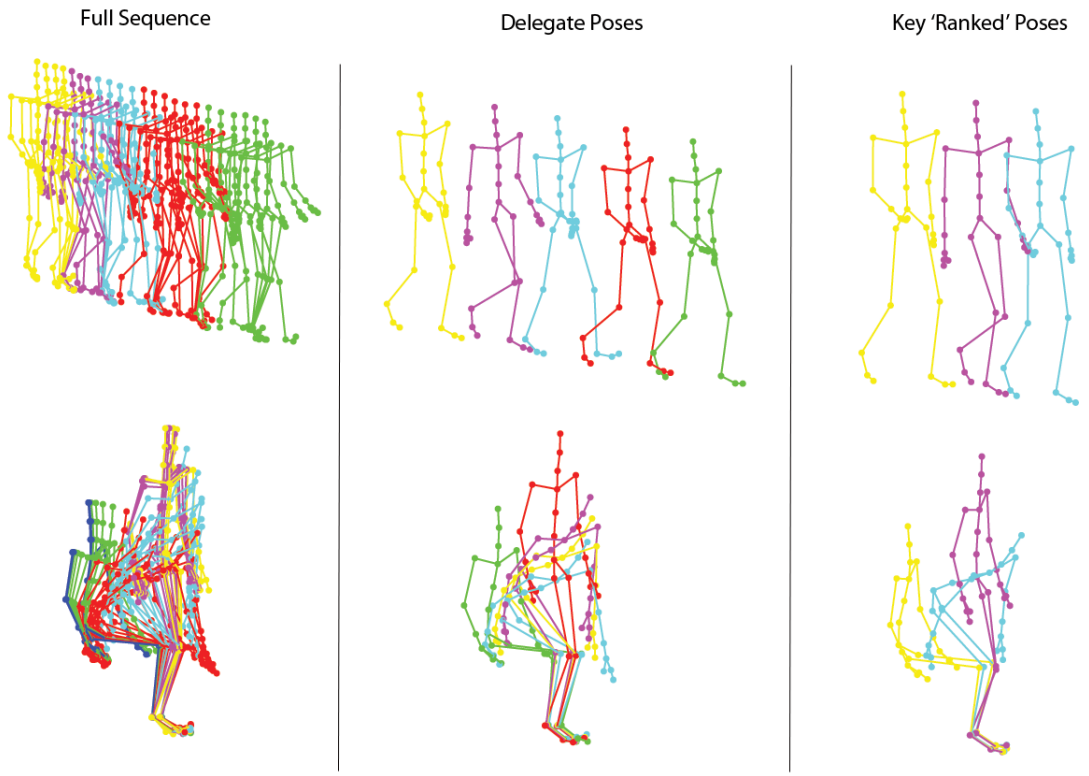

FIGURE 4.6: Decomposition of a human participant. Top row denotes a person walking in a straight line. Bottom row is a person performing a chair rise (extracted from MoCap). (a) original motion sequence of a human. (b) k-means clustering, where each colour denotes the cluster. (c) delegate pose for each cluster. (d) the selected delegate exemplars for the motion sequence. where $e = 3$.

4.5 Discussion and Conclusion

The techniques presented in this chapter demonstrate the ability of both approaches to handle a large variety of action sequences from different sources for applications in HCI. Further, and more importantly, both approaches presented have removed the need for large-scale training of complex MoCap sequences. For example, if we were to consider the proposal of Barnachon et al. [22], the framework relies on manual selection of k value for clustering which is an incredibly difficult task. When constructing a model with the exemplar paradigm it is important to select the most representative action sequences, as the model will be generated based on the observed sequence. Any irregular, or poor performance in the activity would otherwise be modelled. In the approach of [22], sequences are manually selected and large-scale testing does not seem feasible.

The DIS framework selects delegate postures using a statistical ranking and joint discrimination power. A generative exemplar-based framework for human action classification,

with MoCap as an input, is introduced and discussed in the subsequent sections. By ranking and then selecting the most informative poses based on statistical significance, instead of selecting the most informative based on cluster generalisation and placing in time sequential order, proved to be very effective. On the other hand, DKPI determines key poses by assessing the maximum subspace scoring of the dissimilarity space of the star skeleton representation. The approach computes local representations based on joint dissimilarity and mutual joint respect to identify key poses. The approach consistently achieved the higher classification accuracy for all experiments. This may be due, in part, to the model containing more examples of the action class that makes it easier to perform classification. Nonetheless, a challenge still remains; the ability to select the most suitable representation for each action class is very difficult. The exemplar paradigm requires the manual selection of the most typical sequence, future work should seek to address the question; is it possible to determine what is a correct performance of an action sequence free from human interpretation? In addition, work should focus on determining how to select an action sequence to model that represents the ideal characteristics of the motion. This thesis unfortunately does not directly address this question. However, future work should seek to address this.

Several works have sought to reduce motion sequences to it's key representations, such as Barnachon et al. [22], Gowayyed et al. [153] and Bloom et al. [117] have required manual selection of a large number of parameters such as the number of clusters to generate. Both approaches presented in this chapter require user involvement in parameter selection that may have inadvertently affected classification accuracy. Automation takes the subjective decision-making process away from the user and places it within the confides of the machine. Manual selection of the cluster parameter can influence the overall result, as has been found in the works above. This thesis attempts to remove the need for large amounts of user involvement to develop an automated supportive framework for classification, detection and analysis.

This chapter has introduced two approaches for human action recognition in which a larger number of varied action classes can be recognised, as well as reducing the intra-/inter-class variations. The main contribution of this work lies with the ability to detect *key* poses to represent each cluster and in turn action class. The use of an action model to represent each action class has shown strengths compared to traditional data intensive training-based approaches. However, it is important to consider the practical

implications of marker-based MoCap and the proposed approaches. There is a "start up" cost in placing the markers on the participant, setting up the hardware and calibration. It is not feasible for implementation of marker-based systems in the *real world*. While action classification can be performed robustly in an offline approach, it is important to make decisions as quickly as possible. Chapter 5 extends the concepts and principles presented in this chapter to perform human action recognition in *real-time*.

Chapter 5

Exemplar Paradigm: Online Template Matching and Posture Representation with Marker-based MoCap

In this chapter a template-based exemplar approach with recognition performed online is introduced. A set of delegate poses are recovered to represent each class by applying k-means clustering. A dynamic model, based on a novel similarity function is constructed to represent a set of diverse action classes. Robust feature representation is demonstrated using exponential map transformation with a novel real-time recognition framework.

5.1 Introduction

Recognising and classifying human action is a difficult task. This is in a large part due to large variations among subjects, overlap of poses between actions and data noise [6]. It is important to have a sufficient number of training samples to allow for each action to be represented effectively. As a result, a large number of training samples are common for many recognition approaches. However, it is not always practical to utilise a large number of training samples due to the size and complexity in computing classification.

Therefore, this chapter proposes the use of the exemplar paradigm as an effective approach to modelling human action. In the approach proposed, the exemplar paradigm is implemented to reduce an action sequence to its most descriptive *key* elements.

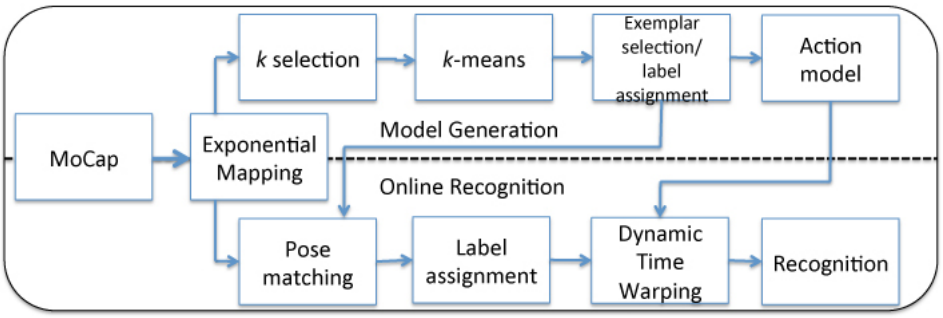

FIGURE 5.1: Flowchart of the approach. Top row denotes the process for generating an action model. Bottom row denotes the online recognition framework.

This chapter presents a combined template-based exemplar approach for online action recognition evaluated by using streamed motion sequences provided by three popular marker-based MoCap datasets (as demonstrated in Figure 5.1). To construct an exemplar-based action model a single action sequence is used to represent each action class, with each pose of the sequence transformed and represented in Exponential Map form. Because of the success of clustering in chapter 4, k-means is employed to group each action based on similarity. For each k cluster, a delegate exemplar is selected based on a novel ranking scheme. An interesting by-product of the clustering process (as discussed in Section 4.4.4), is that specific phases of the action sequence can be represented, with each delegate exemplar placed into a time-ordered sequence to reflect the differences phases. For recognition, each pose is transformed into Exponential Map form and compared against the delegate exemplar model using the City Block metric, which returns a prediction of the most similar exemplar. Finally, DTW is performed to match and recognise human action by analysing the labels associated with each classified pose and a corresponding exemplar.

The main contributions of this chapter are as follows:

1. The use of the exemplar paradigm to model "key" descriptive elements of action sequences enabling the overall reduction of the training sample (Section 5.2).

2. The integration of k-means clustering and automatic selection of the k criteria for segmenting MoCap (Section 5.2).

3. The use of a template-based label matrix to represent each action classes and Dynamic Time Warping for online real time recognition (Section 5.3).

5.2 Exemplar-based Template Model Definition

Recall that human motion, typically captured by a marker-based MoCap system, is modelled using a *kinematic chain*. A kinematic chain consists of *body segments* that are connected to various body *joints*. A sequence can be seen as a time-sequential sequence of 3D joint coordinates that relate to the fixed kinematic chain. In this thesis, a motion sequence is a series of frames (otherwise denoted as poses), with each frame specifying the 3D coordinates of the joints at a certain time period. Recall, in Section 3.1.1, a sequence is denoted as $\mathcal{P} = \{p_t^j | t = 1, \ldots, T; j = 1, \ldots, J\}$, where t denotes the time and j is the joint index.

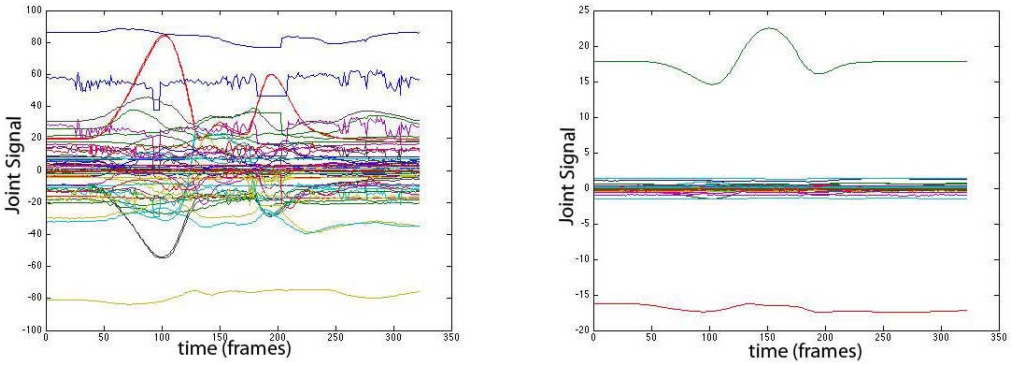

FIGURE 5.2: Joint Representation: A signal for the action *Jump* from the CMU Dataset. a) Euler Angle signal. b) Exponential Map signal.

As highlighted in Section 3.1, there is no single solution to parameterisation of MoCap rotation that are suitable for all application domains. The approach presented in this chapter relies on parameterising the 3D Euler Angle (refer to Section 3.1.3 for a detailed discussion) into Exponential Map form (see Eq. 3.6). The representation and transformation from Euler Angle to Exponential Map can be visualised in Figure 5.2. This representation has been selected over other approaches as it avoids gimbal lock, discontinuities and ball-and-socket joints complications that are associated with using marker-based MoCap data.

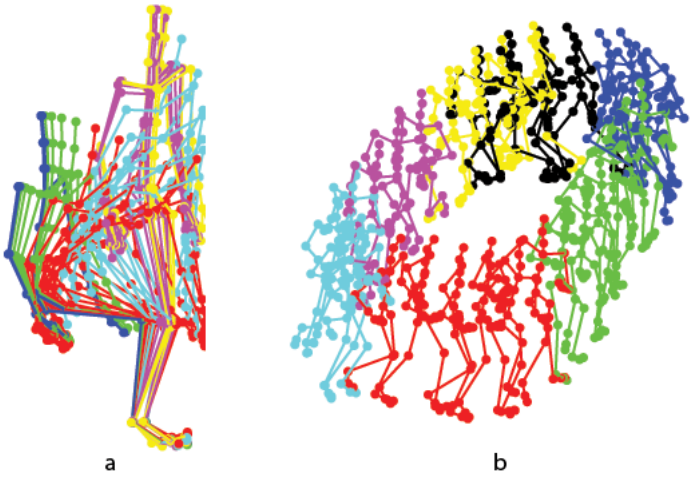

FIGURE 5.3: Cluster Visualisation: Example of a MoCap sequence represented by clusters. Left: *Chair Rise*. Right: *Running in a circle*.

5.2.1 Delegate Selection

The k-means clustering algorithm underlines the proposed approaches in this section (presented and discussed in Section 3.1.7). k-means is iterative in nature, starting with an initial estimation of the centroid for each cluster which continues until convergence of a motion sequence into an assigned number of k clusters. It has been shown to be efficient in segmenting and clustering MoCap data, as demonstrated by Zhou et al. [136]. As highlighted earlier in chapter 4, k-means is computationally faster when dealing with a large number of observations (such as poses) when compared to hierarchical clustering methods making it ideal for clustering marker-based or marker-less MoCap data.

The objective is to cluster an action sequence in to k clusters, with each cluster containing similar poses. Hence, as a by-product of k-means process, each cluster characterises a phase of the action. Figure 5.3 demonstrates the clustering process for two distinctively difference action sequences. Observe that for each cluster, similar poses are grouped together, making it ideal for the application of pinpointing similar poses. However, the difficulty with k-means is the selection of k. Refer to Section 3.1.7.2 for a detailed discussion on selecting the optimum k.

To select a delegate for each cluster, a ranking scheme is applied for each cluster pose according to the City Block metric (also referred to as Manhattan distance). The equivalence D between any two poses p_t^j in a single cluster is measured using the total distance amongst corresponding joints, defined as:

$$D(p_t^j, p_t^j) = \sum_{j=1}^{J} |p_t^j - p_t^j| \tag{5.1}$$

A delegate exemplar, denoted as $E_{e,k}$, is a pose which has the lowest distance average between a cluster grouping of poses. A visual representation of the data at each stage is demonstrated in Figure 5.4.

5.2.2 Temporal Pose Ordering

With each delegate determined for each cluster it is possible to determine where in the temporal sequence the pose has occurred. Therefore, each delegate is ordered based on appearance in the original sequence and assigned a unique label. This label is utilised by the recognition framework to perform template matching. Thus, for each action model, k number of delegate exemplars are retained and a time-ordered unique label sequence to represent each action class. For simplicity, let $C = \{c_e | e = 1, \ldots, E\}$ be the action set, where $c_e = \{E_{e,k} | k = 1, \ldots, K_e\}$ be the action model which is composed by K number of exemplars for action class e.

5.3 Recognition: *real-time* classification

The approaches presented in chapter 4 operated in an offline manner, meaning that they were not able to provide results in *real-time*. In this section, a real-time classification framework is proposed to enable recognition to performed as the action sequence unfolds.

Given a test stream of MoCap, $A = (p_1^J, p_2^J, \ldots, p_T^J)$, where T can be of any length and J is the number of joints. Each t-th pose is treated as an individual entity, with the assumption that the evolution of the sequence would result in an observed action sequence in temporal order. The classification of each system is as follows.

For each t-th pose of A, a parameterisation process is undertaken, each pose is parameterised in Exponential Map form (as mentioned previously, see Section 3.1.3). With the pose represented, the distance function defined in Eq. 8.6 is employed to determine the similarity between the t-th pose and the delegate exemplar model representing. The comparison between each pose and the exemplar model is given as:

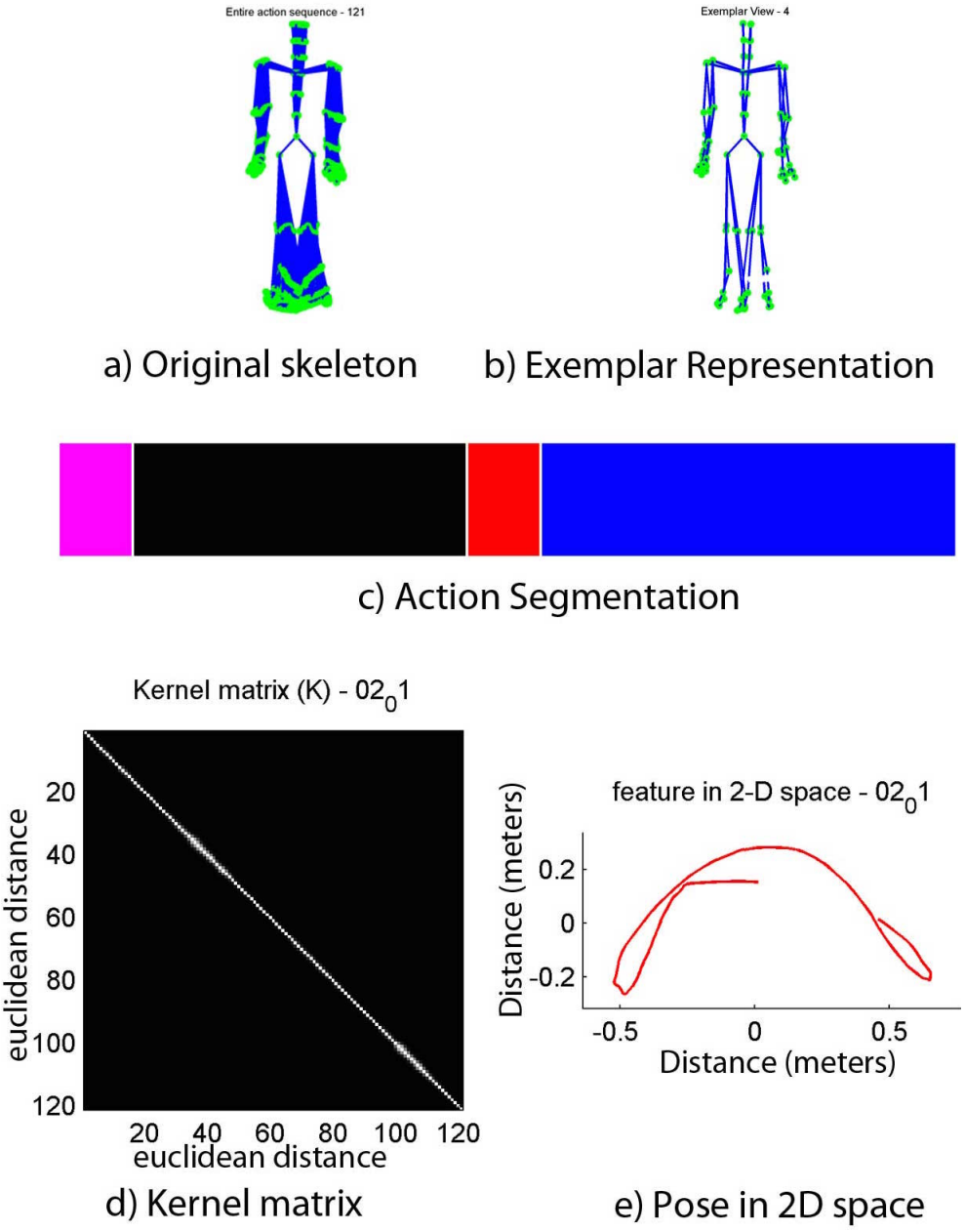

FIGURE 5.4: Example of the "walk" sequence from the CMU Dataset. a) Original skeleton sequence. b) Delegates representing the action sequence. c) Visual segmentation bar of the sequence. d) Kernel matrix representing the agreement. e) Pose projected and visualised in 2D space.

$$L_t := min\{D(\bar{a}_t, E_{e,k}) | e = 1, \ldots, E; k = 1, \ldots, K\} \tag{5.2}$$

where L_t is a class indicator matrix. The L_t for each pose is determined by the label assigned to the delegate pose. Recall, for each delegate exemplar a unique label is assigned. The defined exemplar, and the associated label form a unique string that

represents the action being considered. A version of DTW is implemented to determine the action class of the current t-th observation based on the history if assigned labels. Thus, over time a sequence of unique labels describing the action is constructed. It is possible to then consider the temporal domain, as classifying poses individually will not provide the context to allow for robust recognition. An observed sequence can be of any length, to handle this DTW is employed as the action unfolds to match the observed label sequence to one of our previously learnt unique label sequences which describe the phases for each action class.

Given an observed label $L_1, \ldots, L_t$ which is derived using the approach above, a class indicator matrix is compared with each action model c_e to find the best pattern match. Figure 5.5 demonstrates how the mapping between the model and a test sequence occurs. The winning class $\mathcal{L}_t$ is determined by the minimum DTW cost with respect to the Itakura constraint, given as:

$$\mathcal{L}_t := min\{DTW(L_{1\sim t}, c_e)|_{e=1,\ldots,E}\} \tag{5.3}$$

Finally, by having each pose classified, recognition of the sequence up to any time period is based on the cumulative voting indicated in $\mathcal{L}$. The recognition rate is computed by the total number of correctly classified frames divided by the total number of frames up to point t.

5.4 Experimental Results: *Real-time* Recognition

This section contains action recognition results for detecting typical gaming motions using the approach defined in Section 5.2. The same experimental protocol as employed in Chapter 4 is used. Further, to simulate a real world application, all experiments presented hereafter were conducted on *unknown participant actions*, meaning that no data from the participant being tested was included in the training set. While it is possible to construct a model with multiple action sequences, this chapter has introduced a framework that requires only a single action sequence to construct an exemplar model. All results presented in this section were computed based on the average accumulation for classification.

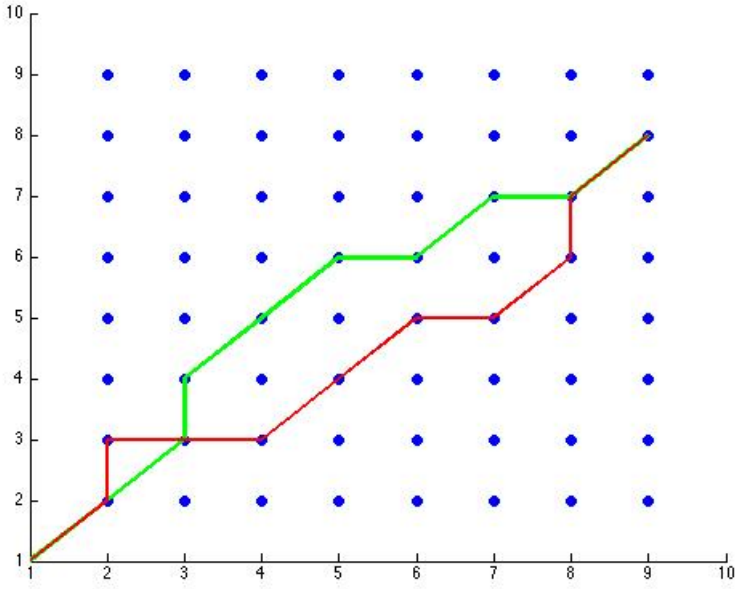

FIGURE 5.5: Example of the best path matching between the action model and a test sequence. Green: Denotes the correct label sequence and structure. Red: Shows the attempted mapping for the test sequence. Note: there is no end point constraint.

5.4.1 Protocol and Machine Learning

Marker-based MoCap data extracted from the CMU [4], HDM05 [85] and TUM Kitchen [86] datasets are described in detail in Section 2.3. These datasets consist of a number of marker-based MoCap sequences providing a range of human action sequences performed by a number of different participants in a variety of recording environments. Processing these datasets offer two very important advantages over other methods of evaluation, (i) a recorded ground truth of the MoCap permits the quantitative evaluation of the proposed framework and (ii) the same participant performing the same action multiple times and that is intended for training. Each dataset was captured at 120Hz. Each dataset contains a number of 3D Euler joint angles extracted from markers (CMU: 41, HDM05: 42, TUM: 46) that are placed on anatomical landmarks of a human subject. In addition to the joint angles, root orientation and translation is provided for each frame. Each dataset was captured and presented at 120Hz.

To be consistent with the current state-of-the-art [22], 9 action sequences, each representing a single action class were extracted and 49 action sequences selected for testing from the CMU dataset. In total, 9 action sequences, each representing a single action class and 144 action sequences for testing were extracted from the HDM05 dataset.

Finally, 9 action sequences, each representing a single action class and 108 segmented action sequences were extracted from the TUM Kitchen dataset. As a pre-processing task, six joint angles contained within CMU, five from the HDM05 and six from the TUM datasets consisted of constant values, so they were removed from the training and testing sequences. Therefore, this reduces model complexity and removes the need for the approaches to model zero values. The remaining joints had between two and three DOF. For testing, the conversion process was undertaken on a per pose basis in a real-time manner.

