REAL-TIME PERSON RE-IDENTIFICATION FOR INTERACTIVE ENVIRONMENTS

By

Mohd Hafizuddin Mohd Yusof

A thesis submitted to

The University of Birmingham

For the degree of

DOCTOR OF PHILOSOPHY (Ph.D.)

Digital Humanities Hub
Ironbridge International Institute for Cultural Heritage
School of History and Cultures
The University of Birmingham
March 2016

Abstract

The work presented in this thesis was motivated by a vision of the future in which intelligent environments in public spaces such as, galleries and museums, deliver useful and personalised services to people via natural interaction, that is, without the need for people to provide explicit instructions via tangible interfaces. Delivering the right services to the right people requires a means of biometrically identifying individuals and then re-identifying them as they move freely through the environment. Delivering the service they desire requires sensing their context, for example, sensing their location or proximity to resources.

This thesis presents both a context-aware system and a person re-identification method. A tabletop display was designed and prototyped with an infrared person-sensing context function. In experimental evaluation it exhibited tracking performance comparable to other more complex systems. A real-time, viewpoint invariant, person re-identification method is proposed based on a novel set of Viewpoint Invariant Multi-modal (ViMM) feature descriptor collected from depth-sensing cameras. The method uses colour and a combination of anthropometric properties logged as a function of body orientation. A neural network classifier is used to perform re-identification. The method was tested using an experimentally acquired dataset comprising sixty-four people performing fourteen individual and four group choreographed and free-form walking activities. The activities ranged from turning on the spot to free-form movement in random directions. This dataset was used to test the classification performance of ViMM feature descriptor. The results show that the ViMM descriptor possesses strong discriminant properties compared with its subsets and that it achieves viewpoint invariant re-identification in real-time. Single-frame re-identification performance of 86.3% rank-1 classification and nAUC of 99.4% was achieved, increasing to 92.4% rank-1

classification and nAUC of 99.7% when multiple frames were combined. Application of the method in multi-person re-identification scenarios is explored and its integrated application with the person-sensing tabletop is also explored as a context-aware intelligent environment scenario. In addition, prototype context-provider software applications are developed and are used in the generation of presented results.

Dedication

To my family.

Acknowledgments

I am deeply thankful to my main supervisor Dr Eugene Ch'ng for giving me the opportunity to perform this research under his supervision at the Digital Humanities Hub of the University of Birmingham. The hub houses a digital prototyping facility called Chowen and Garfield Weston Foundation Digital Prototyping Hall, a unique resource for user testing of research outputs that is well equipped with a number of large modern multi-touch displays and tracking sensors. He always encouraged me to read, explore and experiment and this has gained me invaluable knowledge and experience. I am very grateful for Dr Ch'ng's valuable continuous support and guidance that have made this research possible.

I would like to thank my second supervisor Dr Tim Collins for his help and guidance in keeping my research direction and focus. His technical expertise in digital signal processing and programming has helped me get through many stages of the research. His dedication and interest in this research have kept me motivated to produce the highest quality output possible.

My thanks must also go to Dr Sandra I. Woolley whose advice has been critical to the content and structure of this thesis and also to the philosophical aspects of this project. She never failed to compliment the work, for this has kept my spirit high at all times. She is the kindest person I have ever known and I am very grateful and honoured to have her as one of my supervisors.

I also want to thank my other supervisor Professor Chris Baber for his valuable help and insights during the first year of my study, for letting me have his students' theses to help me learn more about PhD research. This has helped me in various stages during my thesis writing period.

I would like to thank my colleagues and friends from the Digital Humanities Hub – namely Gido Hakvoort, Simon Hartley, Andrew Lewis, Henry Chapman, Richard Clay, Chris Creed,

John Sear, Lara Ratnaraja, Nadia Wood, Joseph Sivell, Juliet Chikore, Anthony Hughes, Johannah Latchem, Louise Woodall, Vanessa Rouse, Vincent Gaffney, Philip Murgatroyd, Eamonn Baldwin, for the kind atmosphere, good chat and discussions, the support and the great fun. My thanks also goes to Henry Chapman from the Digital Humanities Hub and Sue Bowen from the College of Arts and Law's finance office for all the help with purchasing of research equipment.

I would like to also thank my friends, colleagues and the many students of the University of Birmingham who have given their time to help me collect recordings of human appearance in various activities, without these recordings this research would not have been possible.

Finally, special thanks goes to my beloved wife for her unconditional love, undying support and understanding, my kids for their always cute actions that have helped in keeping me cheerful, and my family in Malaysia for their encouragement and support before and during my academic life.

Table of Contents

List of Abbreviations

List of Figures

List of Tables

CHAPTER 1:

Introduction

People interact with computers through various means. The technologies supporting these interactions are evolving and expanding to include more natural forms. As shown in Figure 1.1, tangible means of interaction such as keyboards and mice, and touch-screens, are being supplemented by less tangible means such as voice commands, and, more recently, hand gestures, head and eye movements and whole body actions.

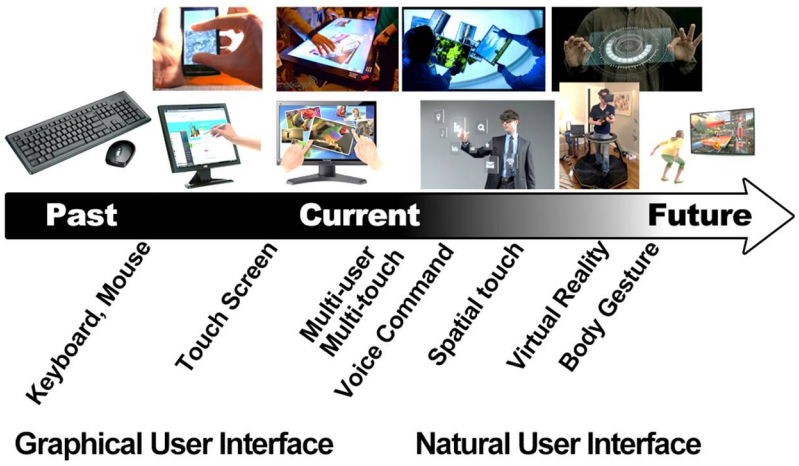


Figure 1.1 The trend in development direction of user interfaces. Adapted from (Ahn et al., 2008)

Gesture-based inputs are made possible with the advent of more functional sensors such as the Leap Motion hand/finger sensor[1], the Microsoft Kinect depth-sensing camera[2] and the radar-based sensor prototyped by Google's Project Soli[3]. Voice recognition technologies have improved over time and have become more popular, finding applications in intelligent personal assistant applications such as Apple's Siri, Google's Now, Microsoft's Cortana, Amazon's

[1] https://www.leapmotion.com/
[2] https://dev.windows.com/en-us/kinect
[3] https://www.google.com/atap/project-soli/

Echo (Alexa), and Facebook's M. Sensors used by gesture and voice sensing technologies are either attached to the person or placed externally while they continuously sense for triggers.

The trend toward more natural human-computer interaction is accompanied by ambitions for the delivery of personalised services. To achieve these, computers must be able to robustly identify individuals as well as determine their needs depending on the context of their interaction at the time. Identifying individuals in uncontrolled settings is a challenging task.

1.1 Background and Motivation

Research in computer vision involving depth sensing cameras has advanced in recent times. The advancements in computer systems and peripherals such as increased computing power and more efficient and sophisticated sensors have given ways for more effective pattern recognition and machine learning techniques to be used in real time. Consequently more advanced and sophisticated human computer interfaces can be developed.

Intelligent environments are evolving, for example, smart homes with embedded sensing and smart cars with in-vehicle ambient intelligence by Rakotonirainy and Tay (2004). A vision of the future sees ambient intelligence integrated into our daily lives where devices help to support us in everyday activities in an easy and natural way. Providing personalised services in these environments necessarily requires some form of identification mechanism. There are two general approaches to identification, 1) an active, voluntary, user-initiated approach, for example, when a person provides their thumb print on a scanner to gain access to a room, and 2) a passive, involuntary, system-initiated approach such as when a person approaches a door and the system immediately detects the person, performs identification, recognises him/her as a registered user and opens the door.

Biometrics is the term used to describe the measurement and analysis of human attributes; the main application of biometrics being for person identification. A distinction is made between "soft" and "hard" biometric types by Zewail et al. (2004). Hard biometrics include physiological attributes such as fingerprints and retina patterns, and, to a lesser extent, behavioural attributes such as voice that can strongly, if not uniquely, distinguish between individuals. Soft biometrics are the broad set of descriptors comprising physical, behavioural and adhered characteristics that, collected together, can discriminate between individuals but, individually, cannot. For example, physical characteristics could include height and hair colour, behavioural characteristics could include typing rhythm, and adhered characteristics could include tattoos and clothing.

1.1.1 A Future Smart Museum Scenario

This thesis envisions a scenario such as described below.

A museum visitor walks to a tabletop display in a museum. He is greeted with a welcoming message and a "wizard" tells him what is on the tabletop display and what he can do on the tabletop. An image gallery application running on the display allows him to digitally collect objects that he likes into a personal virtual folder that shall persist throughout the visit. These objects can be retrieved for further inspection at any time on any of the displays in the museums. He collects a few objects for later inspection as the area near the tabletop is currently crowded with other people. He then moves on to another exhibit that displays physical artefacts. There is a sign beside the physical artefact that says "Touch here to collect this object". He touches the sign. After a while he proceeds to another exhibit with a wall-mounted display. The display greets and welcomes him and provides him with a brief story about the application on the wall display. The display also suggests that he visit other exhibits that he has missed. But he chooses

to skip that. There is also an icon on the display that says “My Folder” right in front of where he is standing. This folder contains the digital objects that he has collected so far. He decides to retrieve one of the objects and inspects it in more detail. He performs a pinch-to-zoom gesture to enlarge the digital object, revealing very fine, high resolution, detail. He is done now with the object and turns away from the display and move onto another exhibit. The folder icon and images that are left open disappear automatically. After spending a few hours in the museum, he goes on to a three-dimensional (3-D) printed souvenir station that has a 3-D printer with a touch-screen user interface and a payment terminal. The screen greets him and presents a list of objects in his virtual folder. He chooses an item, sends it to the 3-D printer and makes a payment. Once the printing is complete, he collects his souvenir and leaves the museum taking with him sweet memories. When the museum closes for the day, a museum curator logs into the system and is given options of information such as maps of visitors’ whereabouts during their visits, duration spent at each section or exhibit, which exhibits get skipped a lot, and what objects are “collected” most. The curator can even drill down to details of activity at an individual level. With this information, the curator will be able to re-design and implement the exhibition that offers better educational value and enjoyable experience to visitors.

The scenario above demonstrates how an intelligent environment can use natural user interaction to interact with people. People do not need to wear any devices or initiate an action to interact with the environment. Valli and Linari (2008) emphasised that interaction should be easy, natural and attractive for everyone. The key component in making the above scenario a reality is to employ effective person identification as part of the intelligent environment. Person identification systems for intelligent environments have previously relied on mechanisms that require people to adhere to certain rules such as looking at a camera for face recognition, carrying a device such as Radio-frequency Identification (RFID) enabled device, putting their

finger print on a fingerprint scanner, etc. These mechanisms of identifying people are not natural in that they require people to first initiate the interaction process. This can hinder participation from people as they are not always active or willing users, they can simply walk by and passively enjoy the encounter (Valli and Linari, 2008). Person identification method based on a natural processes are much more desirable for the scenario presented above. A person once identified at one place and will need to be re-identified again and again as they move around the environment and this identification and re-identification needs to be accurate for the environment to function correctly.

1.2 Person Identification and Re-identification

The means of robustly identifying individuals is a challenge which spans millennia. From the use of seals and signatures through to the use of identification cards, PIN numbers and passwords, the approaches have evolved with the technologies of the day; motivated by the means of ready access to privileged services and also informed by repercussions of mistaken identity.

Biometric person identification is the process of establishing the identity of an individual by comparing their biometric features such as facial features, gait, fingerprint, hand geometry and iris pattern to those registered in a database. Person re-identification is the process of recognising a person that has previously been observed, for example, at different locations, at different times and by different cameras.

1.2.1 Identification and Re-identification Applications

Person identification and re-identification are active research topics in computer vision and have wide potential applications in surveillance and security as well as in human-robot

interaction and in the personalisation of services in smart environments, for example, for interactive services in retail environments, museums, art galleries and public spaces.

i. Surveillance in security. For example, to detect if a person appears at a particular location at particular times (Bedagkar-Gala and Shah, 2014) and predict the location where the person is heading based on the walking direction.
ii. Retail or shopping (Gong et al., 2014b). For example, to provide useful information for improving customer service such as personalised product recommendation, and also data for shopping space management.
iii. Healthcare (Gong et al., 2014b). For example, to track patients when they move about in a hospital for caring and monitoring purposes (Ma et al., 2014b).
iv. Personalisation of services in smart home (Baltieri et al., 2010). For example, to track the whereabouts of a person in a house and activate personalised TV or radio station when one enters a living room, and offer coffee making service when a coffee drinker enters a kitchen.
v. Personalisation of services in public spaces such as museums (Schulman et al., 2008). For example, to provide personalised services to visitors such as personal guides providing narratives of an exhibition, and personal virtual collection, for more engaging experiences.
vi. Human-robot interaction (Bedagkar-Gala and Shah, 2014). For example, to provide more natural interaction similar to the ones existing between humans. Robots will be able to interact and communicate with humans by means of social behaviours and rules through speech, gestures, and facial expressions (Olivera, 2012). Other human-robot interaction application areas include Entertainment, Education, Field robotics, Home and companion robotics, Rehabilitation and Elder Care, Hospitality and Robot Assisted Therapy.

1.2.2 The Identification Process

In identification, a system typically tries to establish the identity of a person by collecting their biometric features and matching them to records in a database. Examples of biometric features used include facial features, gait, fingerprints, iris pattern, hand geometry and voice. The identification process typically involves three tasks as described by Apostoloff and Zisserman (2007) which are:

i. Detection – the area of detection depends on the biometric features which are being used. For example, in the case of facial recognition, detection is best performed when a frontal face view is available. For gait recognition, human body detection is needed before accurate features can be extracted from a video sequence of full body views. Other biometrics can require cooperation from users, for example, for achieving fingerprint detection with a fingerprint scanner and iris detection with an iris scanner.
ii. Tracking – tracking the area of interest is normally required for non-cooperative (passive) methods such as facial recognition and gait recognition, to allow extraction of features as the person moves.
iii. Identification – person recognition is the process of identifying a person by submitting features of the person to a pre-trained classifier which will then return the probabilities of identities belonging to the N top highest match. Typically the identity of the person is the one with the highest probability although there are strategies that take into account the results from previous frames before making the decision on the current frame. This is also called multi-shot identification.

1.2.3 Person Re-identification

Achieving person re-identification (Gong et al., 2014a) involves:

i. Processing of raw data and the extraction of features.
ii. Constructing a descriptor or representation, e.g. a histogram of features, capable of both describing and discriminating individuals.
iii. Matching observations of individuals in acquired images against a gallery of persons in another camera view by measuring the similarity between the images, or using some model-based matching procedure. A training stage to optimise the matching parameters may or may not be required depending on the matching strategy.

1.2.4 Challenges in Person Re-identification

Designing effective feature representations for humans is a challenging task. The features should, ideally, be illumination invariant, that is, robust to changes in lighting which can be caused by observation from different viewpoints, location of subject, and uneven lighting. Changing viewpoints can also make the appearance of a person differ, especially when the clothes they are wearing are not uniformly coloured. A person carrying a backpack appears different when observed from front and back. Other challenges are associated with occlusion, background clutter, and image quality/resolution. The use of RGB-D (depth sensing) cameras such as Microsoft's Kinect camera can assist in these challenges by making use of their depth data to provide automatic person detection and background removal via their Software Development Kits (SDK)s.

Research addressing the challenges in person re-identification have primarily focused on two main tasks (Gong et al., 2014a):

1) Designing stronger discriminant feature descriptors of a person that are less susceptible to factors such as viewpoint and lighting (Gray et al.,2007), (Farenzena et al., 2010), (Prosser et al., 2010).
2) Designing approaches for effective learning methods so that parameter optimisation of a re-identification model (Wei-Shi et al., 2013) can be achieved.

1.3 Research Questions

This study addressed the following research questions.

i. Can the performance of previously proposed human aware multi-touch tabletop systems be achieved at much lower cost (i.e. low in construction price and computational power) using a reduced number of sensors?
ii. Can person re-identification accuracy be improved by supplementing clothing appearance descriptors with 3-D anthropometric parameters extracted from depth data, using RGB-D cameras, in unconstrained settings?
    a. Can a person re-identification be performed accurately when individuals are observed from unconstrained viewpoints?
    b. What combination of features can best improve re-identification?

1.4 Thesis Statement

Personalisation of services through natural user interaction in smart environments can be made possible with the use of an accurate appearance based person re-identification system.

1.5 Contribution

This thesis deals with the problem of person re-identification by employing methods identified in the taxonomy diagram in Figure 1.2 below.

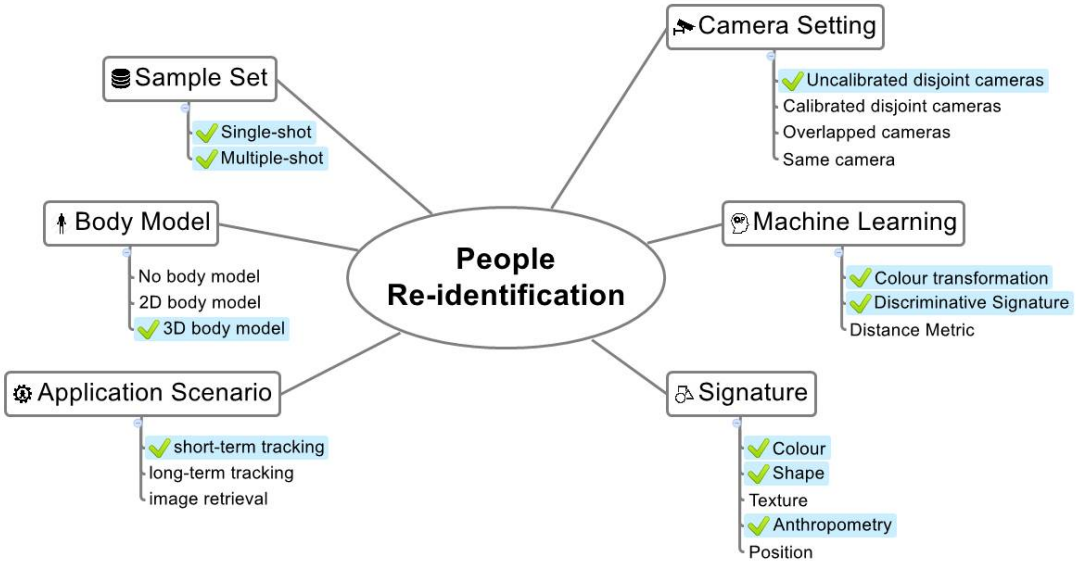


Figure 1.2 Multidimensional taxonomy for people re-identification problem.
(Diagram reproduced from Vezzani et al. (2013)). Ticked items are addressed in this thesis.

The main contribution of this thesis is the proposed Viewpoint Invariant Multi-Modal (ViMM) feature descriptor for person re-identification. The ViMM feature descriptor possesses strong discriminant properties and achieves viewpoint invariant re-identification in real-time. ViMM in itself makes new contributions to the re-identification field, including, 1) providing a robust method to estimate body orientation by applying ellipse fitting to partial depth data at the shoulder level, 2) providing a complete feature descriptor ViMM which uses multi-modal colour and anthropometric properties which are logged as a function of body orientation, and enable person re-identification from unconstrained viewpoints, 3) creation of a new RGB-D dataset of sixty-four people performing sixteen walking activities acquired using the higher-resolution Microsoft Kinect version 2 RGB-D camera, 4) comparisons of performance analysis of ViMM feature descriptor and its subsets comprising of anthropometric and colour features, and also comparisons with results from relevant methods from the literature, 5) an evaluation

of the ViMM performance at runtime via prototype development and demonstration, and classification results demonstrating strong discriminant properties of the ViMM descriptor, 6) an alternative configuration for a simple and inexpensive human-sensing multi-touch tabletop display.

1.6 Thesis Organisation

The thesis is organised as follows.

Chapter 2 reviews the literature relevant to natural interaction and person identification and re-identification. This chapter also discusses the research methodology that aimed to address the research questions described in the earlier section. The methodology is divided into two sections, describing processes in 1) designing and building human sensing system on a multi-touch tabletop, and 2) designing and building person re-identification system for a short-term scenario.

Chapter 3 presents the proposed context-aware human sensing system around multi-touch tabletops. Previous works on interactive displays are reviewed and state of the art systems are compared with the proposed system. The tracking system used is outlined in detail and technical setup and procedure are presented. Prototype applications are developed to demonstrate the system accuracy and functionality.

Chapter 4 contains the significant work of this thesis. This chapter starts with reviews of previous work on person re-identification covering 2-D and 3-D appearance based feature descriptors. Challenges faced by re-identification research are explained in detail. This chapter explores how the view invariant problem was tackled by proposing a set of multi-modal features, a combination of 2-D and 3-D soft biometric features called the ViMM feature

descriptor. In this work, a new 3-D dataset containing sixty-four people performing various walking activities was created and used to train and test classifiers built using ViMM descriptors and its subsets. The classification method and matching technique used are discussed. The experimental results follow and discussion on the results conclude the chapter.

Chapter 5 extends the work in Chapter 4 where the ViMM feature descriptor is tested with multi-person re-identification scenario. Experimental results are presented and challenges faced in multi-person re-identification are discussed in this chapter.

Chapter 6 proposes a context-aware system utilising person re-identification to work with context aware system utilizing a person's identity and location information to allow continuous tracking of the person in disjoint spaces. Two scenarios are given as test cases and system design for both scenarios are proposed.

Chapter 7 discusses the results of the experimental chapters and presents the thesis conclusions, contributions and recommendations for future work.

1.7 Publications

The following journal and conference paper have been produced as part of outcomes of this research:

1.7.1 Journal

i. Yusof, H., Collins, T., Ch'ng, E., and Woolley, S. (2016), "Viewpoint Invariant Multi-modal Person Re-identification Using RGB-D Cameras". Submitted to the journal of the IEEE Transactions on Consumer Electronics on 3rd January 2016.

1.7.2 Book Section

i. Yusof, H., Eugene Ch'ng, E., and Baber, C. (2014), "Human Sensing for Tabletop Entertainment System." *Context-Aware Systems and Applications* (pp. 283-292). Springer International Publishing, 2014.

CHAPTER 2:

Literature Review

Ambient intelligence (AmI) is the enabling of physical spaces to sense the presence of people and respond (Han et al., 2012), for example, by providing services and information to people in the environment while requiring minimal interactive effort (Chavira et al., 2004). Context awareness is an essential component when building an intelligent environment. The context-aware system has been defined by Anind Dey as a "system that uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task" (Dey, 2001). In this chapter the concept of natural interaction is introduced, followed by reviews of the literature relevant to human sensing systems and person identification and re-identification methods. The chapter concludes with a summary of the direction of the research work in this thesis.

2.1 Overview of Natural Interaction

Natural interaction has gained much attention recently. It is a concept of modern interaction that breaks away from traditional keyboard and mouse, replacing them with a "natural user interface" (NUI). There are several definitions of NUI, for example:

i. "*Natural User Interfaces are just a way of explaining the method you interact with machines. Some machines require tools, like a remote control, keyboard, or a mouse. People who specialize in designing natural user interfaces challenge themselves with designing methods of interacting with machines that require no tools other than the ones you were born with.*" – Ron George, Microsoft NUI/UX Designer.

ii. "*An NUI is a type of user interface that is designed to feel as natural as possible to the user. The goal of an NUI is to create seamless interaction between the human and machine, making the interface itself seem to disappear.*" (Tech Terms, n.d.)

iii. "*An NUI is an emerging paradigm shift in man machine interaction of computer interfaces to refer to a user interface that is effectively invisible, or becomes invisible with successive learned interactions, to its users. The word natural is used because most computer interfaces use artificial control devices whose operation has to be learned. An NUI relies on a user being able to carry out relatively natural motions, movements or gestures that they quickly discover to control the computer application or manipulate the on-screen content.*" (NUI-Group, n.d.)

NUIs based on voice recognition have received much attention lately especially in the category of personal digital assistants. It was revealed in the Apple Worldwide Developers Conference (WWDC) 2015 keynote presentation, that Apple's Siri receives over 1 billion requests per week. Unlike Apple's previously released mobile phones, the latest phone from Apple, iPhone 6s, has a built-in chip dedicated to voice command and can be activated via voice without the need to touch a button, even when the phone is in sleep mode. Amazon's Echo is a standalone artificial-intelligence cloud powered speaker, listed as the "most admired" digital assistant product by Adobe Social's report (Zdnet.com, 2015). Facebook's M (Wired.com, 2015) is a hybrid virtual assistant powered by artificial intelligence as well as humans, called "M trainers" intended to perform the "trickier judgment calls" and other tasks that software can't. At the time of the writing of this thesis, it is available within the Facebook's Messenger app in a limited beta form for selected locations in the US. Google Now by Google is available within the mobile application for Android and iOS, and Google Chrome web browser on personal computers. Google utilises its Search and Knowledge Graph project to recognise repeated user

actions and constructs more detailed results from the meaning and connections graph. Microsoft being another search giant and computer operating manufacturer, employed its Bing search technology into Cortana, its personal assistant which is available within its mobile and personal computer operating systems, as well as Android and iOS devices.

Research work in this field aims to design systems that improve the way people communicate (i.e. through gestures, voice commands, expressions and movement), and engage people in activities, while they naturally interact with each other and the environment. "*People don't need to wear any device or learn any instruction, interaction is intuitive*" (Valli and Linari, 2008).

Ahn et al. (2008) proposed an interactive home control system called Ubi-touch to display an NUI via Holo-screen (holographic transparent film), location and status of home appliances via rear projection, tracking user motion via vision cameras with infrared-pass filter, and to direct audio sound via sound-beam. Their implementation used a standard PC with ZigBee interface to control home appliances via wireless commands. This work demonstrates a next generation UI, i.e. holographic touch, based on natural interaction that moves away from traditional input devices such as mouse, keyboard and remote control, with the aim of achieving practicality and enjoyment through immersive emotions.

There has been a trend toward use of natural user interfaces for public displays. The central goal is to improve the quality of viewer experience by engaging them in an interactive exchanges that can lead to better understanding and promote excitement about the content. One of such works by Swartout et al. (2010) used two life-sized virtual humans acting as information agents to engage dialogues with visitors using natural language interaction (i.e. speech). Other work by Motta and Nedel (2013) used gesture based input (i.e. navigation, selection and manipulation of objects, as well as panning and zooming on the screen) as a means of interaction

with the public display. Kinect RGB-D cameras were used to capture gestures and a 55 inch LED TV was used as the display or interaction station.

2.2 Human-sensing Tabletop Displays and Context-Awareness

There has been an increasing number of research projects focusing on ambient intelligence (Wakkary et al., 2005) and context awareness (Toninelli et al., 2009) utilising a variety of technologies such as computer vision, infrared (IR) sensors, Radio Frequency Identification (RFID), and software agents. Ambient intelligence deals with the issue of how we can create context-aware, electronic environments which encourage the development of seamless human-computer interaction. Ambient intelligence makes use of these computer technologies combined with artificial intelligence, to respond to and reason about human actions and behaviours within the environment (Wakkary et al., 2005). Context-aware systems adapt their operations to the current context without explicit user intervention and thus aim at increasing usability and effectiveness by taking environmental context into account. These systems make use of external context factors as they provide useful data, such as location information. Furthermore, external attributes are easy to sense by using off-the-shelf sensing technologies (Baldauf, 2007).

Next, research on user positional sensing and tracking around tabletop displays using proximity sensors are reviewed. Alternative sensing technologies are also covered.