To assess the performance of the approach, the following experiments were performed. Firstly, machine learning techniques SVM [141] and RF [144] and ANN were trained for each approach with a two-folder cross-validation undertaken to fine-tune the algorithms. Secondly, a comparison using the DTW-based label sequence matching is undertaken.

5.4.2 Recognition

Using the benchmark datasets, defined in Section 5.4.1, the approach has been compared to those introduced in chapter 4 and these results are summarised in Table 5.1. In addition, the approach has been compared to state-of-the-art machine learning techniques, these results are summarised in Table 5.2.

TABLE 5.1: Exemplar-based template matching. Recognition accuracy and recognition time (in milliseconds) for each dataset when compared with chapter 4 approaches.

Dataset	Proposed Approach	DIS (Section 4.2.1)	DKPI (Section 4.2.2)
CMU	91.89% (17ms)	89.01%	**92.96%**
HDM05	**97.95%** (9ms)	93.19%	94.45%
TUM	**93.93%** (11ms)	86.29%	91.30%

The desire to perform recognition in real-time is the main contribution of this chapter, observed in Table 5.1 that recognition has been performed in real-time and all fall under 17ms. The difference in recognition times is, in part, due to the type of datasets and their dimensionality.

The CMU dataset achieved a recognition rate of **91.89%** using the exemplar-based approach introduced in this chapter. However, this falls behind the proposal in Chapter 4. This may be in part due to only a single action sequence being trained, further, unlike the DPKI approach, the objective is to perform recognition in real-time. Therefore, the

delegate representations contained a more generalised representation that may have impacted the recognition results. The approach was capable of distinguishing between *Walking* and *Running* sequences due to relatively small variation amongst subjects. However, it was observed, that for sequences such as *Boxing* and *Punching*, the temporal variations caused inter-/intra-class variations which impacted the results. A second set of experiments was performed, in which state-of-the-art machine learning techniques were trained using the exemplar model. As can be observed in Table 5.2, each technique was capable of recognising the human action sequences to a high confidence, yet the proposed recognition framework has been able to achieve a higher accuracy as it takes into account prior history.

TABLE 5.2: Exemplar-based template matching. Recognition accuracy when compared to state-of-the-art machine learning techniques.

Dataset	Proposed Approach	SVM	RF	ANN
CMU	**91.89%**	78.85%	83.67%	72.84%
HDM05	**97.95%**	82.45%	86.81%	78$\dot{2}$1%
TUM	**93.93%**	77.79%	89.28%	79.56%

In the HDM05 dataset, a recognition rate of **97.95%** (see Table 5.1) was achieved using the exemplar-based framework. For the HDM05 dataset, the approach has outperformed both approaches defined in chapter 4. However, it must be noted that the other two approaches performed very well. This may be due to the way in which the dataset has been collected, in a rigid manner resulting in higher recognition results. As with the CMU dataset, *Walking* and *Running* classes were clearly identifiable. However confusion was observed for *Sit* and *Squat*, this may be due to similar poses being performed between the two classes. To overcome this, a consideration may need to be made to factor the temporal domain. Tentatively, interclass confusion remained limited reflecting the strength of the approach to correctly model action sequences using the exemplar-based paradigm. As can be observed in Table 5.2, each technique was capable of recognising the human action sequences to a high confidence, however, there is a 12 percentage point difference between the state-of-the-art and the proposed recognition framework.

Finally, for the TUM dataset, a recognition rate of **93.93%** (see Table 5.1) was achieved using the exemplar-based framework. Using the TUM dataset, the approach outperformed, albeit marginally, both approaches introduced in chapter 4. This may be due to the limited action types within the dataset, with several distinguishably different making the recognition process less of a challenge. The actions contained within the

dataset reflect a kitchen environment, class confusion was observed for those that had overlapping actions such as *placing an item in a high cupboard* and *placing an item in a low (floor level) cupboard*. Once again, the temporal domain may aid in improving recognition accuracy further. As can be observed in Table 5.2, each technique was capable of recognising the human action sequences with several machine learning techniques obtaining the highest accuracy results.

The approach presented in this chapter was capable of outperforming the current state-of-the-art, and the approaches proposed in chapter 4. This is significant advance on current approaches, with the added benefit of being a more straightforward in analysing highly complex datasets and also the small number of exemplars retained (average $k = 15$). For recognition in real-time it is important to obtain the classification as quickly as possible, the approach is capable of providing results in under 17ms, matching several state-of-the-art real-time techniques.

5.5 Discussion and Conclusions

The exemplar-based template approach for human action recognition appears to support robust recognition from a number of marker-based MoCap data. Selecting informative delegates proved to be very important as evidenced by high recognition results. However, it is important to highlight the class confusion between similar action sequences, which is widely seen throughout experiments. Nevertheless emphasis is placed on the importance of selecting *ideal* actions for representing each action class. Modelling inconsistent, or incorrect performance of an action could cause a reduction in performance. This challenge is further discussed in chapter 6.

The use of an action model to represent each action class offers an advantage over traditional approaches in terms of characterising each class by a small number of exemplars, which has reduced the need to use whole motion sequences for action representation by an average of 98%. This in contrast to using full motion sequences to train machine learning techniques, with performance inevitably suffering as the quality of training data degrades due to confusion. Chapter 4, explores this concept further and extends the use of feature selection and ranking to improve recognition results for online application. The use of Exponential Map representation enables characterisation of the posture in a

more discriminative usable form, but also handles singularities and discontinuities. The approach presented in this chapter has improved on the approaches defined in chapter 4 and enabled real-time recognition.

For recognition, this chapter has focused on pose classification and template matching using *real* number sequences, which has presented a number of challenges. It is difficult to distinguish certain actions due to poses being indistinguishable for a periods of time, such as similar gait cycles between *walking* and *jogging*, creating inter-/intra-class confusion. A solution, proposed by this chapter, is to focus on adjacent poses in the temporal domain. Focusing on temporal difference aids in identification of the correct action class as the sequence unfolds.

So far only marker-based approaches to human action recognition have been considered. Invasive marker-based MoCap have a number of advantages, the placement of markers on the body allows for accurate and reliable 3D feature extraction with no background subtraction requirements. However, they have a number of drawbacks, they are expensive and only function within a defined area (usually a laboratory). They further require trained users who post-process the data, manually identify and label markers. This leads to the following questions: is it feasible and practical to utilise marker-less tracking technology for use in feature representation, detection and classification? Chapter 6 presents a feasibility study for the ability of marker-less technology for use in classification, specifically focusing on health related tasks. With the concept extended further for age-related health implications and real world deployment in chapter 7 and chapter 8.

Chapter 6

Feature Representation with Marker-less MoCap data using Machine Learning

In this chapter the Microsoft Kinect 360 and the underlying skeletal tracking algorithm is explored to assess the reliability in a recognition framework. While marker-based MoCap has clear advantages, one of its major disadvantages is the hardware and markers which make it difficult to apply for real world application. Marker-less technology offers significant advantages in terms of cost and deployable applications. This chapter extracts the kinematic location, velocity and energy of each skeletal joint at each time period to form a feature representation. Principle Component Analysis is applied as a pre-processing step to reduce dimensionality and identify significant features amongst action classes. The resulting algorithm is demonstrated for recognition to analyse the ability of machine learning techniques to accurately classify action sequences.

6.1 Introduction

Thus far, this thesis has focused on marker-based MoCap datasets, and as discussed in chapter 5, these are expensive and difficult to deploy for real-time application. Amongst researchers, there has been a shift towards marker-less technology due to its low cost,

adaptability and accessibility. A detailed discussion is presented in Section 2.2.2. This section utilises the Microsoft Kinect 360 (Kinect 360), which is a low-cost peripheral accessory initially released for use with the Xbox 360 gaming console. The Kinect 360 allows for real-time body detection and tracking of human activities and gestures. By incorporating infra-red and RGB camera technology, the underlying body detection algorithms create a three-dimensional (3D) depth map of the area in front of the device, randomised decision forest algorithms are then used to automatically detect and determine anatomical joints on the body of the user and stream the 3D coordinate location for each joint [24].

This chapter builds upon the shortcomings of Patsadu et al. [157], where the Kinect 360 was evaluated by placing its on the ceiling facing vertically down at the participant. The conclusions presented in chapter 5 and the limited use of Restricted Boltzmann Machines [19] to explore the feasibility of using machine learning to perform recognition on a set of marker-less action sequences captured using a MoCap dataset. This type of feasibility study, at present, does not exist. While traditional frameworks seek to recognise gaming action sequences, this thesis intends to analyse human motion for health-related quantification. A health-related dataset comprising of 10 participants was recorded using the Kinect 360.

The approach uses skeletal information (see Section 3.1.5 and Section 3.1.1 for a detailed discussion), obtained from the Kinect 360 to form a feature vector comprising of the vertical location (y axis), velocity (over time) and energy of each skeleton joint for each t-th time period. Section 3.2 presents the machine learning techniques employed in this chapter. Each technique is assessed for the accuracy in which it recognises unseen test sequences. This chapter provides a feasibility and suitability of using Kinect 360 data for feature representation and recognition.

The main contributions of this chapter are as follows:

1. Novel feature representation of Kinect 360 skeletal information which represents the spatial-temporal information and encodes it into a single representation (Section 6.2).

2. Discussion and evaluation for the feasibility of machine learning techniques for recognising human action sequences using Kinect 360 derived features (Section 6.3).

6.2 Data Capture and Feature Encoding

Thus far, this thesis has focused on marker-based MoCap, which provides angle representations in Euler Angle form, however marker-less technology such as Kinect 360 provide Axis-Angle (coordinates), where x extends from the left to right, y indicates vertical position and z extends in the direction in which the Kinect 360 is facing. The approach utilises data obtained via the Kinect 360, transforms it into a suitable representation and trains several machine learning algorithms. The subsection will discuss the basic principles of the algorithm, and present terminology and methodology important to the overall framework.

6.2.1 Action Sequences and Data Collection

The data used in this chapter was acquired by Kinect 360 and Kinect for Windows Software Development Kit [158]. The Kinect 360 acquired the 3D coordinates of 20 fixed anatomical landmarks, as defined by Microsoft Corporation, at a rate of 30Hz. A visualise representation of a rendered skeleton is presented in Figure 6.1).

The Kinect 360 (tilt 0 degrees) was placed on a tripod at a height of 0.70 meters (m) with the participant standing 2m from the device in a defined movement area of 0.5m×0.5m. Participants were asked to perform actions periodically (to characterise temporal variations) within the defined movement area for a 10 second period directly facing the Kinect 360. These activities were chosen to reflect activities of daily living, as well as actions a person would perform while undertaking health-related training programs. These actions are typically used at home, but also in a motor-control rehabilitation setting. In addition, consideration was given to the type of actions being performed and the possibility that they may produce noisy skeleton data. Participants were asked to assume a neutral standing pose at the start and end of the activity, in which they stood still with legs fully extended and arms extended and relaxed by the side of the body. The aim was to ensure consistency between the training and testing datasets and, to

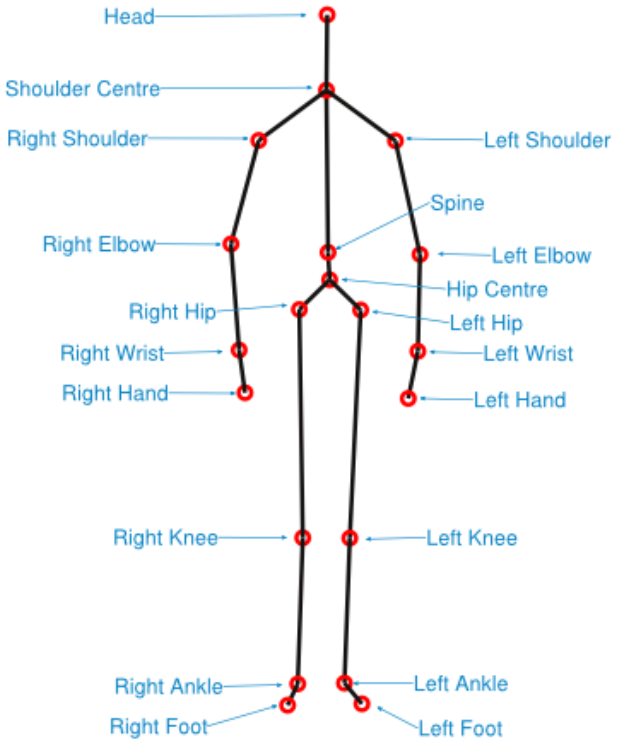

FIGURE 6.1: Visual representation of Kinect human pose and associated joints.

limit the anatomical variance between the participants. A group of twenty participants (12 men, 8 women) performed a set of ten activities (defined in Table 6.1), resulting in the capture of 200 sequences with 60,225 frames of skeletal data.

6.2.2 Feature Encoding

There are inherent challenges associated with marker-less technology, a consideration is the environmental factors, such as occlusion or loss of the skeleton while tracking the participant. To overcome these challenges, the data used in this chapter was manually visualised to ensure accurate recording of the MoCap for each actions. No issues were detected with the data. A row vector of 20 body-joint coordinates represents each frame. Each action captured is aligned to the "hip-centre" joint to create a coordinate system relative to the "hip-centre" of the frame. Where the original coordinate $\mathbf{P}_j^t$ of the j-th joint at time t is subtracted by "hip-centre" $\mathbf{P}_t^{hipcentre}$ of the each corresponding frame, defined as:

TABLE 6.1: Detailed capture protocol and test descriptions for Kinect 360 feasibility analysis.

Test	Capture Protocol	Instructions or Constraints
Jump (maximum power)	The participant stood with their legs fully extended and slightly less than shoulder width apart. When instructed, they produced a counter movement jump by bending at the knees and then performing a maximal-level jump	Perform a maximal-effort jump
Arm Movement	The participant stood with their legs fully extended and slightly less than shoulder width apart. When instructed, arms extended along the frontal plane moving to a side-by-side position	Test terminated after 10 seconds
Pickup Object	The participant stood with their legs fully extended and slightly less than shoulder width apart. When instructed, from a standing position bending down to pick up an object off the floor with the right hand	Test terminated after 10 seconds
Squats	The participant stood with their feet as close together as possible side-by-side. When instructed, the participant bent down so that gluteals approximately 10cm off the ground	Test terminated after 10 seconds
Walk towards (towards Kinect)	The participant started from a standing position and walked forwards in a straight line towards the sensor at their usual walking speed	Walk at 'usual' walking speed
Jogging towards (towards Kinect)	The participant started from a standing position and jogged forwards in a straight line towards the sensor at their usual jogging speed	Jogging at 'usual' jogging speed
Bending to Toes	The participant started from a standing position and bent forward with there arms extended to touch their toes	Test terminated after one attempt
Chair Rise	The participant started from a seated position. When instructed, they had to stand up so that the legs were fully extended, and then sit down again. This was repeated five times with the aim to complete five complete stand/seat cycles. The arms were held across the chest so that all of the power needed to stand and sit was produced by the legs muscles	Perform five chair rises as quickly as possible. Test terminated after 60 seconds.
Upper Body Twist	The participant started from a standing position, when instructed they raised both arms vertically in front of the torso and twist from left to right	Test terminated after 10 seconds
Arm Stretch	The participant started from a standing position, when instructed they raised both arms vertically as high as possible	Test terminated after 10 seconds

$$\hat{p}_t^j = p_t^j - p_t^{hipcentre} \tag{6.1}$$

where $\hat{\mathbf{P}}$ is a set of aligned poses, which represents a local coordinate system. Axis-angle location, velocity (over time) and energy are representative kinematic features. An outcome from chapter 5 found that providing temporal information may aid in the recognition process. Therefore, in this approach they are used to represent the dynamic variation of each action sequence, observed over time.

The vertical location (y axis) is discriminative compared to horizontal left to right (x) and forward to backwards directions (z) because intrinsically most motion exhibits some form of vertical motion. Chapter 4 and chapter 5 highlighted the importance of the exemplar paradigm and encoding features that contain low dimensionality. Therefore, y-position (y), y-velocity (v_y) and energy (e) are computed to form the feature vector $\mathbf{F}_n^t$ of the n-th joint at time t. The y-position has been extracted to describe the action in the y-plane. Velocity and energy are computed for each joint position are given as:

$$\begin{aligned}\mathbf{F}_t^j = \{(y_{j,t}, v_{y(j,t)}, e_{j,t}) | v_{y(j,t)} = y_{j,t} - y_{j,t-\sigma}, \\ e_{j,t} = (v_{x(j,t)}^2 + v_{y(j,t)}^2 + v_{z(j,t)}^2)\}\end{aligned} \tag{6.2}$$

where $y_{n,t}$ is the aligned y-axis in $\hat{\mathcal{P}}$, as shown in Eq.6.1, energy is computed as the sum of energy in x, y, z of each joint. The velocity v_y and energy (e) are calculated over the previous σ frames. σ is a user defined parameter to allow for increased tolerance for the joint tracking error presented by the Kinect 360 and any subtle variance that may be present [27].

6.2.3 Kinematic Reduction and Pre-Processing

Principle Component Analysis (PCA) is applied to the data for each class to reduce dimensionality of the aligned activity feature vector $\mathbf{F}$ further by projecting the data into a lower-dimensional space. Refer to Section 3.1.8.1 for a detailed description of PCA functionality. Transposing to low-dimensional feature space provides a number of benefits, such as reduced computational complexity, stabilisation of data noise and

improved accuracy. The variance was set as 98%, meaning the retention of the columns (dimensions) the variance for the projected datasets. This resulted in a projection that contained 13 (± 4) eigenvectors, to represent the entire class. The projected data is defined as:

$$\mathbf{S} = \{(\mathbf{F}_{n,t}, l_a) | n = 1, \ldots, N,\ t = 1, \ldots, T, l_a = 1, \ldots, L\} \tag{6.3}$$

where $\mathbf{F}_{n,t}$ is the projected activity feature vector defined in Eq. 6.2 for the j-th joint at time t and l_a is the numerical class label for each activity (e.g. Jumping class: 1, Walking class: 2).

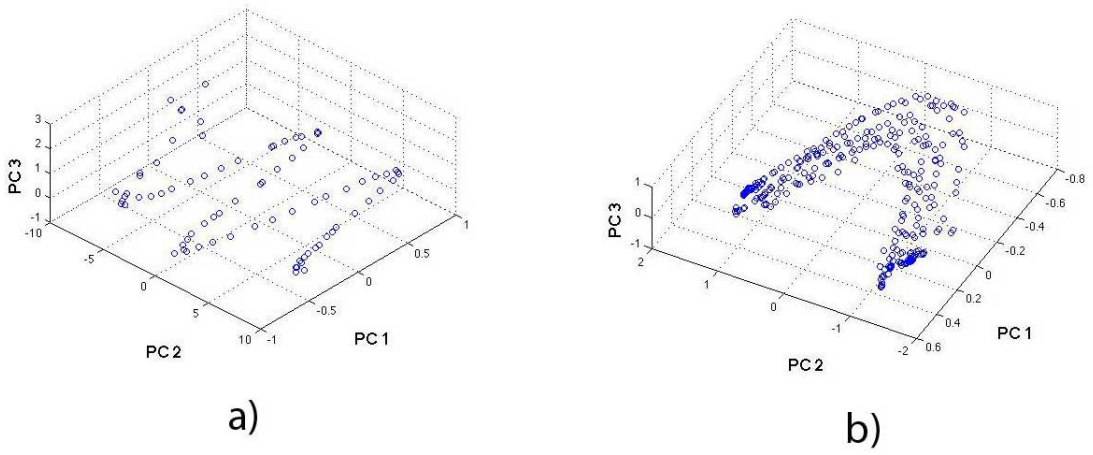

FIGURE 6.2: Visualisation of three Principal Components representing the features derived in Eq. 6.2. a) Walking b) Chair Rise. Where PC represents the Principal Component dimensions.

6.3 Experiments

This section contains action recognition results for detecting health-related motions using the proposed feature representation technique defined in Section 6.2.1. A brief introduction to the experimental process is presented hereafter, followed by presentation of the training protocol and experiment results in the subsection. Marker-less MoCap data collected from the Kinect 360 for use in this chapter represents a large number of health-related actions that you would find in the health domain. A set of action

sequences for each motion class are processed to generate the feature representation presented in Section 6.2.2. By encoding the velocity and energy information the approach removes subtle variations that machine learning algorithms may attempt to model.

As with other experiments presented in this thesis, to simulate a real world application, all experiments were conducted on *unknown participant actions*, meaning that no data from the participant being tested was included in the training set.

6.3.1 Protocol and Machine Learning

SVM, RF, ANN and GRBM are popular machine learning techniques used in a number of domains, however in the field of 3D action recognition, GRBM have had limited use within the action recognition, however it has shown promise in detecting human actions [6, 7, 159, 160]. In other domains several studies have sought to investigate the performance differences in classification confidence of the machine leaning techniques. Nitze et al. [161] sought to provide a comparison for crop type classification (for agriculture) by use of image representations of different crop fields. Statnikov et al. [162] sought to compare microarray-based cancer diagnosis and prediction based on gene profiling. Finally, Tang et al. [163] assessed for spam detection based on IP addresses. SVM was deemed by two of the studies [161, 162] to be the most accurate with its predictions, with one study finding RF more applicable [163]. Brennan et al. [164] utilised ANN for assessment of motor control for upper arm function, the authors found ANN provides string reliable results. A theme apparent in many studies is that SVM was more accurate due to it being less sensitive to the choice of input parameters than RF and ANN.

In order to assess recognition accuracy, model training and recognition time, the data was randomly split into two subsets, ten participants formed the training set and the remaining formed the testing set. With each experiment introducing another participant from the training set until all the participants were used. By training the classifiers in this way, the study is able to determine the suitable number of participants for training to achieve a stable classification rate. In simple terms, experiment 1 included one training participant per class which was tested against ten testing participants as demonstrated in Table 6.2.

A detailed description is provided in Section 3.2 for the machine learning techniques described. To ensure optimal performance of each classifiers, parameter optimisation was

TABLE 6.2: Optimum machine learning parameters for each model trained based on the participant iteration.

x participants/	2	4	6	8	10
Parameter	Support Vector Machines				
C	28	28	32	28	30
γ	8	8	8	6	6
	Random Forests				
n_{tree}	400	700	800	800	800
	Artificial Neural Networks				
layer	4	4	4	4	4
	GRBM (Stacked SVM)				
C	12	14	14	14	14
γ	4	4	4	4	4

performed as demonstrated in Table 6.2. To enable comparison, the following decisions were made; an ANN had a set number of layers, set as $layer = 2$, this was to ensure cross-validation comparison. A GRBM model was trained and then learnt using an SVM, this stacked approach is common within the community.

For C and γ used in SVM and GRBM, the selection was undertaken according to the cross-validation method [143]. To perform cross-validation, the training set was segregated into two subsets of equal size. Then the classifier was trained on one subset (training data) and accuracy is tested with the introduction of the second subset. The optimisation process was repeated for each of the possible parameter in exponential steps for both C and γ between 10^{-4} to 10^{5} and 10^{-6} to 10^{3} respectively.

In RF, n_{tree} represents the number of trees to be generated for RF requires optimisation. To perform optimisation, the range of trees has been tested with incremental increases of 100 between 100 up to 1000 trees. The optimised number of trees required for each experiment are shown in Table 6.2. The results suggest a consistent number of 800 n_{tree} was sufficient for training.

6.3.2 Recognition: Confidence in Detection

The use of PCA needs to be validated, the first experiment conducted on the data illustrated the benefits of using the projected PCA space to train the machine learning algorithms (Figure 6.4). The comparison between the latent space (Figure 6.3) and the ambient "projected" space (Figure 6.4) support the use of PCA. Observe that the accuracy results are generally more confident when PCA is applied than not. It can also

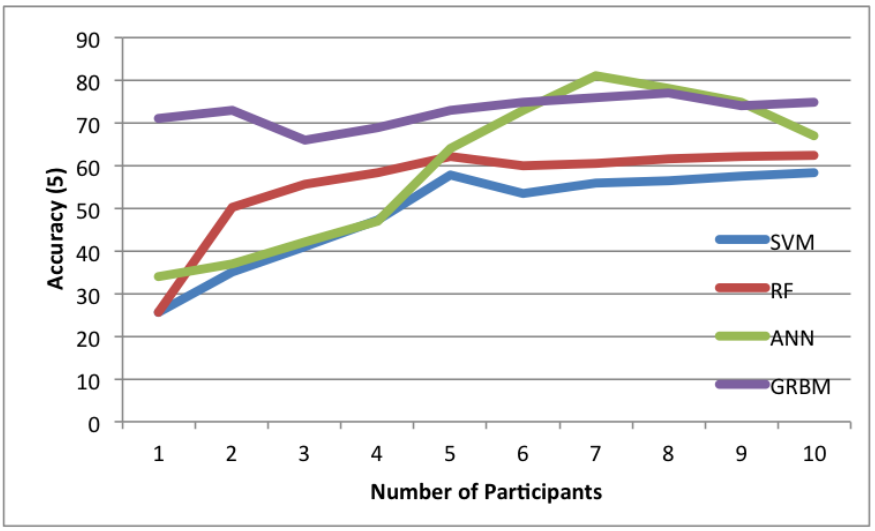

FIGURE 6.3: Overall action recognition rate for SVM, RF, ANN and GRBM trained from an iterative number of participants without using PCA as a dimension reduction technique.

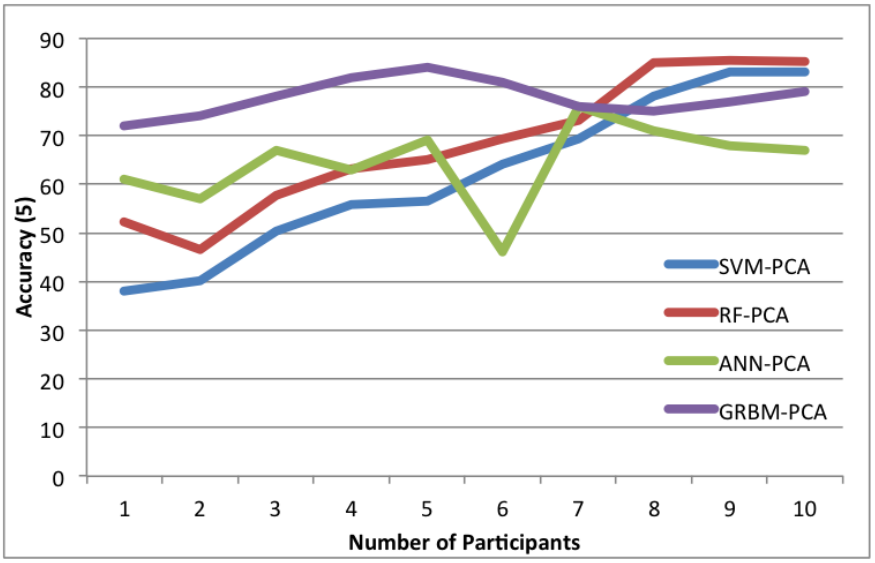

FIGURE 6.4: Overall action recognition rate for SVM, RF, ANN and GRBM trained from an iterative number of participants with using PCA as a dimension reduction technique.

be observed, that while the number of participants increases, the SD of accuracy error decreases. This is represented by the SD presented in Figure 6.5.

The standard procedure for calculating SD is the deviation of the average recognition accuracy for each participant's actions. Observing Figure 6.5, the deviation of error reduced, represented by the error bar, for each machine learning technique when participants were increased from 1 to 10. Initial findings demonstrate that recognition accuracy improved and error reduced when the number of participants in the training

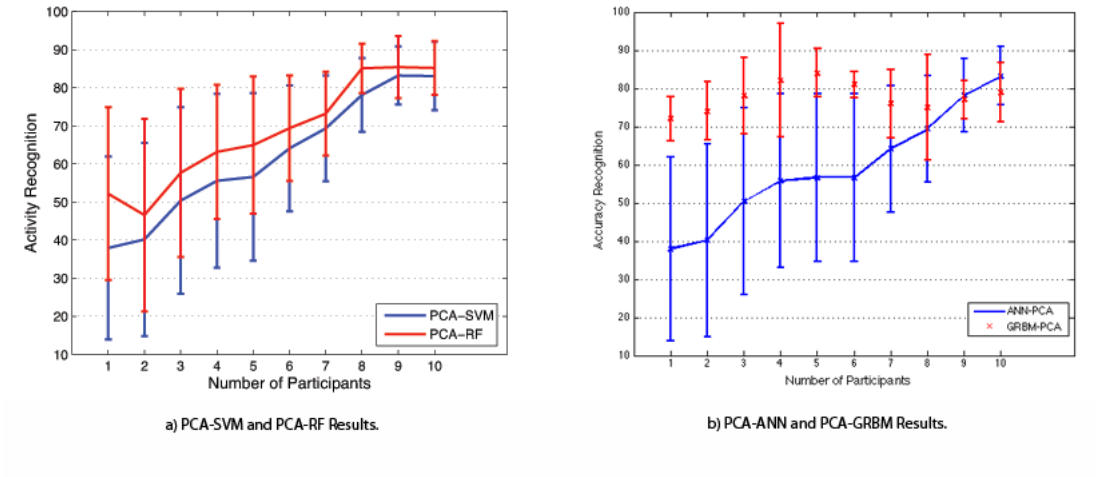

FIGURE 6.5: Action recognition and its standard deviation of accuracy results for each participant iteration and machine learning technique.

set was incrementally increased (Figure 6.4 & Figure 6.5). Table 6.3 summarises recognition results for each action class. As demonstrated, there is variation between each of the machine learning techniques with the number of training participants affecting the average recognition accuracy. RF exhibited the highest average overall accuracy with 85.17%, outperforming the other machine learning techniques. With both SVM and RF, increasing the number of training participants improved classification accuracy considerably, with SVM and RF having a similar linear increase in accuracy for between 6 to 10 participants. However, ANN and GRBM both had varied accuracy results with respect to the number of participants. This may indicate that potentially the importance of data quality for the training process.

TABLE 6.3: Recognition accuracy per action based on increasing the number of participants for each action class.

x participants	2				4				6				8				10			
Overall Accuracy [%]	SVM	RF	ANN	GRBM	SVM	RF	ANN	GRBM	SVM	RF	ANN	GRBM	SVM	RF	ANN	GRBM	SVM	RF	ANN	GRBM
Jumping	19.23	36.19	68.98	92.98	41.97	55.22	83.09	92.09	60.14	69.49	62.87	83.97	76.56	85.24	81.96	79.76	85.07	86.45	73.95	84.56
Arms Movement	41.03	48.64	67.37	87.87	66.42	68.64	64.98	93.07	62.67	73.93	53.67	84.58	80.5	85.76	78.75	74.96	92.25	85.23	75.96	86.59
Pickup Object	36.64	55.79	51.09	14.46	39.61	58.47	44.09	85.76	58.25	57.93	42.09	87.48	74.85	79.71	84.06	72.56	78.47	76.24	74.91	82.34
Squats	29.45	26.88	40.97	87.89	34.36	46.08	56.976	80.01	57.92	55.09	39.87	82.47	78.86	85.92	75.81	72.84	82.22	84.03	54.09	79.11
Walking	8	13.37	49.76	54.04	40.96	54.99	59.86	79.37	59.75	68.1	35.38	93.97	74.16	84.3	76.96	81.09	79.59	86.65	58.94	70.55
Jogging	42.2	61.16	56.87	78.534	73.08	78.81	68.03	82.48	70.48	79.12	46.1	89.16	78.41	87.42	56.91	78.38	80.95	85.84	61.47	73.45
Bending to Toes	49.35	42.48	73.98	75.98	64.84	63.61	66.87	78.34	67.37	71	47.34	84.06	77.59	86.02	63.74	76.6	81.13	85.86	61.29	77.47
Chair Rise	46.86	41.97	69.98	79.75	59.03	59.75	71.13	77.19	64.25	69.19	41.49	86.73	76.57	83.09	76.76	79.54	85.15	84.58	73.85	72.04
Upper Body Twist	74.33	78.18	64.78	83.97	82.02	79.77	64.07	79.41	80.13	81.62	51.37	74.87	81.58	84.09	76.95	81.09	87.71	87.55	63.65	83.49
Arm Stretch	54.64	61.03	34.08	87.65	55.18	66.17	51.07	75.87	80.06	85.04	45.07	58	90.03	91.02	48.09	71.26	89.16	89.32	75.18	81.95
Average [%]	40.17	46.57	57.78	74.31	55.75	63.15	63.01	82.35	64.09	69.38	46.52	82.52	78.1	85.09	71.99	76.80	83.05	85.17	67.32	79.15

The recognition results fluctuated due to the number of training participants used and the machine leaning technique employed. Observing the results in Table 6.3, using two, four and six training participants resulted in low accuracy across the activities, the use of eight and ten training participants saw a major improvement, and the levelling off in accuracy across the range of action sequences captured. Yet, for ANN and GRBM the results varied meaning that this same conclusion cannot be drawn.

Due to anatomical similarities between *Upper Body Twist* and *Arm Movement*, class confusion was observed, with both SVM and RF over classifying Upper Body Twist. Recognition accuracy differed for two, four and six training participants. For tasks such as *Walking* and *Arm Stretch* both ANN and GRBM demonstrated class confusion, with the inter-/intra-class variations of the sequences causing the machine learning frameworks to struggle to correctly identify the correct action class. An example of class difference between the machine learning techniques is observed in Figure 6.6, this figure demon rates the class probability for Arm Movement. Observe that RF and SVM both experience confusion at different points in the activity sequence for *Arm Movement*, this inconsistency is observed throughout the study. However, for eight and ten training participants the classifiers stabilised, with SVM producing the highest accuracy for Arm Movement with 92.25% for all activities throughout the study yet for ANN and GRBM the results fluctuate making it difficult to make an accurate determination of its suitability.

Further confusion between *Walking* and *Jogging* was encountered due to the similarity in limb rotation and joint motion. As observed in Table 6.3 and Figure 6.7, recognition accuracy for the aforementioned actions had a notable difference in recognition accuracy and confusion for all techniques employed. For each participant iteration increase, SVM struggled to correctly classify *Walking*, with an over confidence in *Jogging* observed. RF observed a similar over confidence in *Jogging*, however for both SVM and RF, when ten training participants were used results were improved and levelled off. Throughout all the trials ANN demonstrated large amounts of variability depending on the number of training participants, for example *Walking* had significant confusion throughout, with varied results. Finally, GRBM performed very well overall, with consistency in several action classes, however the technique struggled when actions required the analysis of the temporal domain, such as for *Walking* and *Jogging*.

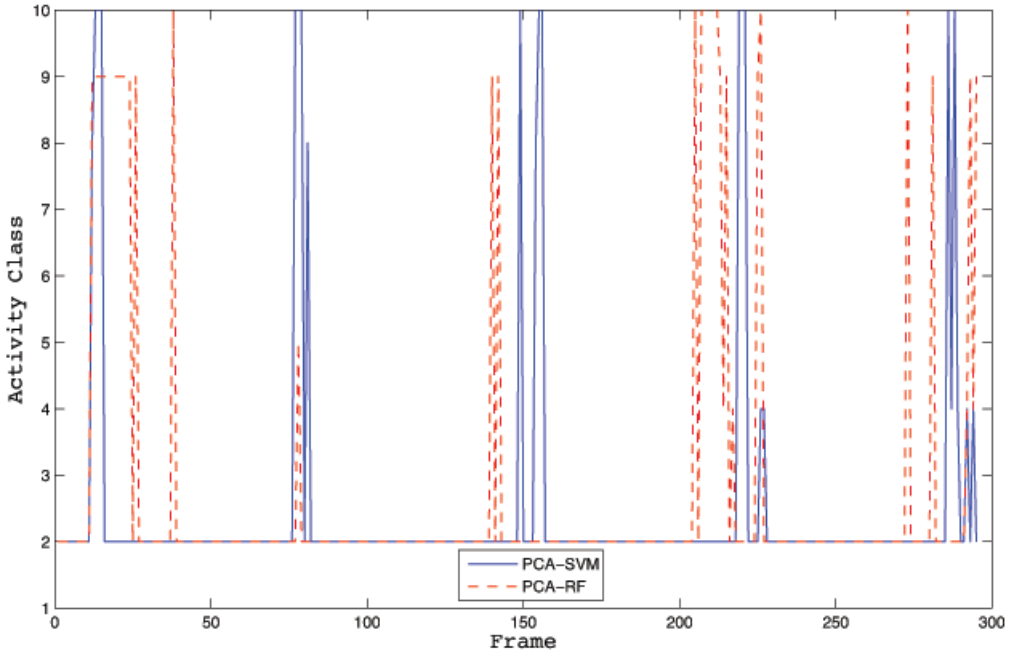

FIGURE 6.6: An example class estimate for SVM and RF by a participant performing *Arm Movement*. Expected class is 2.

Finally, the *Pickup Object* action suffered consistent misclassification throughout the experiments. Due to the similarity and overlapping of posture with a number of other action classes, it was observed to show misclassification for two, four and six training participants. This should be addressed in future work. To conclude, RF provided the highest average recognition accuracy for each participant iteration when compared to the other techniques. However, both ANN and GRBM were affected by the temporal element, such as it struggled to distinguish between action classes because the context was not present or understood. The accuracy results changed depending on the number of training participants and variance observed. Conversely, SVM provided improved recognition results for several action classes, namely, *Arm Movements*, *Pickup Object*, *Chair Rise* and *Upper Body Twist*, with *Arm Movements* seeing a 7.02% improved difference on RF.

6.3.3 Computational Model Training and Recognition Rate

SVM overall was the quickest to train, while GRBM took considerably longer, as demonstrated in Table 6.4. The training time for each machine learning technique was affected by the number of training participants, data complexity and parameters selected (Table 6.2). With the increase in the number of training participants, it was observed that

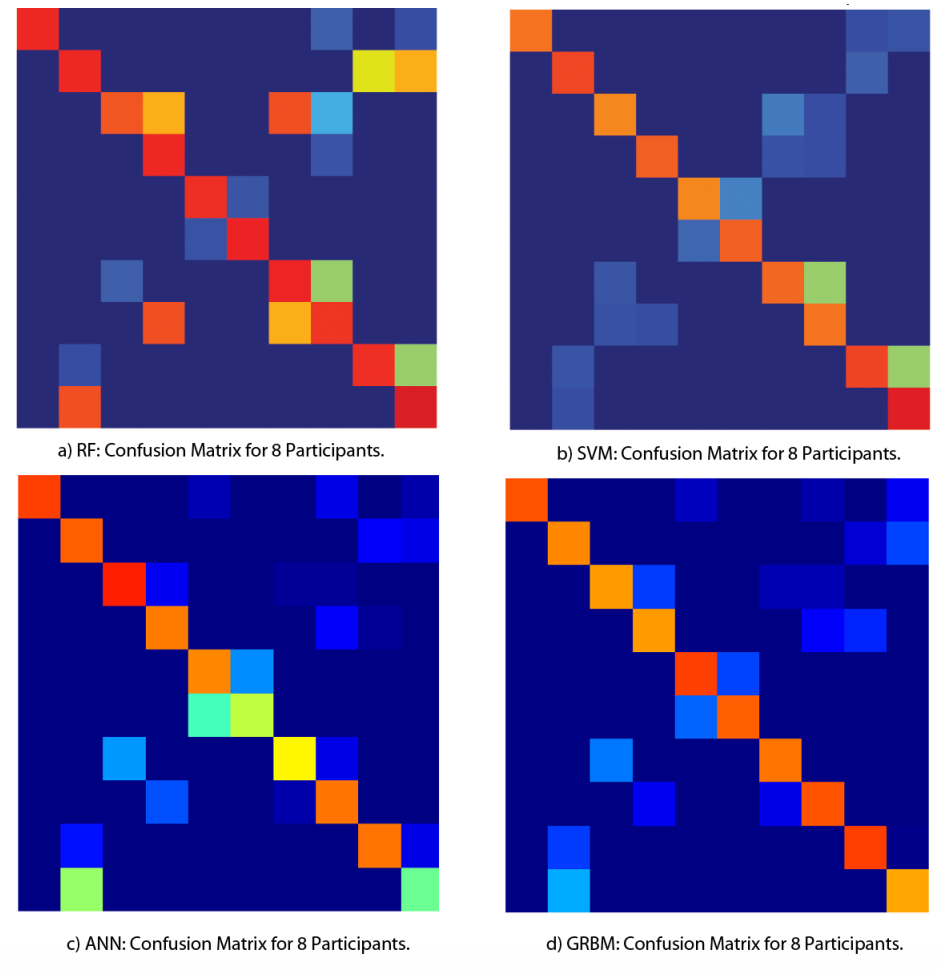

FIGURE 6.7: Confusion Matrix between action classes for SVM, RF, ANN and GRBM where 8 participants in the training sets were modelled.

TABLE 6.4: Computation time for each mode based on incremental increases of the participant number.

x participants/	2	4	6	8	10
Classifier	Training time [Sec]				
SVM	2.381	14.69	30.39	42.24	55.96
RF	5.08	19.35	49.06	79.21	123.17
ANN	10.98	18.42	35.28	50.95	59.73
$GRBM$	34.87	65.64	80.43	126.42	160.36

training of the techniques became more complex; consequently the training time of the classifiers tends to increase exponentially.

Average recognition time was significantly reduced to millisecond predictions, when compared with training time, with GRBM on average performing faster than the other techniques as demonstrated in Table 6.5. ANN was computationally more expensive than the others, however the duration for recognition is directly linked to the n_{tree}, C, γ.

TABLE 6.5: Computational time to perform recognition per action sequence.

x participants/	2	4	6	8	10
Classifier	Classification time [Sec]				
SVM	0.036	0.07	0.114	0.146	0.179
RF	0.007	0.022	0.031	0.034	0.037
ANN	0.09	0.14	0.27	0.48	0.56
$GRBM$	0.002	0.013	0.018	0.019	0.022

The average classification time of each action also increased with the introduction of new participants, albeit on a lower scale (100 up to 300 milliseconds) as demonstrated in Table 6.5.

6.4 Discussion and Conclusions

The innovation of the Kinect 360 enables real-time HCI through recognition of user's gestures and body movements to, for example, control a character or gameplay elements. However, despite advances in vision-based HCI, the main constraint is machine understanding of the gesture, action and behavioural context, which still remains an open and ambitious problem to solve. Motivated by current limitations, this chapter has focused on the promising application of the Kinect 360 for analysis of human motion, specifically health-related motions for full-body tracking and gestured-based evaluation by an intelligent HCI system [26, 165]. Chapter 8 extends the framework of recognition to include the analysis of human motion in the attempt to detect subtle variations between participant groups using informative feature representations.

Normalising with the "hip-centre" joint of the first frame, before computing $y_{n,t}, v_{y(n,t)}, e_{n,t}$ and applying PCA reduces anatomical differences and aids in improved classification accuracy (see Figure 6.4 and Table 6.5). In addition, the computed $y_{n,t}, v_{y(n,t)}, e_{n,t}$ is high-dimensional, containing twenty body joints and while the techniques are capable of handling high-dimensional data, feature space reduction by PCA has aided in providing higher accuracy results, improved model training and classification times. PCA has demonstrated its potential for use in noise reduction and reducing computational rate. However, complexity of performing PCA may cause problems for future computational tasks - most notably when attempting to decouple spatial temporal relationships.

It must be noted that parameter selection, both in terms of C, γ for SVM and n_{tree} for RF has had an impact on the training and recognition time due to the introduction of additional complexity (Table 6.2 and Table 6.4). In addition, the number of layers selected for ANN could also have impacted the result. For SVM, C affected the number of support vectors, leading to an increase in classification time, whereas RF, n_{tree} increased model training times when a high number of decision trees needed generating. The requirement to fine-tune these parameters may cause limitations in future work, chapter 8 proposes an approach that moves away from traditional machine algorithms.

The Kinect 360 has sensitivity of joint rotation, with the Kinect 360 designed to detect activities of a participant who is standing face forward to the Kinect 360, large rotation could prove difficult to detect. Recognition between the range of activities was reliable, even rotation and subtle posture changes between similar activities of *Walking* and *Jogging* could be recognised accurately. The *Chair Rise* action presented a further challenge, due to occlusion of the chair and natural limb movement, yet each classifier was capable of achieving acceptable classification results. Nevertheless, misclassification was an issue for a number of activities that have similar movements, with both classifiers finding it difficult to classify individual frames without any information about past frames. Finally, while the Kinect 360 is robust in tracking the human body, there are still issues with robust classification, with multiple poses presenting confusion.

This chapter has explored the use of the Kinect 360 in detecting, tracking and recognising human motion. It has found that while the Kinect 360 is capable, there are skeletal tracking issues that need to be overcome. Nevertheless, the skeletal tracking issues are acceptable for real world deployment. A concern that extends from utilising the Kinect 360 MoCap data is its reliability and accuracy, Chapter 7 explores the possibility of using the Kinect 360 to detect age-related changes. This research domain has yet to be explored and remains unresolved; this thesis will present a solution to this challenge by uniting computer vision and health research.

Chapter 7

Detection of Age-related Changes between Young and Old

This chapter assessed the ability of the Microsoft Kinect One to detect age-related changes between the young, athletic old and old adults using a novel digital analysis framework. This chapter presents typical routines of clinical movements based on standardised tests such as the Short Physical Performance Battery, Timed-Up-And-Go, Three-Meter Walk and Balance (e.g. Tinetti [1986]).This chapter has found that the Microsoft Kinect One and the framework introduced in this chapter is capable of detecting subtle age-related differences between participant groups. Given the benefits of this chapter, the Microsoft Kinect One could therefore become a useful tool for assessing age-related changes in a clinical setting.

7.1 Introduction

Automatic methods for detection, recognition and quantification of human movements have become more accessible due to increased availability of low-cost multi-modality marker-less capturing devices. This provides potential to develop applications suitable for use in healthcare settings to detect problems that participant have in coordination of movements [14, 167–170]. Chapter 6 introduced the ability to using Kinect 360 for recognising human action. With rapid technological developments, this chapter utilises the latest version of the Microsoft sensor, the Microsoft Kinect One (Kinect One).

Throughout comparison, the sensor has been found to be far superior to the Kinect 360 sensor, providing a more robust pose estimation and skeleton tracking [170, 171]. Briefly, the Kinect One provides a higher quality depth map image, which results in a higher number of tracked joints (Kinect 260 tracks 20 joints whereas the Kinect One tracks 25). In addition, for other data modalities the Kinect One offers further advances such as audio capability and High Definition RGB images.

Movement problems experienced by a diverse group of participant include slow and altered gait, difficulties changing from standing-to-sitting or sitting-to-standing, and balancing. These problems increase the risk of disability and falls that have major consequences for quality of life and healthcare provision. Specialist nursing staff, physiotherapists and geriatricians routinely assess movements using standardised tests such as the Short Physical Performance Battery (SPPB) [172], Timed-Up-And-Go (TUG) [173], Three-Meter Walk [174] and Balance (*e.g.* Tinetti [166]). However, manual-assessments require trained staff and variations between assessor ratings and experience may cause problems. Computer-based analysis of these movements can standardise the assessments and may be more resource-effective. Automated assessment requires algorithms to detect joint angles in different body segments, stride length and foot positioning, whilst also accounting for the diversity that exists across populations in terms of body size and shape.

Depth-sensors have been used several times for the assessment of balance (*e.g.* [111, 113, 116, 169, 170, 175]) extracting simple gait-based vectors from a skeletal stream to provide basic stability-based single value scores. Using the Microsoft Kinect for Xbox 360, Zhou et al. [176] showed that detection of the Centre-of-Mass (CoM) was correlated with standard assessments performed on a force platform. This chapter demonstrates the use of the Kinect One in a analysis framework for health-related movements such as balance, walking, sitting and standing from a diverse population of young and older adults. The movements were based on common clinical assessments used to assess movements in disease and frailty. This chapter confirms the suitability of the Kinect One in quantifying movement between young and older populations.

The main contributions of this chapter are as follows:

1. Digitalisation of the Short Physical Performance Battery to enable quantitative

automatic analysis using maker-less technology free from human subjectiveness (Section 7.3).

2. Validation of the Microsoft Kinect One ability to detect jump height and Centre-of-Mass using markrt-based correlation (Section. 7.5.2).

3. The ability of Kinect One to detect age-related changes between different participant groups (Section 7.5).

7.2 K3Da: A clinically relevant dataset

Alankus et al. [177] and Wang et al. [14] devised techniques to characterise movements in stroke and musculoskeletal patients, respectively. However, both utilised publicly available datasets that were intended for use in gaming populations, restricting their broader application. Indeed, existing datasets (*e.g.* [74–76]) were captured for specific purposes, such as daily living, first person or gestures, principally for use in the entertainment and gaming industries. Currently, none of the available datasets and works related to health application specifically includes movements based on common clinical assessments of participant groups and this limits the development of tools for use in healthcare settings.

Developing systems to automatically detect and classify movements is of great importance in healthcare, and related applications. For instance, early identification of people most at risk of deterioration of physical function gives more time for remedial interventions, such as lifestyle or physical rehabilitation, before the impairments are irreversible. Such systems also give the opportunity for long-term monitoring of participant to observe effects of illness or ageing or monitor effectiveness of rehabilitation programmes. While datasets exist to benchmark daily living and gaming actions/activities systems, according to the literature there does not exist a dataset to provide clinically supported motion sequences from both the young and elderly using depth sensor technology. In this section, a new clinically-supported and relevant dataset is established. While a dataset was captured in chapter 6 is lacks the clinical validity required for many applications discussed in this thesis.