Human sensing systems for display purposes is not new (Ballendat et al., 2010). Sensing systems that take advantage of proximity sensors have been developed for tabletop displays to make them aware of user positions. Some of these systems have features that are able to detect touch and identify the users who initiated them. Examples of such systems are Bootstrapper by Richter et al. (2012), Maximally Stable Extremal Regions (MSER) by Ewerling et al. (2012),

See Me See You by Zhang et al. (2012), Carpus by Ramakers et al. (2012), Medusa by Annett et al. (2011) and adaptive tabletop by Klinkhammer et al. (2011).

DiamondTouch by Dietz and Leigh (2001) was one of the earliest systems that tracked users' positions around a multi-touch tabletop. The static position of receiver pads placed on the chairs is a direct indicator of where users are located at that instant of time. However, DiamondTouch requires users to remain in contact with their pads all the time without which tracking is lost. Tănase et al. (2008) uses twelve infrared sensors; three sensors on each side facing away from the table's edges. The system tracks users' positions at a very low resolution with five discrete user positions on each side of the table. An improved system, Medusa by Annett et al. (2011) was constructed to sense users' presence and track their positions around a tabletop display using an array of 138 infrared sensors. Medusa has extra functionalities which tracks hands and arms above the display. Out of the total 138 sensors, 34 are used to sense and track users' positions around the table. Both of these setups require additional sensors for larger sized tables. Having a massive array of sensors can be costly both in production and in computing resource. Furthermore, portability will be an issue if duplication is to be carried out on alternate tables due to its crowded arrangement of sensors and cables. Klinkhammer et al. (2011) implemented a successful but rather costly setup of sensing tabletop display capable of tracking users' positions around the table by placing 96 infrared sensors around the 65-inch table in a similar configuration to Medusa. Klinkhammer et al's system focuses on an interaction design using visually separated space for each user serving them as a personal territory. While there were 96 infrared sensors used in Klinkhammer et al's implementation, Medusa's design used 34 infrared sensors for similar functions because Medusa used a much smaller 40 inch display. It can be concluded that, the number of sensors needed, increases significantly for larger displays as long as the arrangement of sensors used is similar to Figure 2.1 (Yusof et al., 2014).

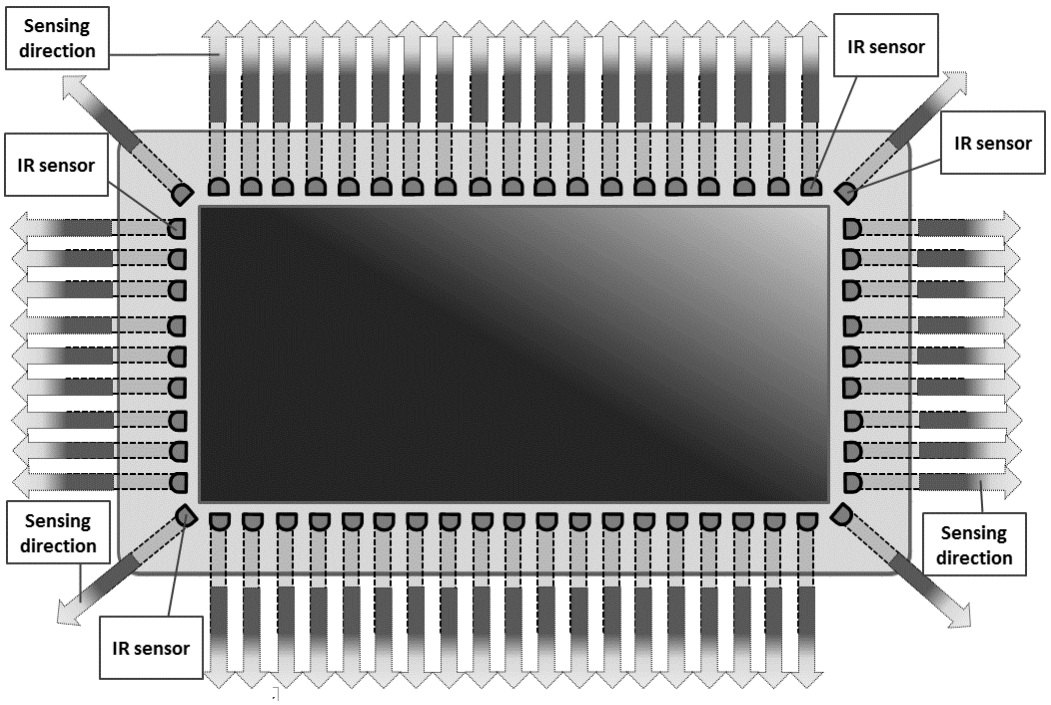


Figure 2.1 Common arrangement of tabletop display sensors described in the literature where all sensors are facing outwards

2.3 Person Identification Techniques

Person identification is the task of recognising an individual. It may be achieved in real-time or by offline processing. As this thesis deals with real-time interactions between human and computer, only real-time methods will be reviewed.

A number of early works attempted to solve identification problems for surveillance (Chien et al., 2003) and human-computer interaction (Song and Chen, 2004) using facial features (Pentland et al., 1994), gait (Han and Bhanu, 2006), (Goffredo et al., 2010), thermal imagery (Arandjelović et al., 2006) and clothing (Sivic et al., 2006). Frequently these methods work under certain constraints such as stationary cameras, fully frontal faces and consistent lighting (Apostoloff and Zisserman, 2007).

Generally, the task of person identification can be categorised into either 1) vision-based or 2) non-vision, sensor-based methods. Examples of these are provided in the following sections.

2.3.1 Vision-based Methods

Real-time vision-based person identification systems typically use image frames from a video stream. These frames are then pre-processed by image processing algorithms. Initially, features of a person extracted from these images are used for training of classifiers, and typically the same acquisition and pre-processing are performed on images of an unknown person before they are sent to the classifiers.

Face recognition is one of the popular methods of real-time person identification and under controlled settings such as stationary cameras, fully frontal faces and consistent lighting, it can perform very well. The work by Apostoloff and Zisserman (2007), one example of such systems, performs real-time identification using random-ferns classifier on the facial features extracted from a video stream from a standard web cam as the person approaches. In live experiment, up to 5 persons can be tracked at 15fps and it takes 5 seconds to identify a person.

Nakatani et al. (2012) proposed a method using top-view image from a depth camera aiming to remove altogether the occlusion problem in the person identification process. Features such as body height, body dimensions, body size, and depth histogram are computed from a depth image consisting of pixels location in 3-D space. They evaluated their method on only eight persons and with the constraint that they needed to stop walking when directly under the camera.

Gait-based features are generally invariant to illumination, shape and clothing. John et al. (2013) proposed features that are based on height dynamics of a person extracted using a top-down depth camera and combined with an RGB-height histogram where colours are collected at different heights. It is however reported that the method achieved “good” accuracies only for small population (≤ 10) and limited to offline applications because of the supervised feature

selection step. Goffredo et al. (2010) proposed a reconstruction method to rectify and normalised gait features recorded from various viewpoints into a side-view plane. Their method achieved high recognition rate of 73.6% using the kNN classifier on a large gait dataset with over 2000 video sequences.

Person identification based on a "bag of soft biometrics" has been proposed by Dantcheva et al. (2011). The "bag" consists of a set of facial soft biometrics (i.e. skin colour, hair colour, eye colour, beard presence, moustache presence, and glasses presence), body weight (build) and colour of clothes of the torso area and legs area. The authors reported that the method overcomes the limitation of single soft biometric traits by using multiple traits. It is also claimed that there is no other similar work on person identification that is based solely on soft biometric features.

Viewpoint invariant person identification is characterised by the ability of a method to identify the person when seen from different viewpoints such as front, side and up-angled views. Jaha and Nixon (2015) proposed the use of soft-clothing biometrics as viewpoint invariant (side- and front-view) method for person identification. Human descriptions via manual human labelling (i.e. head coverage, hat, sleeve length, leg length, belt presence, neckline size, and heel level) were used in the experiment and good performance has been reported. Fully automated clothing label detection and description is envisioned to open up a way for effective viewpoint invariant person identification.

User identification in a public space via biometric recognition was demonstrated by Schulman et al. (2008) where hand geometry recognition was used as the identification method. Virtual agents, also called relational agents were designed to interact with visitors and were required to distinguish individual users from others with high accuracy. These agents remember their interaction and dialogue history with each user so that conversations can be continued in the

future dialogue. The system requires users to "log in" to an agent before starting an interaction. Avoiding the requirement to carry devices such as ID cards and RFID tags, the authors used hand geometry-based biometric user identification system to establish an interaction. A similar identification method using hand geometry analysis was proposed by Schmidt et al. (2010) to allow users to access personal data on a shared surface such as interactive displays. The method called "HandsDown" could be used to identify users and access personal picture collections. A hand could be put down at any location on the display, and once identified, a personal picture collection would be displayed and automatically oriented towards the user. This method however is inconvenient as a user needs to continuously place their hand flat on the display to activate identification, without which leads to a loss of tracking.

Reid et al. (2014) proposed a soft biometrics method based on relative measurements of human descriptions (i.e. "much shorter", "shorter", "same", "longer", "much longer", "thinner", "male/female", etc. for various body parts) as opposed to conventional description based on absolute labels (i.e. "very short", "short", "average", "long", "very tall", "thin", etc.). Comparisons between subjects were performed on the Soton gait database (Shutler et al., 2002) requiring annotators to compare a single target to multiple subjects. Comparative descriptions are reported to contain more discriminative information achieving 90% accuracy at rank 19.

2.3.2 Sensor-based Methods

The ORL (Oracle Research Laboratory) Active Floor by Addlesee et al. (1997) and The Smart Floor by Orr and Abowd (2000) were two floor systems based on footsteps profiles that attempted to overcome problematic situations faced by other methods such as occlusions and illumination problems by face recognition, noise problem by voice recognition, and requirement to carry item for ID-based identification. Both systems used load cell sensors to

detect changes of weight and inertia called ground reaction force (GRF) within a specific time and store this as profile features of a person. Relying on uniqueness of footsteps, these systems are therefore only suitable for discrimination amongst small groups of people (up to about 15) (Orr and Abowd, 2000).

A custom-built wearable ring by Vu et al. (2012) is capable of identifying a person wearing it by transmitting electrical signals when the ring comes in contact with a touch-screen device installed with the identification software. This technique is limited for use on devices with capacitive-based touch-screens in addition to a person having to carry the ring. Rofouei et al. (2012) proposed a user identification method in multi-user interactive display setting by first associating phones to users, and then touches to users. Users will need to perform motion actions such as shaking to trigger an association. Their algorithm then cross-correlates acceleration data from a phone with hand acceleration of a hand which is holding the phone. The hand acceleration is captured by a Kinect RGB-D camera. User identification is performed using the two-step association method for the duration of the interaction with the display. A phone representing an identity of a person, is assumed to have been paired to the system earlier, so effectively the system will know the person if the system can associate the hand to the phone.

“intuit” proposed by Wiethoff et al. (2011) investigated simple identification techniques on digital surfaces using tangible objects called TUIs (Tangible User Interface), which were placed on the displays, to grant users access to the application and their personal content. TUI was a cube built to allow five different identification techniques to be performed using it, such as fingerprint scanning, handwriting recognition, spatial gestures, tapping signals and virtual keyboard. Tapping method was concluded as the most preferred method of identification because if its ease of use, however an accuracy of each technique was not reported.

2.4 Person Re-identification Techniques

Previous research reported in the field of human re-identification has used colour (RGB), depth (D) and RGB-D cameras (Han et al., 2012). RGB approaches have included the use of colour, shape and texture descriptors, interest points and image regions, and, often, combinations of these (Ma et al., 2014b). The recent availability of depth cameras has led to approaches including point-cloud (Munaro et al., 2014b) and anthropometric (Han et al., 2012), (Munaro et al., 2014b), (Barbosa et al., 2012) methods. Reviews of person re-identification methods based on 2-D appearance collected by RGB cameras and 3-D information gathered by RGB-D cameras are presented in the next sub-sections 2.4.1 and 2.4.2.

2.4.1 Person Re-identification Based on 2-D Appearances

Many appearance-based object recognition methods, including person re-identification, use colour and texture as primary cues. Colour is an established feature that is widely used in object recognition (Geusebroek et al., 2001) where its distributions or histograms have been used as models for human appearance. For example, Lin and Davis (2008) incorporated spatial information into colour features to preserve vertical colour structure in appearances. Illumination changes were dealt with by using normalized colour feature and colour rank feature. Nearest neighbour classification was used on pairwise dissimilarity profiles between individuals. In Schwartz and Davis (2009), texture, gradient and colour were combined and projected into a low-dimensional latent space by Partial Least Squares (PLS) reduction. In recent years, Kviatkovsky et al. (2013b) proposed a colour (i.e. illumination) invariant re-identification method using a log chromaticity colour space with a region based covariance descriptor and demonstrated that colour can be a powerful cue for person-re-identification, when used properly.

It has been shown that there is a trade-off between illumination invariance and its discriminative power (Geusebroek et al., 2001). Stokman and Gevers (2007) proposed a fusion of colour models for feature detection to achieve an optimal balance between illumination invariance (repeatability) and discriminative power (distinctiveness). The method improved discriminative power compared to standard weighting schemes.

In addition to colour, descriptors based on interest point have been explored by Hamdoun et al. (2008) where they matched signatures based on interest-points descriptors collected from short video sequences. 'Camellia' key-points detection and characterization functions, a quick variant of SURF (Speeded Up Robust Features) (Bay et al., 2008) was used to detect interest points and compute the descriptor. Matching was done by a function from the Camellia image processing library which implements a Best Bin First (BBF) search in a KD-tree containing all models. Gheissari et al. (2006) also extracted key-points similar to Hamdoun et al. (2008) but here key-points were invariant in the spatio-temporal domain. They used colour and structural information around each key-point to generate a discriminative and robust signature. The use of spatio-temporal object alignment has improved matching performance.

Image region based descriptors together with colour cue were used in methods that are mostly based on Multiple Part Multiple Component (MPMC) representations that subdivided the human body into parts to accommodate its non-rigid nature (Pala et al., 2015). Symmetry-Driven Accumulation of Local Features (SDALF) (Bazzani et al., 2014) subdivided the human body into left and right torso and legs. HSV colour histograms for each body part were extracted together with descriptors for maximally stable colour regions and recurrent highly-structured patches (RHSP). These descriptors were shown to have improved robustness to appearance variations (Bazzani et al., 2014). MSCR and RHSP were extracted from several randomly sampled image patches, which were clustered to find the most significant ones. It is said using

a quad- core Intel Xeon E5440, 2.83 GHz with 30 GB of RAM, "*Partitioning of the silhouette in symmetric parts takes 56 milliseconds per image. SDALF is then composed by three descriptors WCH, MSCR and RHSP that take 6, 31 and 4843 milliseconds per image, respectively. It is easy to note that the actual bottleneck of the computation of SDALF is the RHSP. Matching is performed independently for each descriptor and it takes less than 1 millisecond per pair of images for WCH and RHSP and 4 milliseconds per image for MSCR. In terms of computational complexity, the computation of the SDALF descriptor is linear in the number of images, while the matching phase is quadratic*". In Martinel and Micheloni (2012) and Farenzena et al. (2010), head to torso and legs data were also included as features. Weighted Gaussian colour histograms were used to describe each body part in addition to pyramid of histograms of oriented gradients and texture description using Haralick features. Bak et al. (2010a) proposed a method based on Haar-like features and dominant colour descriptors (DCD). A body was subdivided into basic upper and lower parts; each part was described using an MPEG7 dominant colour descriptor. Multiple Component Matching (MCM) (Satta et al., 2011) subdivided the body into torso and legs, randomly extracted rectangular overlapping patches from each component and used HSV colour histograms to represent the patches. Illumination changes were accommodated by generating samples of varying brightness and contrast from the patches. In Bak et al. (2010b) an appearance model based on spatial covariance regions extracted from body parts using Histograms of Oriented Gradients (HOG) was proposed. Covariance descriptor was used to find out the similarity between corresponding body parts. Dissimilarity measure based on spatial pyramid matching was used for recognition. Illumination invariance was achieved using histogram equalisation before computing covariance regions to generate a "person signature".

Most of the works in person re-identification have involved combinations of the methods discussed in the previous reviews. The work by Wang et al. (2007) involved segmenting a person's image into regions and uses a co-occurrence matrix to store the colour spatial relationship with the regions defined. This worked well for a limited range of viewpoints. Gray and Tao (2008) addressed the viewpoint invariance problem by combining spatial and colour information to form a set of discriminating localised features used to train an ensemble of classifiers. The method did not attempt automatic person recognition but did help human operators reduce the search time needed to match pedestrians from a large gallery. In Martinel et al. (2015) features were transformed by non-linearly warping the feature space to produce "warp functions". A body was subdivided into four main body parts and dense colour and texture features were extracted from the body parts. Given frames from two cameras, this method learnt a discriminative model in the warp function space (WFS) to get the probability of a sample feature warp function coming from the same person or not.

Wang et al. (2014) dealt with inconsistency of feature distributions of person images captured by different cameras by using a feature projection matrix where image features of one camera were projected to a feature space of another camera. These inconsistencies were mainly caused by camera view switching that causes lighting and image–scale variations. Tao et al. (2013) improved Keep It Simple and Straightforward (KISS) metric learning by smoothing and regularising KISS to make estimation of covariance matrices more robust and hence resulting in improved performance. They also introduced incremental learning to RS-KISS (Regularized smoothing KISS) to reduce computation cost. An et al. (2015) proposed a reference-based re-identification method across different cameras. Reference space was used for matching where colour or texture descriptors were translated to similarity measures between a person and the others in the reference set. A subspace was created where training and test data were projected

into this space and reference descriptors (RDs) of the training and test data were generated by computing the similarity between them and reference data. A test person was determined by comparing the RD of the test data and RDs of the training data.

Wu et al. (2015) emphasised that a normal method would likely fail when a pair of images of the same person from different viewpoints were trying to be matched, especially in a situation where the person was carrying a backpack, or wearing clothes with logos at the front and not at the back. So the authors proposed a "pose prior" method, used to make descriptor distance (between probe and training) invariant to viewpoint changes. They applied six horizontal strips to the body, recording colour and texture as a function of pose. Orientation angle was estimated from the trajectory of movement which was computationally simple but inaccurate when people move in "unusual patterns" or in static position.

Eisenbach et al. (2012) used an automatic online feature selection to obtain candidate features based on appearances, with a purpose to reduce dimensions of a feature space. Multi-frames matching technique was used to improve classification results. A combined Mahalanobis distance and average distance decision methods was employed instead of majority vote. Mean colour of a pre-defined region was used as the feature, which could cause problems when lighting changes. It was reported that the method might not be able to distinguish a person when many people (quoted as >100) were present as a result of using small number of appearance-based features.

An et al. (2015) proposed a reference-based re-identification method. In the reference spaces, colour and texture descriptors were translated into similarity measures between each individual and the members of a reference set (a selected subset of the training data). The colour and texture features, however, were not explicitly designed to be illumination invariant. A spatio-

temporal segmentation algorithm was used by Gheissari et al. (2006) that combined normalised colour and salient edge histograms to obtain a colour and pose invariant identity signature. The algorithm demonstrated partial viewpoint invariance using a limited set of viewpoints. Ma et al. (2012) developed a representation that used filtering, "biologically inspired" by the human visual system, and covariance descriptors (BiCov). An "enriched" BiCov descriptor (eBiCov) that incorporated SDALF has been reported to have improved robustness to illumination changes (Ma et al., 2014a).

2.4.2 Person Re-identification Using Depth Information

Depth information of a scene has recently become widely available through the use of low-cost depth-sensing cameras such as the Microsoft Kinect. The Kinect API provides functions to extract three-dimensional anthropometric measures providing real-time body pose estimation (Shotton et al., 2013, Taylor et al., 2012) via the position of body joints (twenty body joints for Kinect V1 (Microsoft, n.d.) and twenty-five for Kinect V2 (Microsoft, n.d.). To address the long-term re-identification problem, Barbosa et al. (2012) eliminated clothing appearance descriptors and used only anthropometric cues consisting of ten descriptors from the Kinect V1. Using their own dataset of seventy-nine people, results as high as normalized Area Under the Curve (nAUC) = 91.76% were reported but with rank-1 performance below 20%, and rank-10 approximately 60%.

Colour-based re-identification methods have also been reported using RGB-D cameras. Albiol et al. (2012) subdivided the body into horizontal stripes at different heights. The mean colour of each stripe was calculated to form a "body print" for each person. Depth measurements were used to locate pixels at the correct 3-D coordinates. However, the method is not viewpoint invariant and is susceptible to error when, as is often the case, people have differing appearance

from different viewpoints. A similar 3-D-based appearance descriptor model was proposed by Gandhi and Trivedi (2006). Re-identification was performed using a panoramic map where the appearance (i.e. colour) information of a person extracted from multiple cameras is projected onto the surface of a cylinder. The method requires a surrounding multi-camera array for re-identification. Oliver et al. (2012) also based their feature vector on a cylindrical structure, but used depth information to improve the assignment of colour pixels into orientation bins. Another similar approach, by Baltieri et al. (2011), maps a 2-D image onto a pre-defined 3-D body model. The method requires individuals to be moving because body orientation is estimated using a tangent to their trajectory. Another colour based approach by Southwell and Fang (2013) extracted colour information of a shirt through shirt segmentation using depth information. HSV colour space was used to handle lighting variations, and nine colour bins were defined from the HSV colour space to represent the shirt colours as red, green, blue, yellow, cyan, magenta, white, black and grey. The method was tested on 8 different shirts in two lighting conditions. 98% identification performance has been reported, although tested on small dataset, it is claimed to be suitable for real world applications.

None of the colour-based methods described in Albiol et al. (2012), Gandhi and Trivedi (2006), Oliver et al. (2012), Baltieri et al. (2011) and Southwell and Fang (2013) makes use of the anthropometric data available from depth-cameras and so can be confused by different individuals wearing similarly coloured clothing.

Multi-modal implementations have been reported by Pala et al. (2015) to achieve better first-rank recognition results by fusing appearance descriptor with anthropometric measures using an MCD (Multiple Component Dissimilarity) framework. They also implemented a multi-modal method using SDALF and eBiCov fused with anthropometric descriptors, and reported improved performance of the multi-modal method compared to the appearance descriptor

alone. However, these are not viewpoint invariant methods; re-identification is only possible for subjects viewed directly from the front, back or sides.

Viewpoint invariant re-identification has only been reported using either colour or anthropometric data alone. In Chapter 4, a novel method of person re-identification was proposed in which multi-modal approach combining both colour and anthropometric shape properties of people was formulated such that they vary as a function of viewpoint.

2.5 Body Orientation Estimation Methods

Human body and head orientation estimations, in addition to location tracking, have been used in human visual focus of attention analysis research which have gained much interest recently among researchers in human behaviour understanding (Ozturk et al., 2011) and in human-computer interaction, robot-human interaction, and surveillance applications (Chen et al., 2012). A review of the literature shows only a few works of person re-identification that have used body orientation estimation methods to address the viewpoint invariant person re-identification problem, the most recent being the work by Wu et al. (2015) reviewed earlier in Section 2.4.1. Body orientation estimation methods based on 2-D and 3-D are reviewed in the following sub-sections.

2.5.1 Body Orientation Estimation Based on 2-D Appearances

There are generally three categories of 2-D body orientation estimation methods which are based on static cues, motion cues, and a combination of static and motion (Liu et al., 2013). Chen et al. (2011) proposed a method that estimates body and head poses (orientations) using features characterising the body pose and the head pose. Body orientation would be assigned to one of eight directions i.e. N, NE, E, SE, S, SW, W, and NW after applying a multi-level

Histogram of Oriented Gradients (HOG) feature and sparse representation method on the image from the human detection output. The body/head orientation estimation would fail if bounding boxes were not detected correctly.

Body and head orientation change estimation method from low-resolution images was demonstrated by Ozturk et al. (2011) to detect gaze direction using a top-view single camera setup. Shape Context matching of the head-shoulder region was used to estimate body orientation. Head orientation was estimated by calculating changes in motion flow vectors of Scale Invariant Feature Transform (SIFT) features around the head region. The method was tested on 17 participants and might not work for cases where people carry backpacks, or wear hats which would occlude head-shoulder region. SIFT features could not be tracked when face is not visible.

A method that used a 3-D human body model to estimate body orientation was proposed by Chen et al. (2012) in multi-view scenarios. The 3-D model was composed of stacked elliptical cylinders and a 3-D coloured point cloud that utilised multiple 2-D templates for matching using a likelihood function, before orientation estimation could be obtained. 12 discrete orientations covering 360^{O} were defined.

2.5.2 Body Orientation Estimation Using Depth Information

3-D based methods for body orientation estimation are emerging from the introduction of consumer RGB-D sensors such as the Kinect. Using RGB-D sensors obviates some of the challenges from 2-D based sensors such as cluttered environment, illumination change, and partial occlusion, in orientation estimation. Glas et al. (2007) proposed a method of estimating body orientation using information from a network of laser scanners. Lasers from different sources are projected onto a human body at a fixed height from the floor resulting in large

variations in cross-sectional contour shape between subjects because scan was performed on roughly at wrist-height for taller subjects, and above elbow height for shorter subjects. Several factors reportedly contribute to the difficulty in developing a precise generalizable human model such as more arm (wrist-height) movement for taller subjects, type of clothing – a loose shirt or heavy coat causing torso's appearance look large and asymmetrical, and the same problems for backpack carrying subjects. A three circle model, representing a human's torso, and arms was proposed to deal with challenges mentioned. The model is unable to give a correct estimate in scenarios such as a body pose with arm movement (wrists) higher than the scan height, a sitting person, and children (also really tall people) because of the fixed scan height. The model also requires at least two sensors positioned far apart to be able to fit the scan data to the model.

A method proposed by Liu et al. (2013) used static and motion cue extracted from RGB-D cameras. The human body orientation is quantised into eight direction classes i.e. S, SW, SE, W, E, NW, NE, and N. The authors proposed a superpixel-based viewpoint feature histogram (SVHF) method that improved viewpoint feature histogram (VHF) method by Rusu et al. (2010). VHF is extracted from all 3-D points, which is time consuming and more sensitive to noise while SVHF are reported to be more robust and sparser than original RGB-D points. Dynamic Bayesian network system (DBNS) was employed combining both the static (SVHF) and motion cues of superpixels-based scene flow (SSF) method.

2.6 Summary

The chapter explored the literature and state-of-the-art in five research areas relevant to the research in this thesis. It provided insights into the trends in natural interaction which motivate this research.

Natural interaction systems of the future may integrate many components such as physical objects, projected light, screens, computers, cameras or sensing devices, speakers and other digital media and devices. These components must work together as a whole and appear as a single, integral and fluid system. This illusion can be achieved with the real-time sensing, interpretation, behaviour simulation and information rendering. A successful natural interaction system requires very short response times, because "*latencies on any part of the feedback loop can easily make the interaction unfeasible and frustrating.*" (Valli and Linari, 2008). Hardware and software are necessary to have very good overall performance due to having to process multiple video streams from cameras (or data streams from sensors), analysing system's behaviour, and generating output such as images, videos and audio, at the same time. Human sensing and context-aware systems can be made to work hand-in-hand with natural interaction system by providing an extra personalisation layer to the system. However targeted personalisation of services are currently only explored with person identification.

It is observed that the existing methods of person re-identification generally work well when anthropometric features are used. Improved performance has been demonstrated when clothing colour information is used because it has been proven to be a powerful cue for person re-identification (Kviatkovsky et al., 2013b). However, colour descriptors can vary drastically due to illumination variations, pose (or view-point) variations and scale changes in a multi-camera setting (Bedagkar-Gala and Shah, 2014). The performance of most re-identification methods is

currently highly dependent on the view-point and can degrade significantly when appearance changes, as it often does, with viewpoint, for example, as caused by clothing which has different colours on the back and the front, by the presence of backpacks, long hair or opened jackets Albiol et al. (2012). Wu et al. (2015) expressed a concern that most person re-identification methods perform matching across images using the same descriptors, regardless of viewpoint or human pose, which can induce serious errors in the matching of target candidates. Recent work by Jaha and Nixon (2015) demonstrated re-identification's improved performance when using clothing analysis via manual human labelling for view invariant person re-identification. The tests were performed on a Soton database (Shutler et al., 2002) in which each of 115 individuals was labelled by multiple users describing 23 soft bodily traits. The results could be used as benchmarks for re-identification methods with automatic feature extractions. There were very few works on viewpoint invariant person re-identification that utilised body orientation, such as that of Wu et al. (2015), Jungling and Arens (2011), and Eisenbach et al. (2012), but only Wu et al. (2015) incorporated their own body orientation estimation algorithm in their re-identification method.