7.2.1 Participants and ethical approval

Data collection was approved by the Research Ethics Committee at Manchester Metropolitan University (approval SE121308). All participants gave signed informed consent to take part in data collection and for their depth and skeleton data to be published. The sample size was limited by the narrow pool of suitable candidates and volunteers. The acquisition sessions consisted of 13 tests based on the SPPB [172], TUG [173] and additional tests of balance and power output. A detailed description can be found in Table 7.1. Fifty-six participants (characteristics shown in Table 7.3) were recruited with a mean age of 48.2 (SD of 21.5) and minimum/maximum of 18/89 year and a diverse range of body compositions. The participants were divided into three groups: young; those aged between 18 and 59; old: those aged 60 or more with no physical training; athletic old: those who are British Masters' athletes undertaking at least one session of training per week (an average of 39.1 years competing in sports).

7.2.2 Data collection and storage

In all data capture, the Kinect One depth sensor was fixed horizontally to a tripod at a height of 0.7 m and all assessments were confined to within range of the sensor. Room furniture was removed to ensure maximum visibility and room lighting was standardized. The participants were provided with a maximum of three attempts to complete each short task. A countdown timer was created to prompt the participant to start each test and sessions were recorded and stored automatically.

The Kinect One sensor coupled with the Microsoft Windows Software Development Kit [158] synchronised capture of depth and skeleton streams at 30 fps. Each data stream was retrieved and stored in a unique file for each time period with a unique millisecond timestamp. The raw storage format was selected for the depth stream; the raw information contains the depth of each pixel in millimetres. The 16-bits of depth data contain 13 bits for depth and 3 to identify the person-index. A text format was selected for storing the skeleton information with participants position, pose and relative depth map coordinates. The pose includes 25 joints and two action states as defined by Microsoft. This is an improvement on the Kinect 360 that only recorded 20 joint positions. The participants overall and joint positions are given as x, y and z coordinates

TABLE 7.1: Detailed Capture Protocol and Test Descriptions for the K3Da Dataset

Test	Capture Protocol	Instructions or Constraints
Balance (open eyes)	The participant stood with their feet as close together as possible side-by-side. They balanced with their eyes open and arms extended parallel to the floor	Test terminated after 10 seconds
Balance (closed eyes)	The participant stood with their feet as close together as possible side-by-side. They balanced with their eyes closed and arms extended parallel to the floor	Test terminated after 10 seconds
Chair Rise	The participant started from a seated position. When instructed, they stand up so that the legs were fully extended, and then sit down again. This was repeated five times. The arms were held across the chest so that all of the power needed to stand and sit was produced by the legs muscles	Perform five chair rises as quickly as possible.
Jump (low power)	The participant stood with their legs fully extended and slightly less than shoulder width apart. When instructed, they produced a counter movement jump by bending at the knees and then performing a low-level jump	Perform and low-level jump
Jump (maximum power)	The participant stood with their legs fully extended and slightly less than shoulder width apart. When instructed, they produced a counter movement jump by bending at the knees and then performing a maximal-level jump	Perform a maximal-effort jump
One Leg Balance (closed eyes)	When instructed, the participant balanced with one leg (participant preference) 6 inches off the ground with their eyes closed and arms extended horizontally	Test terminated after 10 seconds or when the second leg touched the ground
One Leg Balance (open eyes)	When instructed, the participant balanced with one leg (participant preference) 6 inches off the ground with their eyes open and arms extended horizontally	Test terminated after 10 seconds or when the second leg touched the ground
Semi Tandem Balance	The participant was asked to place one foot behind the other so that the big toe of the back foot was touching the side of the heel of the front foot. Their arms were fully extended horizontally	Test terminated after 10 seconds
Tandem Balance	The participant placed one foot directly behind the other so that the big toe of the back foot was touching the back heel of the front foot. The arms were fully extended horizontally	Test terminated after 10 seconds
Walk towards (towards Kinect)	The participant started from a standing position and walked forwards in a straight line towards the sensor at their usual walking speed	Walk at 'usual' walking speed
Walk away (from Kinect)	The participant started from a standing position very close to the sensor and walked away from the sensor in a straight line at their usual walking speed	Walk at 'usual' walking speed
Timed Stand Up and Go	The participant started in a seated position. They had to rise from the chair, walk 3 meters, turn around and walk back to sit on the chair again	Walk at 'usual' walking speed
Hopping	The participant was asked to hop with one leg (participant preference) on the spot multiple times	Test terminated after 10 seconds

in meters, otherwise referred to as MoCap. These positions are also mapped into depth coordinates. The skeleton data includes a joint tracking state, shown as "tracked", "not tracked" and "inferred".

Depth map and skeletal streams were extracted from the Kinect One data stream while the participant performed the movements. The Kinect One sensor provided a 512×424 depth image up to 30 frames-per-second (fps). Skeletal time series consisted of 25 3D orthogonal (x, y, z) locations. An example representation is shown in Figure 3.2. Frame data were extracted in real time using the technique of Shotton et al. [3], which is part of the Microsoft SDK [158]. This resulted in a dataset that comprises of 525 tests from 56 participants. Resulting in over 200,000 frames of depth and skeleton data. An example of the extracted depth map image and MoCap can be observed in Figure 7.1.

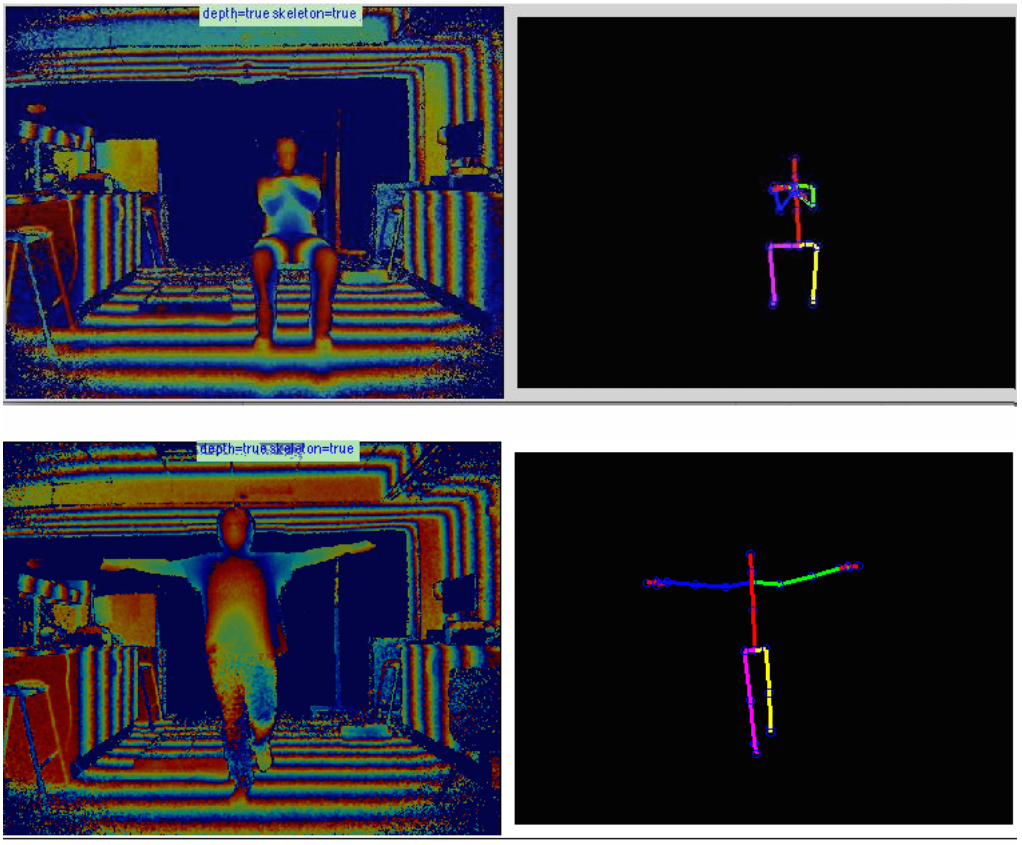

FIGURE 7.1: Skeleton visualisation: Left-to-right: raw depth image (512 x 424) and a MoCap skeleton representation (25 tracked joint locations).

The movements captured were designed by healthcare professionals and the data collection sessions were conducted according to standardised protocols (described in Table 7.1). The movements are commonplace and necessary parts of typical daily living, such

as walking, sitting, standing and balancing. These same movements become problematic in disabled and in older people, leading to frailty that affects around 9% of the population [10]. Thus, the movements derive from common clinical assessments. Due to the large inter-individual variability in age and physical capabilities, the dataset contains large motion variation in the skeletal pose. For example, some participants could easily perform five chair rises very quickly without losing balance or performance, while others (mainly older people) experienced a deterioration of their performance throughout the test. It is also possible to see intra-individual variations where the same subjects performed the same test more than once, resulting in slight differences in movements and timings. These features make this dataset a unique addition to publicly available marker-less datasets and a considerable improvement on that used in chapter 6.

7.3 Kinect Data Extraction, Interpretation and Feature Representation

The raw axes coordinate data (x, y, z orthogonal coordinates) are sufficient to accurately describe an observed motion. However, as has been presented throughout this thesis, it is possible to extract and represent descriptive kinematic features to enable a more informative representation feature to be formed, computation of these descriptive features are presented hereafter for clarity. The features extracted are based on the tests presented in Table 7.1, where necessary specific measurements based on the test are discussed.

Total time is defined as the absolute time taken to perform a test in seconds. This is defined by the time taken to pass the test or failing to complete the test. Each test has been manually annotated to define an "end of tests" point and this is used a reference for computing the total time.

CoM extracted from Kinect One data [116, 168] is vital for identifying age-related changes. In order to evaluate and measure stability, it is necessary to measure the movement of the body's centre-of-motion. The spatial parameter, CoM is derived from multiple joints of the Kinect One skeletal stream at each time period t to represent the motion characteristics. Let *com* be the centre-of-motion at time t computed from three joints (*hip left*, *hip right*, *spine*) is given by:

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{3} v'_{t,x_i}}{i} \\ \bar{y} &= \frac{\sum_{i=1}^{3} v'_{t,y_i}}{i} \\ \bar{z} &= \frac{\sum_{i=1}^{3} v'_{t,z_i}}{i} \\ com &= [\bar{x}, \bar{y}, \bar{z}]^t \end{aligned} \tag{7.1}$$

where $\bar{x}$, $\bar{y}$, $\bar{z}$ is the mean, i is the joint index of frame t and *com* is the concatenation of the mean values.

Medial-Lateral (ML) and Anterior-Posterior (AP) movement is the directional movement along specific axis of motion. Utilising information obtained for the CoM defined above, movement in ML and AP are characterised as the x and z coordinates axis respectively. The change in position between consecutive frames is computed and is considered as the ML and AP directional change over time.

Tests such as *Chair Rise*, *Jump* and *Walking towards Kinect* were represented with additional features to enable a more descriptive feature representation. These descriptive features are discussed hereafter for clarity.

The estimation of the number of chair rises per test was undertaken automatically, enabling greater automation of motion analysis framework. This is undertaken per test using spectral analysis as follows; for each sequence a number of local peaks in the data are extracted based on a maximum peak threshold. A local peak is defined as a data point that are separated by a minimum distance of 20 frames or greater than the standard sequence mean (computed as: sequence mean + 90%). An inversion of this process is undertaken to define the starting and end point of each rise. Therefore, the time taken for each rise is computed.

Walking towards Kinect is represented by an Upper-Body CoM representation. Unlike other tests in which the CoM is defined as the centre of the hip joints, the shoulders and the spine (middle) for this test it is defined as the shoulders and the spine (middle) - the same equation in Eq. 7.1 can be used. This selection allows for a greater understanding and representation gait characteristics associated with walking. Distance travelled is computed based on the directional z motion towards the Kinect One that is represented in meters. A participant is deemed to have "passed" the test if they have been able to

walk the required 3 meters. Speed is computed based on distance travelled and total time taken to walk the required distance. Upper body sway, either Medial or Lateral is computed as follows; the Upper-Body CoM is defined and an average Upper-Body CoM is computed from the entire sequence. Sway in either direction is computed on a per frame basis as the difference between the average and the current frame.

Estimating the jump height is defined as the maximum change observed from the y axis of the CoM with regards to the first frame of the test sequence. Actual height is obtained by the Leonardo Mechanography force platform (Novotec Medical Group GmbH, Pforzheim, Germany).

7.4 Statistical Analysis

Participant group data (young, old and athletic old) was compared using a paired Students t-test. For a comparison of distance a two-sample equal variance sample t-test was used. Whereas a comparison between total time for each participant group a two-sample unequal variance sample t-test was performed. For all statistical comparisons the significance level was 5% ($p < 0.05$). To determine the difference between participant groups, a one-way ANOVA is used to determine whether there are any significant differences between the means of the three groups.

7.5 Detection of Age-related Change

In this section, the ability of the Kinect One in measuring balance is compared to a force platform. Then, the performance of the Kinect One in detecting age-related differences in balance and jump height is presented.

7.5.1 Kinect Sensor Validation

The ability of the Kinect One to identify movements of the CoM during balancing was validated against measurements taken from a force platform. The results are provided in Table 7.2. No significant differences were found between the various measurements of

TABLE 7.2: Measurement results for validating the Kinect One for Centre-of-Mass evaluation compared to the Force Platform measurements.

Test Sequence (Measurement)	Kinect Sensor (SD)	Force Platform (SD)	p-value
Two-leg (Open Eyes) ($n = 10$)			
Total Time (s)	10 (0)	9.91 (0.51)	0.637
ML-CoM SD (cm)	0.29 (0.10)	0.52 (0.33)	0.066
AP-CoM SD (cm)	0.40 (0.22)	0.43 (0.15)	0.744
CoM SD (cm)	0.69 (0.22)	0.95 (0.41)	0.131
One-leg (Open Eyes) ($n = 10$)			
Total Time (s)	9.86 (0.40)	9.86 (0.39)	0.978
ML-CoM SD (cm)	0.44 (0.24)	0.61 (0.15)	0.016
AP-CoM SD (cm)	0.49 (0.19)	0.67 (0.18)	0.006
CoM SD (cm)	0.93 (0.40)	1.29 (0.31)	0.005
Artificial Sway ($n = 10$)			
Total Time (s)	9.64 (0.99)	9.85 (0.40)	0.599
ML-CoM SD (cm)	1.37 (0.42)	1.31 (0.15)	0.702
AP-CoM SD (cm)	2.54 (1.17)	2.20 (0.76)	0.508
CoM SD (cm)	3.92 (1.43)	3.52 (0.82)	0.507

balance during normal *two-leg standing* and during a period of artificial sway in which the participants deliberately performed large deviations of the CoM.

During the *one-leg standing with eyes open*, the Kinect One reported movements to be around 28% smaller than was found using the force platform. The differences in stability between the *two-leg standing* and the *one leg standing* were further compared using the outcome measurement of CoM SD (which represents the sum of AP and ML movements to reflect overall stability). The Kinect One showed 35% less stability during *one-leg standing* compared with *two-leg standing*. This difference was very similar to the 36% less stability during one-leg standing compared with two-leg standing when measured on the force platform.

It can be observed that in Figure 7.2 that a very strong correlation between jump height measured by the force platform and by the Kinect One (r^2=0.932; P<0.0005). Figure 7.2 shows a Bland-Altman plot that demonstrates very good overall agreement between the two different measurements of jump height.

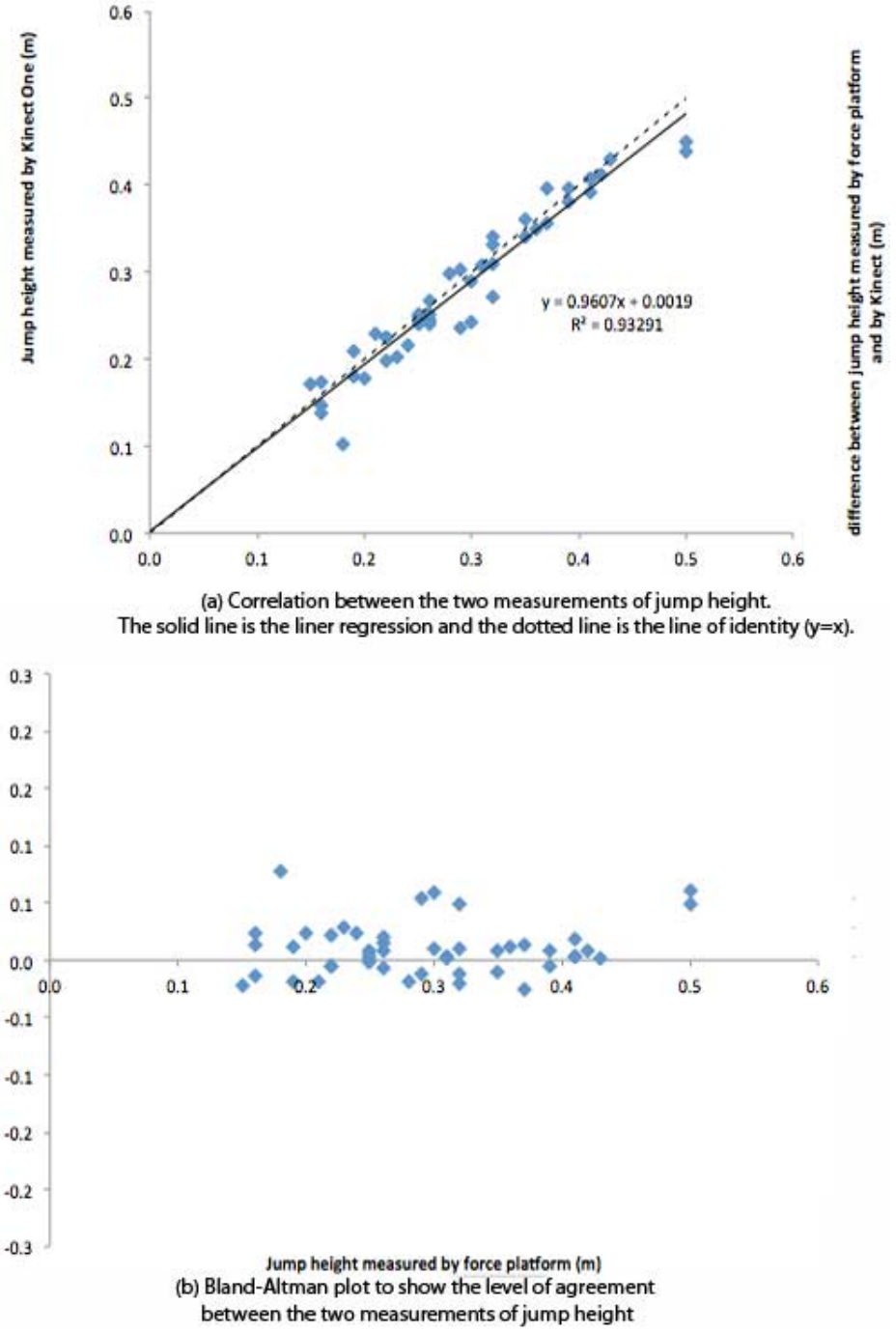

FIGURE 7.2: Vertical jump height measured by the force platform and by the Kinect One.

7.5.2 Use of the Kinect One to Detect Age-related Differences in Balance and Jump Height

Table 7.3 shows the results from the balance and assessments of physical function in young, highly active and frail older men and women.

Everyone was able to complete the full 10 seconds for balance with *two legs eyes open*, groups did not differ significantly for the amount of AP ($p=0.667$) but there was a significant effect of ML movement ($p<0.005$). There was no significant difference between the young and athletic old ($p=0.299$), but old had significantly more AP movements than both young ($p=0.001$) and athletic ($p<0.005$).

When standing in a *semi-tandem* position there was no difference between groups for total time ($p=0.246$), ML-CoM ($p<0.0005$), AP-CoM ($p<0.0005$) and CoM SD ($p=0.001$). When standing in a tandem position there was no difference between groups for total time ($p=0.385$), ML-CoM ($p=0.573$), AP-CoM ($p<0.560$) and CoM SD ($p=0.579$).

There is no difference between the groups total time ($p=0.165$), ML-CoM ($p=0.425$), AP-CoM ($p=0.271$) and CoM ($p=0.381$) when standing on one leg with eyes open.

For *one leg standing* with eyes closed groups differed significantly for total time ($p<0.0005$), ML sway ($p=0.010$), AP sway ($p=0.036$) and overall sway ($p=0.011$). Athletic old ($p=0.048$) and old ($p<0.0005$) had significantly less balance time. Old had less standing time than athletic old ($p=0.009$). Athletic old ($p=0.006$) and old ($p=0.009$) had more ML sway than young; there was no difference between athletic old and old ($p=0.929$). Athletic old ($p=0.045$) and old ($p=0.0.012$) had more AP sway than young; there was no difference between athletic old and old ($p=0.462$). Consequently, overall movement differed between both groups of old compared with the young (all: $p<0.01$) with no difference between the groups of old ($p=0.738$).

For the *Chair Rise* there were no differences between groups for ML movements of the upper body ($p=0.103$). A significant difference existed between groups of the AP and overall movement (both: $p<0.0005$). Compared with the young, both athletic old and old had significantly less AP movements and overall movements (all: $p<0.0005$). Athletic old and old did not differ significantly. There was no difference between the groups in the total time taken to perform five chair rises ($p=0.361$), but the standard deviation of individual standing attempts was significantly different ($p=0.002$). The young and athletic old did not differ significantly ($p=0.58$) but old had higher standard deviation than young ($p=0.001$) and higher than athletic old ($p=0.004$), indicating more variability in the time taken between standing attempts.

The Kinect One recorded no significant difference in distance travelled for all groups (p=0.962). *Walking* speed differed significantly between groups (p=0.022), as did ML sway (p=0.002). Walking speed was similar between young and athletic old (p=0.265), but old walked significantly slower than young participants (p=0.006). There was no difference between athletic old and old participants (p=0.145). For ML sway when walking both athletic old (p=0.012) and old participants (p=0.001) had more sway than young. There was no difference between the groups of old (p=0.508).

7.6 Discussion and Conclusions

This chapter assessed the ability of the Kinect One for detecting age-related differences between the young, athletic old and old. There was a clear lack of health-relevant datasets for benchmarking clinical applications. To address this, this chapter introduce the Kinect 3D active (K3Da) dataset (Section 7.2). The tests include *balance*, *walking* and *chair rise*, which indicate musculoskeletal function and are relevant to the development of frailty, mobility and stability [178]. The introduction of this dataset allows for benchmarking against clinically relevant scenarios. This removes the need to use datasets that were not set-up, or lack the necessary clinical protocol and/or accuracy to be used to assess algorithms associated with human motion analysis techniques in healthcare. The movements are common place and necessary parts of typical daily living, such as walking, sitting, standing and balancing. These same movements become problematic in those with disabilities and in older people, leading to frailty that affects around 9% of the population [10]. Due to the large inter-individual variability in age and physical capabilities, the dataset motions contain large motion variation in the skeletal pose. For example, some participants could easily perform five chair rises very quickly without losing balance or performance, while others (mainly older people) experienced a deterioration of their performance throughout the test.

The Kinect One has enabled new capacities for innovation within the healthcare sector. In this chapter, the Kinect One has been employed to provide detailed measurements of clinically relevant motions. These motions such as balance and walking are typically utilised in a clinical setting to measure stability, mobility and general well-being of a participant. Standard approaches, such as SPPB [172] provide a single score

TABLE 7.3: Computed results across all tests and participant groups using ANOVA test.