It can be concluded from the literature survey that automatic 360° unrestricted view-point invariant, appearance-based methods have not been reported in previous publications hence the contribution of this thesis in this area.

The next section looks at the research methodology and experimental approach used to address research questions. It also provides an overview of the experimental aspects of work presented in the following chapters.

2.7 Research Methodology

The research methodology adopted throughout the work presented in the thesis was, broadly, an experimental prototyping approach with systems and software evolved through iterative testing. With the general research motivation being to resource natural interaction in an intelligent environment, a stage and a means for human-human and human-computer interaction was sought. The shared interactive display space of a tabletop computer system was selected as a naturally collaborative demonstration vehicle, providing an opportunity to sense multiple re-identified users and support their interactions with each other and with their environment.

2.8 Experimental Overview

A tabletop prototype sensing system is presented in Chapter 3, designed with the aim of tracking the positions of multiple users as they interact with the tabletop display. This functionality allows individualised and personalised modes of interaction to be supported within the context of a multi-user session. A further aim of this prototype was to achieve accurate tracking of user positions at a much lower cost in terms of price to setup, low computational power and setup complexity than previously proposed equivalent systems.

Person re-identification, necessary to the provision of personalised services in intelligent environments, is developed in Chapter 4. As shown in the Literature Review in Chapter 2, methods of re-identification reported in the literature have limited performance in re-identification of individuals in unconstrained environments, and often require specific viewpoint observations for both training and re-identification.

A fundamental performance limiting factor in person re-identification is the quality, functionality and resolution of the acquisition sensors. At the commencement of this research commercially-available RGB-D cameras were at an earlier stage of their development. For example, Microsoft's Kinect v1 RGB-D camera and SDK for Windows had been available for less than a year at that time. The ambition was to use the best sensor possible to achieve the best acquisition of data. The new Kinect v2 was not commercially available until July 2014, but was pre-released in November 2013 to selected researchers on submission of acceptable project proposals to the Kinect for Windows v2 Developer Preview Program. This research was successful in obtaining the pre-released Kinect v2. While providing an excellent opportunity to work at the earliest moment with the very latest new technology, it also meant there was no equivalent compatible datasets available for recorded testing or comparison using the same Kinect v2 camera.

The method proposed in Chapter 4 achieves viewpoint invariant performance in real-time, and its performance is demonstrated experimentally with a Kinect V2 RGB-D dataset of 64 people performing 14 individual activities and four group activities that included accessory-carrying activities. The performance results are reported via the average Cumulative Matching Characteristics (CMC) curve whose key properties are summarised by the normalised Area Under the Curve (nAUC) and rank-1 recognition performance. The proposed method uses a novel View Invariant Multi-Modal (ViMM) feature vector to characterise users. The complete ViMM descriptor consists of 18 features although subsets of the vector were also tested for comparison and the relative performances are reported.

Chapter 5 explores multi-person re-identification and again the performance is demonstrated experimentally with Kinect V2 RGB-D dataset and results are reported via CMC curves and nAUC.

Finally, in Chapter 6 the tabletop sensing and person re-identification are brought together to demonstrate how a re-identification system could complement the tabletop display as a sensing module and context provider to aid delivery of targeted personalised services to people.

CHAPTER 3:

Human Sensing around Multi-touch Tabletops

3.1 Introduction

Modern interactive displays are not always content-contextual in the location in which they are placed. Furthermore, these devices do not have the capability of detecting users' presence within the range of interaction distance. One of the ways in which contextual contents can be formed prior to and during user interaction is by making displays aware of their presence before they even interact with the screen.

Multi-touch multiuser (MTMU) tables provide many promising application areas and, looking at the trend of market adoption (Ch'ng, 2012), MTMU can potentially be pervasive in the near future. MTMU computing opens up possibilities where collaboration is transformed from sequential to parallel – all users work on a task together, and at the same time (Ch'ng, 2012). Tabletop computers such as Microsoft's Surface, PQLabs, Ideum and a collection of emerging tabletop computers are targeting public spaces such as Galleries, Libraries, Archives and Museums (GLAMs). Other manufacturers have also included multi-touch capabilities on desktop PC displays. Large Full High Definition (Full HD) displays of up to 65" supporting up to 32 touches and 3-D Stereographics are also being explored at the Chowen and Garfield Weston Foundation Digital Prototyping Hall, the Digital Humanities Hub, at the University of Birmingham (Ch'ng, 2012).

Tabletop displays support touch gestures on their surfaces, with content reacting only to particular types of gestures. These displays were created for human use but are unfortunately not human-aware. Microsoft Surface's PixelSense display allows a 10 cm interaction distance

beyond the surface of the touch screen so is at most, finger-tip aware. It would be beneficial if these tables were able to recognise human presence and provide a more satisfying interaction by establishing social bonds with users and enhance user engagement (Schulman et al., 2008). If displays are able to sense and track individual users, the multiuser coordination process could be greatly enhanced, for example, by virtually partitioning space for each user on the display, and as a consequence, the physical space belonging to each user around the display is made clear. Therefore, screen contents can be channelled to each space as a result. Such a separation of space is termed personal territory (Klinkhammer et al., 2011). As a proximity-aware system, it is important to indicate to users that the system recognises their presence during a session of use (Vogel and Balakrishnan, 2004).

The research reported in this chapter leads to a simple, low cost and robust user sensing system using proximity sensors which could be attached to a range of tabletop displays of sizes (diagonally measured) 40 inches (similar to Microsoft Surface) up to 65 inches (current largest available via multi-touch overlay). The system is able to detect approaching users and can continuously track and maintain user positions around the table. Infrared distance sensors were used instead of alternative sensors such as ultrasonic range detectors due to their higher temporal resolution. The choice was a key factor in demanding interactive settings involving users' movements (Walther-franks et al., 2008).

The chapter describes in detail the architecture of the proposed human sensing and tracking system. It is then followed by the sensor accuracy test, with the aim of finding and verifying optimum and ideal positions where users should be standing when interacting with the table. Next, accuracy and performance tests will be evaluated in a number of interaction scenarios demonstrating two prototype applications performing functions related to the current application areas. Finally, the chapter is concluded with discussion and summary of the work.

3.2 Designing the Human Sensing and Tracking System

Inspired by the previous works using infrared sensors on a multi-touch table; Medusa by Annett et al. (2011) and the work by Klinkhammer et al. (2011), a new system was built upon these previous works, extending human sensing systems using different sensor design and strategy. The aim was to build a simple proximity aware system that uses very minimal sensors, capable of tracking up to six simultaneous users around a multi-touch table ranging from 40 to 65 inches of size. A maximum of six users around a 65 inch table is a natural and comfortable space. Display size smaller than 40" would be too small for multiple users. It was hypothesised that it was possible to have a minimal array of sensors to provide similar or better performance in comparison to larger arrays of sensors.

Recent developments around human sensing and tracking include vision-based top-view tracking systems using ceiling-mounted 3D camera such as Kinect by Hu et al. (2014) and Rusňák et al. (2014). This kind of system requires that the camera (i.e. Kinect) be placed higher than average ceiling (3 meters) because of the field of view of the camera, in order to achieve a tracked region larger than 2 m^2. The camera cannot also be placed too high as it needs to compute a depth image with high resolution.

3.2.1 Technical Setup and Procedure

The prototype setup was built around a 65" multi-touch table supplied by Mechdyne Corporation. The measurement of the table is 172 cm (width) x 108 cm (height) with a screen width of 138 cm and height of 76 cm. A total of twelve Sharp infrared distance sensors are used, 8 of which are for medium range detection 10-80 cm (**2Y0A21)**, and the remaining 4 are for long range detection 20-150 cm (**2Y0A02)**. The reason for using two types of range sensors was to combine the optimum working distance from both types of sensors. These sensors were

chosen for their good performance in ambient light with objects of arbitrary colour. All the sensors produce an analogue voltage related to the range of the object detected and are connected to data acquisition boards, **PhidgetInterfaceKit 8/8/8**, providing eight channels of analogue-to-digital conversion at 10-bit resolution and a rate of up to 500 samples per second. Figure 3.1 illustrates the sensors setup.

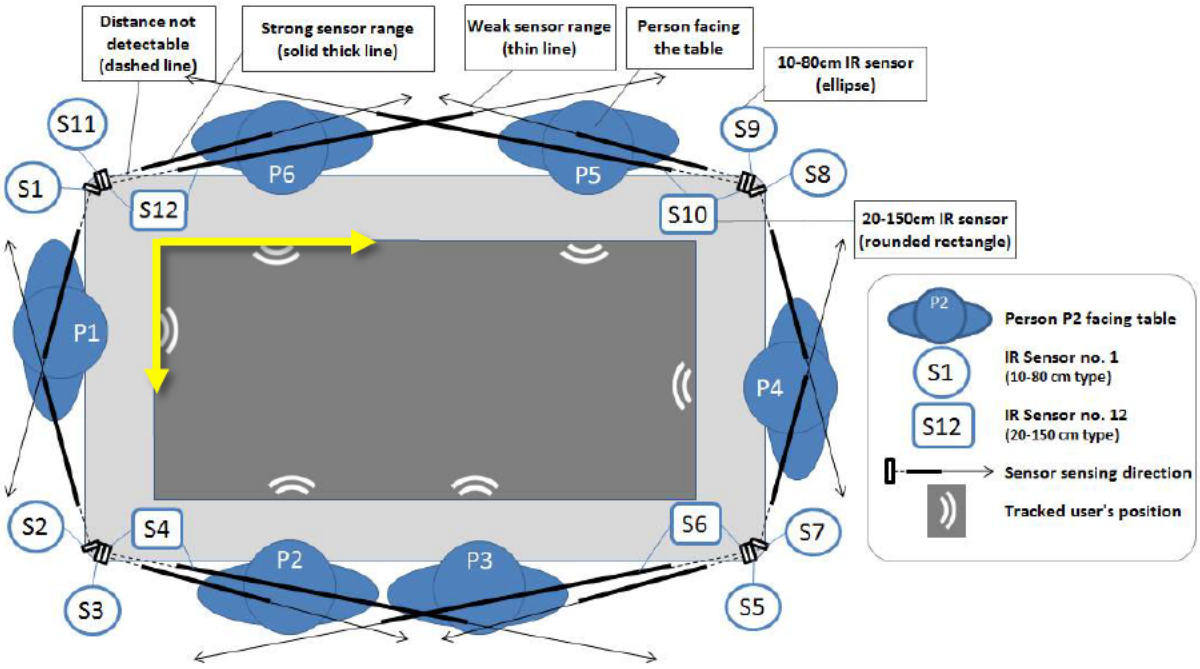


Figure 3.1 The proposed tracking infrared sensors design and setup for multi-touch table top displays. The coordinate origin is at the top left corner of the table as indicated by the yellow arrows.

Initially, the eight medium distance (10-80 cm) sensors were placed at the corners of the table, with each sensor aligned vertically to the table's edge as shown in Figure 3.2.

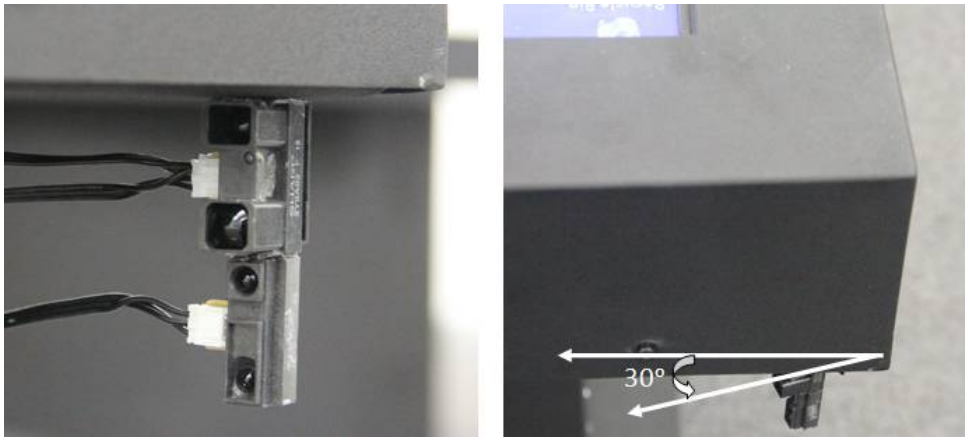


Figure 3.2 Infrared sensors placed at the end of table's edge facing slightly outward about 30^{O} from the table's edge.

3.2.2 Limitations of Standard Sensors

Initial testing showed that the sensitivity of the 10-80 cm infrared sensors degrades drastically when the detection distance exceeds 30 cm and above (see Figure 3.3). Performance of both types of infrared sensors were carefully studied and optimum ranges were then identified and chosen based on the rate of change of output voltage as a function of object distance. They are depicted in the graph in Figure 3.3 as a shaded rectangle with dotted borders for the 10-80 cm sensor and a shaded rectangle with dashed lines for the 20-150 cm sensor. Both optimum ranges are overlaid to give a clear view on how the tracking algorithm chooses the best of both sensor types. To overcome the limited range of the 10-80 cm sensor, long range (20-150 cm) infrared sensors were introduced as a complement for the long sides of the table. The 10-80 cm sensor (e.g. S3 in Figure 3.1) performs the tracking of objects in the range of 10 cm to 30 cm, and the 20-150 cm sensor (e.g. S4 in Figure 3.1) takes over tracking for objects beyond 30 cm. This results in reliable tracking of users when they move along the long sides of the table.

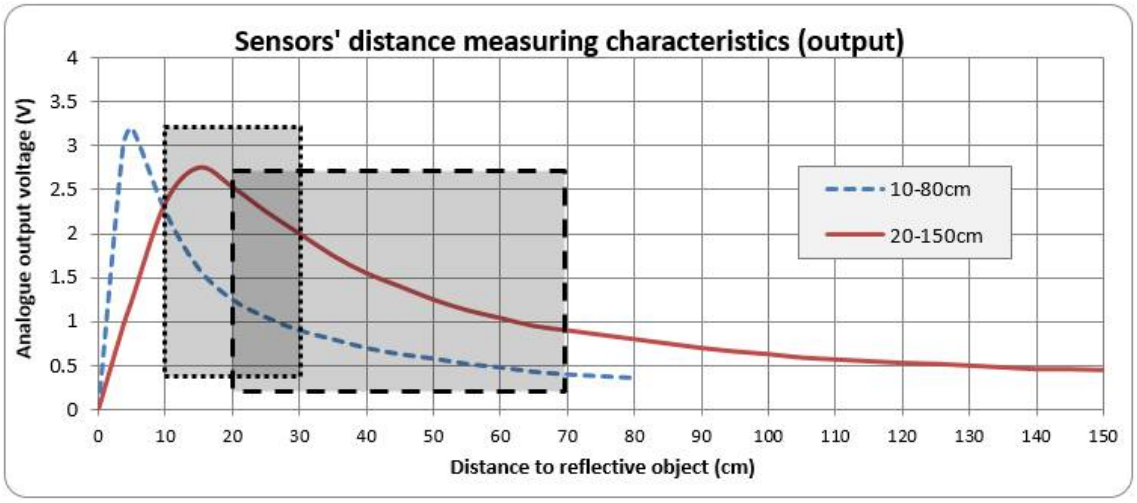


Figure 3.3 Optimum performances of 10-80 cm and 20-150 cm infrared sensors. Shaded rectangle with dotted borders belongs to 10-80 cm sensor and shaded rectangle with dashed lines belongs to 20-150 cm sensor. (Graph shows sensors' measuring characteristics adapted from Sharp's datasheets with added range of sensors' optimum performance).

3.2.3 Acquisition and Pre-processing of Sensor Data

The PC configuration used in the setup was Windows 7 Professional operating system with Intel quad core processors (2.7GHz), 8 Gigabytes of Random Access Memory (RAM) and 1 Gigabyte of Nvidia Quadro 1000M discrete graphic card. Two Phidgets I/O boards (PhidgetInterfaceKit 8/8/8) were used to connect all the sensors. The eight medium distance infrared sensors were connected to the first interface kit board and the remaining four long-range sensors were connected to the second interface board. The sensors were sampled at 20 Hz and a simple Moving Average filter (Smith, 1999) was applied to the readings with a window size of 250ms in order to suppress noise and stabilise tracking. A Moving Average filter was found to give acceptable performance (compared to other more complex filters) and was used because of its simplicity. A window size of 250 ms was selected as a good compromise between noise suppression and system responsiveness.

3.2.4 Tracking Algorithm

The algorithms for processing signals from the sensors are programmed in C# with the Phidgets library for the Windows environment. The tracking algorithm below shows how readings from sensors are processed and calculated to determine users' positional information.

Referring to Figure 3.1, taking S_1 to S_{12} as the distance values returned by sensors 1 to 12, $P_n(x)$, $P_n(y)$ as the estimated *x* and *y* co-ordinates of the *n*-th person (from 1 to 6), and *W*, *H*, as the width and height of the screen, the algorithm is as follows:

$$P_1(x) = 0 \tag{1}$$

$$P_1(y) = \begin{cases} S_1 + C/2, & S_1 < H/2 \\ H - S_2 - C/2, & S_1 \geq H/2 \end{cases} \tag{2}$$

where *C* is an average person's width and has a value of 50 cm.

$$P_2(y) = H \tag{3}$$

$$P_2(x) = \begin{cases} S_3 + C/2, & 10 < S_3 \leq 30 \\ S_4 + C/2, & 30 < S_4 \leq W/2 \end{cases} \tag{4}$$

$$P_3(y) = H \tag{5}$$

$$P_3(x) = \begin{cases} W - S_5 - C/2, & 10 < S_5 \leq 30 \\ W - S_6 - C/2, & 30 < S_6 \leq W/2 \end{cases} \tag{6}$$

$$P_4(x) = 0 \tag{7}$$

$$P_4(y) = \begin{cases} S_8 + C/2, & S_8 < H/2 \\ H - S_7 - C/2, & S_8 \geq H/2 \end{cases} \tag{8}$$

$$P_5(y) = 0 \tag{9}$$

$$P_5(x) = \begin{cases} W - S_9 - C/2, & 10 < S_9 \leq 30 \\ W - S_{10} - C/2, & 30 < S_{10} \leq W/2 \end{cases} \tag{10}$$

$$P_6(y) = 0 \tag{11}$$

$$P_6(x) = \begin{cases} S_{11} + C/2, & 10 < S_{11} \leq 30 \\ S_{12} + C/2, & 30 < S_{12} \leq W/2 \end{cases} \tag{12}$$

The tracking algorithm is designed to have the ability of maintaining the user's presence in the space even when the sensor loses signals, e.g., when a user moves away from the table but comes back immediately. This ensures robustness and reliability in situations of erratic and unpredictable movements. A lost signal of > n seconds (typically where n=5s) will suggest that a user has left the table. The system reliably tracks two people on each of the long sides of the table and one person on each of the short sides making it able to track up to six people at any particular time. The method for deciding the number of users on the long side of the table is described next.

Referring to Figure 3.4, the two points labelled as A cm and B cm can be obtained from the sensor setup. Estimating the width of an average person's body[4] as approximately 50 cm, then if *B* - *A* is approximately equal to 50 cm, it can be assumed that there is only one user present on that side of the table. On the other hand, referring to Figure 3.4 (right), if $B - A$ is greater than or equal to twice the width of a person's body, then two users are assumed to be present on that side of the table. The threshold between one and two people is 100 cm. $B - A < 100$ cm tells the table that one person is detected.

[4] The bideltoid shoulder breadth of the 95%ile adult male is 51cm (c. 20") (Stephen Pheasant and Christine M. Haslegrave, 2005). Assuming some comfort distance relating to personal space of 16.5cm (6.5") provides a minimum distance of 67.5cm (26.5") between two individuals.

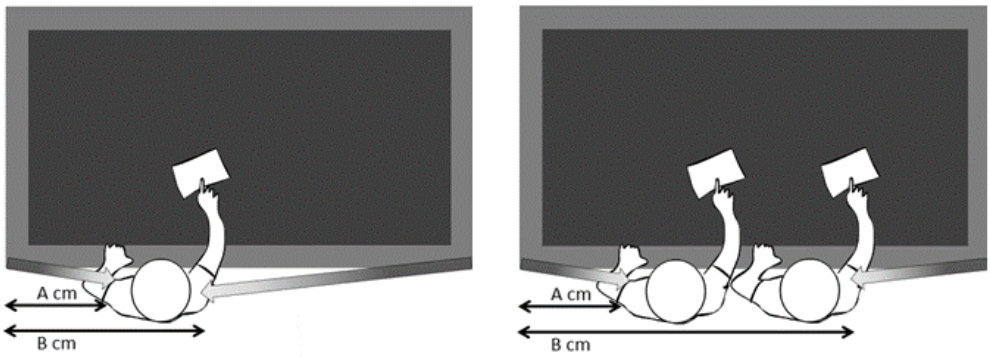


Figure 3.4 When one user is present, B-A should equal to approximate value of an average width of a person's body.When two users are present, B-A equals to more than twice the width of an average person's body

3.2.5 Differentiating Touch

In this section, the ideas of space partitioning and touch differentiation are developed as concepts but not fully implemented. With regards to multi-touch interaction and computer assigned spaces, the distances between spatial touches shall be read by the system as separate touches from different users. These touches allow the system to partition the table into appropriate spaces for each user in relation to the sensor system. This touch differentiation can be approached by implementing two methods. The first method uses a heuristic approach, which is, it can be assumed that a user will touch and perform interaction gestures only to objects within the radius of his reach. Hence a radius of around 40 cm from the body[5] can be defined, so any touches within the radius belong to the user's space (see Figure 3.5 left). For the second method, the touch points outside of the radius can also be mapped to users (see Figure 3.5 right) using a method described next.

[5] A normal working area is typically 35cm to 45cm for a standing operator Stephen Pheasant and Christine M. Haslegrave (2005), so the mid-point of 40cm is chosen.

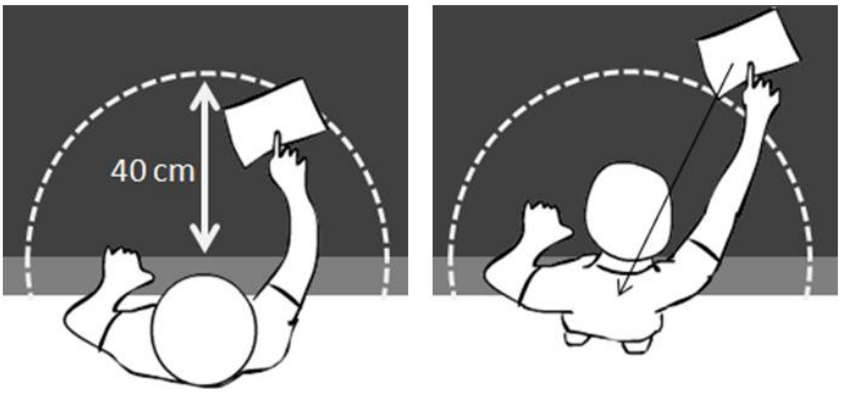


Figure 3.5 Left: Touch within the 40 cm radius from user is assumed to belong to the user.Right: Touch outside the 40 cm radius from user. A method is proposed to compute a touch dragging vector where the owner of the touch can be identified by matching the vector to the user's position.

Previous observations of people interacting with multi-touch tables made during various events and visits at the Chowen and Garfield Weston Foundation Digital Prototyping Hall (people aged 20-60 years old), suggest that users will normally drag an object of interest into their 'personal territory' prior to interacting with the object as a gesture of ownership. This dragging action can be anticipated by assuming a vector from the object towards the table's edge where the user is located. If the vector intersects with the user's position, the system assigns the touch to the user. The accuracy of this identification process can be further increased by measuring the change in user's posture. If the object's position is on the right side of the user, then they would normally move the body a little to the right to reach the object. This small change of positional information can be used to increase the system's performance.

3.2.6 Capabilities and Limitations

The tracking system is currently able to track up to six people simultaneously as illustrated in Figure 3.1. Users are labelled P1, P2, P3, P4, P5 and P6. There is a maximum of 2 persons on each of the long sides of the table, and a maximum of one user on each of the shorter sides. The sensors sense human presence and initiate the tracking algorithm when a new user approaches the table, up to 10 cm in distance.

If more people enter the sensor space the system will track only users nearest to the table (within the casting beams of the sensors). Two persons should be able to work comfortably within the 65” display with 138 cm screen width. Three persons or more will overcrowd the display in terms of working on the table with some comfortable space between users.

We are only interested in users coming close to the table because they are more likely to be the ones who will be interacting with the tabletop display. This was the reason the system was designed in such a way that users who are more than 10 cm away from the table are not detected. From observations during open exhibitions, it can be safely concluded that users standing more than 10 cm away from the table are normally bystanders and not the ones interacting with the table. Users approaching the table are detected from 10 cm away giving ample time for the system to react before they touch the display.

In summary, the tracking performance was very smooth, fast and accurate as long as the users were within 10 cm of the table’s edges. This was demonstrated by a video (Yusof, 2013) showing users interacting with the display. The system continuously tracked users’ positions and was aware when users left the table. It was observed that the tracking system only used 2% of the CPU resource as the result of using a simple algorithm with a small number of sensors.

3.3 Evaluation

The system evaluation is performed by measuring the tracking accuracy. The tracking accuracy of the system has been tested using the method explained in the following section.

3.3.1 System Accuracy Test

The purpose of this test is to confirm that the algorithm is working at its best for all user positions.

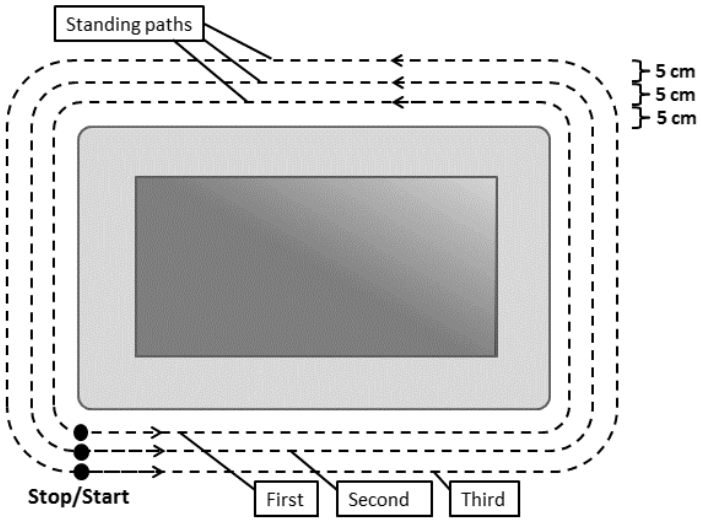


Figure 3.6 The paths where user's moves are indicated by dashed lines.The user will be standing at incremental positions starting at 'Start' along the first path until reaching 'Stop'. This is repeated on the second path and lastly on the third path.

The test began with the user standing at the bottom left corner of the table. The user was then asked to move right on the side by an increment of 5 cm each step. The accuracy of the positioning on the tabletop display in relation to the markers was determined by visual inspection as the error, in cm, between the tracked and actual position of the user. After the first round of measurement, a second test with similar procedure was performed (Figure 3.6) this time at 10 cm away from the table. The process was repeated for the third test, with distance of 15 cm. The whole process was repeated three times and readings were averaged and rounded to nearest integers.

The result of the accuracy test is illustrated in Figure 3.7. It is safe to conclude from the heat map generated that the white areas are the ideal positions where users should be standing in order to get the best tracking accuracy. The positions at the corners of the table indicated by black colour are blind spots for the sensors. This was designed this way so that onlookers standing at the corners will not be confused as active users of the table.