Test Sequence (Measurement)	Young ($n = 15$)	Athletic Old ($n = 15$)	Old ($n = 13$)	ANOVA p value
Participant Characteristics				
Age (years)	25.5 (6.4)	67 (5.2)	74.6 (3.9)	-
% male	68	47	50	-
Height	173.2 (8.5)	165.7 (10.1)	170.9 (6.1)	-
Body mass	77.1 (16.3)	61 (9.5)	26.4 (5.8)	-
Balance (Open Eyes)				
Total Time (s)	10 (0)	10 (0)	10 (0)	-
ML-CoM SD (cm)	0.27 (0.11)	0.22 (0.09)	0.44 (0.15)	0.001
AP-CoM SD (cm)	0.32 (0.20)	0.38 (0.17)	0.36 (0.21)	0.667
CoM SD (cm)	0.58 (0.23)	0.60 (0.24)	0.81 (0.31)	0.057
Semi-Tandem (Open Eyes)				
Total Time (s)	9.64 (1.09)	10 (0)	10 (0)	0.246
ML-CoM SD (cm)	0.29 (0.08)	0.29 (0.11)	0.49 (0.16)	0.001
AP-CoM SD (cm)	0.21 (0.07)	0.28 (0.14)	0.36 (0.14)	0.008
CoM SD (cm)	0.50 (0.12)	0.58 (0.18)	0.86 (.27)	0.001
Tandem (Open Eyes)				
Total Time (s)	9.66 (0.96)	10 (0)	9.74 (0.82)	0.408
ML-CoM SD (cm)	0.41 (0.20)	0.30 (0.12)	1.87 (3.86)	0.116
AP-CoM SD (cm)	0.27 (0.11)	0.30 (0.16)	1.33 (1.86)	0.016
CoM SD (cm)	0.68 (0.23)	0.61 (0.20)	3.20 (5.62)	0.060
One-leg Balance (Open Eyes)				
Total Time (s)	9.74 (0.72)	9.51 (1.96)	8.47 (2.42)	0.165
ML-CoM SD (cm)	0.28 (0.09)	3.51 (12.72)	3.85 (4.62)	0.425
AP-CoM SD (cm)	0.41 (0.21)	1.56 (3.51)	1.78 (2.12)	0.271
CoM SD (cm)	0.68 (0.25)	5.06 (16.22)	5.63 (6.52)	0.381
One-leg Balance (Closed Eyes)				
Total Time (s)	9.47 (1.24)	8.12 (2.96)	5.09 (1.70)	0.001
ML-CoM SD (cm)	1.50 (1.78)	11.93 (14.86)	12.66 (9.10)	0.010
AP-CoM SD (cm)	1.47 (1.16)	5.48 (7.68)	7.07 (5.57)	0.036
CoM SD (cm)	2.97 (2.49)	17.41 (22.08)	19.73 (12.45)	0.011
Chair Rise				
Estimated # Chair Rise	4.47 (0.51)	4.94 (0.68)	4.90 (0.31)	0.938
Actual # Chair Rise	5 (0)	4.88 (0.61)	5 (0)	0.608
ML-Upper CoM SD (cm)	1.35 (0.58)	1.15 (0.30)	1.67 (0.90)	0.102
AP-Upper CoM SD (cm)	17.07 (4.60)	8.97 (3.08)	10.83 (3.57)	0.001
CoM SD (cm)	18.42 (4.75)	10.12 (3.22)	12.50 (4.12)	0.001
Time Rise Average (s)	1.43 (0.27)	1.54 (0.23)	1.55 (0.27)	0.361
Time Rise SD (s)	0.53 (0.11)	0.58 (0.42)	0.79 (0.16)	0.002
Walking				
Total Time (s)	2.27 (0.52)	2.37 (0.34)	2.64 (0.51)	0.144
Distance Travelled (m)	3.07 (0.27)	3.06 (0.19)	3.04 (0.33)	0.962
Velocity (m/s)	1.41 (0.24)	1.31 (0.17)	1.17 (0.15)	0.021
CoM SD (cm)	2.69 (1.11)	4.83 (2.32)	5.39 (2.36)	0.002
Jump (Maximum Power)				
Est Jump Height (cm)	34.47 (6.93)	23.52 (7.71)	21.95 (4.49)	0.002
Act Jump Height (cm)	36 (7.15)	26.47 (7.53)	22.69 (4.73)	0.002

measurements with no contextual information whereas the Kinect One is able to provide quantified kinematic information. This information has been extensively utilised in Kinect-based rehabilitation frameworks such as [177, 179, 180] but yet to be used for age-related change detection.

This chapter has used the centre-of-mass as a key indicator for detecting age-related changes. The features extracted were carefully validated against a force platform, which is typically used in research. No significant differences were found between the various measurements extracted, with a strong correlation found between the Kinect One and force platform in jump height. These results demonstrate the suitability of the Kinect One in detecting motion differences between young and old participant groups.

Under stable balance tests of *two-leg (open eyes)* and *semi tandem (open eyes)*, all groups did not differ significantly in stability which is consistent of other studies of balance such as [108, 170, 181]. These results suggest that the Kinect One is not suitable in detecting subtle stability differences in static balance [168], this hypothesis is supported by validity of the Kinect One sensor hardware [27]. When the difficulty in balance tests increased there was a significantly greater amount of postural sway in both ML and AP directions for the older adults than with the healthy young. For *tandem (open eyes)* balance greater movement was identified for old adults, analysis of the test suggests that foot position in relation to the torso causes stability impairment. The older groups demonstrated significantly more sway and stability impairment than young for tests related to single foot balancing (with eyes open or closed). The retest rate for both old groups compared with young in relation to these tests was 75%.

The older participants, most notably the old participants were very cautious in undertaking tests that required either speed or a requirement to complete the task within a specific time frame. Conversely, younger participants were able to undertake tests confidently, even though it reduced stability. For example, in *Chair Rise*, participants were asked to perform five repetitions as quickly as possible. While average time per rise and standard deviation were similar, upper-body stability significantly varied between the young and old. In this chapter, the results demonstrate that the Kinect One discriminates well between different participant groups and is feasible for a clinical environment. Ejupi et al. [11] further supports the conclusion that the Kinect One device was capable of detecting subtle changes in a clinical setting for five times chair rise.

Using features extracted from the Kinect One, balance control during walking in older adults compared to their younger counterparts are significantly different for sway and CoM change. Stability was varied between the young and old, with little difference in time taken and velocity. This is supported by other studies such as [13, 111]. However, other researchers have found [12] that older participants exhibited a more conservative gait patterns, which is characterised by a slower velocity. The inconsistency in results may in part be due to the Kinect One providing more detailed information of the observed motion.

It can be hypothesised that the use of sensor technology has enabled researchers to extract motion indicators with relative ease. Older adults exhibited greater medial-lateral sway in their gait when walking. That said, they are still capable of maintaining a similar velocity to that observed by the younger participants, in line with other works [12].

This chapter comprised of data compiled from 56 participants, however, amongst the population of older people, few of them had serious mobility limitations. Despite this, there were some very clear differences between young and older people, for example in balance and walking. The older participants were matched with young participants, yet using the feature list defined previously has identified differences in motion performance between the groups. Using marker-less technology, such as the Kinect One can aid in the quantifiable detection of age-related mobility differences. The framework in this chapter demonstrated the use of a commercial, low-cost product to provide accurate motion information and analysis robustly. This work can aid in the development of software solutions capable of supporting and directing clinical provision to aid mobility enhancement of the participant. Chapter 8 extends the concepts presented in this chapter to create a framework that is capable of automatically determining age-related changes between participant groups.

Chapter 8

Application: Analysis and Automated Quantification of Human Mobility

In this chapter, a solution of automated quantitative evaluation of motor-skeletal control disorders using the Microsoft Kinect One is presented. The application is divided into two parts. Firstly, the ability to robustly detect a set of standardised tests (e.g. sit-to-stand, walk 4 meters) from a depth sensor. Secondly, analyse and evaluate the test sequence to identify the changes in kinematic features by comparing the mobility of the young and old. This chapter introduces novel analysis and quantification framework that has proven successful in quantifying human mobility.

8.1 Introduction

In Chapter 7, the problem of detecting age-related changes between the young and old was discussed. With the population ageing, an important factor in providing health and social care services is to quantify and continuously assess participant. Frailty is an indicator of general health and well-being, and is usually assessed by asking the person to perform several standardised tests (*e.g.* walk back and forth, sit to stand which are components of the Short Physical Performance Battery (SPPB) [172]) during which a clinician observes the activity for stability, duration, coordination and posture

control. Although the person-clinician assessment method is common, there is room for improvement to construct a more efficient and reliable framework for the following reasons [179, 182–185]. First, clinician-led assessment is mostly subjective instead of objective quantification. Second, clinical scales, such as the SPPB lack objectivity and correlation when person variance is taken into account. Further, clinicians are required to interpret the results to make a decision. Third, the entire process can be time consuming considering the participant need to travel to the appointment, prepare, and undertake the assessment. Fourth, it is clear that the majority of participants would prefer to undertake the assessment at home instead of travelling. Fifth, participants may exhibit different behaviour as a result of examined which may alter the outcome.

Inspired by the success of [14, 186], there is growing support in the literature which indicates, if we were able to reliably detect and/or identify a person who has "poor motor-control", it can be a predictor of a general decline in health and intervention can be undertaken [10, 178]. Several attempts have been made to develop home-based monitoring and quantification systems for assessment and rehabilitation [187–189]. While these systems have been clinically validated and have potential to solve home-based monitoring and quantification task, they fall short of assessing frailty and mobility. Further, in the majority of cases, these systems provide a single indicator instead of a detailed analysis, which would provide a more detailed measure to clinicians [185]. In addition, as highlighted in chapter 7, many of the existing frameworks have been evaluated using game-orientated datasets, and not clinically supported. This chapter uses the K3Da dataset introduced in chapter 7 to propose an application framework to monitor, assess and quantify mobility using depth sensor technology.

This chapter unites human action recognition techniques presented in chapter 5 and chapter 6 with motion analysis presented in chapter 7 to develop a reliable, accurate monitoring and evaluation system that is capable of measuring mobility between the young and old. The system acquires the skeletal stream from a single depth sensor. The skeletal stream is decomposed into novel joint-group features that are used for recognition and quantification. A novel framework for evaluating and analysing mobility to aid in clinical invention is further introduced.

The main contributions of this chapter are as follows:

1. Proposal of a non-invasive recognition framework using MoCap feature selection and representation for real-time recognition (Section 8.4).

2. Proposal of a clinically reliable quantification and analysis framework to provide joint-level feedback indicating the mobility of the participant (Section 8.5).

3. Empirical experimental evaluation and comparison between young and old to obtain quantitative objective outcome measures for recognition and human mobility (Section 8.6).

8.2 Application Framework

The general framework for the method presented in this chapter is divided into three main parts. Firstly, feature encoding to provide a rich powerful compact representation (Section 8.3). Secondly, identification and recognition of human motion (Section 8.4). Finally, motion analysis and quantification (Section 8.5). The dataset captured and presented in chapter 7, obtained using a depth sensor with participants performing standardised clinical tests is used to validate the proposed framework. For clarity, recall that a Microsoft Kinect One skeleton is represented by a stream of MoCap skeletons, with up to 25 joints tracked at a rate of 30 frames-per-second (Section 7.2 provides more detail on the functionality of the Kinect). The sensor is low-cost and able to be operated in a wide variety of locations, making it ideal for the application presented in this section.

8.3 Feature Encoding

Identification and recognition of gestures, motions and activities is not a trivial task. In chapter 6, an evaluation of the ability to detect human action using a Microsoft Kinect sensor yielded promising results. However, the same feature set would not provide the abstract level of detail required for in-depth quantitative analysis, evaluation and outcomes to be determined. This, in part, is due to the inherit way in which actions between humans differs slightly, making a single top-level feature vector generalised to a high degree [16, 117]. Rich and informative features have been shown to provide an improved feature representation for recognition [104, 106, 190, 191]. Du et al. [122]

proposed a hierarchical recurrent Neural Network for human action recognition, at the core was the concept of dividing the skeleton into joint groups, based on anatomical significance to the action sequence. Using this knowledge, a novel joint level group feature which is informative, representative of multiple action types and capable of encoding subtle variations is introduced.

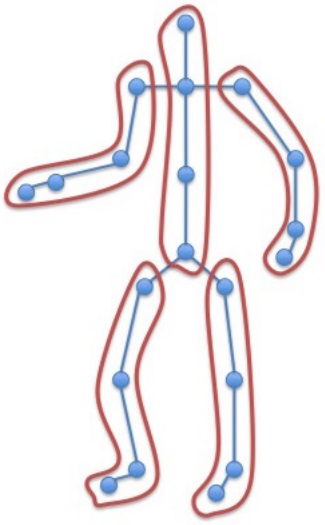

FIGURE 8.1: An illustration of the whole skeleton divided into five joint groups. Each joint group represents a key motion area which is capable of representing all types of human motion.

The framework for feature encoding is presented in Table 8.1, and a visual representation of the joint group decomposition of the skeleton is presented in Figure 8.1. For recognition, the given joint groups are merged into a single feature vector for training, whereas for motion analysis each joint group is encoded and modelled individually.

There are multiple measurements that are capable of being extracted from skeletal stream [110, 142, 191]. This dilemma requires the selection of the most appropriate features independent of human subjectivity. Specifically, only those features that are anthropometric, style invariant and can easily and quickly be extracted from the stream should be utilised. Table 8.1 provides a summary of the joint decompositions plus derived features that form each joint group. Alongside the introduction of new features, features derived from the MoCap data itself are also provided. Specifically, the x and y are extracted for each joint to describe the posture in MoCap coordinates [104]. Further, several of the features mentioned in Table 8.1 have been discussed previously in this thesis, for clarity a summary is provided.

TABLE 8.1: Summary of joint decompositions and derived features that form joint group representations and the corresponding dimensionality of the final feature vector.

Joint Group	Features	Length	Notation (where I is the number of features)
Left Arm ($LeftShoulder$, $LeftElbow$, $LeftWrist$, $LeftHand$)	Left arm Euler Angle (between left shoulder and left wrist), Euclidean distance between the left shoulder and left hand, x and y axis vectors.	10	$F_{LeftArm} = \{1 \ldots, I\}$
Left Leg ($LeftHip$, $LeftKnee$, $LeftAnkle$, $LeftFoot$)	Left leg Euler Angle (between left hip and left ankle), Euclidean distance between the left hip and left foot, x and y axis vectors.	10	$F_{LeftLeg} = \{1 \ldots, I\}$
Right Arm ($RightShoulder$, $RightElbow$, $RightWrist$, $RightHand$)	Right arm Euler Angle (between right shoulder and right wrist), Euclidean distance between the right shoulder and right hand, x and y axis vectors.	10	$F_{RightArm} = \{1 \ldots, I\}$
Right Leg ($RightHip$, $RightKnee$, $RightAnkle$, $RightFoot$)	Right leg Euler Angle (between right hip and right ankle), Euclidean distance between the right hip and right foot, x and y axis vectors.	10	$F_{RightLeg} = \{1 \ldots, I\}$
Torso ($SpineBase$, $SpineMid$, $Neck$, $Head$, $SpineShoulder$)	Torso Euler angle (between the spin base and neck) relative to the body, Euclidean distance between the spine base and head, Body lean angle (relative to the floor with torso as a reference), Centre-of-Mass (between left shoulder, right shoulder, spine mid), x and y axis vectors.	16	$F_{Torso} = \{1 \ldots, I\}$

Euler Angle: Recall in chapter 3, that any rigid body can be described as some angle around three mutually orthogonal coordinates in fixed space. However, obtaining Euler Angles from marker-less MoCap is difficult. The Euler Angle is useful as it provides a subject invariant feature that is indiscriminate to body size or performance style. The Euler Angle between a set of joints (reference joints provided in Table 8.1) is computed by calculating the Coordinate Matrix (other known as the Rotation Matrix), discussed in Section 3.1.4, and performing Euler transformation into Euclidean 3-space ($\mathbb{R}^3$) [192].

Euclidean Distance: An important characteristic of human motion is the way in which

the participant transitions over time in relation to a fixed point. For example, in the torso group, the Euclidean distance is computed between the base of the spine and head. While this value will remain stable for actions such as walking, when the participant performs a bend, or sit-to-stand motion the distance between the two joints differs. This change in distance is modelled by the Euclidean distance between joint features. The Euclidean distance between each reference joint is defined as:

$$distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \tag{8.1}$$

where x_1, y_1, z_1 and x_2, y_2, z_2 are the respective joint locations.

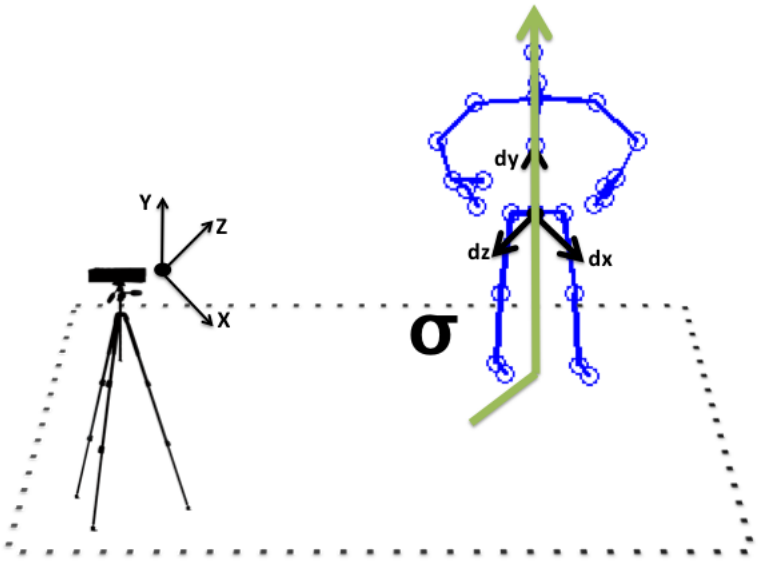

FIGURE 8.2: Visual representation of the body lean angle in relation to the Microsoft Kinect sensor. The angle is computed by the intersection between the ground plane and spine on the skeleton.

Body Lean Angle: Recall in chapter 3, that it is possible to represent an **SO**(3), which is a rotation in Euclidean space as a pair of vectors (unit vector $\hat{\mathbf{e}}$ indicating the direction of the axis rotation, and an angle θ representing the magnitude of rotation about the axis). The body lean angle represents the body orientation in relation to the ground plane, see Figure 8.2 for a visual representation. The angle is computed by the flexion of the spine in relation to ground floor plane, defined at the centre of the feet. The lean angle between the spine and the floor is defined as:

$$\theta = \arccos\left(\frac{S \cdot Q}{\|S\|\|Q\|}\right)^t \tag{8.2}$$

where S is the spine vector (x, y, z) and Q is the floor vector - the middle of the feet (x, y, z).

Centre-of-Mass: The CoM is computed for the torso, extracted from Kinect One data, describes the directional movement of the participant. Figure 8.3 demonstrates a visual example for the CoM for two specific action sequences. Subtle direction movements, such as *Chair Rise* (see Figure 8.3), are identified by the Kinect One due to its ability to track millimetre postural changes [3]. Chapter 7 introduced the CoM feature, and provides further discussion on the computational process (Eq. 7.1). However, a key outcome from the chapter, supported by the literature [27, 116, 168], is the ability of the Kinect One to robustly and with significance capable of tracking the CoM accurately.

8.4 Recognition: Motion Identification

The general framework for recognising tests is shown in Figure 8.4, which is derived from the contributions presented in chapter 4 and chapter 5. The framework presented in the Figure 8.4 is split into two aspects: offline training of multiple state-of-the-art machine learning techniques based on an exemplar-based pose selection. Online detection and identification of motions in real-time to provide motion analysis. While the skeleton stream can be an informative representation (as found in chapter 6), the features presented in Section 8.3 are used instead of the MoCap skeleton. The recognition framework is presented hereafter.

8.4.1 Feature Reduction and Selection

This section describes the feature ranking and selection-based method used to classify actions (tests) from Kinect One MoCap (visualised in Figure 8.4). For generalisation and consistency, let $\mathcal{P}$ be the set of all skeletal poses, ordered in a time sequential manner. The human skeleton obtained from the Kinect One is descriptive, however as discovered throughout this thesis it lacks the depiction of the integral details of the motion. To overcome this, the skeleton is encoded with the features summarised in Table 8.1. Then, all actions that belong to the action class are grouped, based on an automatic k-means clustering technique. This top-level clustering process generates a generalised k which is an atypical representation of the action phases. With this knowledge, the

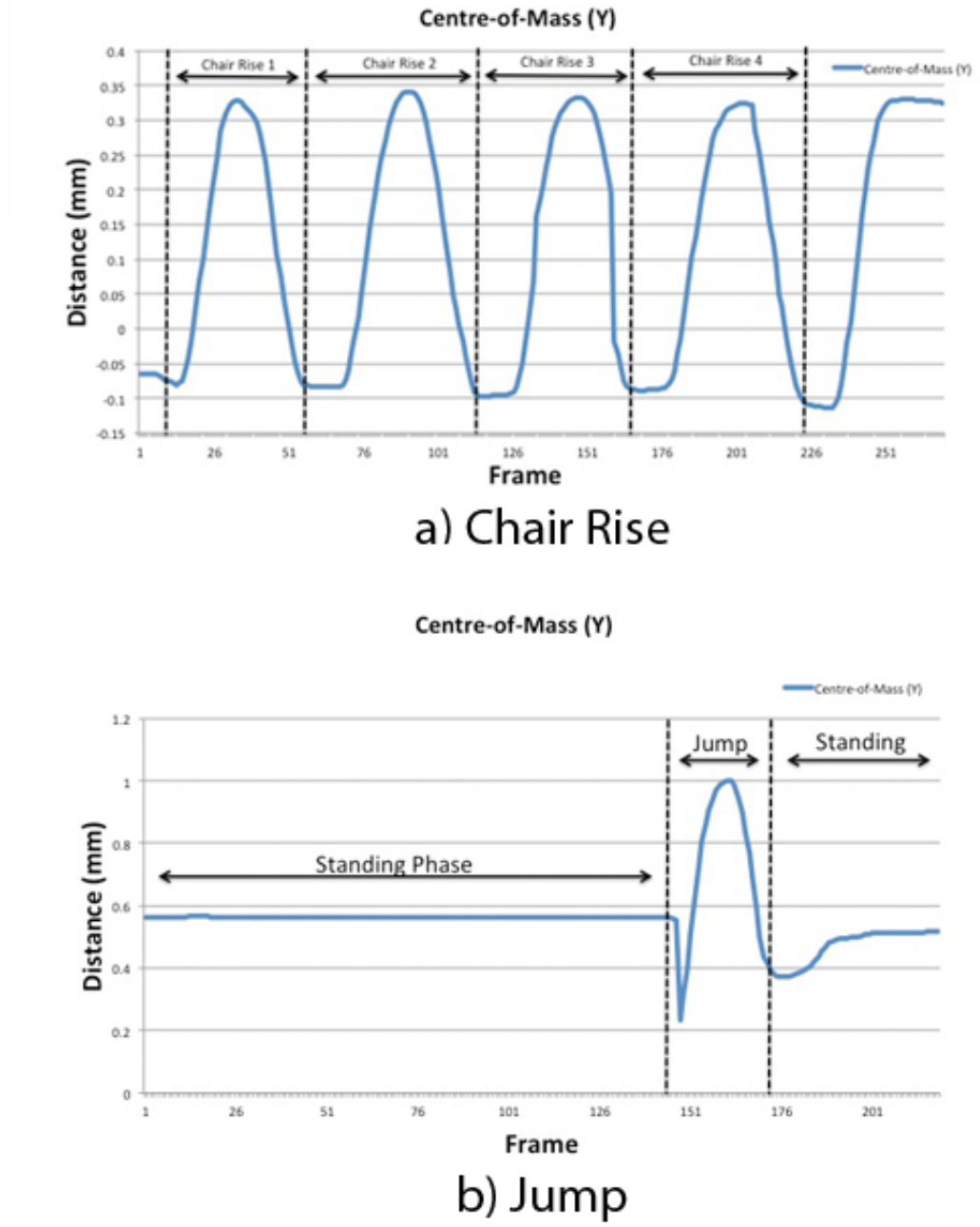

FIGURE 8.3: Visual representation of the Centre-of-Mass. a) Example of the Centre-of-Mass (y) for *Chair Rise*. b) Example of the Centre-of-Mass (y) for *Jump*.

feature encoding for a single action sequence is recovered. Then, a sub-level k-means is performed to group the features. To finalise the framework, firstly key clusters are identified and retained. Secondly, within those key clusters features are identified and selected based on their informative representation to the cluster using a novel equivalence function.

The encoding of $\mathcal{P}$ is performed as described previously. Each feature encodes a specific

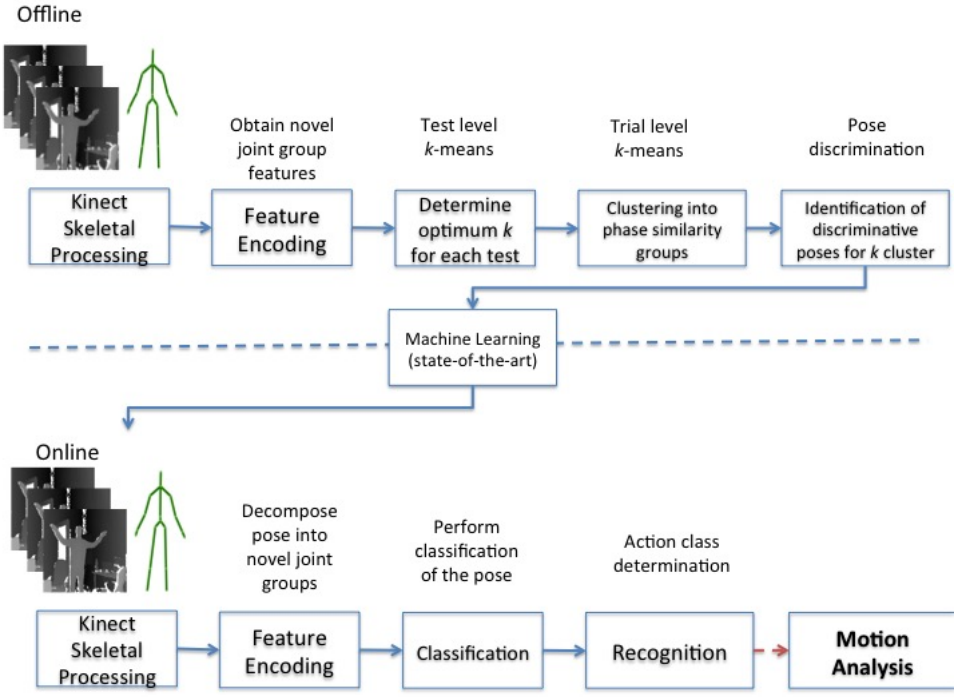

FIGURE 8.4: Recognition Overview: An overview of the recognition framework. The top row illustrates the training process to training state-of-the-art machine learning techniques. The bottom row illustrates the online recognition process utilised to perform classification of a motion.

motion type, such as gait, or action style. To train any machine learning, a unified training sample needs to be formed. Therefore, the features for each joint group are concatenated together, given as:

$$\hat{F} = \{F_{LeftArm}, F_{LeftLeg}, F_{RightArm}, F_{RightLeg}, F_{Torso}\} \in \mathbb{R}^{56} \tag{8.3}$$

where $\hat{F}$ is a combined vector consisting of the features derived for each joint group.