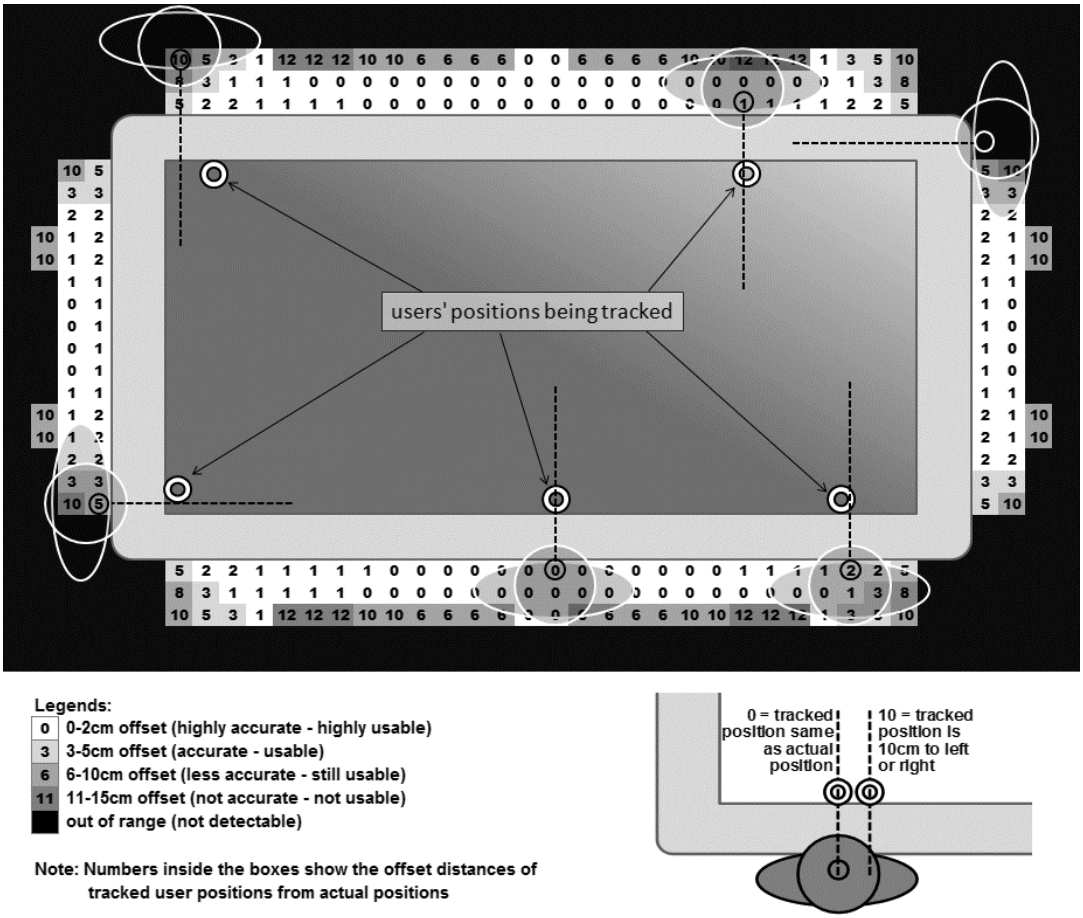


Figure 3.7 Accuracy test results show the optimum and ideal positions (indicated by white area)where users are encouraged be standing on when interacting with the table.

Figure 3.8 illustrates the accuracy of the computed positions when compared to the actual physical body position on one of the table's long sides. The test confirmed that the accuracy of the tracking system was robust when the distance between table's edge and the user was less than 15 cm. Tracking accuracy dropped when a user was at a distance of 15 cm and beyond because of the design arrangement of the sensors.

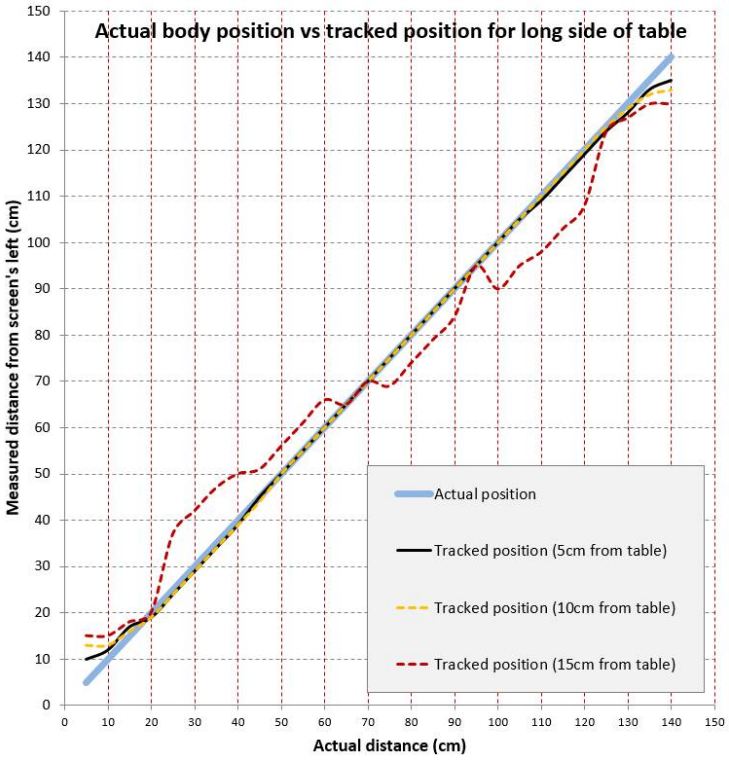


Figure 3.8. Graph showing distance of actual body positions (at hips level) from screen's leftmost, indicated by a straight diagonal line, versus tracked (computed) positions when user moves from left to right on the long side of table.

3.3.2 Prototype Applications

To demonstrate the robustness and reliability of the system, three demo applications were developed. The first one displayed "orb" like objects on the screen in the same position as the users standing at the side of the table. The second application was a personal image gallery application that would ask for a login name to retrieve the personal collection and present it right in front of them. The third application was a derived Pong game. The original Pong game (Wikipedia, n.d.) released in 1972 by Atari Incorporation simulates a table tennis where a player moves a moveable paddle vertically on one of the vertical sides of the screen. The player compete against a computer-controlled opponent or another player controlling a second paddle on the opposite side of the screen. The paddles are used to hit the ball to the opponent's side. Points are earned when the opponent fails to return the ball. The playable online version is made available by Atari on its website (Atari.com, n.d.).

Using users' positional information from the tracking system, paddles were moved to follow users' body positions as they manoeuvred. Up to 6 users were able to play simultaneously. Extracted images from a video recording of the application in action are shown in Figure 3.9.

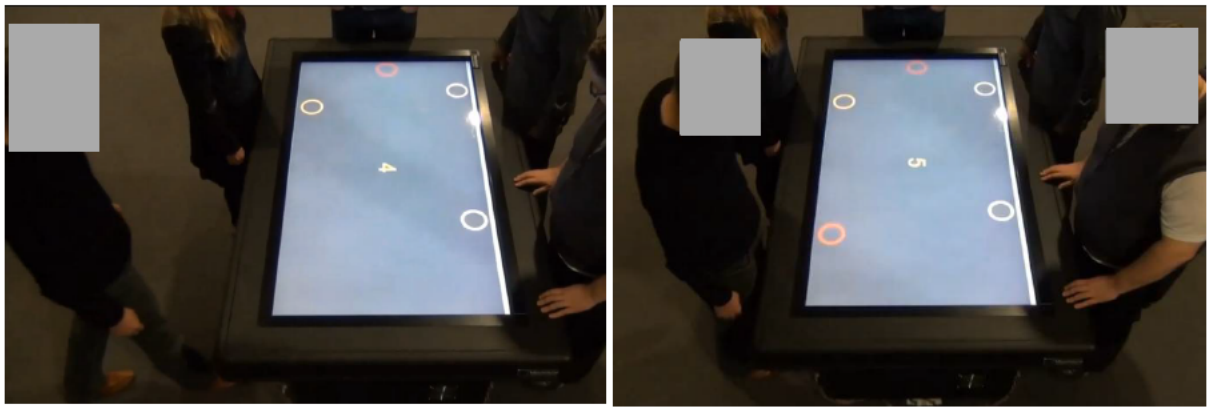


Figure 3.9. Users' positions are being tracked and are indicated by circles.Four users are currently being tracked (*left*). Fifth user joins in and is now being tracked (*right*). This host application is created as a middleware that broadcasts users' positions to other applications

The application in Figure 3.9 initiated the tracking when a user approached the table. Each user was given an orb in front of them indicating their current position. The total number of users currently standing around the table was also displayed.

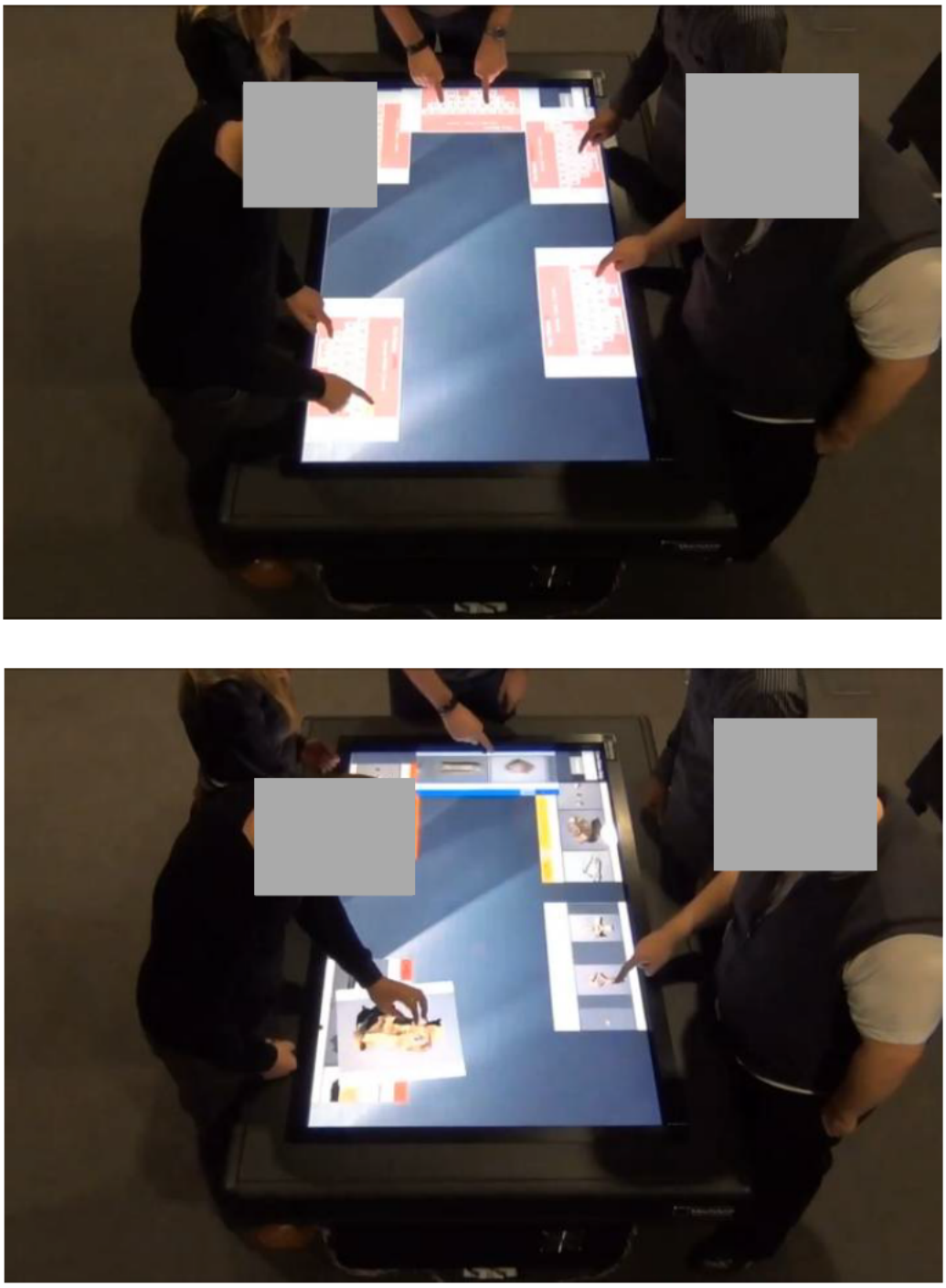


Figure 3.10. Toolboxes appear at users' positions.They enter name information (top) and personal image galleries are presented (bottom) in front of them. They can pull out images from the toolboxes and view details of the images.

When a user was detected (Figure 3.10 (top)) the table responded by displaying a user login dialog box in front of them. The user could shift position to unoccupied areas and the gallery window would follow the user along the edge of the table based on the positional information received from the tracking system. The application demonstrated in Figure 3.10 was a collaboration output with another researcher in the department. It must be mentioned that the interface and content were of his work and the backend communication with the human sensing middleware was of my work. This application used the human sensing system to provide context aware services to the users.

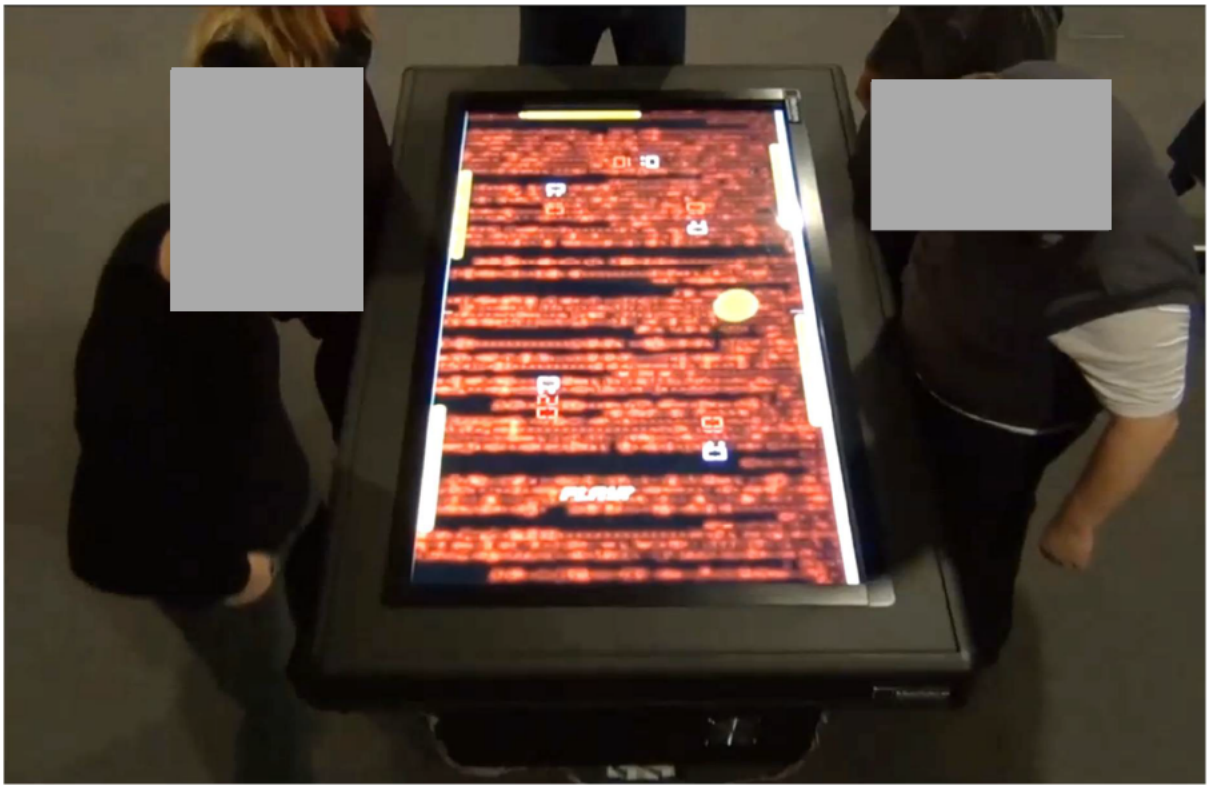


Figure 3.11 "Body Pong" application on the table was a derived Pong game where paddles followed the players' positions.

The robustness and responsiveness of the system was demonstrated by a derived Pong game customised for the system as shown in Figure 3.11. Six players could play simultaneously and each player was given an individual scoring; rewards (based on number of times the ball was saved from hitting the table's edge) and penalties (for number of times the ball hits the edge).

All players were given equal space partitions on the table during play (Figure 3.12: Top left). Individual scores were displayed in front of each player (Reward scores were coloured white, and penalties were red. Rewards and penalties were multiplied by 10, hence "***10:20"*** means 1 reward and 2 penalties).

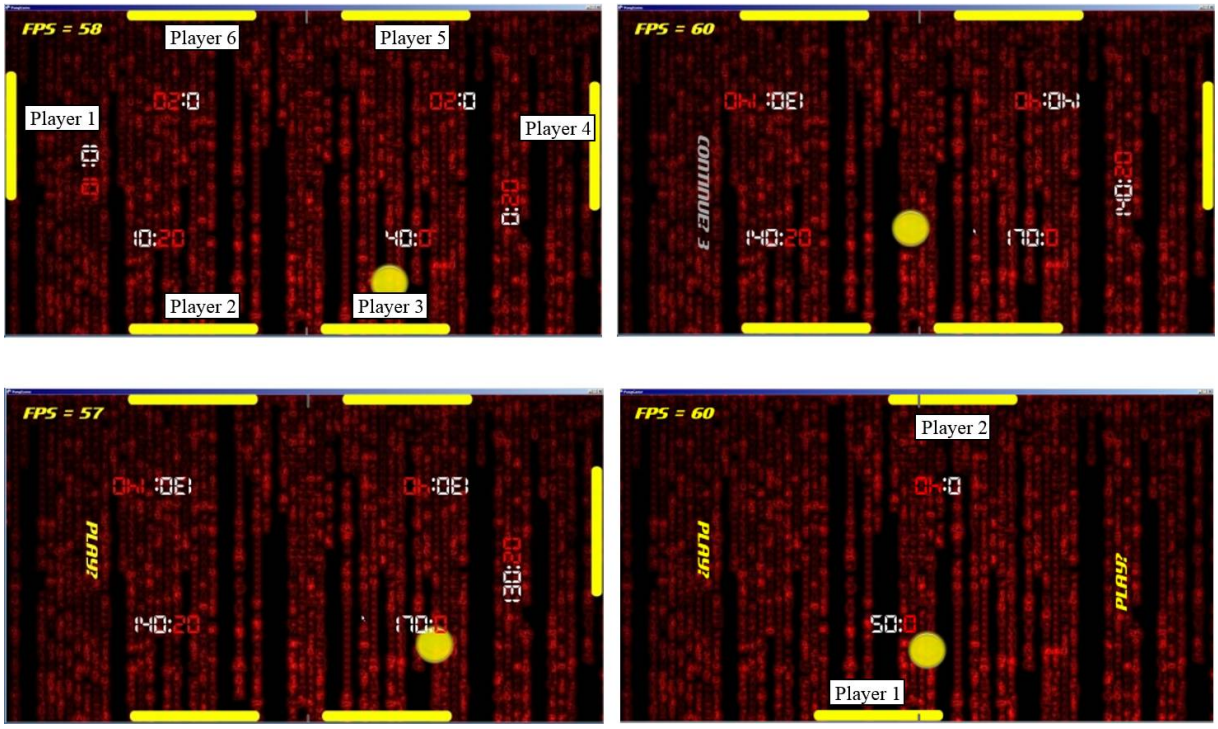


Figure 3.12: (Top left): Six players standing around the table, playing. (Top right): Player 1 stepped back from the table, a "*Continue?*" message was displayed at his position. (Bottom left): 10 seconds has elapsed, a "*Play?*" message was displayed to invite new player to join. (Bottom right): At different time, two players were playing. The game automatically kept scores for Player 1 and Player 2, one on each long side of the table. The same condition applied if they were each on the left and right side of the table.

When a player joined the table, a new paddle and score would be created and positioned in front of the player. New paddles were created when additional players joined in. There was a maximum limit of six players. Each player would interact with the game by moving their body left and right and the game would move the paddle in front of the player's body position. The tracking was instantaneous and continuous throughout the game. If a player stepped out of the game, the message "Continue?" prompted the player for action to continue the game with his

current scores. After 10 seconds had elapsed the message would switch to a "Play?" prompt inviting new players. Figure 3.12 illustrates some of the game scenarios.

3.3.3 Informal Observation

This section presents informal observations and evaluation of behaviours of participants from several "in the wild" events where the Pong game application was demonstrated. A formal and quantitative assessment of user behaviour is beyond the scope of this thesis, but could be one of the aims of future research. The tabletop's Pong game was co-located with the events listed below:

i. The Digital Humanities Hub's Open Day 2013 (28 people aged 20-50 years old).
ii. Birmingham Science and Art Festival 2013 (16 people aged 20-45 years old).
iii. CAKE (Collaboration and Knowledge Exchange) event at Chowen and Garfield Weston Foundation Digital Prototyping Hall (11 people aged 30-45 years old).

It was found that the users' past experiences with touch-based interfaces were transferred. When visitors first played the game, there was a tendency for some to attempt to move the paddle by touching it even though they had been briefed earlier on the body-movement interactions. After a short period of time (2-3 minutes), they eventually got used to the new style of interaction. It was also observed that they swayed their hips left and right to move the paddles instead of sidestepping. Once they got used to the interaction mechanism, excitement and discussions among them followed.

All initial users were surprised at the new interaction modality and expressed excitement at the possibility of application areas. The responses were positive such as: "This is cool!", "This could be turned into an exercise game", "I want this game at my home", "This is weird, swaying

my hip left and right like this, but it's pretty cool", "This is so much fun", "I wish I could have more time playing this", "This is very clever", and "Wow, I like it!". Most comments were made by different people during their initial experiences with the game. On average, the users played for 4-5 minutes before moving on to check out other exhibits in the hall.

3.4 Conclusions

In this chapter, a human-aware display for sensing and tracking users around a multi-touch tabletop was proposed. The setup used a small array of sensors coupled with algorithms developed for the sensing system. Methods were formulated to associate touches with users who initiated them during the interaction with the tabletop display based on anthropomorphic measures and the characteristic movements of a user's body.

The implications of the robustness, reliability, accuracy, and cost effectiveness that have been demonstrated showed many advantages in the system's potential application areas. The system is simple to setup and relatively portable, hence could be replicated easily, allowing exploration of new interaction modalities on tabletop displays.

Potential application areas include the personalisation of services or content with which users could have their personal space on the tabletop, with customised user interfaces when they 'sign in' with their profile. This will be useful in public learning and teaching spaces such as GLAMs. Digital contents (images, videos, 3-D objects) explored on the table by users could be assigned individual ownerships within their 'personal territory', and contents could follow as users move around the table. Objects owned by a user can be locked to the current user's personal space and will not be accessible unless they deliberately hand ownership to another user. This human sensing system could also be used for greeting people as they approach the table. If there is an empty personal 'territory' on the display, users observing from a distance could be coerced to

come forward to use the space, with follow-on tutorials, such as available interaction styles and gestures. This is similar with how modern computer games integrate a walk-through tutorial for new users. Other users will learn these gestures from earlier users through observation.

Other potential application areas include games on a multi-touch table that uses the body's relative position on the sides of the table as the controller. For example, an obstacle avoidance game could use body's position to steer left and right to avoid obstacles. Other example such as a "breakout" game could use body's position to control the paddle's movement, a method similar to that used in the derived Pong game. A breakout game released in 1976 influenced by the 1972 Atari arcade game Pong was built by Steve Wozniak aided by Steve Jobs. The game features a layer of bricks arranged in the top third of a screen. As a ball bouncing off the top and side walls of the screen, it destroys a brick that it hits. A player must prevent the ball from touching the bottom of the screen by moving a paddle across the bottom of the screen. The ball bounces upward when it hits the movable paddle (Wikipedia.com, n.d.). The playable online version of the game is available at Atari's website (Atari.com, n.d.).

The configuration of the proposed system is very cost effective both in construction price and computational power (i.e. resource usage) when compared to existing systems reviewed in the literature. The simple configuration of the system allows easy replication and setup on other tabletop displays. It can be said from experience that this system can be installed on a new tabletop display within an hour. This is believed to be unlikely possible with other systems that are considerably more complicated as reviewed earlier. The proposed system is capable of tracking users' positions at higher resolution when compared to systems by Annett et al. (2011) and Klinkhammer et al. (2011). The resolution is defined as the number of discrete positions on the side of the table. The proposed system performs very responsive tracking while maintaining very low CPU usage of 2% allowing use of the spare computing power for other

multimedia-based tasks, for example, rendering rich media such as hi-resolution images, high definition videos and 3-D objects.

CHAPTER 4:

Person Re-identification

4.1 Introduction

Person re-identification is the process of matching observations of individuals across multi-camera spaces. Unlike biometric techniques, re-identification relies on appearance information alone. It has many applications in surveillance and security as well as in human-robot interaction and in the personalisation of services in smart environments (Han et al., 2012). For example, interactive services in retail environments, museums, art galleries and public spaces. However, because different views of the same person can appear substantially dissimilar, viewpoint invariant re-identification is a challenging problem.

The performance of re-identification methods using anthropometric features has been shown to improve with the inclusion of colour information (Kviatkovsky et al., 2013a). However, this performance is dependent on viewpoint and can degrade significantly when colouring is non-uniform. For example, patterned clothing, backpacks, long hair and opened jackets can all produce significant differences in colour between front and rear views (Albiol et al., 2012). Viewpoint invariant, multi-modal methods have not been reported in the literature hence the research contribution in this area.

The proposed viewpoint invariant multi-modal (ViMM) feature descriptor is comprised of parameters describing colour, anthropometric properties and body orientation. ViMM vectors estimated from observations of different body orientations form the training set for a neural network classifier. These vectors have strong discriminant properties, regardless of viewpoint.

4.1.1 Biometrics and Soft Biometrics

The identity of a person is often described by their appearance measures. Alphonse Bertillon introduced, in 1883, a method of identification based on biometrics such as the colours of the eyes, hair, beard and skin, as well as shape and size of the head Rhodes (1956). He also used anthropometric measures like height or weight, and description of permanent marks such as birth marks, scars or tattoos (Dantcheva et al., 2010). Jain et al. (2004b) first described soft biometrics as a set of characteristics that gives clue about the appearance of an individual but is not enough to authenticate the person due to lack of distinctiveness and permanence. They later added that soft biometrics require low computational power, can be acquired from a distance involuntarily, and can help to narrow down a search from a population of people (Jain et al., 2004a). Soft biometry have also been defined as human physical and behavioural characteristics classifiable in pre-defined human compliant categories where traits are created in a natural way and used for differentiating individuals (Dantcheva et al., 2010).

4.1.2 Short term and Long term Re-identification

Short term re-identification deals with the problem of re-identifying people using a camera after they have left the field of view and then require re-identifying when he or she comes back into sight. The person is assumed not to have changed clothes and the time span between the first and second sights typically is within a period ranging from a few seconds to as long as few hours. For example a person visiting a museum, library, gallery space or shopping mall falls into this category.

Long term re-identification is a more challenging problem, described as the ability of a system to re-identify a person usually days after the first sight. It is more difficult than “short term” re-identification due to the likelihood of the clothes changing hence the inability of the system to

use colour information which has been the most used feature in re-identification because of its highly discriminant property.

4.2 The Challenges in Person Re-identification

There are generally two main categories of challenges when modelling and designing a person re-identification system (Gong et al., 2014b): Feature Representation, and Model and System Design.