The main objective is to identify and extract only those features that provide the most information about the action (motion). To this end, a two tier clustering process (See Section 3.1.7) is undertaken, presented in Algorithm 1. Top-level clustering process combines each feature encoding for each action into a single matrix. An automated clustering approach is then employed to identify the optimum number of clusters. With this knowledge, a sub-level clustering is undertaken on each feature group with the derived optimum clusters. Therefore, for a feature group $\hat{F}$ is represented by k clusters.

Algorithm 1: Automatic k selection and grouping with k-means clustering

Input: $\mathcal{A}_x = \{\hat{F}_1, \hat{F}_2 \ldots, \hat{F}_N\}$ - training instances for all $\hat{F}_n$ in the action class in concatenated form
$MaxIt$ - maximum number of convergence iterations
Output: $L = \{l(e)|1, 2, \ldots, E\}$ - set of cluster associate labels for $\hat{F}_n$

For a set of features $x \in \mathcal{A}$ **do**
foreach $k = 2 : N$ **do**
  randomly initialize k centroid location, C_i, for each cluster
  foreach $a_i \in \mathcal{A}$ **do**
    $l(e) \leftarrow argminDist||a_x - c_i||^2, i \in \{1, \ldots, k\}$
  end
  $it \leftarrow 0$
  repeat
    foreach $a_x \in \mathcal{A}$ **do**
      $minDist \leftarrow argminDist||a_x - c_i||^2, i \in \{1, \ldots, k\}$;
      if $minDist \neq l(e)$ **then**
        $l(e_n) \leftarrow minDist$
      end
    end
    $it++$;
  until $it \leq MaxIt$;
  $wcss_k \leftarrow argminDist||\mathcal{A}_{\hat{n}} - C_i||^2, i \in \{1, \ldots, I\}$
end
$est_k = E_I^*\{\log(wcss_k)\} - \log(wcss_k)\}$
then
foreach $\hat{F}_n \in \mathcal{A}$ **do**
  randomly initialize est_k centroid location, C_i, for each cluster
  do classify $\mathcal{F}_n$ samples according to nearest C_i
  recompute C_i
  until no change in C_i
  end
  return cluster identifications for each feature (n)
end

After the clusters have been identified, the next stage is to identify and extract key features. A product of the clustering process is k, but because of the dynamic process of clustering, they could contain only a few relevant features. Therefore, only those clusters that contain a number of poses are identified as "key clusters", and key features are only extracted from these clusters. A "key" cluster is identified if it contains more than N/K, the average size of the clusters. For each "key" cluster, the key features are those that are the most representative and informative, this is determined by an equivalence function that will be presented hereafter.

The similarity between two features, a and b from $\hat{F}$ is computed as:

$$Similarity(\hat{F}_a, \hat{F}_b) = \min ||a_i - b_j||^2 \tag{8.4}$$

A Self-Similarity Matrix S for a "key" cluster KC from $\hat{F}$ can be computed using Eq. 8.4 and defined as:

$$S := (s_{i,j})_{N_z x N_z} = \{Similarity(\hat{F}_i, \hat{F}_j)\}_{N_z x N_z} \in KC \tag{8.5}$$

where S is the computed Similarity-Matrix with a dimensionality of $N_z x N_z$ for a cluster KC. The Self-Similarity Matrix provides an insight into the relation between features, it is now possible to rank and extract those features that are the most informative. The median element of the Similarity-Matrix S_{median} is selected, and a cost function is defined to identify those features that are within a threshold, denoted as *hold* are retained. This is computed as:

$$D(S_{median}, S_i) = hold \nleqslant \sum_{i=1}^{I} ||S_{median} - S_i||^2 \tag{8.6}$$

where $i \in I = \{1, 2, \ldots, I\}$ is the number of poses for the key cluster and D are those features that fall within the threshold *hold*. This results in the extraction of only those features that are informative and representative of the key clusters. Furthermore, this provides a more compact representation than the original feature.

8.4.2 Recognition

For recognition, machine learning techniques are employed, see Section 8.6 for the experimental protocol. For training, each action class is represented by a set of key clusters, derived from each motion associated with the class. These are modelled using machine learning. To classify and identify the action, the skeletal stream in real time is encoded using the features summarised in Table 8.1. These for each time period, they are passed to the machine learning approach to determine the associated class.

8.5 Motion Analysis and Evaluation

The general framework for analysing human mobility is presented in Figure 8.5, which is supported by the clinical validation presented in chapter 7. The framework is split into two aspects. Firstly, the data is assigned a ground truth marker identifying if it contains "good" or "poor" mobility, then multiple SVM's are trained to detect mobility changes based on joint groups. Secondly, detection, identification and analysis of participants mobility is given with clinically supportive outcomes. While the skeleton stream can be an informative representation (as found in chapter 6), the features presented in Section 8.3 are used instead of the MoCap skeleton. The motion analysis and mobility framework is discussed hereafter.

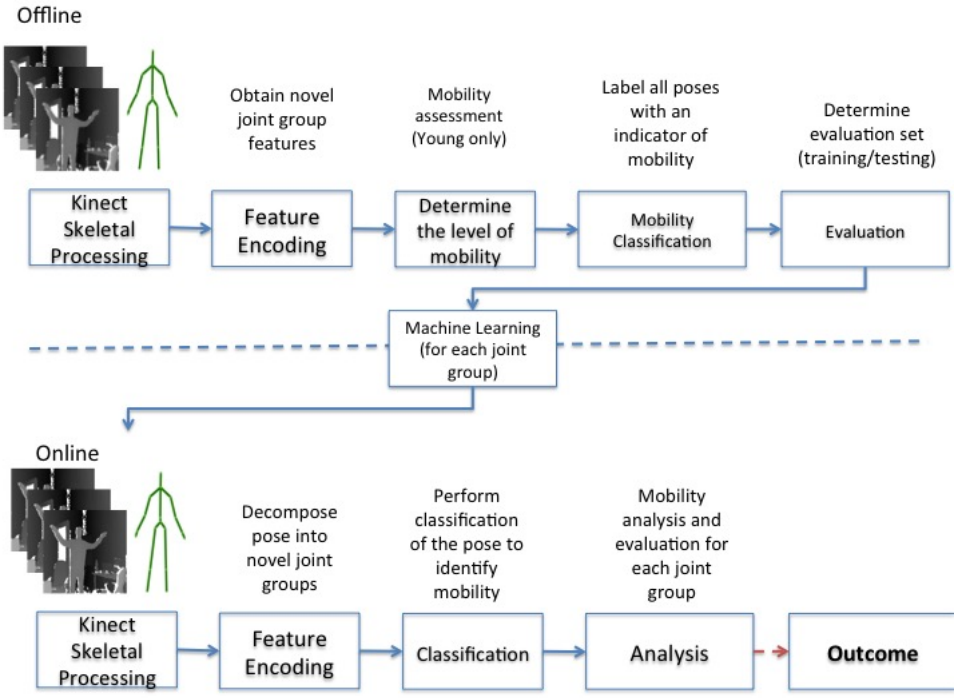

FIGURE 8.5: Motion Analysis Overview: An overview of the analysis framework. The top row illustrates the process undertaken to label, group and train a set of SVM models. The bottom row illustrates the quantification and analysis approach utilised to provide clinically supportive feedback.

The framework for feature encoding is presented in Table 8.1, and a visual representation of the joint group decomposition of the skeleton is presented in Figure 8.1. For recognition, the given joint groups are merged into a single feature vector for training, whereas for motion analysis each joint group is trained into a separate model. There is a dilemma in which features to retain to enable efficient mobility representation. For

motion analysis, only those features that are anthropometric, style relevant and contain directional movement should be extracted and utilised. Table 8.1 provides a summary of the joint decompositions and derived features that form each joint group. Further, Section 8.3 provides a detailed explanation for computing these features, and their relevance to action representation.

8.5.1 Labelling and Computation of Human Mobility *Score*

For a typical recognition task, prior knowledge of the class label is required. This is usually very straight forward to determine, for example a person *walking* or *jumping* can easily be defined with a single label [139]. However, the task becomes very difficult to identify and label in the context of different styles of the same motion. For example, attempting to group different types of gait manually can result in subjective grouping and bias [113, 128, 171]. There have been several approaches proposed to obtain clinically supportive outcomes, yet they have been manually annotated with no clinically supportive reasoning for how labelling was undertaken [14, 16, 128]. On the other hand, in the clinical literature there are methods for objectively identifying human motion. Baumgartner et al. [193] introduced a normal distribution of motion values to derive the SD of the mean as a way for defining groups of sarcopenia (loss of muscle mass with ageing). This methodology has been used extensively within the medical community [184, 194], yet to the authors' knowledge it has not been utilised within the computer science community as a form of labelling.

In this section, the methodology proposed in Baumgartner et al. [193] is used to define "good" and "poor" mobility using a novel digitalised labelling framework. This framework is free from human interpretation, bias or subjectiveness. The labelling can be summarised as follows: Those frames that contain a value greater than the $\pm$ 1.5 SD of the mean are identified as "poor" mobility', whereas those within $\pm$ 1.5 SD of the mean are identified as "good" mobility. In [193], $\pm$ 2 SD from the mean was used to identify groups, however in this chapter a $\pm$ 1.5 SD from the mean has been selected to represent the limited data sample used for evaluation. Each joint group (and associated features) is labelled individually. Meaning, that for each joint group, of each motion, each frame is labelled as having "good" mobility or "poor" mobility. The labelling is summarised as follows:

1. Using only **young tests**, for each action class, each joint group is combined into a single matrix. This results in five matrices representing each joint group of the class.

2. The ± 1.5 SD from the mean is computed for each joint group.

3. Using the SD of the mean values computed at item 2, all tests including **young and old** are labelled for each joint group.

4. Those frames that that are within the ± 1.5 SD are labelled as having "good" mobility.

5. Those frames that lie greater than ± 1.5 SD are labelled as having "poor" mobility. Which presents is a concern in relation to the mobility observed.

The mean ± 1.5 SD threshold value are computed from the young only, to represent the general population. As this work seeks to identify mobility, using the elderly may create a bias within the model and increase the rate of false positives. Table 8.2 provides a summary for the number of frames labelled as having "good" or "poor" mobility for each participant group.

TABLE 8.2: Summary of associated frame labels assigned for "good mobility" and "poor mobility" for each joint group for the young and old.

Joint Group	Young		Old	
	Good (%)	Poor (%)	Good (%)	Poor (%)
Left Arm	31,382 (87)	4,131 (13)	12,516 (51)	12,851 (49)
Right Arm	32,145 (84)	5,146 (16)	13,728 (54)	11,639 (46)
Left Leg	30,367 (89)	3,368 (11)	18,089 (60)	7,278 (40)
Right Leg	31,344 (87)	4,169 (13)	19,725 (72)	5,642 (28)
Torso	31,355 (87)	4,158 (13)	16,089 (57)	9,278 (43)

The labelling of each frame of a motion provides an overview of the state of mobility for the participant performing the motion; this section introduces a "mobility score" metric. This metric indicates the level of mobility the participant has compared to the population and computed using the number of frames identified as having "good" and "poor mobility". The mobility score is an aggregate of the number of frames identified as "good" mobility versus "poor" mobility for each joint group, the final score the average of all joint group scores.

Automated labelling enables a ground truth to be derived from the data itself, free from human interpretation or subjectiveness. Furthermore, computing the mobility

score provides another insight to the state of mobility for each participant. Using this information, it is possible to quantify human mobility and validate it using the ground truth-values obtained in this section.

8.5.2 Analysing Mobility using Multiple SVMs

Throughout this thesis, a number of machine learning techniques have been utilised. Consistently throughout these experiments, SVM have yielded consistently high accuracy results and they are computationally less expensive to train, and provides a low latency for classification (see Section 6.3). The objective of this chapter is to provide detailed insights into the level of mobility of a test participant. To that end, a sample of the dataset is extracted, and each joint group is modelled using an SVM with 10-fold cross-validation, Figure 8.6 demonstrates the training and evaluation pipeline. While it is possible to train a single SVM, indeed chapter 6 obtained high accuracy results for the task of recognition, however these approaches model subtle motion variations resulting in over generalisation (over fitting) leading to inter-/intra-class confusion between "good" mobility and "poor" mobility. Training an individual SVM for each joint group enables the modelling of subtle changes in motion, providing a greater contextual understanding, which leads to improved classification accuracy. Furthermore, it enables the framework to identify specific joint groups that may be of concern.

To obtain an outcome, test data is decomposed into the feature set defined in Table 8.1 and fed into the corresponding SVM. Each corresponding SVM provides a feature-level classification of "good" mobility or "poor" mobility, detailing the level of mobility being observed. Using this classification, detailed analysis of the motion is undertaken, resulting in a decision of level of mobility being observed, Figure 8.7 demonstrates a sample output from the framework.

For sample output observed in Figure 8.7, the level of mobility is determined at a group level. Each joint group is assessed based on the number frames classified as having "good" mobility or "poor" mobility. If any joint group has more than a predefined number of frames labelled as "poor" mobility, an outcome is generated highlighting that further investigation is required. In this example, the *Left Leg* has been highlighted as a concern. This is due to 43% of the frames in this group being identified as having poor mobility. Otherwise, if the joint group is below the threshold, such as *Left Arm, Right*

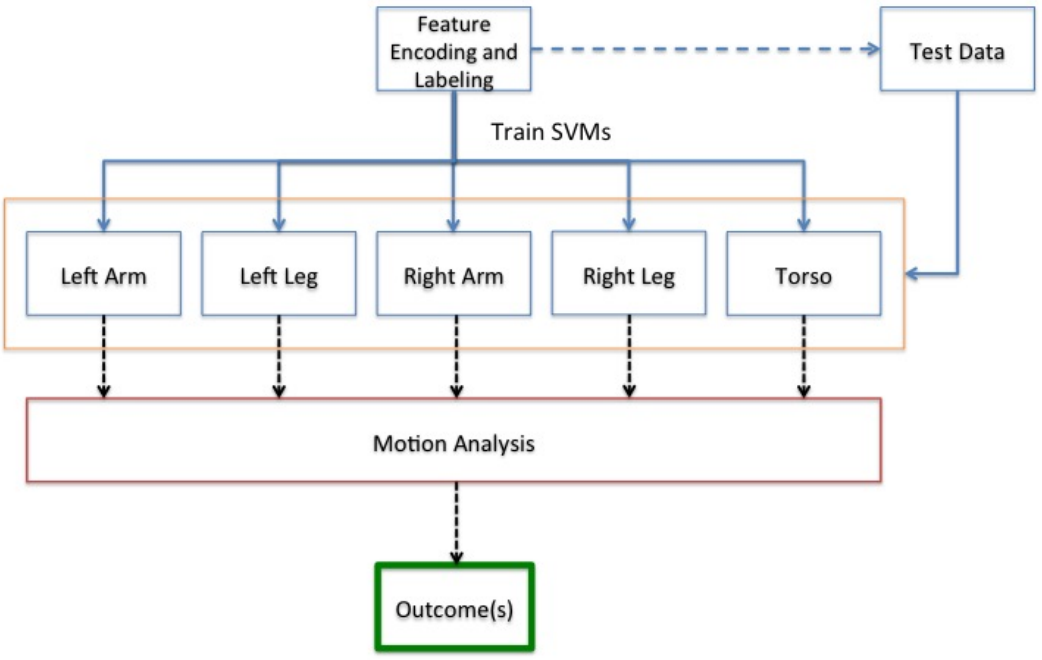

FIGURE 8.6: Summary of the training and evaluation (testing) approach for analysing and evaluating human mobility.

```
Subject Name:  09_JR_8_1 Subject Group: 1
Left Arm is a acceptable (mob=1 acc=1)
Left Leg is a concern (0.43017 acc=0.91061)
Right Arm is a acceptable (1 acc=1)
Right Leg is a acceptable (1 acc=1)
Torso is a acceptable (1 acc=1 )
Estimated Mobility 0.88603 Actual Mobility 0.87263
```

FIGURE 8.7: Example output from the proposed mobility analysis framework.

Arm, Right Leg, Torso, the joint group is noted as having acceptable - meaning that no mobility issues have been detected. The Mobility Score provides a snapshot, single-level value that quantifies the overall level of mobility the participant has (for an observed test).

8.6 Experimental: Motion Detection and Quantification

This section presents the quantitative in-depth analysis of the proposed framework for real-world detection and analysis to support mobility-related clinical outcome measures. The evaluation of the framework is decomposed into two tasks; Firstly, Section 8.6.1

evaluates the ability of the framework to robustly perform action recognition. Secondly, Section 8.6.2 evaluates the ability of the framework to detect age-related mobility concerns. As a consistent theme throughout this thesis, the evaluation is performed on "unseen" test sequences, meaning that no test data has been "seen" by the modelling. A number of standardised test scenarios for assessing mobility were proposed in Section 7.2 (see Table 7.1), as part of the K3Da dataset, which have been clinically validated. In this work, eight tests scenarios have been selected for the focus of this evaluation, namely; balance (eyes open), chair rise, semi-tandem balance, tandem balance, walk (4 meters).

8.6.1 Evaluation: Motion Detection

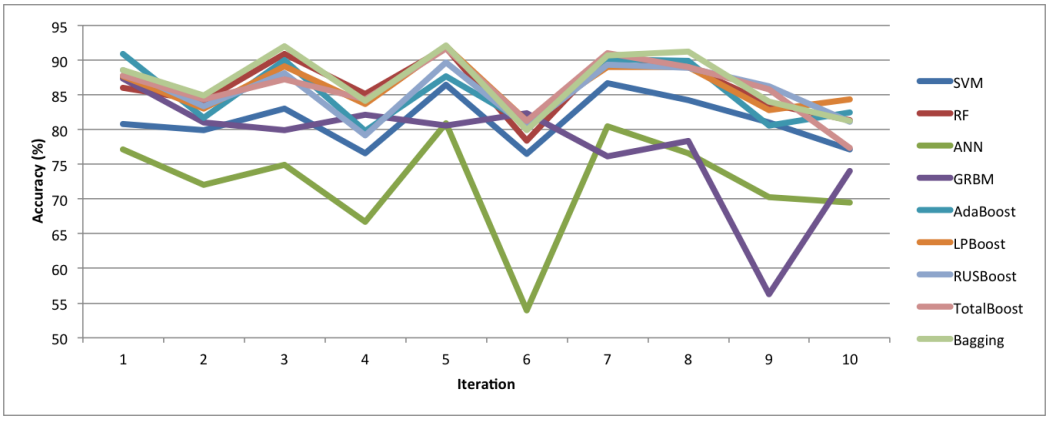

FIGURE 8.8: Visual representation of the motion detection and recognition rate for each technique across all iterations.

In order to assess the validity of the feature representation framework, a large number of state-of-the-art machine learning techniques were employed for motion detection and recognition. These techniques, along with required parameters and parameter selection methodology are presented in Table 8.3. For each technique, a 10-fold cross-validation using leave-one-out was implemented to compute the recognition results. Table 8.4 and Figure 8.8 show the results from using each machine learning technique for motion detection and recognition.

Each machine learning technique was capable of obtaining acceptable recognition rates (see Figure 8.8), most notably when factoring in the similarity in motions and the variations contained within the motion. Overall Bagging produced the highest average recognition rate of 86.88% with ANN producing the lowest average result of 72.25%.

TABLE 8.3: Summary of the machine learning techniques utilised to validation the recognition framework, including parameter selection approach.

Machine Learning	Parameters Required	Parameters Selection
Support Vector Machines [141]	C and γ	Parameter selection undertaken to cross-validation method to determine required parameters [143].
Random Forests [144]	n_{tree} and m_{try}	The number of trees, n_{tree} was set at a default of 2000, and m_{try} set at a default of 3.
Artificial Neural Networks [144]	n_{layer}	The number of layers, n_{layer} was set to a default of 2.
Gaussian Restricted Boltzmann Machines [19]	$h_{variable}$	The number of hidden units, $h_{variable}$ was set at a default of 500.
Adaptive Boosting [195]	N/A	Default Matlab parameters.
LPBoost [196]	N/A	Default Matlab parameters.
RUSBoost [197]	N/A	Default Matlab parameters.
Total Boost [198]	N/A	Default Matlab parameters.
Bagging [146]	N/A	Default Matlab parameters.

The motion detection and recognition results presented in Table 8.4 fluctuated due to the leave-one-out validation framework, and the machine learning technique employed. Robustly, across the spectrum of results, motion detection of specific tests, such as *semi-tandem* and *tandem* balance, were high, with little inter-/intra-class variation.

Low motion detection and recognition rates were observed for several iterations (see Table 8.4), this may be due, in part, to the formation of the training and testing sets for the specific iteration. Or, cross-validation and parameter selection may have struggled due to inter-class similarity. However, overall it is clearly demonstrated that several of the state-of-the-art techniques for motion detection and recognition performs very well on the K3Da motions encoded using the feature representation and recognition framework presented in Section 8.4. Another consideration is the time in which is takes to perform motion detection, across all iterations an average recognition per frame was below 1ms, it is clear that *real-time* recognition is viable based on the framework introduced in this chapter. Being able to correctly identify a motion is of critically important to ensure the correct model is applied for motion analysis and quantification.

8.6.2 Experimental: Motion Analysis

This section presents the ability of the proposed framework to detect mobility concerns between a group of participants using the framework presented in Section 8.5. It is incredibly difficult to assign a classification to identify clinical motions, however Section 8.5.1 provides a methodology for identifying ground-truth labels. These ground-truths are used throughout this section to evaluate the proposed framework for the task of detecting mobility issues across the participant range. Figure 8.9 provides an overview for the accuracy of the overall framework in identifying features of concern in relation to the ground-truths.

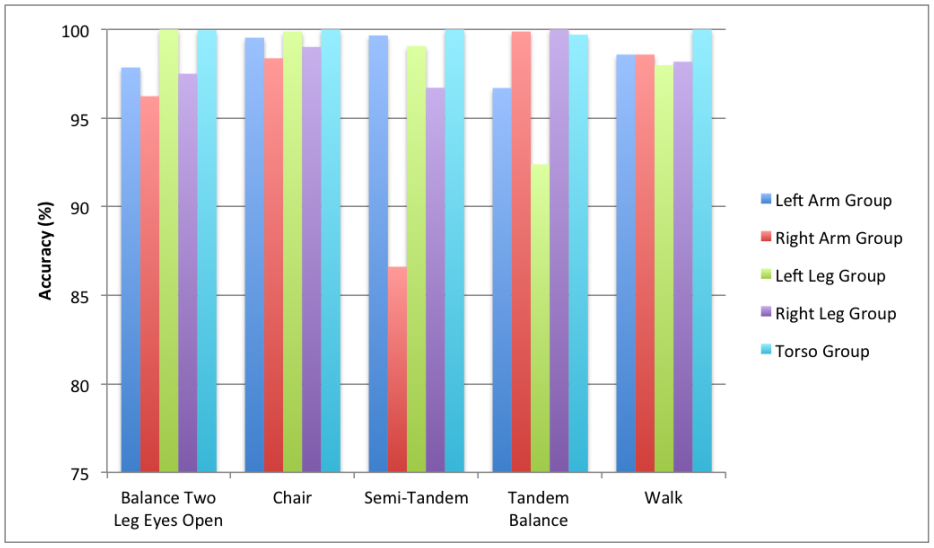

FIGURE 8.9: Visual representation of the motion mobility evaluation compared to ground-truth labels.

For evaluation, random samples of participants are selected, a portion is used for training and the remainder used for testing (80/20 split). The success of the framework is presented hereafter. Observe that across all the clinical tests assessed, a high true-positive rate is obtained. This indicates that the framework is capable of correctly identifying mobility concerns, in relation to the ground truths.

Balance - Two Legs (Eyes Open): The framework was capable of detecting if a participant required any intervention based on ground-truth labels. Table 8.5 provides an overview of the confusion matrix for each joint group. Each joint group could be correctly identified, with only *Left Leg* providing the lowest "accuracy" of 96.23%. The framework was also capable of detecting large amounts of mobility concern, most notably in the *Left Arm* for individual features. Additional analysis demonstrates the success of the

framework with a sensitivity of 0.99, a specificity of 0.96 and a Matthews Correlation Coefficient (MCC) of 0.96. Overall mobility was in line with expectations, with the framework performing reliably across all joint groups using the multiple measurements of accuracy, sensitivity, specificity and MCC.

Chair Rise: The framework was capable of detecting if a participant required any intervention based on ground-truth labels, Table 8.6 provides an overview of the confusion for each joint group. Each joint group could be correctly identified, with only *Left Leg* providing the lowest "accuracy" of 98.38%. Further analysis yielded promising results, with a sensitivity of 0.99, specificity of 0.96 and an MCC of 0.96. Overall, mobility across the participant range was good, with only a minority of features falsely classified as having "concern" across the joint groups.

Semi-Tandem Balance: The framework was capable of detecting if a participant required any intervention based on ground-truth labels, Table 8.7 provides an overview of the confusion for each joint group. Each joint group could be correctly identified, with only *Right Arm* providing the lowest "accuracy" of 86.60%. A large number of features were identified as having a "concern" in *Right Arm*; this may be due to the incorrect classification of features. This false classification is furthered observed with a relatively low sensitivity of 0.95, a specificity of 0.99 and a MCC of 0.91.

Tandem Balance: The framework was capable of detecting if a participant required any intervention based on ground-truth labels, Table 8.8 provides an overview of the confusion rate for each joint group. Each joint group could be as being correctly identified, with only *Left Arm* providing the lowest "accuracy" of 96.69%. For *Right Leg*, all features were classified correctly being of good health. This also may be in part due to most participants using their left leg for the tandem balance resulting in the sensor being obscured; therefore the Right Joint Group may be obscured for duration of the motion. Furthermore, a number of features for *Right Arm* were also misclassified, this may be due to the subtlety of the motion. A high sensitivity of 0.99 was achieved, however low scores for specificity of 0.91 and MCC of 0.92 support the confusion that for this type of balance is can be difficult to obtain a correct classification.

Walk (4 meters): The framework was capable of detecting if a participant required any intervention based on ground-truth labels, Table 8.9 provides an overview of the confusion rate for each joint group. Each joint group could be correctly identified,

with only *Right Arm* providing the lowest "accuracy" of 98.10%. Across the Joint Groups, features were classified correctly, with only a few features being identified as a concern requiring further invention, this was further supported with a sensitivity of 0.99. However, when considering specificity of 0.94 and a MCC of 0.95, the framework does yield low results for this type of test when compared to the others.

With the proposed framework we have implemented a framework to identify any clinical intervention. A threshold of 70% was selected through experimentation, if a participant joint group contain more than 70% of frames classified as "concern" it would be identified as requiring investigation by a clinical professional. Of the participants used in these experiments, 16 were highlighted as having at least one joint group of concern. In accuracy terms, this is a 94% success rate in detecting mobility concerns between participant groups.

8.7 Discussion and Conclusion

With the Kinect One, it has enabled new capacities for innovation within the healthcare sector. The ability to deploy the sensor in a wide range of locations, as well as its low-cost are important highlights. Further, the Kinect One has provided detailed measurements of clinically relevant motions and features. Standard approaches, such as SPPB [172] provide a single score measurements with no contextual information whereas the Kinect One is able to provide finite kinematic information. This information has been extensively utilised in Kinect-based rehabilitation frameworks such as [177, 179, 180] but yet to be used for mobility evaluation, analysis and quantification. The extraction of joint groups provides an abstract level of detail and insights to how the joint group is operating in relation to the motion as a whole, this leads to an improved insight for clinicians to make a recommendation.

While analysis is an important aspect of this work, it is important to highlight the importance of detecting motions as they occur to ensuring the correct outcome measure. Several state-of-the-art machine learning techniques have been computed, presenting a detailed summary of their ability to recognise the motions. They have all done incredibly well at detecting subtle differences between action classes, for example *semi-tandem* and

tandem balance. The features have shown to be the main factor in enabling improved recognition features on a novel clinical dataset.

The data used in this chapter is derived from the K3Da dataset, but it does contain only a limited number of participants. If the number of participants increase in size, a more reliable and representative population model can be computed.

This chapter proposes an application framework which unites human action recognition techniques presented in chapter 5 and chapter 6 with motion analysis presented in chapter 7. The framework has been shown to be reliable, accurate at monitoring and evaluation mobility between the young and old. By utilising low-cost depth sensor technology the application framework is deployable in a large number of scenarios and environments, resulting in real world practical benefits.

TABLE 8.4: Summary of the machine learning recognition results for each iteration per classifier.

Iteration	SVM	RF	ANN	GRBM	AdaBoost	LPBoost	RUSBoost	Total Boost	Bagging
1	80.78%	85.99%	77.10%	**87.83%**	**90.85%**	87.64%	88.62%	87.82%	88.55%
2	79.94%	84.17%	72.09%	81.01%	81.65%	83.06%	83.38%	84.32%	84.88%
3	82.99%	90.90%	74.96%	79.87%	90.02%	89.14%	88.11%	87.20%	92.05%
4	76.59%	85.12%	66.68%	82.10%	79.77%	83.70%	79.11%	84.37%	84.17%
5	86.43%	**91.77%**	**80.89%**	80.56%	87.66%	**92.02%**	**89.69%**	**91.73%**	92.14%
6	76.49%	78.40%	53.95%	82.37%	80.87%	81.26%	80.38%	81.26%	79.89%
7	**86.65%**	90.72%	80.49%	76.10%	90.12%	89.07%	89.33%	91.00%	90.73%
8	84.24%	88.89%	76.83%	78.34%	89.92%	88.99%	88.87%	88.98%	91.22%
9	81.05%	83.64%	70.24%	56.28%	80.54%	82.78%	86.28%	85.77%	**93.99%**
10	77.14%	81.37%	69.50%	74.09%	82.47%	84.34%	81.15%	77.33 %	91.21%
Average (SD)	81.23%	86.11%	72.25%	73.08	85.39%	86.20%	85.50%	85.98 %	86.88%

Left Arm - Predicted Outcome

actual value	p	n
p′	992	49
n′	0	1241

Left Leg - Predicted Outcome

actual value	p	n
p′	1045	0
n′	86	1151

Right Arm - Predicted Outcome

actual value	p	n
p′	1657	0
n′	0	625

Right Leg - Predicted Outcome

actual value	p	n
p′	1372	0
n′	57	853

Torso - Predicted Outcome

actual value	p	n
p′	1891	0
n′	1	390

TABLE 8.5: Balance - Two Legs (Eyes Open): Confusion matrix highlighting the performance of the framework for each joint group in identifying motor control groups of concern by an SVM per feature (pose in time). Where true positive indicates health participants with good mobility and true negative indicates participants with mobility of concern.

Left Arm - Predicted Outcome

actual value	p	n
p′	1569	0
n′	8	145

Left Leg - Predicted Outcome

actual value	p	n
p′	1420	9
n′	19	274

Right Arm - Predicted Outcome

actual value	p	n
p′	1592	1
n′	1	128

Right Leg - Predicted Outcome

actual value	p	n
p′	1510	9
n′	8	195

Torso - Predicted Outcome

actual value	p	n
p′	1489	0
n′	0	233

TABLE 8.6: Chair Rise: Confusion matrix highlighting the performance of the framework for each joint group in identifying motor control groups of concern by an SVM per feature (pose in time). Where true positive indicates health participants with good mobility and true negative indicates participants with mobility of concern.

Left Arm - Predicted Outcome

actual value	p	n
p′	1115	1
n′	5	641

Left Leg - Predicted Outcome

actual value	p	n
p′	1645	3
n′	4	110

Right Arm - Predicted Outcome

actual value	p	n
p′	886	236
n′	0	640

Right Leg - Predicted Outcome

actual value	p	n
p′	1685	58
n′	0	19

Torso - Predicted Outcome

actual value	p	n
p′	1123	0
n′	0	639

Table 8.7: Semi-Tandem Balance: Confusion matrix highlighting the performance of the framework for each joint group in identifying motor control groups of concern by an SVM per feature (pose in time). Where true positive indicates health participants with good mobility and true negative indicates participants with mobility of concern.

Left Arm - Predicted Outcome

actual value	p	n
p′	1128	50
n′	6	510

Left Leg - Predicted Outcome

actual value	p	n
p′	1231	2
n′	0	461

Right Arm - Predicted Outcome

actual value	p	n
p′	1175	0
n′	129	390

Right Leg - Predicted Outcome

actual value	p	n
p′	1694	0
n′	0	0

Torso - Predicted Outcome

actual value	p	n
p′	1448	0
n′	5	214

TABLE 8.8: Tandem Balance: Confusion matrix highlighting the performance of the framework for each joint group in identifying motor control groups of concern by an SVM per feature (pose in time). Where true positive indicates health participants with good mobility and true negative indicates participants with mobility of concern.

Left Arm - Predicted Outcome

actual value	p	n
p′	426	3
n′	4	61

Left Leg - Predicted Outcome

actual value	p	n
p′	335	3
n′	4	152

Right Arm - Predicted Outcome

actual value	p	n
p′	407	2
n′	8	77

Right Leg - Predicted Outcome

actual value	p	n
p′	397	1
n′	8	88

Torso - Predicted Outcome

actual value	p	n
p′	472	0
n′	0	22

TABLE 8.9: Walk (4 meters): Confusion matrix highlighting the performance of the framework for each joint group in identifying motor control groups of concern by an SVM per feature (pose in time). Where true positive indicates health participants with good mobility and true negative indicates participants with mobility of concern.

Chapter 9

Conclusions

In this thesis a number of methods have been proposed for human action recognition, motion analysis and quantification. In this chapter, the contributions introduced throughout this thesis are presented, reviewed and future work proposed.

9.1 Feature Selection, Representation and Recognition

In chapter 4, chapter 5 and chapter 8, methods for representing, ranking and selecting marker-based and marker-less MoCap are described. The approaches seek to generate efficient representation to provide features for recognition. The approach is discussed below and future work highlighted.

9.1.1 Contributions

The DIS (see Section 4.2.1) and DKPI (see Section 4.2.2) entail the use of marker-based MoCap to rank, identify and extract informative poses for use in the recognition process. The approaches have the ability to handle a large variety of action sequences from data sources. The DIS framework selects delegate postures using a statistical ranking and joint discrimination power. Ranking and then selecting the most informative poses based on statistical significance, instead of selecting the most informative based on cluster generalisation and placing in time sequential order, proved to be very effective. On the other hand, DKPI determines key poses by assessing the maximum subspace scoring

of the dissimilarity space of the star skeleton representation. The approach computes local representations based on joint dissimilarity and mutual joint respect to identify key poses.

The DKPI approach consistently achieved the higher classification accuracy for all experiments (see Section 4.4.3). This may be due in part to the model containing more selective examples of the action class which provide more training to enable an improved recognition. Further, by selecting the most informative poses, the approaches have removed the need for large-scale training sets to capture the essence of the action. For example, considering the proposal of Barnachon et al. [22], the framework relies on manual selection of k value for clustering which is an incredibly difficult task. Conversely, this thesis utilises an automatic k selection framework (presented in Section 3.1.7.2). The observed improvements are in line with other work on the exemplar paradigm and human action recognition.

The drawback that unites these two approaches is the ability to select the most suitable representation for each action class. Further, it is important to consider the practical implications of marker-based MoCap and the proposed approaches. There is a "start up" cost in placing the markers on the participant, setting up the hardware and calibration. It is not feasible for implementation of marker-based systems in the *real world*. While action classification can be performed robustly in an offline approach, it is important to make decisions as quickly as possible. It is this limitation that motivates the work presented in chapter 5 and chapter 6.

The ability to detect actions in *real-time* is the desire for many recognition systems. A exemplar-based template system is introduced in chapter 5. The use of an action model to represent each action class offers an advantage over traditional approaches in terms of characterising each class by a small number of exemplars, which has led to the deduction of the size of the training sequences by an average of 98%. This is in contrast to using full motion sequences to train machine learning techniques, with performance inevitably suffering as the quality of training data degrades due to confusion. Chapter 5 explores this concept further and extends the use of feature selection and ranking to improve recognition results for online application. Positively, the use of exponential map representation enables characterisation of the posture in a more discriminative usable

form, but also handles singularities and discontinuities. This has improved on the DIS and DKPI (see chapter 4) and led to real-time recognition.

Conversely, marker-based MoCap is not without its challenges. for example, the cost of hardware and time constraints to place anatomical significant markers on the participants body. But also, the technical knowledge and expertise to utilise the system. Therefore, it led to the question of is it possible to utilise marker-less tracking technology for use in classification by using feature representation techniques? The answer is found in chapter 6, that presented a feasibility study for the ability of marker-less technology for use in classification, specifically focusing on health related tasks. With the concept extended further for age-related health implications and real world deployment in chapter 8.

The use of marker-less technology, notably the Microsoft Kinect One, has created the possibility of creating a low-cost human action recognition framework that is deployable in a wide range of scenarios. Part of the Human Mobility framework presented in chapter 8 relies on robust recognition, with Kinect data as an input. Identification and recognition of gestures, actions and activities is not a trivial task. In chapter 6, an evaluation of the ability to detect human action using a Microsoft Kinect 360 sensor yielded promising results. Rich and informative features have been shown to provide an improved feature representation for recognition [104, 106, 190, 191]. To this end, a rich framework that encodes the spatial-temporal variations of the motion is introduced (see Section 8.3). By utilising these features, a framework inspired by chapter 4 and chapter 5 is proposed to model and detect clinical trials. The framework extracts only those features that are informative, disregarding those that are not - reducing the framework size and improving training times. The performance are in line with other work on human action recognition, see for example Section 8.6.

9.1.2 Future Work

Depth sensor technology presents a new data modality for the application of human body extraction and MoCap extraction. This is important, as it has the ability to remove the need for marker-based systems, resulting in the ability to deploy tracking solutions to a wide range of scenarios (see chapter 8). An emphasis should be placed on improving the accuracy of extracting MoCap from depth sensor technology for use

in the exemplar paradigm. However, using this data in the exemplar paradigm further complicates matters due to the participant validation and data noise. The exemplar paradigm requires the manual selection of the most atypical sequence, future work should seek to address the question; is it possible to determine what is a correct performance of an action sequence, free from human interpretation?

The use of features, such as the temporal and spatial domain would provide a more efficient representation (see Section 6.3). For example, complex activities and interaction between participants are extremely difficult to model. Here, a more detailed, "sub-level" feature encoding framework may provide better results. Further, while MoCap is informative, uniting multi-modality output such as depth and RGB may yield a better understanding of the behaviour and contextual setting of the performance.

9.2 Motion Analysis

In chapter 7 and chapter 8 a framework for analysing human motion was presented. This approach is discussed further below and future work highlighted.

9.2.1 Contributions

Chapter 7 assessed the ability of the Microsoft Kinect One to detect age-related changes between the young, athletic old and old adults using a digital analysis framework. The chapter presented typical routines of clinical movements based on standardised tests such as the Short Physical Performance Battery, Timed-Up-And-Go, Four-Meter Walk and Balance. An important note, these frameworks (and actions) within the science community have not been utilised up until this point. The method was supported with a detailed quantitative analysis for detecting subtle age-related differences between the participant groups. The method used centre-of-mass as a key indicator for detecting age-related changes that was carefully validated against a force platform, which is typically used in research. No significant differences were found between the various measurements extracted, with a strong correlation found between the Kinect One and force platform in jump height. These results demonstrate the suitability of the Kinect One in detecting motion differences between young and old participant groups.

The method was evaluated using data compiled from 43 participants, however, amongst the population of older people, few of them had serious mobility limitations. Despite this, there were some very clear differences between young and older people, for example in balance and walking. Using marker-less technology, namely the Kinect One can aid in the quantifiable detection of age-related mobility differences. The method demonstrated the use of a commercial, low-cost product to provide accurate motion information and analysis robustly. The finding of chapter 7 enables the creation framework that is capable of automatically determine age-related changes between participant groups.

The knowledge that it is possible to quantify human action into different age groups based on mobility is extended to develop an application framework in chapter 8. The problem of automated quantitative evaluation of motor-skeletal control disorders using the Microsoft Kinect One in presented and a solution presented. The application enables non-invasive monitoring and analysis of a participant to provide clinical feedback to aid in the decision process. Conversely, the framework does not seek to remove the clinician from the process, but provide clinically relevant feedback to support the decision making process. The participant performed clinically validated standardised tests (e.g. sit-to-stand, walk 4 meters), extracted from the K3Da Dataset (introduced in this thesis). The application is split into two parts. Firstly, the ability to robustly detect the test obtained from the sensor (see Section 9.1). Secondly, analyse and evaluate the test sequencer to identify if any mobility issues exist. A multilevel approach to detecting human mobility is presented, which relies on the feature representation frameworks discussed in Section 9.1, to identify those poses that may have a mobility issue. Finally, a quantitative framework for determining the mobility "score" of the participant is introduced, and validated.

9.2.2 Future Work

With the benefits of depth sensor technology, the Microsoft Kinect One could become a useful tool for assessing age-related changes in a clinical setting. However, while this thesis has introduced a novel dataset comprising of clinically relevant motions obtained from the sensor, more varied datasets need to be introduced. It is important for the community, which are developing health-related approaches to benchmark against clinically valid datasets (see Section 7.1) to assess their impact. Further, due to the large

inter-individual variability in age and physical capabilities, methods need to be developed that take this into account. For example, some participants could easily perform five chair rises very quickly without losing balance or performance, while others (mainly older people) experienced a deterioration of their performance throughout the test.

Although Section 8.5 contains a framework for analysing human mobility and quantifying it is difficult to generalise the results to a large proportion of the population. Given a large enough dataset of examples this can be addressed experimentally.

9.3 Closing Remarks

This thesis has presented a collection of work aimed at bridging the gap between human action recognition and human motion analysis, by using feature selection and extracting approaches. This has included: (i) Defining approaches for feature extraction/representation (ii) Combining techniques to enable real-time recognition (iii) Detecting age-related mobility issues between participant groups. Each of these contributions has been tested within experimental frameworks used by the community, and various functions have been proposed for recognising and quantifying motion. These techniques have permitted the reduction of the training size to only key informative features, to improve efficiency and reduce latency. With an action class containing sufficiently rich samples, it is possible to identify actions with a high confidence. Further, with the action known, a quantitative framework for determining the level of mobility, in relation to the sample is defined, leading to a clinically viable framework.

Bibliography

[1] L Gorelick, M Blank, E Shechtman, M Irani, and R Basri. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12): 2247–2253, 2007.

[2] X Yang, C Zhang, and Y Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACM International Conference on Multimedia*, number 4, pages 1057–1060, 2012.

[3] J Shotton, R Girshick, A Fitzgibbon, T Sharp, M Cook, M Finocchio, R Moore, P Kohli, A Criminisi, A Kipman, and A Blake. Efficient human pose estimation from single depth images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99:1, 2012.

[4] Carnegie Mellon University Motion Capture Dataset. The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

[5] V Bloom, D Makris, and V. Argyriou. G3d: A gaming action dataset and real time action recognition evaluation framework. In *IEEE Conference on Computer Vision and Pattern Recognition (Workshop)*, 2012.

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

[7] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16:1–16:43, April 2011.

[8] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. A survey on human motion analysis from depth data. In *Lecturer Notes in Computer Science*, pages 149–187, 2013.

[9] J. K. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 18:70–80, 2014.

[10] R. M Collard, H Boter, R. A Schoevers, and R. C Oude Voshaar. Prevalence of frailty in community-dwelling older persons: a systematic review. *Journal of the American Geriatrics Society*, 60:1487–1492, 2012.

[11] A Ejupi, M Brodie, Y. J. Gschwind, S. R. Lord, W. L. Zagler, and K Delbaere. Kinect-based five-times-sit-to-stand test for clinical and in-home assessment of fall risk in older people. *Journal of Gerontology*, 2015.

[12] H. B. Menz, S. R. Lord, and R. C. Fitpatrick. Age-related differences in walking stability. *Age and Ageing*, 32(2):137–142, 2002.

[13] N. Shkuratova, M. E. Morris, and F. Huxham. Effects of age on balance control during walking. *Archives of Physical Medicine and Rehabilitation*, 85(4):582–588, 2002.

[14] R. Wang, G. Medioni, C. J. Winstein, and C. Blanco. Home monitoring musculo-skeletal disorders with a single 3d sensor. In *IEEE Conference on Computer Vision and Pattern Recognition (Workshop)*, June 2013.

[15] D Prochnow, S Bermudez i Badia, J Schmidt, A Duff, S Brunheim, R Kleiser, R.J Seitz, and P.F.M.J Verschure. A functional magnetix resonance imaging study of visuomotor processing in a virtual reality-based paradigm: Rehabilitation gaming system. *European Journal of Neuroscience*, 37(9):1441–1447, 2013.

[16] B. Galnaa, G. Barrya, D. Jacksonb, D. Mhiripiria, P. Olivierb, and L. Rochestera. Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease. *Gait and Posture*, 39(4):1062–1068, 2014.

[17] B. Krausz and C. Bauckhage. Action recognition in videos using nonnegative tensor factorization. In *IEEE Conference on Pattern Recognition*, pages 1763–1766, 2010.

[18] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1290–1297, 2012.

[19] G. W. Taylor, G. E. Hinton, and S. Roweis. Modeling human motion using binary latent variables. In *Advances in Neural Information Processing Systems*, pages 1345–1352, 2006.

[20] H. Fujiyoshi and A. J. Lipton. Real-time motion analysis by image skeletonization. In *IEEE Winter Conference on Applications for Computer Vision*, pages 15–21, 1998.

[21] A. Elgammal, V. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition. In *IEEE Computer Vision and Pattern Recognition*, pages 571–578, 2003.

[22] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou. Ongoing human action recognition with motion capture. *Pattern Recognition*, 47(1):238–247, 2013.

[23] A. Gupta, S. Satkin, A. A. Efros, and M. Hebert. From 3d scene geometry to human workspace. In *IEEE Computer Vision and Pattern Recognition*, pages 1961–1968, 2011.

[24] J Shotton, A Fitzgibbon, M Cook, T Sharp, M Finocchio, R Moore, A Kipman, and A Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297 – 1304, 20 - 25 June 2011.

[25] D. Webster and O. Celik. Experimental evaluation of microsoft kinect's accuracy and capture rate for stroke rehabilitation applications. In *Haptics Symposium*, pages 455–460, Feb 2014.

[26] S. Obdrzalek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel. Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population. In *Engineering in Medicine and Biology Society*, pages 1188 –1193, August 2012.

[27] R. Clark, Y. H. Pua, K. Fortin, C. Ritchie, K. Webster, L. Denehy, and A. Bryant. Validity of the microsoft kinect for assessment of postural control. *Gait and Posture*, 36(3):372 – 377, 2012.

[28] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (3):257–267, 2001.

[29] M. Ziaeefard and H. Ebrahimnezhad. Hierarchical human action recognition by normalized-polar histogram. In *IEEE Conference on Pattern Recognition*, pages 3720–3723, 2010.

[30] H. S. Chen, H.T. Chen, Y. W. Chen, and S. Y. Lee. Human action recognition using star skeleton. In *ACM Conference on Video Surveillance and Sensor Networks (Workshop)*, pages 171–178, 2006.

[31] F. Zhu, L. Shao, and M. Lin. Multi-review action recognition using local similarity random forests and sensor fusion. *Pattern Recognition Letters*, 34:20–24, 2013.

[32] S. Cherla, K. Kulkarni, A. Kale, and V. Ramasubramanian. Towards fast, view-invariant human action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (Workshop)*, 2008.

[33] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, 2006.

[34] L. Liu, L. Shao, and P. Rockett. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition*, 46:1810–1818, 2013.

[35] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):288–303, 2008.

[36] V. Kellokumpu, G. Zhao, and M. Pietikäinen. Human activity recognition using a dynamic texture based method. In *British Machine and Vision Conference*, 2008.

[37] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris. Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding*, 119:27–40, 2014.

[38] W. Zhou, C. Wang, B. Xiao, and Z. Zhang. Action recognition via structured codebook construction. *Signal Processing: Image Communication*, 2014.

[39] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE Conference on Computer Vision*, pages 1365–1372, 2009.

[40] H. B. Zhang, S.Z. Li, S. Y. Chen, S. Z. Su, X. M. Lin, and D. L. Cao. Locating and recognizing multiple human actions by searching for maximum score subsequences. *Signal, Image and Video Processing*, 9(3):705–714, 2013.

[41] D.Y. Chen, H. Y. M. Lioa, H. R. Tyan, and C.W. Lin. Automatic key posture selection for human behavior analysis. In *Multimedia Signal Processing*, pages 1–5, 2005.

[42] X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 58–65, 2009.

[43] I. Laptev and T. Lindeberg. Space-time interest points. In *IEEE Conference on Computer Vision*, pages 432–439, 2003.

[44] C. Harris and M. Stephens. A combined corner and edge detector. In *Fourth Conference on Alvey Vision*, pages 147–151, 1988.

[45] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, pages 65–72, 2005.

[46] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *IEEE Conference on Computer Vision*, pages 1395–1402, 2005.

[47] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *IEEE Conference on Pattern Recognition*, pages 32–36, 2004.

[48] I. Laptev, M. Marszalek, C. Schim, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[49] J. Dou and J. Li. Robust human action recognition based on spatio-temporal descriptors and motion temporal templates. *Light and Electron Optics*, 125(1): 1891–1896, 2014.

[50] Z. Zhao and A. Elgammal. Human activity recognition from frame's spatiotemporal representation. In *IEEE Conference on Pattern Recognition*, 2008.