4.2.1 Feature Representation

The most critical and challenging part when developing a person re-identification system is to design a strong and robust feature representation of persons. The observed appearance of a person can change at any time during the course of a camera's observation, such as:

- changes in illumination causing the person's skin and clothes to appear darker or brighter
- changes in viewpoint causing the person to look different if seen from different angles
- background clutter causing contamination of pixels compared to the original appearance of the person if segmentation is not properly done
- occlusion from other people or objects causing the appearance to be different than the complete appearance
- low image quality/resolution causing loss of information or details

An ideal feature representation for person re-identification should be discriminatively powerful and robust to the changes mentioned earlier, however it is not known if there exists such features (Gong et al., 2014b). Research in this field has been trying to improve the quality of feature

representations for person re-identification, aiming to be both discriminative and robust to changes in illumination, viewpoint, background clutter, occlusion and image quality or resolution.

4.2.2 Model and System Design

During the model and system design phase, challenges that can arise include the following:

i. Inter and intra class variations: A “class” here corresponds to a “person”. Inter class variation is where different persons can look similar across camera views whereas intra class variation is where the same person can look different when viewed under different conditions such as different body pose, body orientation, ambient lighting and clothing appearance. These variations are what make the re-identification problem complex in general and difficult for a model to learn.
ii. Small sample size: Good models generally require multiple training data representing variations of person’s features. A small sample size reduces the ability of a model to handle intra-class variability.
iii. Data labelling requirement. Good models are normally trained using data from two or more cameras to handle cross-camera view variations. However the data collection process can become very expensive for a place with large camera network. It is therefore desirable to have a good model that can be trained using less training data.
iv. Generalisation capability: Intra class variations are normally observed on two or more cameras from different locations. Models trained for a specific camera mostly do not generalise well to other cameras with different viewing conditions (Gong et al., 2014b). Good models, when trained once, should have the ability to handle intra class variations.

v. Scalability: The size of search space for person matching can be said to be directly proportional to the size of areas covered by a camera network since large areas cover more people, therefore matching methods need to compare test data with features of all people in the dataset. Consequently test time will also be increased. It is important, especially for real-time systems, that a good model can avoid long response time resulting from increased search space and having to process more video streams,

vi. Long-term re-identification: This is defined as the capability of a model to correctly re-identify a person after a period of time usually a day or more after the last observation. The challenge is to design feature representations that are robust to appearance changes caused by clothes changing and carried objects.

4.2.3 Data and Evaluation

Person re-identification can be used in two types of scenarios, 1) open-world, and 2) closed-world. Most existing person re-identification methods are designed for closed-world scenarios where the gallery and probe sets are assumed to contain the same people (Zheng et al., 2015). Open-world re-identification represents a more realistic scenario especially for surveillance applications in open environments and are more difficult to solve. The accuracy and effectiveness of re-identification systems are measured with a number of metrics. “Rank-1” accuracy and the “Cumulative Match Characteristics (CMC)” curve are two most commonly used metrics. “Rank-1 accuracy refers to the conventional notion of classification accuracy: the percentage of probe images which are perfectly matched to their corresponding gallery image. High Rank-1 accuracy is notoriously hard to obtain on challenging re-id problems.” (Gong et al., 2014b). The CMC curve summarises the percentage of correct match appearing in the top 1, 2, …, N (i.e. rank-1, rank-2, …, rank-N) of the ranked list.

4.2.4 Benchmark Datasets

Various datasets for person re-identification are publicly available such as VIPeR (Viewpoint Invariant Pedestrian Recognition) (Gray et al., 2007), i-LIDS pedestrians (Zheng et al., 2009), ETHZ (*Eidgenössische Technische Hochschule Zürich)* (Schwartz and Davis, 2009) and many others. The Table 4.1 summarises the details of each dataset.

Table 4.1. Details of various public datasets for Person Re-identification

	Capturing device	Multi camera	No. of people	Short-term / Long-term	Avg. no of samples per person	Size (pixels) W x H	No. of images	View-points	No. of video sequences	Viewpoint variations	Pose Varia-tions	Scale Varia-tions	Ligh-ting varia-tions	Remark
VIPeR (Gray et al., 2007)	RGB Camera	✓	632	Short-term	2	48 x 128	632 x 2 = 1264	2	-	✓	✓	×	-	- Single shot
i-LIDS MCTS (Zheng et al., 2009)	RGB Camera	✓	119	Short-term	4	64 x 128	479	-	-	✓	✓	×	-	- not fit well in multi-shot scenario.
ETHZ 1 (Schwartz and Davis, 2009)	RGB Camera	×	83	Short-term	-	32 x 64	4857	-	-	✓	✓	✓	✓	- pedestrian images captured from head-height moving camera
ETHZ 2 (Schwartz and Davis, 2009)	RGB Camera	×	35	Short-term	-	32 x 64	1936	-	-	✓	✓	✓	✓	- pedestrian images captured from head-height moving camera
ETHZ 3 (Schwartz and Davis, 2009)	RGB Camera	×	28	Short-term	-	32 x 64	1762	-	-	✓	✓	✓	✓	- pedestrian images captured from head-height moving camera
RGBD-ID (Barbosa et al., 2012)	Kinect v1	×	79	Long-term	20	1280 x 960 (colour)	79 x 20 = 1580	2	-	✓	-	-	-	- entering and leaving a laboratory
KinectREID (Pala et al., 2015)	Kinect v1	✓	71	Short-term	-	1280 x 960 (colour)	-	3	7 x 71 = 483	✓	-	-	-	- taken at a lecture hall
BIWI RGBD-ID (Munaro et al., 2014a)	Kinect v1	✓	50	Long-term	-	1280 x 960 (colour)	-	Uncon-strained	2 x 50 = 100	-	-	-	-	- taken in a laboratory space
CAVIAR4REID 1 (Cheng et al, 2011)	RGB Camera	×	22	Short-term	10	17 x 39 72 x 144	-	-	22	-	✓	✓	-	- real scenario
CAVIAR4REID 2 (Cheng et al, 2011)	RGB Camera	✓	50	Short-term	10	17 x 39 72 x 144	-	-	50	-	✓	✓	-	- real scenario - pose variations severe
KinectV2 RGBD-ID * newly created for this thesis.	Kinect v2	✓	64	Short-term	~300	1920 x 1080 (colour)	-	Uncon-strained	14 x 64 = 896	✓	✓	✓	-	- taken in a prototyping hall resembling public space with multi-touch displays.

4.3 Current Hardware and Sensors

4.3.1 2D Colour Camera

The most common capture devices in person re-identification research are two-dimensional colour video cameras. The many types of colour cameras used to create datasets in re-identification research include closed-circuit television (CCTV) or surveillance cameras, web cams, consumer video cameras, and mobile phone cameras. Colour cameras are often characterised by their capabilities to record images in terms of frame rate (fps) and pixel resolution. Depending on the camera's intended purpose, the video resolution and fps can vary widely, for example CCTV cameras typically record VGA quality, 640 pixels wide by 480 pixels tall (640 x 480) at 30 fps or lower such as 15 fps to minimise storage requirement. However as storage is becoming cheaper nowadays, it is not uncommon for CCTV cameras to record full high definition quality at 15 or 30 fps depending on the criticality requirements of the installation. For example traffic monitoring CCTV cameras might only need to record at 15 fps, whereas cameras installed for in-shop surveillance may need to record at 30fps to provide better visuals for analysis in case of robbery. For sports purposes, action cameras are usually built to have a capability to record in high quality images at high frame rate. It is usually preferred to record in Full HD quality 1920 x 1080 pixels at high frame rate such as 60 fps or even 120 fps as quality of images is of more priority.

For the purpose of creating datasets for person re-identification and evaluating re-identification methods, colour cameras have been used to record appearances of person(s), in controlled and uncontrolled settings. A person's appearance may vary subject to illumination changes which can be caused by ambient light change or different camera observation angle. In person re-

identification, images taken by colour cameras are known to be susceptible to illumination variance. This problem demands for effective illumination invariant solutions.

4.3.2 Depth Camera – Kinect V1 vs Kinect V2

The first generation Microsoft Kinect was introduced in November 2010, designed as a motion sensing input device for Microsoft XBOX 360 gaming console. It tracks player's motions via depth imaging using structured light emitted by the sensor. PrimeSense[6] (now a subsidiary of Apple Inc.) is the patent owner of this technology. Kinect's data is streamed via USB 2.0 to the host PC. This technology was much more economical than conventional time-of-flight (ToF) setups at the time. Kinect devices consist of one RGB camera supporting a resolution of 640 × 480 pixels at 30Hz, or 1280×960 pixels at 15Hz, one infrared projector, and one infrared camera with a resolution of 640×480 pixels at 30Hz or 1280×960 pixels at 10Hz (Berger, 2013). The Kinect also has a microphone array and an accelerometer. Kinect operates best indoors because sunlight causes interference to the infrared projected from the sensor. The depth reading also is not reliable for regions that are further than 4 meters away. The Kinect SDK for Windows for Kinect v1 has been made available for free download to allow developers to create applications using C++/CLI, C# or Visual Basic .NET and the latest version is 1.8 (Microsoft, n.d.).

[6] https://en.wikipedia.org/wiki/PrimeSense

Figure 4.1 Kinect version 1 (left) and Kinect version 2 (right)

The proprietary software that operates the Kinect is capable of processing a full body 3-D motion capture, performing facial recognition as well as voice recognition. Up to six people can be tracked simultaneously (i.e. body positional information), but only two persons closest to the sensor will be fully tracked with motion analysis details such as joints information. There are 20 joints features supported by Kinect v1 such as illustrated in Figure 4.2 below.

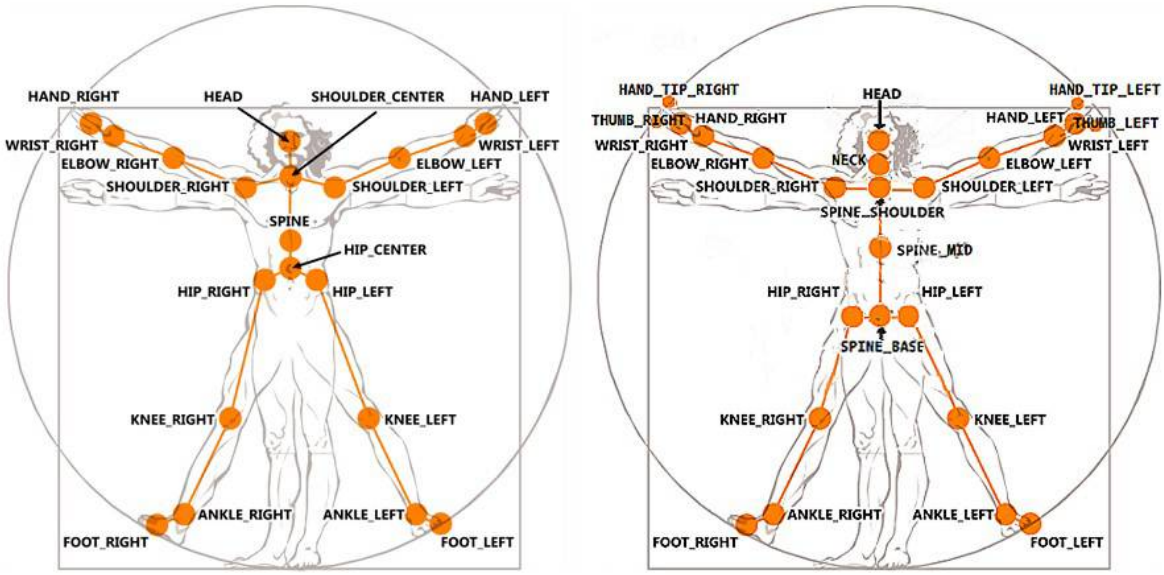


Figure 4.2 Orange blobs represent 20 joints tracked by Kinect v1 (Microsoft, n.d.)(left) and 26 joints by Kinect v2 (Microsoft, n.d.)(right)

The second generation Kinect for XBOX One announced in May 2013 was an upgraded sensor from the version 1. The new Kinect v2 features a wide angle colour camera with full HD resolutions (1920 x 1080) at 30Hz and depth sensor-based on a much powerful time-of-flight technology with a resolution of 512 x 424 at 30Hz. It processes 2 gigabits of data per second to

read its environment thus a PC needs a supported USB 3.0 chipset to be able to work with Kinect v2 because of a huge data transfer bandwidth requirement. The new Kinect is much more accurate than its predecessor; it can track up to 6 bodies simultaneously, all with their own joint information as well as a depth frame for each person. Additional features such as heart rate tracking, facial expression and weights on limbs are available on the new sensor. Table below shows the main difference between Kinect v1 and Kinect v2.

Table 4.2 Features comparison between Kinect v1 and v2

Feature	**Kinect for Windows v1** (Microsoft, n.d.)	**Kinect for Windows v2** (Microsoft, n.d.)
Colour Camera	640 x 480 @30 fps (10 x 10 pixels per degree) 1280 x 960 @15 fps	1920 x 1080 @30 fps (22 x 20 pixels per degree)
Depth Camera Native Resolution: Interpolated Resolution:	 320 x 240 @30 fps (5 x 5 pixels per degree) 640 x 480 @30 fps	 512 x 424 @30 fps (7 x 7 pixels per degree) -
Max Depth Distance	~4.0 meters (3.0 meters in near mode)	~4.5 meters
Min Depth Distance	0.8 meters (0.4 meters in near mode)	0.5 meters
Horizontal Field of View	57^{O}	70^{O}
Vertical Field of View	43^{O}	60^{O}
Tilt Motor	Yes	No
Skeleton Joints Defined	20 joints	26 joints
Maximum Human Recognized	6 persons	6 persons
Full Skeleton Tracked	2 persons	6 persons
USB Standard	2.0	3.0
Supported OS (minimum)	Win 7, Win 8	Win 8

It must be mentioned that the depth images of the Kinect v1 and Kinect v2 cannot be compared directly using resolution figures. The reason is Kinect v2 measures each pixel in the 512 x 424 depth image individually using a high precision measuring device resulting a much more accurate and robust distance estimate (Z-coordinate) from the Kinect sensor (Lachat et al.,

2015) while the depth image of the Kinect v1 uses a structured light mechanism resulting in an interpolated depth image that is based on a lower number of sample points than the native (actual) depth image resolution.

4.4 The Proposed System - Viewpoint Invariant Multi-modal (ViMM) Person Re-identification

The proposed ViMM feature descriptor combines 2-D and 3-D anthropometric measurements, body orientation and colour-based appearance descriptors. The architecture of the proposed system is illustrated in Figure 4.3 below. The parameters making up the ViMM vector are also detailed in this section.

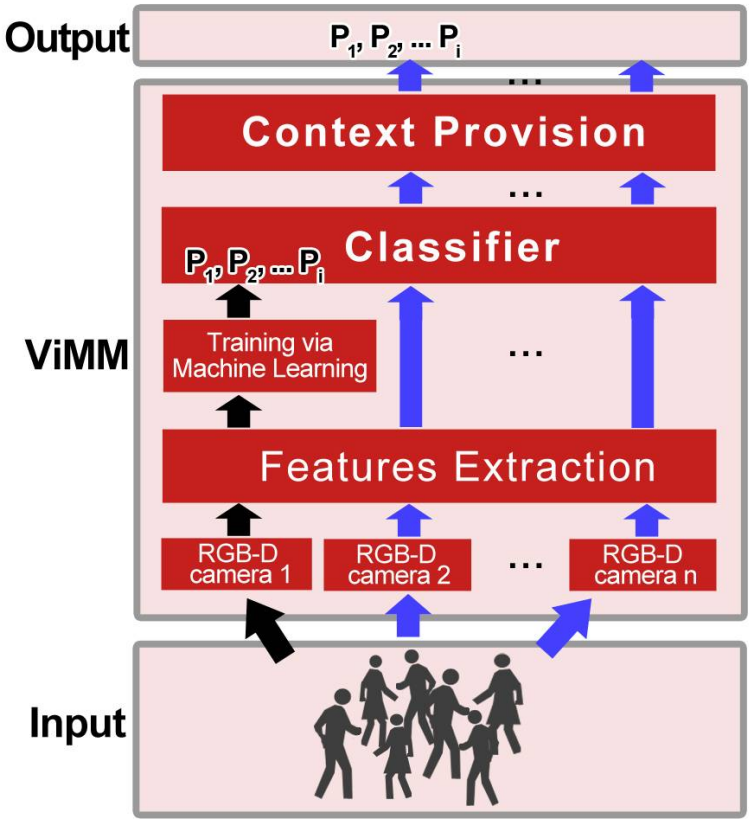


Figure 4.3 Architecture of the proposed re-identification system

4.4.1 Hardware Setup

A 3-D depth sensor/camera system was set up using a Microsoft Kinect version 2 for Windows sensor connected to a notebook computer with Intel quad-core 2.40 GHz i7 CPU and 16GB RAM via a USB 3.0 port. This version 2 sensor was chosen because of its availability and low cost in addition to being superior in terms of colour and depth pixels quality compared to version 1. The Kinect contains a depth camera that allows positions of a person and the body parts in 3-D space, to be recognised when standing or moving in front of it. It also has a full colour camera built in so a full high definition colour image of a person can be extracted from the sensor's video stream.

4.4.2 Human Appearance Model

A standing human body was modelled as a set of vertically stacked elliptical bands perpendicularly arranged as illustrated in Figure 4.4. Other shapes such as super-ellipses which might more accurately model the shape of a person's cross-section were also considered, but were not suited to the fitting of partial sets of points. Ellipses were fitted to shoulder, mid spine, and hip cross-sections. These upper body parts' cross-sections were selected because of their ease of extraction from unconstrained poses and their improved chance of continuous visibility compared to the lower body parts. However the inclusion of the lower body parts such as knees and feet will be discussed in Section 4.8 for further exploration.

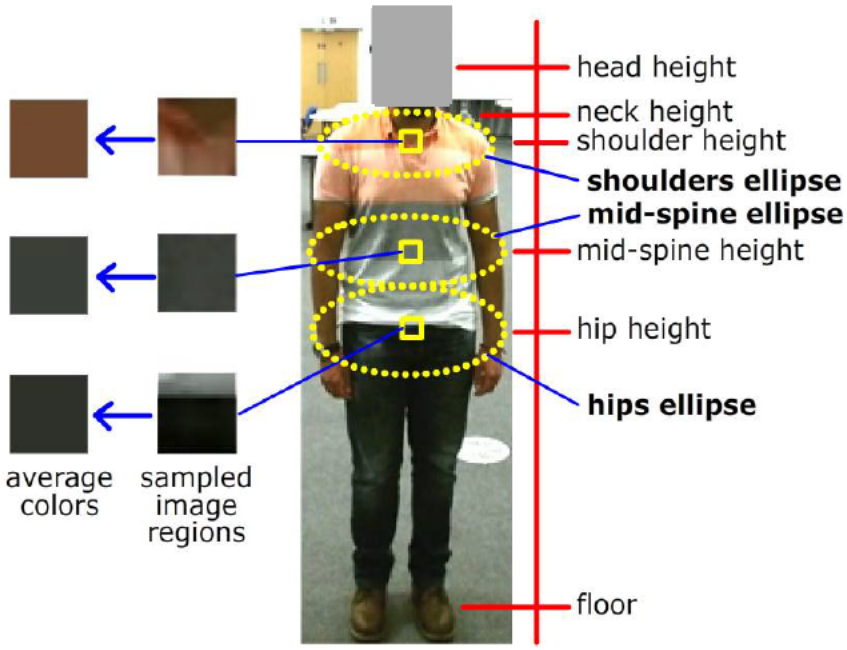


Figure 4.4 The ViMM body model. (Best viewed in colour)

The human appearance model gave rise to a comprehensive list of candidate features such as given in Listing 1. This initially considered feature descriptor was called ViMM v0 (version 0) before the trimmed down version called ViMM was selected for the purpose of the research in this thesis.

Listing 1

ViMM v0 = (θ || d || s || c || c'|| c'')
ViMM = (θ || d || s || c)

where,

θ	= angle of body orientation
d	= vector of distances: (d_1, d_2, d_3, d_4, d_5)
s	= vector of ellipses: (s_1, s_2, s_3, s_4, s_5, s_6)
c	= vector of colours: (r_1, g_1, r_2, g_2, r_3, g_3)
c'	= vector of colours: (r_4, g_4, r_5, g_5)
c''	= R_1, G_1, B_1, R_2, G_2, B_2, R_3, G_3, B_3, R_4, G_4, B_4, R_5, G_5, B_5
d_1	= distance between floor and head
d_2	= distance between floor and neck
d_3	= distance between floor and shoulders
d_4	= distance between floor and mid-spine
d_5	= distance between floor and hips
s_1	= semi-minor axis of shoulders
s_2	= semi-major axis of shoulders
s_3	= semi-minor axis of mid-spine

s_4 = semi-major axis of mid-spine
s_5 = semi-minor axis of hips
s_6 = semi-major axis of hips
r_1 = red component of rg-chromaticity at shoulder height
g_1 = green component of rg-chromaticity at shoulder height
r_2 = red component of rg-chromaticity at mid-spine height
g_2 = green component of rg-chromaticity at mid-spine height
r_3 = red component of rg-chromaticity at hip height
g_3 = green component of rg-chromaticity at hip height
r_4 = red component of rg-chromaticity at knee height
g_4 = green component of rg-chromaticity at knee height
r_5 = red component of rg-chromaticity at ankle height
g_5 = green component of rg-chromaticity at ankle height
R_1 = Red component of RGB at shoulder height
G_1 = Green component of RGB at shoulder height
B_1 = Blue component of RGB at shoulder height
R_2 = Red component of RGB at mid-spine height
G_2 = Green component of RGB at mid-spine height
B_2 = Blue component of RGB at mid-spine height
R_3 = Red component of RGB at hips height
G_3 = Green component of RGB at hips height
B_3 = Blue component of RGB at hips height
R_4 = Red component of RGB at knee height
G_4 = Green component of RGB at knee height
B_4 = Blue component of RGB at knee height
R_5 = Red component of RGB at ankle height
G_5 = Green component of RGB at ankle height
B_5 = Blue component of RGB at ankle height

The list above was reduced to a smaller subset, such that the features are extractable in a targeted scenario commonly found in interactive environments, i.e. presence of multi-touch tabletop displays. In this scenario, lower body parts will be most likely not visible to a camera when a person is behind a tabletop display. This was the reason why only the upper body parts were considered for ViMM, because of their high chance of continuous visibility compared to the lower body parts. However the inclusion of lower body parts will be discussed in Section 4.8 for further exploration. Until then, ViMM will be the main point of discussion in this thesis.

4.4.3 Person Detection

Person detection was accomplished using Kinect SDK v2.0 which automatically detects and tracks a maximum of 6 people within the maximum range and field of view of the sensor. “Skeleton” data from people in the sensor’s view was captured giving the positions of each joint or body part as extracted from the video frames by the Kinect API. The lengths between each joint of body parts were calculated based on their 3-D positions. The colour frames of the video were also recorded.

4.4.4 Body Parameters and Orientation Estimation using Ellipse Fitting Algorithm

The method estimates body parameters such as depths and breadths of body segments and body orientation from an ellipse fitting algorithm.

Listing 2. Collect depth points at shoulder level that belong to a person.

```
for (int y = 0; y < depthFrameHeight; ++y)
{
    for (int x = 0; x < depthFrameWidth; ++x)
    {
        int depthIndex = (y * depthFrameWidth) + x;

        byte player = _bodyData[depthIndex];

        if (player != 0xff) //pixel belongs to a human (0xff is 255)
        {
            if (y == (int)joints_dep[shoulderLevel].Y)
            {
                shoulderDepth[x] = depth;
            }
          }
    }
}
```

To fit ellipses to the body, depth points are obtained from a Microsoft Kinect V2 RGB-D camera. The C# snippet shown in Listing 2 is used to collect depth points at shoulder level. Only the part of the body facing the camera is visible. Consequentially, only a partial elliptical arc can be obtained from a single frame. A complete ellipse can be estimated from a partial arc

using an ellipse fitting algorithm based on Direct Least Squares method by Halir and Flusser (1998). This algorithm is simple, stable and robust making it suitable for real-time application purpose. It is also non-iterative and based on least squares minimization that guarantees an ellipse-specific solution even for scattered or noisy data. This feature is important for the feature extraction process where a strictly elliptical solution is required for every depth frame received from the sensor. The paper provides a MATLAB implementation for the algorithm that fits an ellipse in a Canonical form. The general equation for any conic section is

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$$

where *A, B, C, D, E* and *F* are constants. A conic's shape changes if the values of some of the constants are changed. The conic's type is determined by equations below:

If $(B^2 - 4AC) < 0$, if a conic exists, it will be either a circle or an ellipse.

If $(B^2 - 4AC) == 0$, if a conic exists, it will be a parabola.

If $(B^2 - 4AC) > 0$, if a conic exists, it will be a hyperbola.

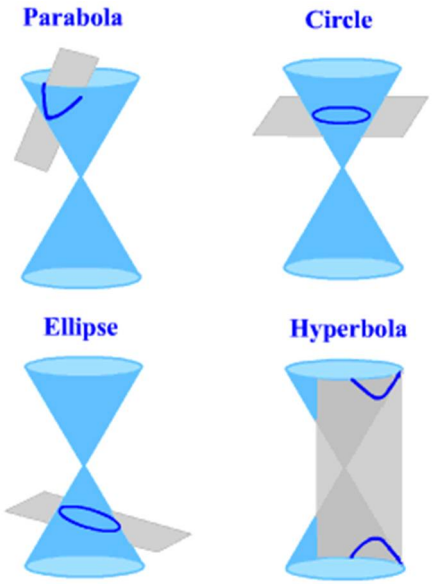


Figure 4.5 Different types of conics can be produced by changing the angle and location of the intersection. Figure reproduced from (Hotmath.com, n.d.)

Listing 3. MATLAB implementation of the direct ellipse-specific fitting algorithm, proposed by Fitzgibbon *et al* in Fitzgibbon et al. (1996)

```
1 function a = fit_ellipse(x, y)
2 D = [x.*x x.*y y.*y x y ones(size(x))];              % design matrix
3 S = D' * D;                                         % scatter matrix
4 C(6, 6) = 0; C(1, 3) = 2; C(2, 2) = -1; C(3, 1) = 2; % constraint matrix
5 [gevec, geval] = eig(inv(S) * C);                   % solve eigensystem
6 [PosR, PosC] = find(geval > 0 & ~isinf(geval));     % find positive eigenvalue
7 a = gevec(:, PosC);                                 % corresponding eigenvector
```

Listing 4. MATLAB implementation of the improved version of Fitzgibbon *et al*'s ellipse-specific fitting algorithm in Listing 3, proposed by Halir and Flusser (1998)

```
1 function a = fit_ellipse(x, y)
2 D1 = [x .^ 2, x .* y, y .^ 2];       % quadratic part of the design matrix
3 D2 = [x, y, ones(size(x))];          % linear part of the design matrix
4 S1 = D1' * D1;                       % quadratic part of the scatter matrix
5 S2 = D1' * D2;                       % combined part of the scatter matrix
6 S3 = D2' * D2;                       % linear part of the scatter matrix
7 T = - inv(S3) * S2';                 % for getting a2 from a1
8 M = S1 + S2 * T;                             % reduced scatter matrix
9 M = [M(3, :) ./ 2; - M(2, :); M(1, :) ./ 2]; % premultiply by inv(C1)
10 [evec, eval] = eig(M);                      % solve eigensystem
11 cond = 4 * evec(1, :) .* evec(3, :) - evec(2, :) .^ 2;      % evaluate a'Ca
12 a1 = evec(:, find(cond > 0));       % eigenvector for min. pos. eigenvalue
13 a = [a1; T * a1];                   % ellipse coefficients
```

It should be mentioned that the implementation in Listing 4 only gives the estimates for parameters A, B, C, D, E and F of a Conic equation. A C# implementation of the MATLAB code by Srikanth (Srikanth Kotagiri, n.d.) has been adapted for use in the feature extraction module. Features such as angle of rotation, semi minor and semi major axes cannot be obtained directly from the parameters A, B, C, D, E and F above. The C# code was then extended by rewriting the MATLAB implementation of (Ohad Gal, n.d.) to finally complete the accumulation of the features. Listing 5 in the Appendix shows the complete ellipse fitting function written in C#.

This algorithm is applied to fit the shoulder, mid-spine and hip ellipses where the semi-minor and semi-major axes for each shoulder, mid-spine and hip, are computed to become part of the feature descriptor.

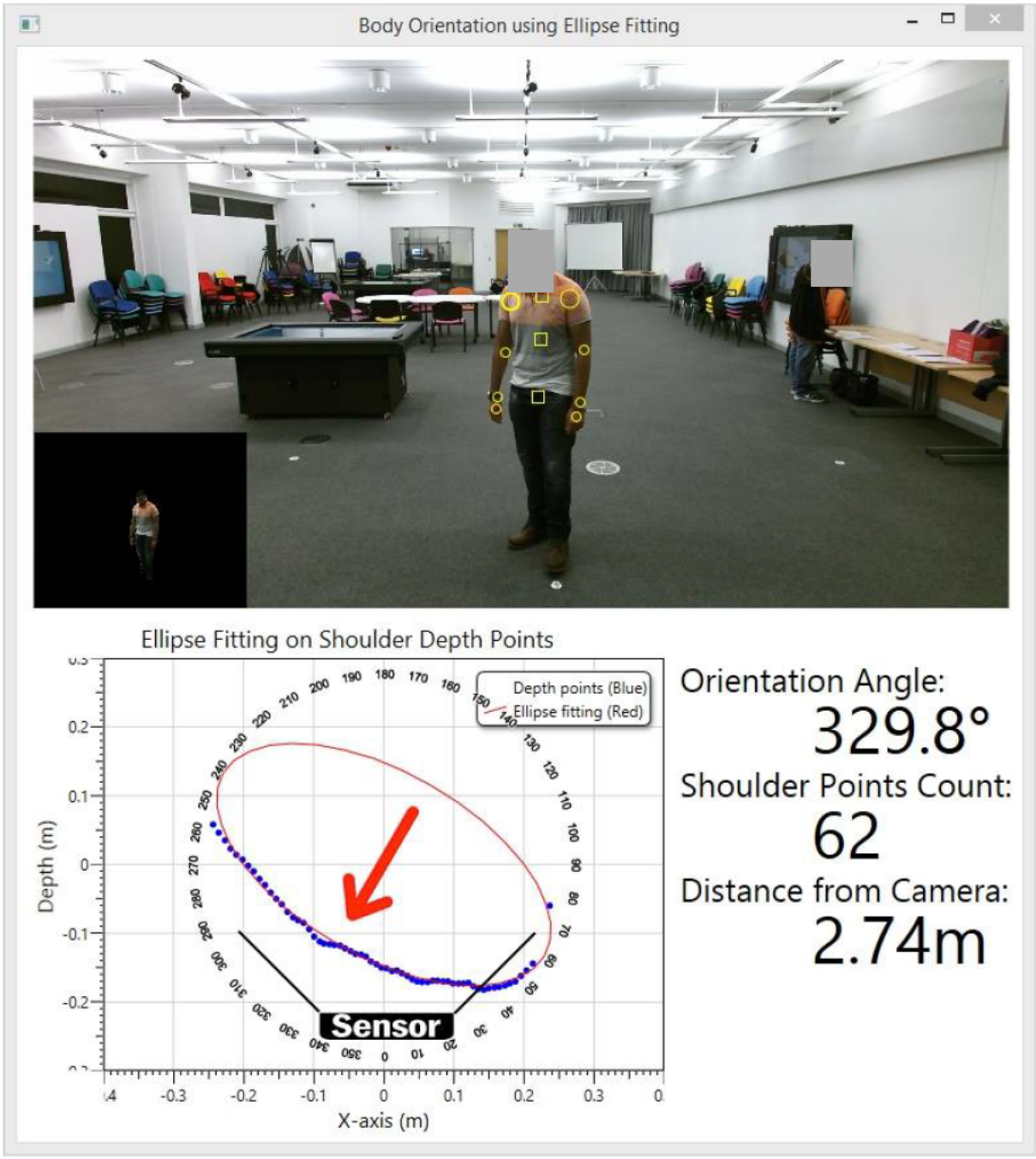
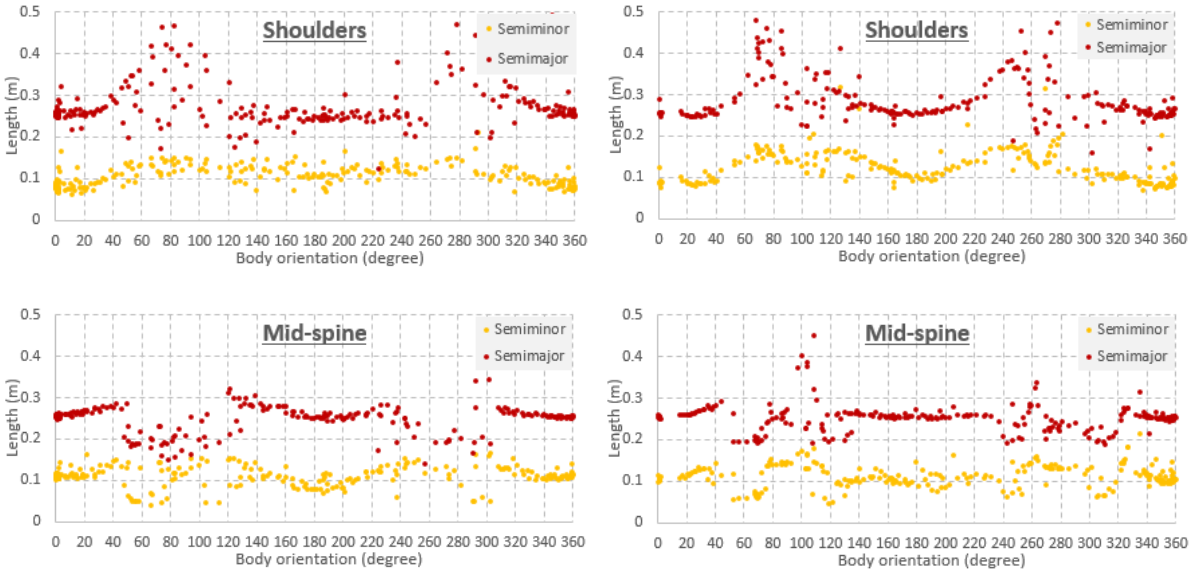


Figure 4.6 Ellipse fitting (shown in red) is performed on partial depth points (shown in blue)at 330° body orientation relative to the camera's viewing angle.

The proposed method estimates body orientation directly from the angle of the major axis of the shoulder ellipse. In testing, the shoulder ellipse provided a consistently more accurate and robust orientation estimate than the other ellipses. The forwards-backwards ambiguity of body orientation is resolved by complementing the ellipse fitting algorithm with the face tracking tools provided by the Microsoft Kinect SDK V2.0 (Microsoft, 2015).

The cross-sections of real people are not perfect ellipses. As a result, the estimated ellipse parameters define an approximation to the cross-section which varies depending on body orientation. To accommodate this variation in the classification process, body orientation is also included in the ViMM feature vector. Viewpoint invariance is achieved by training the classifier to recognise the estimated ellipse parameters of individuals as a function of body orientation. Figure 4.7 shows examples of the different sizes of ellipses estimated at different body orientations for two subjects. The resulting angle defining body orientation becomes one of the features of ViMM.



(a) Person A (b) Person B

Figure 4.7 Seminminor axes and semimajor axes of ellipses fitted around shoulders (S_1, S_2) and mid-spine (S_3, S_4) of (a) person A and (b) person B for all body orientations providing part of a defining body signature. Direct least squares ellipse fitting algorithm is used to estimate the ellipses from depth points. (Best viewed in colour)

Figure 4.8 shows the resulting ellipse dimension of shoulders, from the fitting algorithm which is quite close to an actual person's body dimension. It should also be noted that the estimated measurements with the grid-background wall are subject to small parallax errors. The actual measurements can be assumed to be smaller than those stated in the figure. This makes the

dimension estimation from the ellipse fitting method to be very close to the actual, if not accurate.

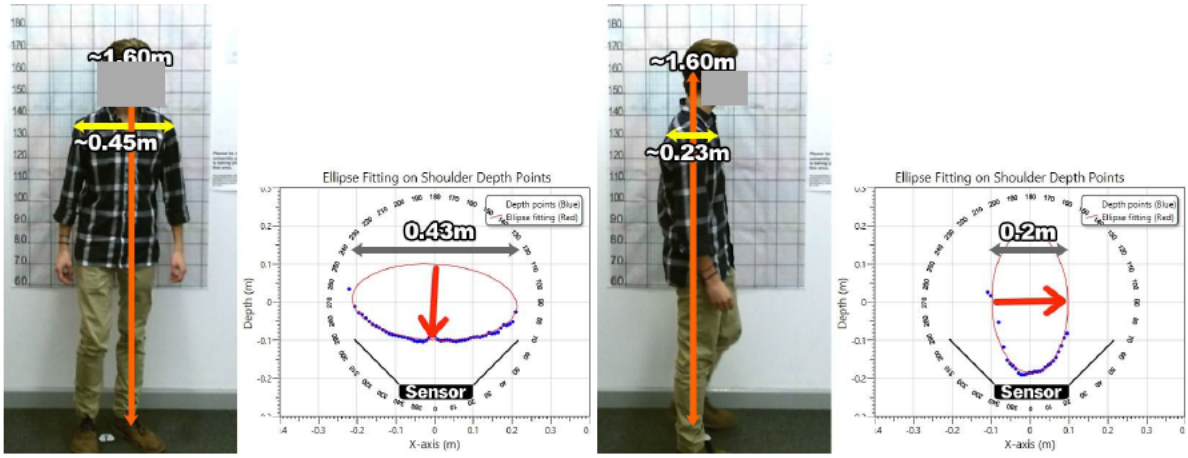


Figure 4.8. The dimension of the estimated ellipse is quite close with the width and thickness of a person's body at shoulder level. Head height is measured from the floor to the centre of the head.

4.4.5 Orientation Estimation using Joint Orientation from Kinect SDK

Alternative method of estimating body orientation is by using joint orientation structure from a body detected by Kinect SDK 2.0. The joint information is provided in the form of a quarternion W, X, Y and Z, W relating to the angle of rotation and X, Y, Z the axis of rotation. Figure 4.9 below illustrates the quarternion.

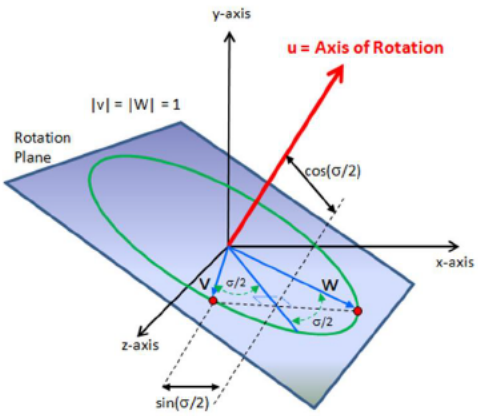


Figure 4.9 Quaternion in the direction of the three axes of rotation (x, y, z) and an angle of rotation (w)

Referring to Figure 4.9, it is clear that the rotation axis of a human body is equivalent to the y-axis, and W value can be taken directly from the joint orientation structure of Kinect SDK, which has values ranging from 0 to 1 representing 0 to 180^{O} counter clockwise, and 0 to -1 representing 0 to 180^{O} clockwise. This can be illustrated with a diagram in Figure 4.10.

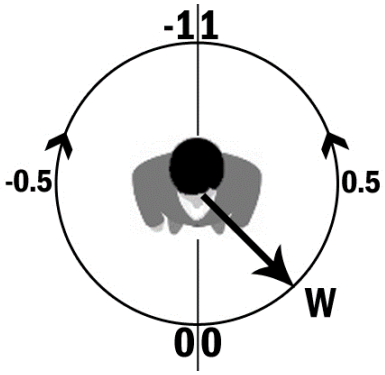


Figure 4.10 Joint orientation of spine-base, mid-spine, spine-shoulder and neck indicated by W component of a quarternion.

To get a body orientation angle in a form similar to the one used in Figure 4.6, the following formula is used.

```
//joint orientation of spine-base at index 2
JointOrientation orientation = body.JointOrientations.Values.ElementAt(2);

if (orientation.Orientation.W >= 0)
{
    bodyOrientationAngle = (orientation.Orientation.W / 1) * 180;
}
else
{
    bodyOrientationAngle = 360 + ((orientation.Orientation.W / 1) * 180);
}
```

Tests showed that this method does not give accurate reading of orientation angle because the joint orientations are too sensitive when occlusion of joints occur. It also does not handle very well when a person is back-facing the sensor. This can be expected because the Microsoft's skeletal tracking was trained with examples of frontal view people only (Munaro et al., 2014a). These conditions are depicted in Figure 4.11.

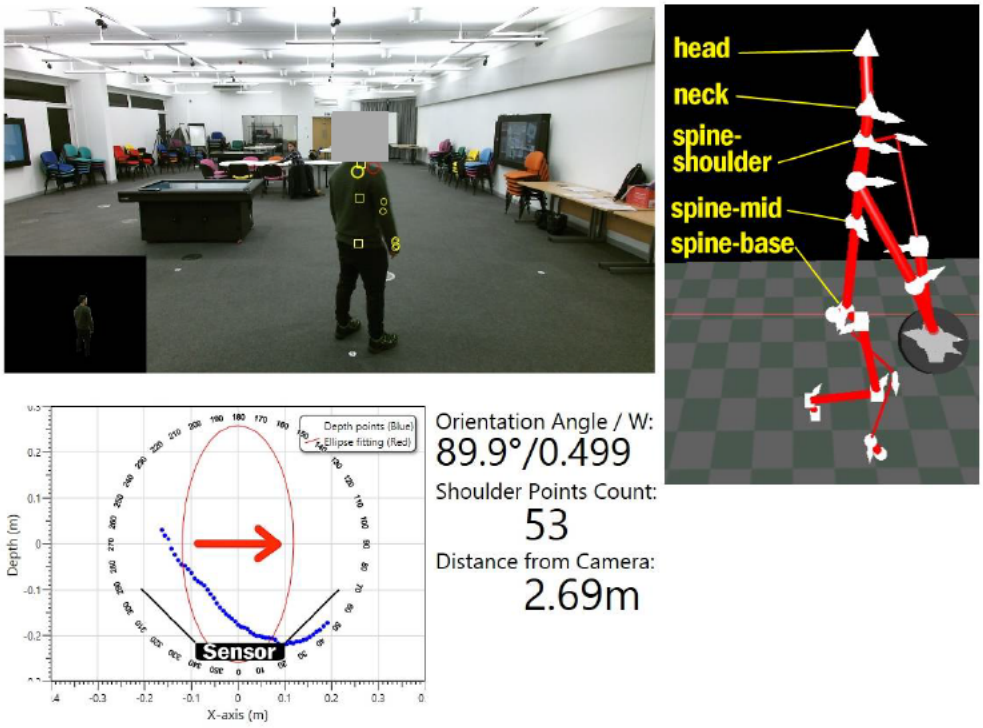


Figure 4.11 **Joint orientations** from Kinect SDK stuck at 90^{0} when a person go back-facing the sensor, causing inaccurate body orientation estimate. Plots in blue are the actual depth points at shoulder level.

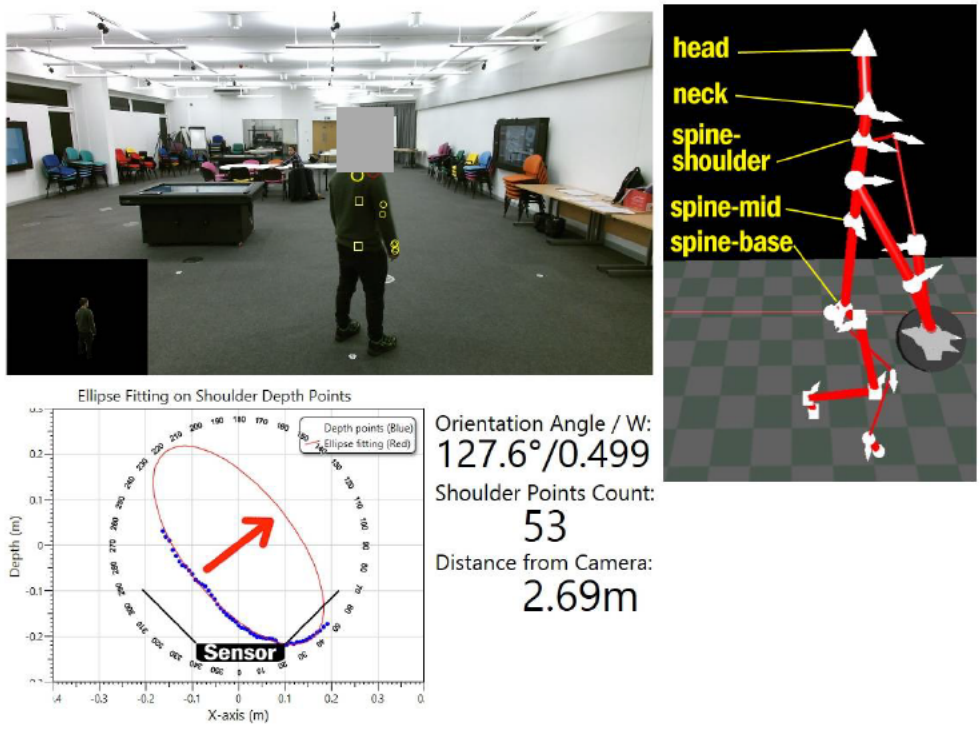


Figure 4.12 **Ellipse fitting** method giving accurate body orientation when a person go back-facing the sensor. This is the same frame as used in Figure 4.11. Plots in blue are the actual depth points at shoulder level.

4.4.6 Accuracy of Ellipse Fitting method on Body Orientation Estimation

To evaluate the accuracy of the ellipse fitting method, the body orientation angle was tracked for the whole duration of a person's movement performing turning in a spot, 360^{O} clockwise and 360^{O} in the opposite direction. This action will later be referred as "Turning 1". The red plot in Figure 4.13 shows the original estimates with some occurrences of error caused by value jumps because of face detection state going from "detected" to "undetected" and from "undetected" to "detected" for certain body orientations. The cause of this error is supported, as can be seen in the figure, by the value difference between the false measurements (red) from the actual (blue) is 180^{O}. A filter is then applied to the original measurement to obtain clean and accurate angle estimation for all body orientations as shown in blue in Figure 4.13.

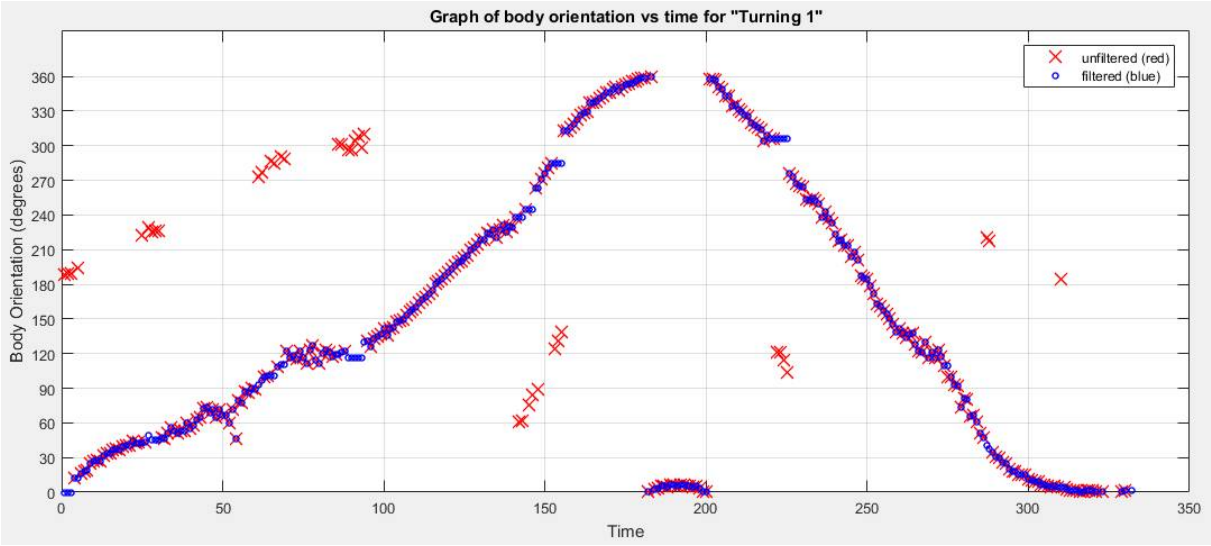


Figure 4.13 Body orientation angle plotted vs time (frame number) during "Turning 1" activity shown in red "x" for unfiltered output and blue circle for filtered output of the body orientation algorithm.

The action "Turning 1" mentioned previously and the ellipse fitting method working can be visually illustrated by screen grabs of the application in Figure 4.14. In this figure, the orientation angle estimates can be seen to work perfectly for all body orientation angles.

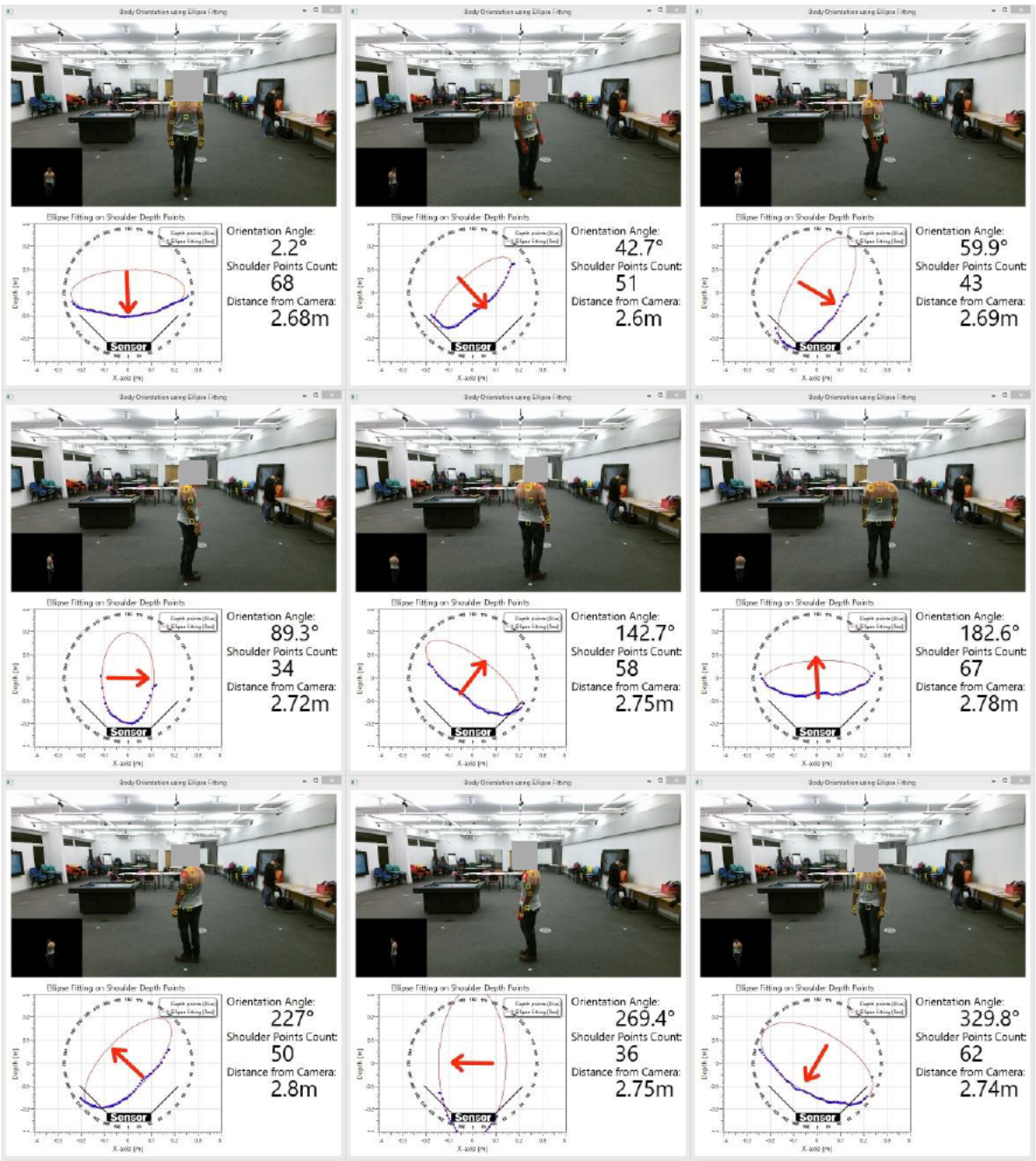


Figure 4.14 Series of frames extracted from a dataset component "Turning 1" for a participant and their corresponding ellipses fitted on shoulders' depth points.

The accuracy of the ellipse fitting method is further evaluated on a more natural action. The orientation angle is tracked on a person walking freely in random directions, within an area, for 15 seconds. This action is called “Free Walking 1”. The tracking plot is shown in Figure 4.15 below. It can be seen that the orientation angle estimates work well for this action too.

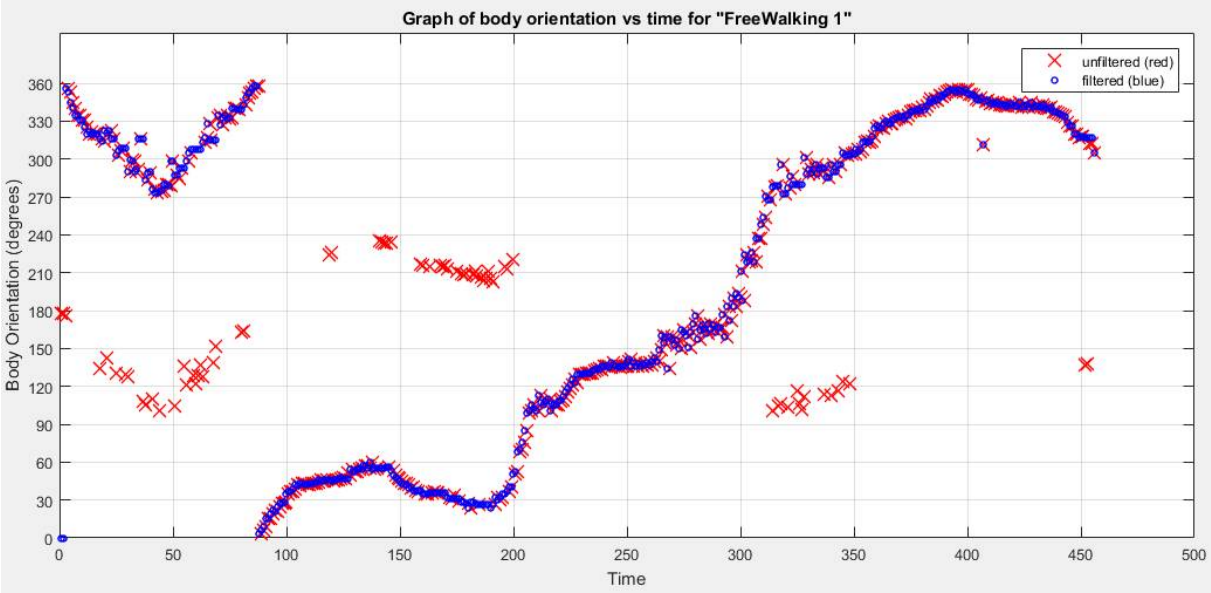


Figure 4.15 Body orientation angle plotted vs time (frame number) during “Free Walking 1”activity shown in red “x” for unfiltered output and blue circle for filtered output of the body orientation algorithm.

Figure 4.16 below illustrates the "Free Walking 1" action and the ellipse fitting method in working.