[51] I. Laptev and P. Pérez. Recognizing actions in movies. In *IEEE Conference on Computer Vision*, 2007.

[52] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *MultiMedia*, pages 357–360, 2007.

[53] A. Patron-Perez and I. Reid. A probabilistic framework for recognizing similar actions using spatio-temporal features. In *British Machine and Vision Conference*, 2007.

[54] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *IEEE Conference on Computer Vision and Pattern Recognition (Workshop)*, 2010.

[55] A. Vieira, E. Nascimento, G. Oliveira, G. Liu, and M. Campos. Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 252–259, 2012.

[56] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *European Conference on Computer Vision*, pages 872–885, 2012.

[57] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *IEEE Conference on Computer Vision (Workshop)*, 2011.

[58] D. Wu, F. Zhu, and L. Shao. One shot learning gesture recognition from rgbd images. In *IEEE Conference on Computer Vision and Pattern Recognition (Workshop)*, pages 7–12, 2012.

[59] J. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *IEEE Computer Vision and Pattern Recognition*, pages 716–723, 2013.

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

[61] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Fifth Berkeley Symposium on Mathematrical Statistics and Probability*, pages 281–297. University of California Press, 1967.

[62] A. Jalal, M. Z. Uddin, J. T. Kim, and T.S. Kim. Daily human activity recognition using depth silhouettes and r transformation for smart home. In *Conference on Toward Useful Services for Elderly and People with Disabilities: Smart Homes and Health Telematics*, pages 25–32, 2011.

[63] S. Fanello, L. Gori, G. Metta, and F. Odone. Keep it simple and sparse: Real-time action recognition. *Machine Learning Research*, 14:2617–2640, 2013.

[64] G. Johansson. Visual motion perception. *Scientific American*, 1975.

[65] M. Ye, X. Wang, R. Yang, R. Ren, and L. Pollefeys. Accurate 3d pose estimation from a single depth image. In *IEEE Conference on Computer Vision*, pages 731–738, 2011.

[66] C. Chen, K. Liu, and N. Kethtarnavaz. Real-time human action recognition based on depth motion maps. *Real-Time Image Processing*, 2013.

[67] F. Lv and R. Nevatia. Recognition and segmentation of 3-d human action using hmm and multi-class adabost. In *European Conference on Computer Vision*, pages 359–372, 2006.

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

[69] Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application for boosting. *Computer and System Science*, 55(1), 1997.

[70] L. Xia, C. Chen, and C. Aggarwal. View invarient human action recognition using histograms of 3d joints. In *Workshop on Human Activity Understanding from 3D Data*, pages 20–27, 2012.

[71] H.S. Koppula, R. Gupta, and R. Saxena. Human activity learning using affordances from rgb-d videos. In *Computing Research Repository*, 2012.

[72] John Darby, Baihua Li, Ryan Cunningham, and Nicholas Costen. Object localisation via action recognition. In *IEEE Conference on Pattern Recognition*, pages 817 –820, Tsukuba, Japan, 2012.

[73] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Workshop on Human Activity Understanding from 3D Data*, pages 14–19, 2012.

[74] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *Conference on Robotics and Automation*, 2012.

[75] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino. Re-identification with rgb-d sensors. In *First International Workshop on Re-Identification*, 2012.

[76] S. Fothergill, H. M. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *CHI*, pages 1737–1746, 2012.

[77] Home Office. Imagery library for intelligent detection systems. *The i-Lids User Guide*, 2011.

[78] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936, 2009.

[79] Carnegie Mellon University Multimodal Activity Dataset. The data used in this paper was obtained from kitchen.cs.cmu.edu and the data collection was funded in part by the national science foundation under grant no. eeec-0540865.

[80] G. Pons-Moll, A. Baak, T. Helten, M. Müller, H. P. Seidel, and B. Rosenhahn. Multisensor-fusion for 3d full-body human motion capture. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 663–670, 2010.

[81] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 36, pages 1325–1339, 2014.

[82] L. Ballan, A. Taneja, J. Gall, L. Van Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision*, pages 640–653, 2012.

[83] L. Sigal, A. O. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. 2006.

[84] N.P. Van Der Aa, X. Luo, G.J. Giezeman, R.T. Tan, and R.C. Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *IEEE Conference on Computer Vision (Workshop)*, pages 1264–1269, 2011.

[85] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber. Documentation mocap database HDM05. Technical Report CG-2007-2, Universität Bonn, 2007.

[86] M. Tenorth, J. Bandouch, and M. Beetz. The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *IEEE Conference on Computer Vision (Workshop)*, 2009.

[87] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. In *AAAI workshop on Pattern, Activity and Intent Recognition*, 2011.

[88] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. In *Robotics Research*, volume 32, pages 951–970, 2013.

[89] Bloom, V and Makris, D and Argyriou, V. G3di: A gaming interaction dataset with a real time detection and evaluation framework. In *European Conference on Computer Vision Workshop*, 2014.

[90] B. Kwolek and M. Kepski. Human fall detection on embedded platform using depth maps and wireless accelerometer. *Computer Methods and Programs in Biomedicine*, 117(3):489–501, 2014.

[91] A. Yao, J. Gall, G. Fanelli, and L. Van Gool. Does human action recognition benefit from pose estimation? In *British Machine and Vision Conference*, 2011.

[92] C. Halit and T. Capin. Multiscale motion saliency for keyframe extraction from motion capture sequences. *Computer Animation and Virtual Worlds*, 22(1), 2010.

[93] S. Sempena and N. Maulidevi. Human action recognition using dynamic time warping. In *International Conference on Electrical Engineering and Informatics*, pages 1–5, 2011.

[94] A. Yao, J. Gall, and L. Van Gool. Coupled action recognition and pose estimation from multiple views. *Computer Vision*, 100(1):16–37, 2012.

[95] S.Z. Masood, C. Ellis, A. Nagaraja, M.F. Tappen, J. LaViola, and R. Sukthankar. Measuring and reducing observed latency when recognizing actions. In *IEEE Conference on Computer Vision (Workshop)*, pages 499–504, 2000.

[96] M. Müller, T Roöder, and M. Clausen. Efficent content-based retrieval of motion capture data. *ACM Transactions on Graphics*, 24:677–685, 2005.

[97] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Conference on Computer Vision and Pattern Recognition (Workshop)*, pages 28–35, 2012.

[98] Y. Sheikh, M. Sheikh, and M. Shah. Exploring the space of a human action. In *IEEE Conference on Computer Vision*, pages 144–149, 2005.

[99] J. Sung, C. Ponce, B. Selman, and A. Sexena. Understanding human activity detection from rgbd images. In *PAIR*, pages 842–849, 2011.

[100] S. Bhattacharya, B. Czejdo, and N. Perez. Gesture classification with machine learning using kinect sensor data. In *Emerging Applications of Information Technology*, pages 348–351, December 2012.

[101] L. Piyathilaka and S. Kodagoda. Guassian mixture based hmm for human daily activity recognition using 3d skeleton features. In *International Conference on Industrial Electronics and Application*, pages 567–572, 2013.

[102] A. W. Vieira, T. Lewiner, W. R. Schwartz, and M. Campos. Distance matrices as invariant features for classifying mocap data. In *IEEE Conference on Pattern Recognition*, pages 2935–2937, Tsukuba, Japan, November 2012.

[103] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou. A real-time system for motion retrieval and interpretation. *Pattern Recognition Letters*, 34, 2013.

[104] A. Sinha and K. Chakravarty. Pose based person identification using kinect. In *IEEE Conference on Systems, Man and Cybernetics*, pages 497–503, Oct 2013.

[105] I. Kapsouras and N. Nikolaidis. Person identity recognition on motion capture data using multiple actions. *Machine Vision and Applications*, 2015.

[106] D. Kastaniotis, I. Theodorakopoulos, G. Economou, and S. Fotopoulos. Gait-based gender recognition using pose information for real time applications. In *DSP*, July 2013.

[107] A.M. Khan, Y. K. Lee, and T. S. Kim. Accelerometer signal-based human activity recogniting using augmented autoregressive model coefficents and artifical neural nets. In *IEEE EMBS*, pages 5172–5175, 2008.

[108] J. E. Condrom and K. D. Hill. Reliability and validity of a dualtask force platform assessment of balance performance: Effect of age, balance impairment, and cognitive task. *American Geriatrics Society*, 50(1), 2002.

[109] Mark Schmeler, richard Schein, Michael McCure, and Kendra Betz. Telerehabilitation clinical and vocational applications for assistive technology: Research, opportunities, and challenges. *International Journal of Telerehabilitation*, 1(1):59 – 72, Fall 2009.

[110] A. Kargar, A. Mollahosseini, T. Struemph, W. Pace, R. Nielsen, and M. Mahoor. Automatic measurement of physical mobility in get-up-and-go test using kinect sensor. In *IEEE Conference on Engineering in Medicine and Biology Society*, 2014.

[111] N. Vernadakis, V. Derri, E. Tsitskari, and P. Antoniou. The effect of xbox kinect intervention on balance ability for previously injured young competitive male athletes: A preliminary study. *Physical Therapy in Sport*, 15(3):148–155, 2014.

[112] S. Gauthier and A.M. Cretu. Human movement quantification using kinect for in-home physical exercise monitoring. In *IEEE Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications*, 2014.

[113] E. Dolatabadi, B. Taati, G. S. Parra-Dominguez, and A. Mihailidis. A markerless motion tracking approach to understand changes in gait and balance: A case study.

In *Rehabilitation Engineering and Assistive Technology Society of North America*, 2013.

[114] A. Gonzalez, P. Fraisse, and M. Hayashibe. A personalized balance measurement for home-based rehabilitation. In *IEEE Conference on Neural Engineering*, pages 711–714, 2015.

[115] J. Hernandez, R. Cabido, A.S. Montemayor, and J.J. Pantrigo. Human activity recognition based on kinematic features. *Expert Systems*, 2013.

[116] B. Lange, Chien-Yen Chang, E. Suma, B. Newman, AS. Rizzo, and M. Bolas. Development and evaluation of low cost game-based balance rehabilitation tool using the microsoft kinect sensor. In *International Conference of Engineering in Medicine and Biology Society*, pages 1831–1834, Aug 2011.

[117] V Bloom, D Makris, and V Argyriou. Clustered spatio-temporal manifolds for online recognition. In *IEEE Conference on Pattern Recognition*, pages 3963–3968, August 2014.

[118] Feng Liu, Yueting Zhuang, Fei Wu, and Yunhe Pan. 3d motion retrieval with motion index tree. *Comput. Vis. Image Underst.*, 92(2-3):265–284, 2003.

[119] M. Reyes, G. Domingguez, and S. Escalera. Featureweighting in dynamic time-warping for gesture recognition in depth data. In *IEEE Conference on Computer Vision (Workshop)*, pages 1182–1188, 2011.

[120] M. Raptis, D. Kirovski, and H. Hopes. Real-time classification of dance gestures from skeleton animation. In *SIGGRAPH*, pages 147–156, 2011.

[121] H. Pazhoumand-Dar, C. P. Lam, and M. Masek. Joint movement similarities for robust 3d action recognition using skeletal data. *Visual Communication and Image Representation*, 30:10–21, 2015.

[122] Y. Du, W. Wang, and L. Wang. Hierarchial recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1110–1118, 2015.

[123] L. Han, X. Wu, W. Liang, G. Hon, and Y. Yia. Discriminative human action recognition in the learned hierarchial manifold space. *Image and Vision Computing*, 17: 836–949, 2010.

[124] J. González, J. Varona, F. X. Roca, and J. J. Villanueva. Automatic keyframing of human actions for computer animation. In *Pattern Recognition and Image Analysis*, volume 2652, pages 287–296, 2003.

[125] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou. Human actions recognition from streamed motion capture. In *IEEE Conference on Pattern Recognition*, pages 3807–3810, 2012.

[126] T. T. Thanh, F. Chen, K. Kotani, and H. B. Le. Extraction of discriminative patterns from skeleton sequences for human action recognition. In *IEEE Conference on Computing and Communication Technologies*, pages 1–6, 2012.

[127] F. Cary, O. Postolache, and P. Silva Girao. Kinect based system and artificial neural networks classifiers for physiotherapy assessment. In *MeMeA*, June 2014.

[128] A. Amini Maghsoud Bigy, K. Banitsas, A. Badii, and J. Cosmas. Recognition of postures and freezing of gait in parkinson's disease patients using microsoft kinect sensor. In *IEEE Conference on Neural Engineering*, pages 731–734, 2015.

[129] M. Müller and T Roöder. Motion templates for automatic classification and retrieval of motion capture data. In *ACM SIGGAPH/Eurographics symposium on computer animation*, pages 137–146, 2006.

[130] F. Sebastian Grassia. Practical parameterization of rotations using the exponential map. *Graphics Tools*, 3(3):29–48, 1998.

[131] U.M. Frayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery*. AAAI/MIT Press, 1996.

[132] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):881–892, July 2002.

[133] B. DeCann, A. Ross, and M. Culp. On clustering human gait patterns. In *IEEE Conference on Pattern Recognition*, 2014.

[134] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in data set via the gap statistic. *Research Statistics*, pages 411–423, 2001.

[135] R. Melfi, S. Kondra, and A. Petrosino. Human activity modeling by spatio temporal textural appearance. *Pattern Recognition Letters*, 2013.

[136] F. Zhou, F. de la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2013.

[137] MichaelJ. B. and C. J.Dennis. A variable-selection heuristic for k-means clustering. *Psychometrika*, 66(2):249–270, 2001.

[138] D. Ketchen and C.L. Shook. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Management Journal*, 17 (6):441–458, 1996.

[139] D. Leightley, J. Darby, B. Li, J. S. Mcphee, and M. H. Yap. Human activity recognition for physical rehabilitation. In *IEEE Conference on Systems, Man and Cybernetics*, Manchester, UK, Oct 2013.

[140] M. Alexa and W. Müller. Representing animations by principal components. *Computer Graphics Forum*, 19(3):411–426, 2000.

[141] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273 – 297, September 1995.

[142] L. Zhang, J. C. Hsieh, T. T. Ting, Y. C. Huang, Y. C. Ho, and L. K. Ku. A kinect based golf swing score and grade system using gmm and svm. In *CISP*, pages 711–715, 16-18 Oct 2012.

[143] C. W. Hsu, C. C. Chang, and C. J. Lin. A practical guide to support vector classification, April 2010. URL `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

[144] L. Breiman. Random forests. *Machine Learning*, 25(1):5 – 32, 2001.

[145] Z. Wei and Y. Xu. The human behavior recognition based on improved r transform and discriminative random fields model. *Computational Information Systems*, 10 (6):2359–2367, 2014.

[146] L. Breiman. Bagging predictors. *Machine Learning*, 26:123–140, 1996.

[147] Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

[148] T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:832–844, 1998.

[149] A. Miller. Gait event detection using a multilayer neural network. *Gait and Posture*, 29(4):542–545, 2009.

[150] K.N. Tran, I.A. Kakadiaris, and S.K. Shah. Part-based motion descriptor image for human action recognition. *Pattern Recognition*, 45(7):2562–2572, 2012.

[151] L. A. Shalabi, Z. Shaaban, and B. Kasasbeh. Data mining: A preprocessing engine. *Computer Science*, 2(9):735–739, 2006.

[152] M. Hoai, Z. Z. Lan, and F. de la Torre. Joint segmentation and classification of human actions in video. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3265–3272, June 2011.

[153] M. A. Gowayyed, M. Torki, M. E. Hussein, and M. El-Saban. Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition. In *IEEE Conference on Artificial Intelligence*, 2013.

[154] E. Pekalska and E. P. W. Duin. The dissimilarity representation for pattern recognition. *World Scientific*, 2005.

[155] Y. Lu, L. Wang, R. Hartley, H. Li, and D. Xu. Gradual sampling and mutual information maximisation for markerless motion capture. In *Lecturer Notes in Computer Science*, volume 6393, pages 554–565, 2010.

[156] M. Müller, A. Baak, and H.P. Seidel. Efficient and robust annotation of motion capture data. In *ACM SIGGRAPH*, pages 17–26, 2009.

[157] O. Patsadu, C. Nukoolkit, and B. Watanapa. Human gesture recognition using kinect camera. In *JCSSE*, pages 28–32, June 2012.

[158] Kinect for Windows Software Development Kit. http://www.microsoft.com/en-us/kinectforwindows. URL `http://www.microsoft.com/en-us/kinectforwindows`.

[159] A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, Feb 2010.

[160] T. Y. Lin, C. H. Hsieh, and J. D. Lee. A kinect-based system for physical rehabilitation: Utilizing tai chi exercises to improve movement disorders in patients with balance ability. In *AMS*, pages 149–153, July 2013.

[161] I. Nitze, U. Schulthess, and H. Asche. Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. In *GEOBIA*, pages 35 – 40, Rio de Janeiro, 2012.

[162] A. Statnikov, L. Wang, and C. F. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Infomatics*, 9, June 2008.

[163] Y. Tang, S. Krasser, Y. He, W. Yang, and D. Alperovitch. Support vector machines and random forests modeling for spam senders behavior analysis. In *GLOBECOM*, pages 2174–2178, 2008.

[164] D. M. Brennan, P. Lum, G. Uswatte, E. Taub, B. Gilmmore, and Joydip Barman. A telerehabilitation platform for home-based automated therapy of arm function. In *IEEE Conference on Telle-Rehabilitation*, pages 1819 – 1822, 2011.

[165] H. Aoki, M. Miyazaki, H. Nakamura, R. Furukawa, R. Sagawa, and H. Kawasaki. Non-contact respiration measurement using structured light 3-d sensor. In *SICE Annual Conference*, pages 614 – 618, 2012.

[166] M. E. Tinetti. Performance-oriented assessment of mobility problems in elderly patients. *American Geriatrics Society*, 34(2):119–126, 1986.

[167] M. S. Cameirão, S. Bermúdez, E. D. Oller, and P. F.M.J. Verschure. Neurorehabilitation using the virtual reality based rehabilitation gaming system: methodology, design, psychometrics, usability and validation. *NeuroEngineering and Rehabilitation*, 7(48), 2010.

[168] A. Gonzalez, M. Hayashibe, and P. Fraisse. Estimation of the center of mass with kinect and wii balance board. In *Conference on Intelligent Robots and Systems*, pages 1023–1028, 2012.

[169] O. Celiktutan, A.C. Burak, C. Wolf, and B. Sankur. Graph-based analysis of physical exercise actions. In *ACM Multimedia Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, 2013.

[170] R. Clark, Y. H. Pua, C. Oliveira, K. Bower, S. Thilarajah, R. McGaw, K. Hasanki, and B. Mentiplay. Reliability and concurrent validity of the microsoft kinect v2 for assessment of standing balance and postural control. *Gait and Posture*, 42(2): 210–213, 2015.

[171] B. F. Mentiplay, L. G. Perraton, K. J. Bower, Y. H. Pua, R. McGaw, S. Heywood, and R. A. Clark. Gait assessment using the microsoft xbox one kinect: Concurrent validity and inter-day reliability of spatiotemporal and kinematic variables. *Biomechanics*, 48(10):2166–70, 2015.

[172] J. Guralnik, E. Simonsick, L. Ferrucci, R. Glynn, L. Berkman, D. Blazer, P. Scherr, and R. Wallace. A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home admission. *Gerontology*, 49(2):85 – 93, March 1994.

[173] D. Podsiadlo and S. Richardson. The timed up and go: a test of basic functional mobility for frail elderly persons. *American Geriatrics Society*, 39(2):142–148, 1991.

[174] P. L. Enright. The six-minute walk test. *Respiratory Care*, 48(8):783–788, 2003.

[175] Y. Yang, F. Pu, Y. Li, S. Li, Y. Fan, and D. Li. Reliability and validity of kinect rgb-d sensor for assessing standing balance. *Sensors*, pages 1633–1638, 2014.

[176] L. Zhou, Z. Lu, H. Leung, and L. Shang. Spatial temporal pyramid matching using temporal spare representation for human motion retrieval. *The Visual Computer*, 2014.

[177] G. Alankus, A. Lazar, M. May, and C. Kellecher. Towards customizable games for stroke rehabilitation. *SIGCHI Conference on Human Factors in Computing Systems*, pages 2113 – 2122, 2010.

[178] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke, and M. A. McBurnie. Frailty in

older adults: Evidence for a phenotype. *Gerontology: Biological Sciences*, 56(3): 808–813, 2001.

[179] D. M. Brennan, S. Mawson, and S. Brownsell. *Advanced Technologies in Rehabilitation*, volume 145, chapter Tele-rehabilitation: enabling the remote delivery of healthcare, rehabilitation and self management. IOS Press, June 2009.

[180] Paula Rego, Pedro Miguel Moreira, and Luis Paulo Reis. Serious games for rehabilitation: A survey and a classification towards a taxonomy. In *5th Iberian Conference Information Systems and Technologies (CISTI)*, pages 1 – 6, June 2010.

[181] V. Hatzitkai, I. G. Amiridis, and F. Arabatzi. Aging effects on postural responses to self-imposed balance perturbations. *Gait and Posture*, 22(3):250–257, 2005.

[182] M.J. Fuhrer. Subjectifying quality of life as a medical rehabilitation outcome. *Disability Rehabilitation*, 22(11):481–489, 2000.

[183] P. Gregory, J. Alexander, and J. Satinksky. Clinical telerehabilitation: Applications for physiatrists. *American Academy of Physical Medicine and Rehabilitation*, 3(7):647 –656, July 2011.

[184] General Practitioners. Fit for frailty: Consensus best practice guidance for the care of older people living with frailty in community and outpatient settings. Technical report, British Geriatrics Society, 2014.

[185] B. Galna, D. Jackson, G. Schofield, R. McNaney, M. Webster, G. Barry, D. Mhiripiri, M. Balaam, P. Olivier, and L. Rochester. Retraining function in people with parkinson's disease using the microsoft kinect: game design and pilot testing. *NeuroEngineering and Rehabilitation*, 11(60), 2014.

[186] D. Leightley, B. Li, M. H. Yap, J.S. McPhee, and J. Darby. Exemplar-based human action recognition with template matching from a stream of motion capture. In *Lecturer Notes in Computer Science*, October 2014.

[187] M. S. Cameirao, S. Bermudez, E. D. Oller, and P. F.M.J. Verschure. The rehabilitation gaming system: a review. *Studies in health technology and informatics*, 145:65– 83, 2009.

[188] M. Golomb, B. McDonald, S. Warden, J. Yonkman, A. Saykin, B. Shirley, M. Huber, B. Rabin, M. AdbelBaky, M. Nwosu, M. Barkat-Masih, and G. Burdea. In-home virtual reality videogame telerehabilitation in adolescents with hemiplegic cerebral palsy. In *Virtual Rehabilitation International Conference*, 91, Tel Aviv, Isreal, June 29 - 2 July 2009. Virtual Rehabilitation International Conference.

[189] J. Bae and M. Tomizuka. A tele-monitoring system for gait rehabilitation with an inertial measurement unit and shoe-type ground reaction force sensor. *Mechatronics*, 2013.

[190] E. Gianaria, N. Balossino, M. Grangetto, and M. Lucenteforte. Gait characterization using dynamic skeleton acquisition. In *International Workshop on Multimedia Signal Processing*, pages 440–445, Sept 2013.

[191] B. Dikovski, G. Madjarov, and D. Gjorgjevikj. Evaluation of different feature sets for gait recognition using skeletal data from kinect. In *International Convention on Information and Communication Technology, Electronics and Microelectronics*, pages 1304–1308, 2014.

[192] M. Lewandowski, J. Martinez-Del-Rincon, D. Makris, and J. C Nebel. Temporal extension of laplacian eigenmaps for unsupervised dimensionality reduction of time series. In *IEEE Conference on Pattern Recognition*, pages 161–164, Aug 2010.

[193] R. N. Baumgartner, K. M. Koehler, D. Gallagher, L. Romero, S. B. Heymsfield, R. R. Ross, P. J. Garry, and R. F. Lindeman. Epidemiology of sarcopenia among the elderly in new mexico. *American Journal of Epidemiology*, 147(8):755–763, 1998.

[194] R. M. Dodds, H. E. Syddall, R. Cooper, M. Benzeval, I. J. Deary, E. M. Dennison, G. Der, C. R. Gale, H. M. Inskip, C. Jagger, T. B. Kirkwood, D. A. Lawlor, S. M. Robinson, J. M. Starr, A. Steptoe, K. Tilling, D. Kuh, C. Cooper, and A. A. Sayer. Grip strength across the life course: Normative data from twelve british studies. *PLoS One*, 2014.

[195] Y. Freund and R. E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 15(4):771–780, 1999.

[196] Hastie T., R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Spriner, New York, 2008.

[197] Seiffert C., T. Khoshgoftaar, J. Hulse, and A. Napolitano. Rusboost: Improving clasification performance when training data is skewed. In *IEEE Conference on Pattern Recognition*, pages 1–4, 2008.

[198] M. Warmuth, J. Liao, and G. Ratsch. Totally corrective boosting algorithms that maximize the margin. In *Proc. 23rd Int'l. Conf. on Machine Learning*, pages 1001–1008, 2006.