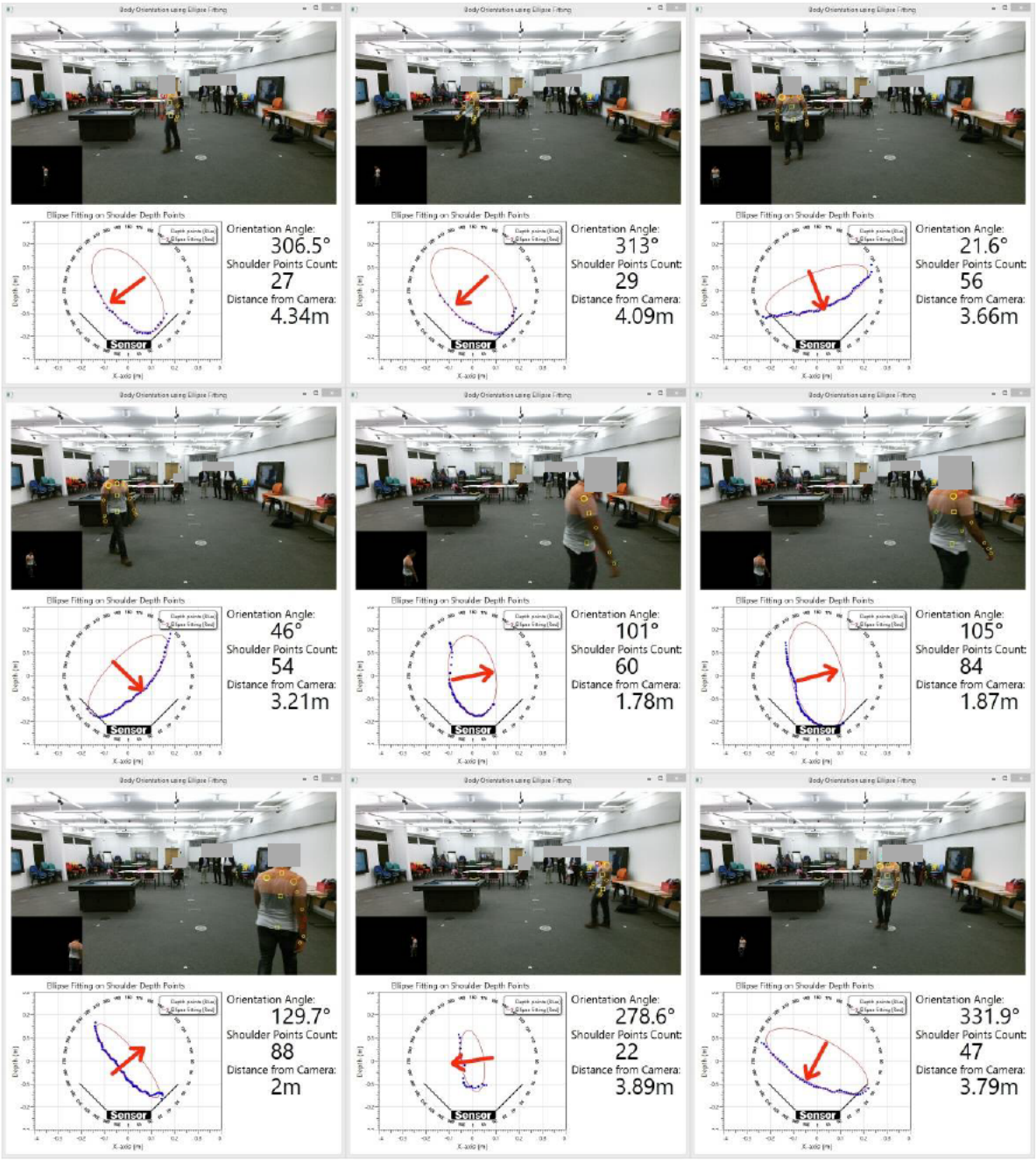


Figure 4.16 Series of frames extracted from a dataset component "Free Walking 1"for a participant and their corresponding ellipses fitted on shoulders' depth points

4.4.6.1 Challenging Cases

The ellipse fitting method works well for most situations giving accurate estimation of body orientation angle especially when smooth depth points at shoulder level, representing partial ellipse are obtained. However there are some conditions where depth points at shoulder level are not smooth and in some cases may lead to the wrong representation of the actual shape of shoulders. Consequently the estimated body orientation angle becomes less accurate or in the worst case, can become completely wrong. Some of the conditions are person wearing a coat with open collars such as illustrated in Figure 4.17, person wearing a sweatshirt with a hood (Figure 4.18) and person carrying a bag (Figure 4.19).

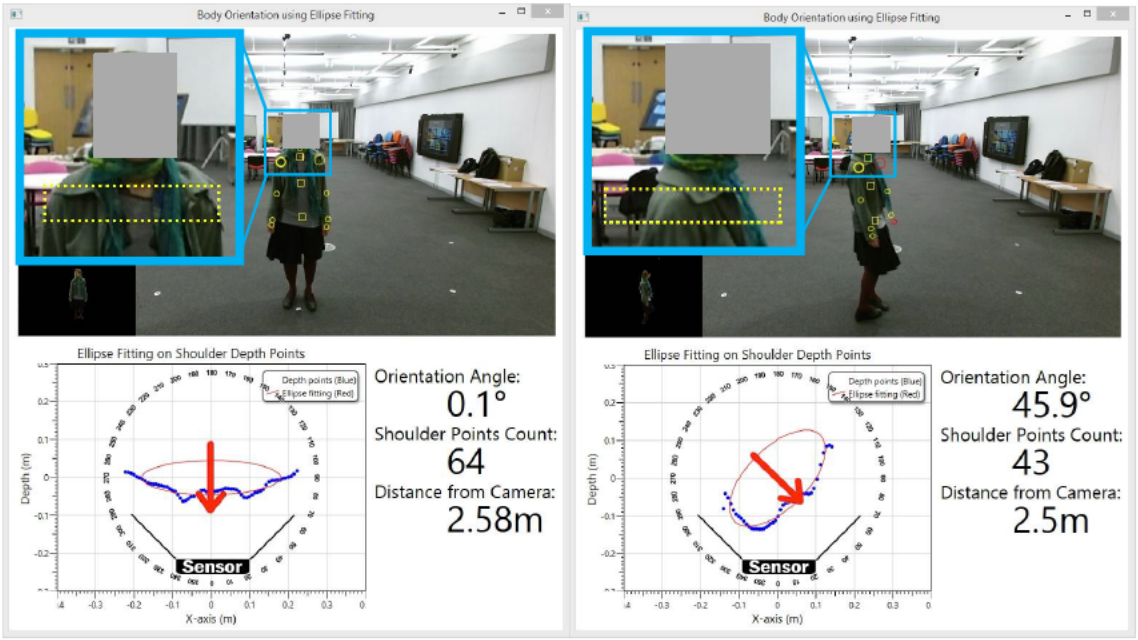


Figure 4.17 Protruding shape caused by open coat collars still give reasonably good ellipse estimate and correct orientation angle.

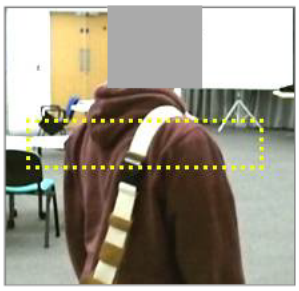


Figure 4.18 A hood from sweatshirt also causes protruding shape to the depth points, visible when a person is facing away from the camera. The depth points collected at shoulder level for this image are shown in Figure 4.19.

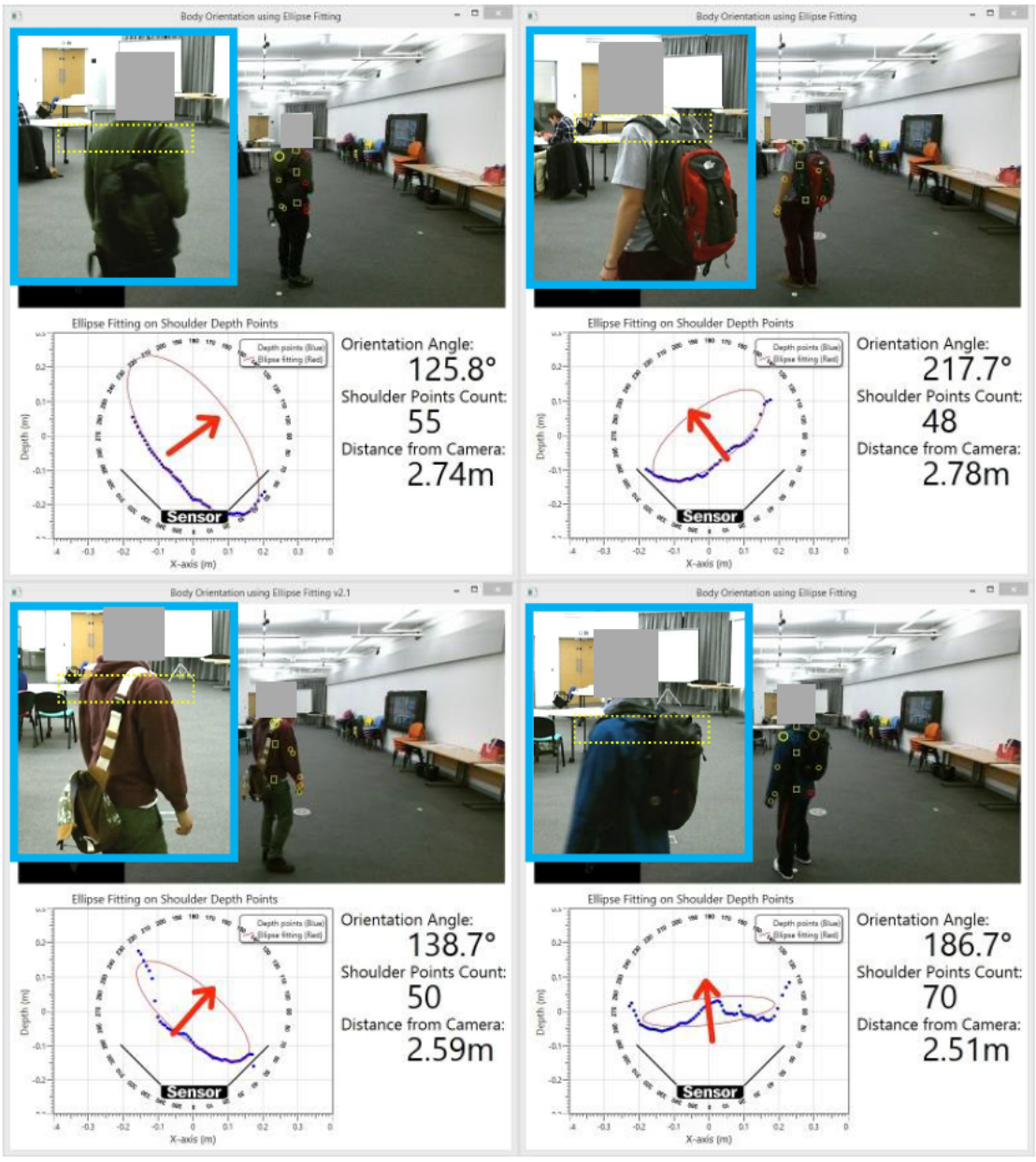


Figure 4.19 Screen grabs showing different common styles of carrying backpacks. The last style affects the performance of the ellipse fitting because the bag is positioned very high and the shoulder's depth points are disturbed too much by the points returned off the bag.

4.4.7 Face Tracking in Darkness

As mentioned in the previous section, the forwards-backwards ambiguity of body orientation was resolved using face detection in the face tracking tools provided by the Microsoft Kinect SDK V2.0 to complement the ellipse fitting algorithm. The new face detection performance is much improved over the previous version for Kinect v1. It stores face location in 3-D world coordinates, face orientation, and basic expressive information such as happy and engaged. It can also tell if the person wears glasses. All this information is computed from the infrared frame hence is available in any lighting conditions, even in complete darkness up to a distance of 3.5 meters (Microsoft, 2015). Figure 4.20 shows how Kinect's face tracking still tracks the face under very low lighting condition with very wide coverage of face orientation.

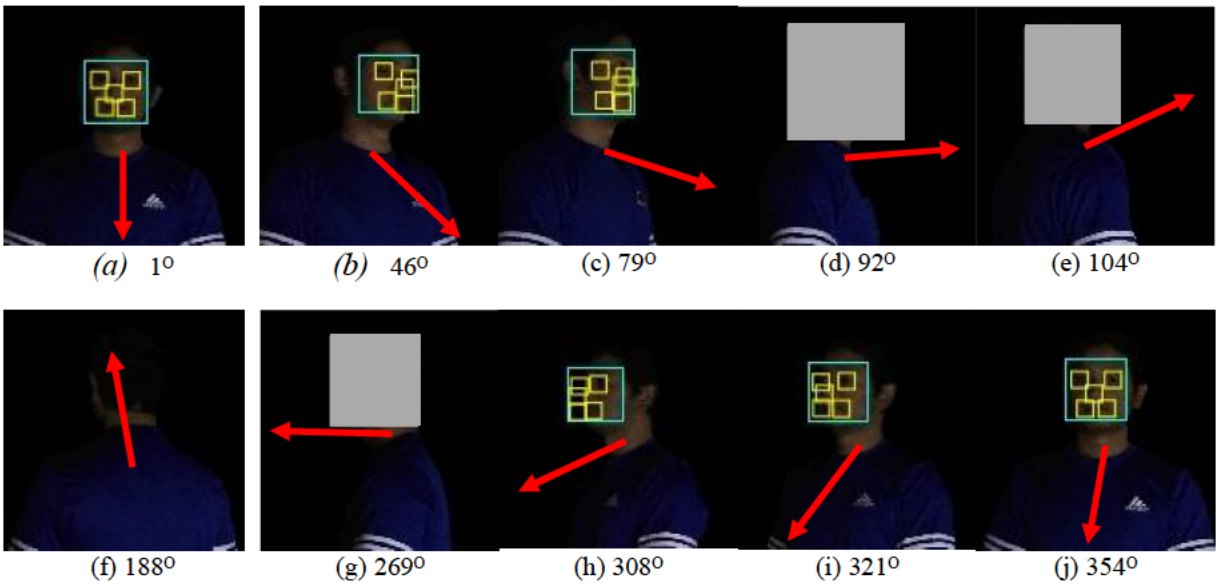


Figure 4.20. Face tracking works in very low lighting conditions. White rectangles indicate faces being tracked. Red arrows indicate the angle orientations of the person.

4.4.8 New RGB-D Datasets

The new datasets were acquired using Kinect V2 cameras and Microsoft Kinect for Windows SDK 2.0. Extraction and processing were performed using applications written in C#. The Kinect V2 is superior in terms of the accuracy of depth data (Clark et al., 2015) compared to

the earlier Kinect V1. The skeletal tracker in Kinect SDK v2.0 for Kinect v2 sensor tracks more joints than Kinect SDK v1.8 for Kinect v1 sensor, 26 vs 20. The new skeletal tracker provides much more accurate and stable joint positions which is very important to allow consistent measures of different observations of the same person. This significant improvement is important for the feature extraction method because the accuracy and reliability of the ellipse estimation depends on the accuracy of the depth images.

Previously published datasets for person re-identification have mostly been acquired using RGB cameras, for example (Gray et al., 2007), (Satta et al., 2012), (Zheng et al., 2009) and (Schwartz and Davis, 2009). There are, as yet, no publicly available datasets that match the thesis' purposes i.e. that are acquired using Kinect v2 cameras, provide both RGB and depth frames, and include all-round views of individuals. For example, the "RGBD-ID" dataset from Barbosa et al. (2012) could not be used because it used the Kinect V1 camera and consists of only 5 RGB-D frames per individual for limited body orientations i.e. facing the camera ±5° and facing away from the camera ±5°. In addition, the blurring of faces in the dataset obviates the use of face detection required by the proposed method.

The "BIWI RGBD-ID" (Munaro et al., 2014b) is another dataset acquired using Kinect V1 cameras. It was designed for long-term re-identification and includes individuals wearing different clothes in different acquisitions. The "KinectREID" dataset developed by Pala et al. (2015) was also acquired with the Kinect V1 and comprises only three viewpoints.

A new dataset is, therefore, needed given the unavailability of existing datasets suitable for the purposes of this research. This new dataset was named "KinectV2 RGBD-ID". Experiments were designed to answer the following questions:

i. How can features of a person be extracted for all viewpoint angles for classifier's training purposes?
ii. What is the best combination of training set that can produce a high performing classifier for person re-identification?
iii. How does the classification perform for simulated activities found in public spaces such as free walking and walking towards and around a tabletop display?
iv. Will a person carrying item such as a backpack, handbag or suitcase affect classification performance of a classifier that has been trained on people without bags?

4.4.8.1 Dataset for Pilot Experiments (Small Size, Limited Activities of 22 People)

Before large scale data collection was conducted, a pilot experiment was arranged for the collection of training and testing data on a small scale. This pilot experiment was useful to evaluate conditions that can be improved prior to the actual final experiment. Such conditions include the suitability of the area for walking activities, strategic location of cameras, time required to setup and complete a session of an experiment, cost incurred in terms of compensation to participants, cost in terms of time taken for processing and analysing of data, and cost of storage of data. With ethical approval, twenty-two volunteers were invited to participate in re-identification experiments via email invitation. All of them are shown (with faces blurred) in Figure 4.21. These pilot experiments took place in the Chowen and Garfield Weston Foundation Digital Prototyping Hall at the European Research Institute. Experiments lasted not more than 30 minutes for a group of four participants, comprising of six activities performed individually and in group of two, such as turning in place, walking on straight path between two points, free walking, and walking up to a tabletop display and interacting with it. There were 144 recordings, 22 of each 6 individual activities and 6 of each 2 group activities.

The details of the activities are listed in Table 4.3 and illustrated in Figure 4.22. The actual view of the experiment area in the hall is shown in Figure 4.23. The participants were also asked to gather around a tabletop display, in group, to test the ability of the system to identify people in similar environment such as found in common interactive space.

Figure 4.21 The twenty-two people in the pilot KinectV2 RGBD-ID dataset

Table 4.3 KinectV2 RGBD-ID Pilot Dataset Activity Components

Dataset Component ID	Description of activity	Individual/ Group
01	Standing on a spot facing a Kinect camera. **Turning body** 360^{O} **clockwise** slowly. Turning the **opposite direction** 360^{O} slowly. (1st run)	Individual
02	Standing on a spot facing a Kinect camera. **Turning body** 360^{O} **clockwise** slowly. Turning the **opposite direction** 360^{O} slowly. (2nd run)	Individual
03	**Walking towards** a Kinect camera, turning back and walking back to origin. (1st run)	Individual
04	**Walking towards** a Kinect camera, turning back and walking back to origin. (2nd run)	Individual
05	**Free walking** in front of a Kinect camera	Individual
06	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.	Individual
07	**Free walking** in front of a Kinect camera	Group
08	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.	Group

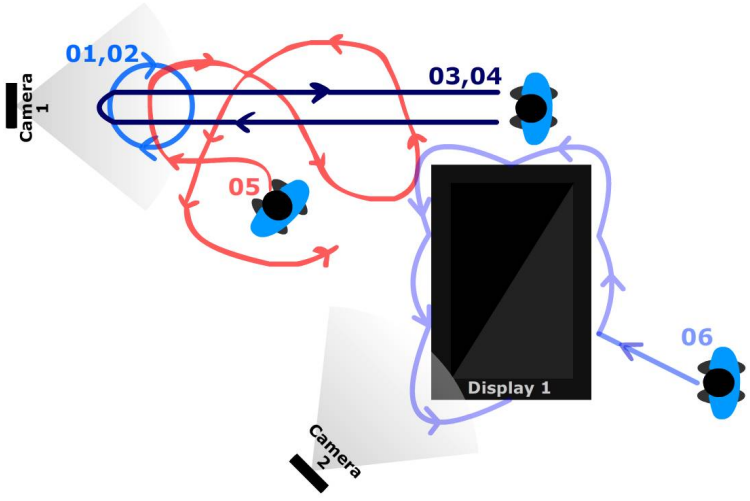


Figure 4.22 The pilot experiment plan layout with six individual activities numbered 01 to 06 and two group activities similar to 05 and 06.

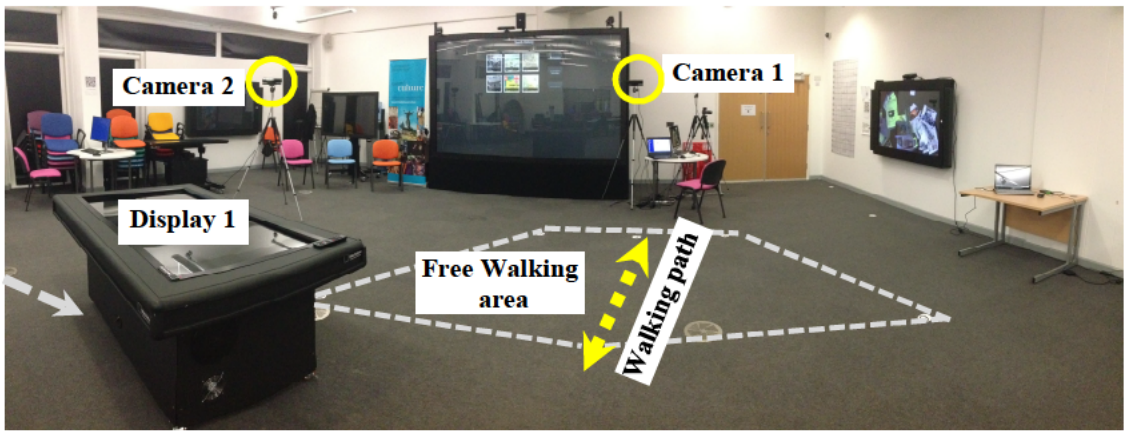


Figure 4.23 The pilot and final experiments took place in the Chowen and Garfield Weston Foundation Digital Prototyping Hall at the University of Birmingham.

4.4.8.2 Dataset of Final Experiment (Larger Size, More Activities of 64 People)

The final experiment improved upon the pilot experiments with the increased number of participants to sixty four, and extra dataset components with accessory-carrying activities such as turning and free walking while carrying a backpack, suitcase or hand bag. The participants were invited via emails and recruited on a first-come-first-served basis. All of them are shown (with faces blurred) in Figure 4.24. The final experiments were conducted in the same place as the pilot experiments. Experiments lasted not more than 50 minutes for a group of four participants. The dataset[7] comprises of sixteen component sets as listed in Table 4.4 and illustrated on a plan layout in Figure 4.25. All recordings were acquired at an average of 25 fps. Colour frames were acquired at a resolution of 1920 × 1080 pixels and depth frames at a resolution of 512 × 424 pixels. For both "Turning 1" and "Turning 2", participants were required to turn 360° clockwise and then 360° anti-clockwise. For "Free Walking 1" and "Free Walking 2" participants were asked to walk freely in random directions for approximately 15 seconds in an area within the field of view of the camera.

[7] *The dataset can be made available to researchers upon request and with a suitable agreement regarding usage conforming to ethical requirements, participant confidentiality and consent.*

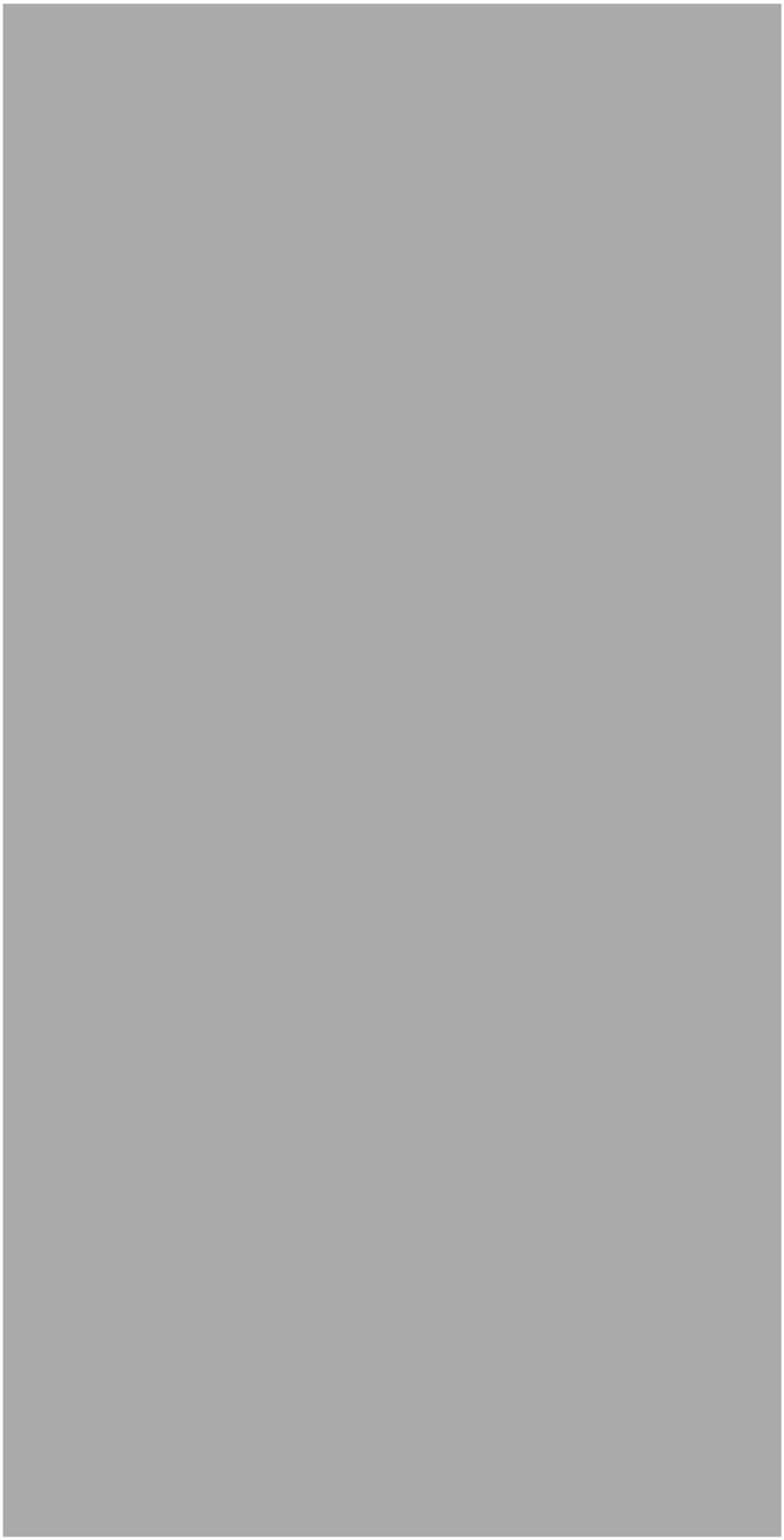


Figure 4.24. The sixty-four people in the KinectV2 RGBD-ID dataset.

Table 4.4 KinectV2 RGBD-ID Dataset Activity Components

Dataset Component ID	Description of activity	Carry bag?	Individual/ Group
01	Standing on a spot facing a Kinect camera. **Turning body** 360^{O} **clockwise** slowly. Turning the **opposite direction** 360^{O} slowly. (1st run)		Individual
02	Standing on a spot facing a Kinect camera. **Turning body** 360^{O} **clockwise** slowly. Turning the **opposite direction** 360^{O} slowly. (2nd run)		Individual
03	Standing on a spot facing a Kinect camera. **Turning body** 360^{O} **clockwise** slowly. Turning the **opposite direction** 360^{O} slowly. (1st run)	Yes	Individual
04	Standing on a spot facing a Kinect camera. **Turning body** 360^{O} **clockwise** slowly. Turning the **opposite direction** 360^{O} slowly. (2nd run)	Yes	Individual
05	**Walking towards** a Kinect camera, turning back and walking back to origin. (1st run)		Individual
06	**Walking towards** a Kinect camera, turning back and walking back to origin. (2nd run)		Individual
07	**Walking towards** a Kinect camera, turning back and walking back to origin (1st run)	Yes	Individual
08	**Walking towards** a Kinect camera, turning back and walking back to origin (2nd run)	Yes	Individual
09	**Free walking** in front of a Kinect camera (1st run)		Individual
10	**Free walking** in front of a Kinect camera (2nd run)		Individual
11	**Free walking** in front of a Kinect camera (1st run)	Yes	Individual
12	**Free walking** in front of a Kinect camera (2nd run)	Yes	Individual
13	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.		Individual
14	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.	Yes	Individual
15	**Free walking** in front of a Kinect camera		Group
16	**Free walking** in front of a Kinect camera	Yes	Group
17	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.		Group
18	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.	Yes	Group

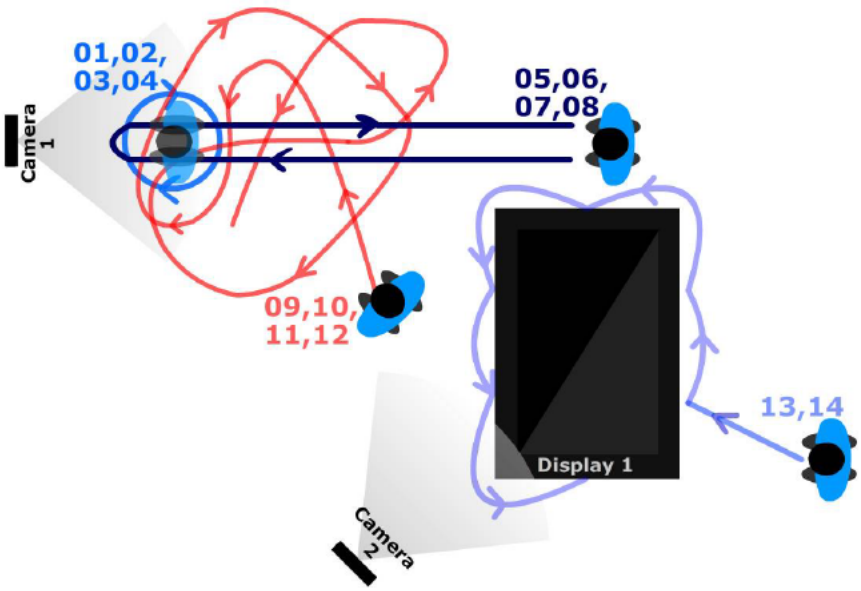


Figure 4.25 The final experiment plan layout with 14 individual activities numbered 01 to 14 and four group activities similar to 09, 11, 13 and 14.

(a) (b)

Figure 4.26 Example cropped frames extracted from (a) "Turning 1" and complete frames from (b) "Free Walking 1" dataset components. (Best viewed in colour)

Example cropped frames extracted from "Turning 1" and complete frames from "Free Walking 1" are shown in Figure 4.26 a) and b), respectively. There were a total of 968 recordings ((14 x 64) + (4 x 18) = 968; 14 individual activities for 64 people, and 4 group activities for 18 groups). The details of these activities can be found in Table 4.4. The dataset includes

synchronized colour images, depth images, infrared images and skeletal data (as provided by the Kinect SDK) complemented with calibration data such as ground plane coordinates.

Re-identification was performed under "closed-set" conditions where the identity of each test individual must match one of the identities in the training set. For each RGB-D frame, a ViMM feature vector was computed, comprising the 18 descriptors defined in equation (1). Frames in the training datasets were labelled with a participant identifier. The total number of frames (i.e. the number of feature vectors) for each dataset are listed in Table 4.5.

Table 4.5. KinectV2 RGBD-ID Training and Testing Dataset Components

Dataset component	**ID from Table 4.4**	**Training/ Testing**	**No. of frames**
Turning 1	1	**Training**	16 980
Turning 2	2	Testing	16 048
Turning-bag 1	3	**Training**	15 295
Turning-bag 2	4	Testing	15 534
Free Walking 1	9	**Training**	26 620
Free Walking 2	10	Testing	27 820
Free Walking-bag 1	11	**Training**	26 192
Free Walking-bag 2	12	Testing	27 726
Turning 1 + Free Walking 1 (halved)	1 + 9	**Training**	21 800
Turning-bag 1 + Free Walking-bag 1 (halved)	3 + 11	**Training**	20 743
Turning 1 + Free Walking 1 + Turning-bag 1 + Free Walking-bag 1 (quartered)	1 + 9 + 3 + 11	**Training**	21 270
Around Tabletop	13	Testing	24 243

Number of frames for each training and testing dataset component. Dataset components labelled "(halved)" such as "Turning 1 + Free Walking 1" was halved by removing alternate frames to reduce number of frames for training. It is labelled "(quartered)" when the components are halved twice.

The example of distributions of body orientation angles for dataset component "Free Walking 1" and "Free Walking 2" in Figure 4.27 can be used to give an indication that the walking direction is fairly random.

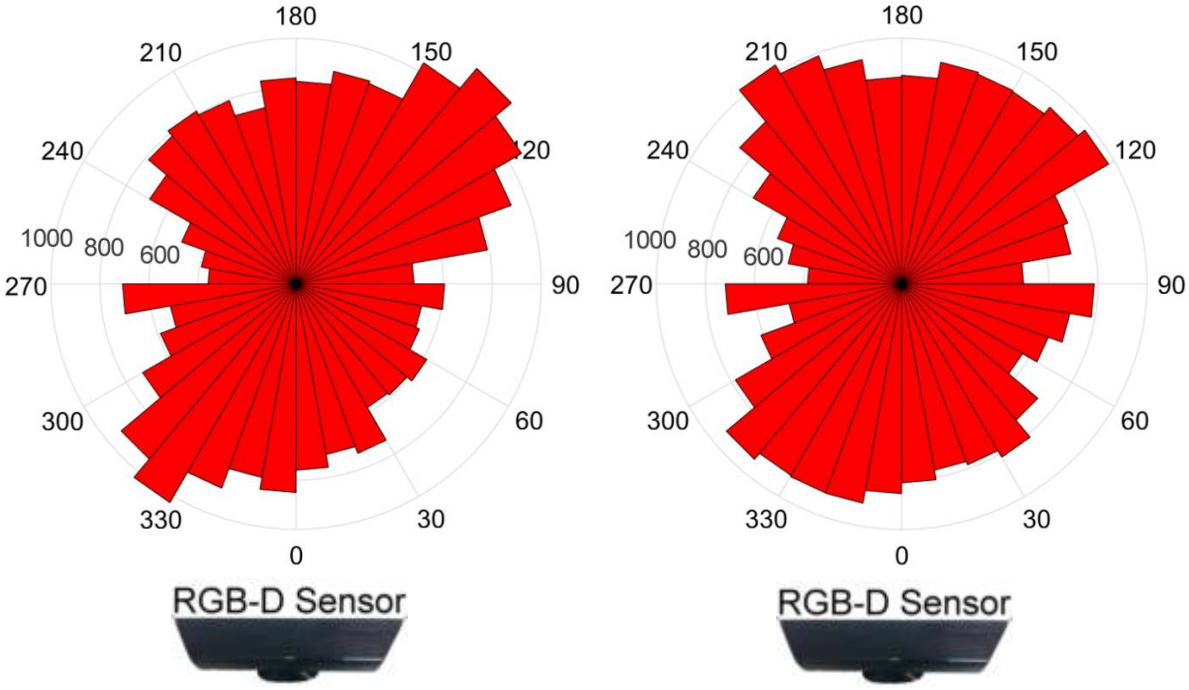


Figure 4.27 Distribution of body orientation for dataset components “Free Walking 1” (left) and “Free Walking 2” (right)

4.4.9 Feature Extraction and Classifier Training

A C# application was written to perform the feature extraction. A distance filtering pre-processor was used for all training and test data, limiting the maximum range to 4.0m. Although the Kinect V2 will return depth data at greater distances, up to 4.5m, the quality and resolution degrades and leads to unacceptable error levels. This maximum range restriction concurs with published analysis reporting the optimum maximum operating distance for Kinect V2 to be between 2.5m and 3.5m (Lachat et al., 2015) and 3.5m of maximum face tracking distance for Kinect V2 (Microsoft, 2015). Orange data mining software was used to train classifiers using the training data. During the base test, classifiers were created from training data from a “Turning 1” dataset component using various standard machine learning methods such as Naïve Bayes, Support Vector Machine (SVM), Neural Network, Logistic Regression, *k*-Nearest

Neighbours (kNN), Random Forest and Decision Trees. Classifiers were tested against the test data from the dataset component “Turning 2”.

The process of extracting features from a recording involves 1) running the C# feature extraction application, and 2) playing back the recorded “.xef” file. The C# application automatically saves the features into an individual file named after the participant ID and activity (component) ID. For example the features file for the first activity of the first participant was given a file name “0101.tab”. The “.tab” extension is just a plain text file prepared in a specific format to be readily readable by Orange software. The file for activity ID “9” of the fifty-second participant has a file name “5209.tab”, and so on. The time taken to do the above process is on average, 30 seconds. If this process was to be manually executed by a human, it would take an average 1 minute inclusive of loading up the 2GB (average) file before playing it. A total of approximately 15 hours non-stop sitting in front of a computer is required to process all the 896 recordings for individual activities. Group activities were processed separately.

To avoid fatigue as a result from the long hours doing the features extraction, an automation method had been employed. An automation software for Windows called AutoIt v3 (AutoIt, n.d.) was used to sequentially perform the feature extraction process on every activity for every participant. The script used is shown in Listing 6 in the Appendices. It should be noted that the Kinect Studio Utility Tool had been used to play the recording instead of Kinect Studio GUI Tool because it is command line friendly and it suits the automation scripting perfectly. The playing of a recording was done in the background and this has an advantage of saving memory consumption and processing power that had been better allocated for the feature extraction process alone. The automation process was left to run overnight. The script used is presented in the Appendix section for reference.

4.4.10 Angle Invariant Anthropometric Measures

The detection of the presence of a person as well as their size and pose can be accomplished by the Kinect v2 camera together with the Kinect SDK v2.0 (Lachat et al., 2015). The SDK provides a real-time estimate of the absolute metric coordinates of twenty-five different body joints. Five heights (also used by Barbosa et al. (2012)) were selected for inclusion in the ViMM feature descriptor: head, neck, shoulders, mid-spine and hips. These heights denoted by d_1, d_2, d_3, d_4 and d_5 are highly reliable against occlusions from other body parts. These heights stay the same hence invariant for all body orientations as it should be unlike the other ViMM features. The heights are estimated by summing the inter-joint distances rather than simply using the *y*-coordinate relative to the floor. For example, the head height, d_1, is the sum of the distances from the floor to hips, hips to shoulder-centre, shoulder-centre to neck, and neck to head. In testing, this method gives more reliable estimates when people are not standing perfectly upright.

4.4.11 Colour Model for Appearance Features

The surface colour of an individual can exhibit significantly different RGB values with variations in illumination as they move around a room. Consequentially, the RGB colour space is unsuitable for colour matching and colour-based object tracking (Southwell and Fang, 2013). To achieve illumination invariance colour descriptor, the *rg*-chromaticity plane was used as described by Balkenius and Johansson (2007) where each colour's hue and saturation are preserved, but the intensity is discarded. Figure 4.28 shows the colour space of *rg*.

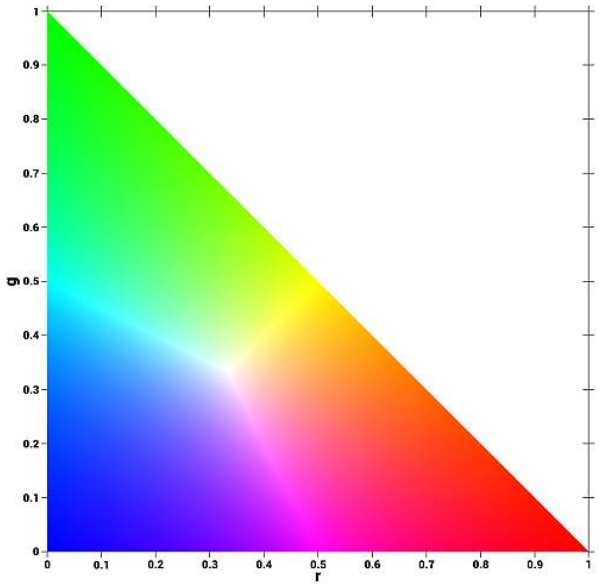


Figure 4.28. Normalised *rg* colour space.

The following transformation is performed on the R, G, and B channels by dividing the red and green components of the pixel coordinates by the sum of the three colour channels.

$$r = \frac{R}{R + G + B} \quad (1)$$

$$g = \frac{G}{R + G + B} \quad (2)$$

Only the colours of the original RGB space are kept in the two dimensional rg-plane but not the intensity of each pixel. The intensity of the colour however can be calculated using this formula:

$$I = R + G + B \quad (3)$$

Three pairs of *rg* samples labelled as r_1, g_1, r_2, g_2, r_3, and g_3 are included in the ViMM descriptor, calculated by taking the average colour of 10-by-10 pixel image regions centred at spine-shoulder (shoulder level), spine-mid (torso level) and spine-base (hip level) joint locations. This is depicted in Figure 4.29.

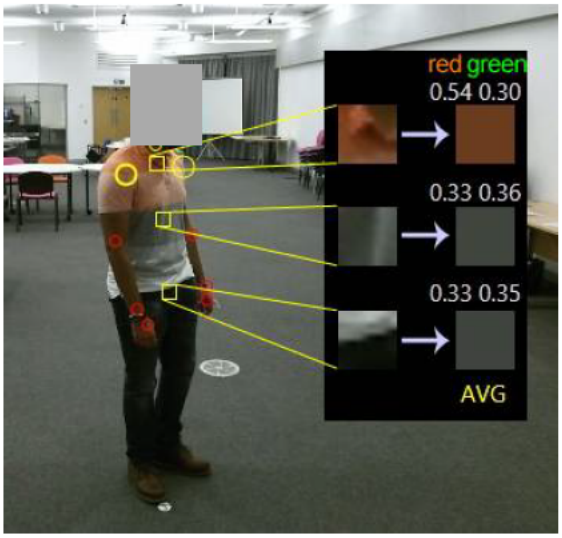


Figure 4.29. Sampled images taken at three different heights for a certain body orientation.Colours for each image are then averaged before converting to *rg*-chromaticity format. The transformed *r* and *g* values are displayed above each average colour rectangle.

4.4.12 The Complete ViMM Feature Descriptor

The features outlined in the previous sections are concatenated into a single feature descriptor which is called the ViMM feature descriptor or vector:

$$\mathbf{ViMM} = (\theta \parallel \mathbf{d} \parallel \mathbf{s} \parallel \mathbf{c}) \quad (1)$$

where,

θ = angle of body orientation
$\mathbf{d}$ = vector of distances: $(d_1, d_2, d_3, d_4, d_5)$
$\mathbf{s}$ = vector of ellipses: $(s_1, s_2, s_3, s_4, s_5, s_6)$
$\mathbf{c}$ = vector of colours: $(r_1, g_1, r_2, g_2, r_3, g_3)$
d_1 = distance between floor and head
d_2 = distance between floor and neck
d_3 = distance between floor and shoulders
d_4 = distance between floor and mid-spine
d_5 = distance between floor and hips
s_1 = semi-minor axis of shoulders
s_2 = semi-major axis of shoulders
s_3 = semi-minor axis of mid-spine
s_4 = semi-major axis of mid-spine
s_5 = semi-minor axis of hips
s_6 = semi-major axis of hips
r_1 = red component of *rg*-chromaticity at shoulder height
g_1 = green component of *rg*-chromaticity at shoulder height
r_2 = red component of *rg*-chromaticity at mid-spine height
g_2 = green component of *rg*-chromaticity at mid-spine height
r_3 = red component of *rg*-chromaticity at hip height
g_3 = green component of *rg*-chromaticity at hip height

4.4.13 Classification Methods

Classification in general is the task of assigning an object to a category using a set of features characterising the object. Examples of classification problems are text categorisation (e.g. spam filtering), fraud detection, optical character recognition, machine vision (e.g. face detection), natural language processing (e.g. spoken language understanding) and bioinformatics (e.g. classify proteins according to their function).

In data mining, classification is a fundamental issue and refers to "*the task of analysing a set of pre-classified data objects to learn a model (or a function) that can be used to classify an unseen data object into one of several predefined classes*" (An, 2008). Referring to Figure 4.30, a set of known objects called the labelled training data is used by a classification program called the classifier to learn how to classify objects. In the training phase, the training data is used to calculate the parameters of a learning model using a machine learning algorithm in order to separate the various classes of objects. In the testing or classification stage, the parameters determined in the training set are applied to a set of unknown objects in order to determine the classes of the objects. Classification belongs to the category of supervised learning (An, 2008), which means the training data consists of pairs of input data and the desired outputs.

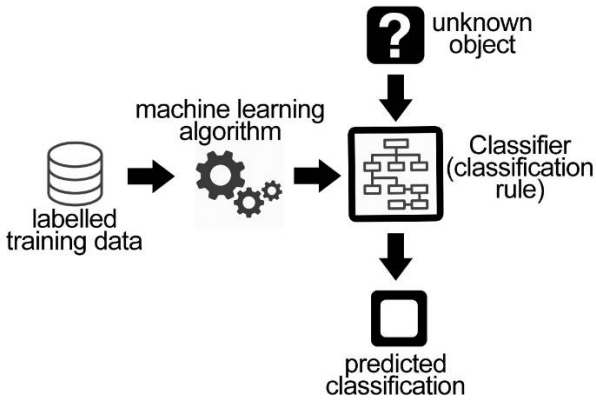


Figure 4.30 Typical classification workflow using machine learning algorithm.

In machine learning, multiclass classification classifies an object into one class out of many classes. On the other hand, binary classification only classifies one object into a class out of two classes. Face detection is an example of binary classification, in which the two classes represent human faces and non-human faces. Unlike face detection, face recognition is an example of a multiclass classification problem that demands an unknown face to be compared with reference images in the training set to obtain the identity of the person.

The person re-identification sharing a similar nature as the face recognition in terms of its classification process, is a problem of multiclass classification, which will be described in the next section.

4.4.14 Multiclass Classification

The proposed ViMM person re-identification system employed supervised multiclass classification method (i.e. number of classes K = 64) to produce a learning model from a labelled training set. The learning model aims to assign a class label for every training input.

Given a set of N training inputs of the form $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ such that x_i is the feature vector of the i-th training input and y_i is the class label, a learning model seeks a function $\mathbf{C}$ such that $\mathbf{C}(x_i) = y_i$ for new test inputs. The problem above can be simulated in the two class case, where class labels y_i are just +1 or -1 for the two classes involved. Machine learning methods have been proposed to solve this problem in the two class case, and some can be naturally extended to the multiclass case, for example decision tree, neural networks, k-Nearest Neighbour, Naïve Bayes classifiers, and Support Vector Machines. Multi-layer Feedforward Neural Networks provide a natural extension to the multiclass problem.

To evaluate the performance of a classifier in multiclass classification, it is necessary to explain about classification ranking (i.e. rank-1, rank-2, etc). In the following example, five classes (i.e. P_1, P_2, P_3, P_4, P_5) are used for simplicity instead of 64. For a test input I_1 (that belongs to class P_1), one test per class is needed, hence five tests (i.e. T_1, T_2, T_3, T_4, T_5) in total will be performed to classify the test input I_1. Each test produces a score when compared to each class. As a general similarity measure, the more score a test produces indicates the more similarity the test input is when compared to a class. If the score between I_1 and class 1 (i.e. P_1) is larger than the other four classes, I_1 is recognised in the first rank. As an example, let us suppose the following similarity scores:

T_1: I_1 vs P_1 = 0.95
T_2: I_1 vs P_2 = 0.8
T_3: I_1 vs P_3 = 0.3
T_4: I_1 vs P_4 = 0.5
T_5: I_1 vs P_5 = 0.85

These scores say that I_1 is more similar to class P_1 than the other classes. So I_1 is correctly recognised in the first rank (i.e. rank-1).

Now let us suppose the following situation:

T_1: I_1 vs P_1 = 0.85
T_2: I_1 vs P_2 = 0.3
T_3: I_1 vs P_3 = 0.95
T_4: I_1 vs P_4 = 0.5
T_5: I_1 vs P_5 = 0.8

These scores now say that I_1 is more similar to class P_3, which is a mismatch because I_1 belongs to class P_1. I_1 is not recognised as the top match, but recognised among top two matches. So I_1 is now said as correctly recognised in the second rank (i.e. rank-2).

The next sections describe selected classification methods out of all the tested methods in this chapter.

4.4.15 Naïve Bayes

The naïve Bayesian classifier is based on Bayes' theorem. It is also the simplest form of Bayesian network (Zhang, 2004) which are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent (Kotsiantis, 2007). It is simple to use and efficient to learn. This classifier makes a naïve assumption that all attributes used in describing an unknown object X are conditionally independent of each other given the class of X (An, 2008). The major advantage of naïve Bayes classifier is that one scan of the training data is usually required, hence short computational time for training.

4.4.16 *k*-Nearest Neighbours

The *k*-nearest neighbour (kNN) classifier classifies an unknown object X to the most common class among its *k* nearest neighbours in the training data. All objects are assumed to correspond to points in an n-dimensional feature space. Object Y from the training data is deemed as the closest neighbour to X if the distance (e.g. in Euclidian) between X and Y is the smallest. When $k = 1$, the object X is classified into the class of its closest neighbour in the training set. This method stores all the training data in their original form and only performs learning/classification when a new unknown object needs to be classified. This type of learning is called instance-based or lazy learning (An, 2008). The *k*-nearest neighbour classifier is intuitive, easy to implement and effective in practice but can be quite costly due to the fact that most computation is done at the classification stage.

4.4.17 Decision Trees

Decision tree is a non-parametric supervised learning method for classification. It is a classifier expressed as a recursive partition of the instance space (Rokach and Maimon, 2005). It is a tree structured prediction model where each internal node represents a test on an attribute, each outgoing branch represents an outcome of the test, and each leaf node is labelled with a class. An object is classified by following a path from the root of the tree to a leaf, taking the edges corresponding to the values of the attributes in the object. Tightly stopping criteria tends to create small and under-fitted decision trees while loosely stopping criteria tends to generate large decision trees that are over-fitted to the training set (Rokach and Maimon, 2005).

4.4.18 Support Vector Machines

The support vector machine (SVM) was developed for multidimensional function approximation. It is a discriminative classifier formally defined by a separating hyperplane. Given labelled training data, the classifier outputs an optimal hyperplane which categorises unknown objects. Its objective is to determine a classifier function which minimises the training error and the confidence interval (which corresponds to the generalisation or test set error) (An, 2008). The advantages of SVM are, effective in high dimensional spaces, memory efficient because of the use of a subset of training points in the decision function (called support vectors) and versatile in terms of possibility to use custom kernels for the decision function. The disadvantages include poor performance when the number of features is much greater than the number of training data, and it does not directly provide probability estimates, instead it calculates the output using an expensive five-fold cross validation. Although SVMs have good generalisation performance, they can be very slow in test phase (Burges, 1998).

4.4.19 Neural Network

A back propagation neural network was chosen over other commonly-used machine learning methods such as Naïve Bayes, *k*-Nearest Neighbors (kNN), decision trees and Support Vector Machines (SVM). The back propagation neural network is a multi-layer perceptron (MLP) with a feed-forward learning algorithms and have been widely and successfully applied in diverse applications, such as pattern recognition, location selection and performance evaluations (Che et al., 2011). This type of network can theoretically be used to approximate any function to arbitrary accuracy, using a finite number of neurons and layers (Mitchell, 1997), however this fact does not address the question of how many neurons and layers such a network will require or what shape the optimal neuron transfer function is. To reduce complexity, no algorithm has been employed to find the optimal answer. Trial and error method has been used instead which yielded acceptable network architecture.

Neural networks are powerful for non-linear classification problems; they handle noisy data well, and have low computational complexity when classifying which is essential for real-time person re-identification. Naïve Bayes classifiers are computationally simple and are fast to train. However, they are only suited for sets of features that are independent from each other and so do not work well with ViMM where most of the features are dependent on angles. kNN classifiers require very little or no processing for training but classification can become slow when the size of the training set is large. SVMs, whilst being powerful non-linear classification algorithms, demand significant computational power to train and to classify. They are also sensitive to noisy data and are prone to overfitting which results in poor generalization.

The four selected methods, i.e. neural network, SVM, Naïve Bayes and kNN were tested with the new dataset consisting of subjects turning on a spot (i.e. “Turning 1” as training and

"Turning 2" as testing). Classification performance observed for the neural network, SVM and kNN were almost identical with 96.87%, 96.95% and 96.42% respectively for rank-1 classification, while Naïve Bayes achieved only 65.63% as shown in Figure 4.31. In addition, the neural network achieved 100% classification by rank-12 while SVM did not achieve this until rank-47 and was very slow during training and testing. kNN was fast during training but very slow during testing. Naïve Bayes was fast but was rejected because of its poor classification performance.

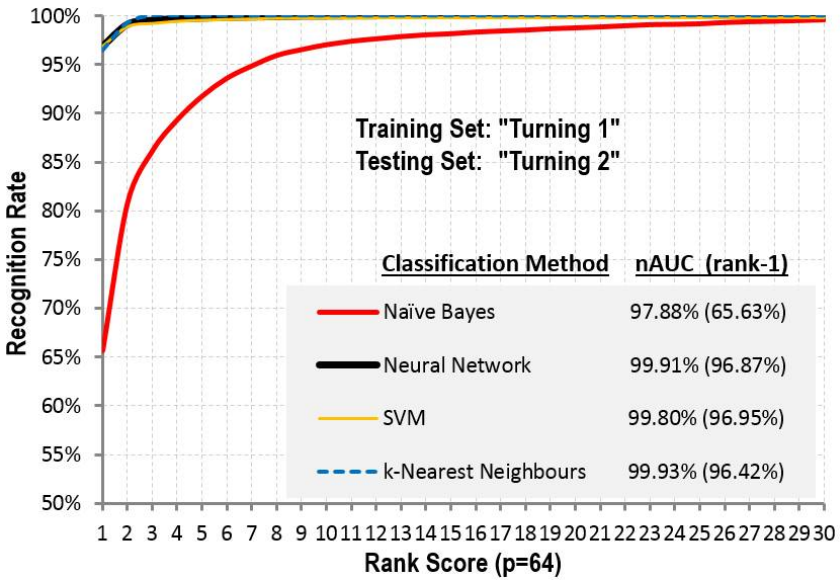


Figure 4.31 Performances observed for the neural network, SVM and kNN are very good and almost identical for rank-1 classification. It can be concluded that ViMM feature descriptor has strong discriminant properties and the method can be implemented using most machine learning methods.

Based on the classification result and performance observation mentioned earlier, the neural network was chosen for the classifier because of its good classification performance and fast performance during testing. All experiments in this chapter and also Chapter 5 were performed using Orange Data Mining software for Windows (Ljubljana, n.d.). Orange is an open source data visualisation and data analysis software for novice and expert, featuring interactive workflows with large selection of toolboxes. The software's NeuralNetworkLearner class implements a back propagation multilayer perceptron learning model. Learning is performed

by minimizing an L2-regularized cost function with scipy's implementation of L-BFGS. A single hidden layer was used by this implementation.

The first step in creating a neural network classifier based on the above described implementation is to select the number of nodes or neurons for the hidden layer. For the pilot dataset, it was chosen to start with 20 nodes which is almost equal to the number of persons or classes. The number of input (features) was 17. In the next step, regularisation factor was selected as 0.2 and lastly, maximum iterations was set to 100. A series of trials on pilot training and test data showed that values 80, 0.4 and 500 for number of hidden layer's neurons, regularisation factor and maximum iterations respectively, gave rise to very good classification performance. However the same values were found to be not suitable for the final dataset having 64 classes. After a series of trials on the data from the final dataset, 140 hidden layer neurons, regularization factor of 0.4 and 600 maximum iterations were chosen to be the best estimate that produced the best classification results. The neural network classifiers required several hundred training vectors which corresponds to only a few seconds of video and hence was suitable for this research's purposes. An example of canvas containing widgets, used to perform classifier learning using training data and testing using separate test data, is shown in Figure 4.32. The results of classification is presented in a form of confusion matrix as shown in Figure 4.33.

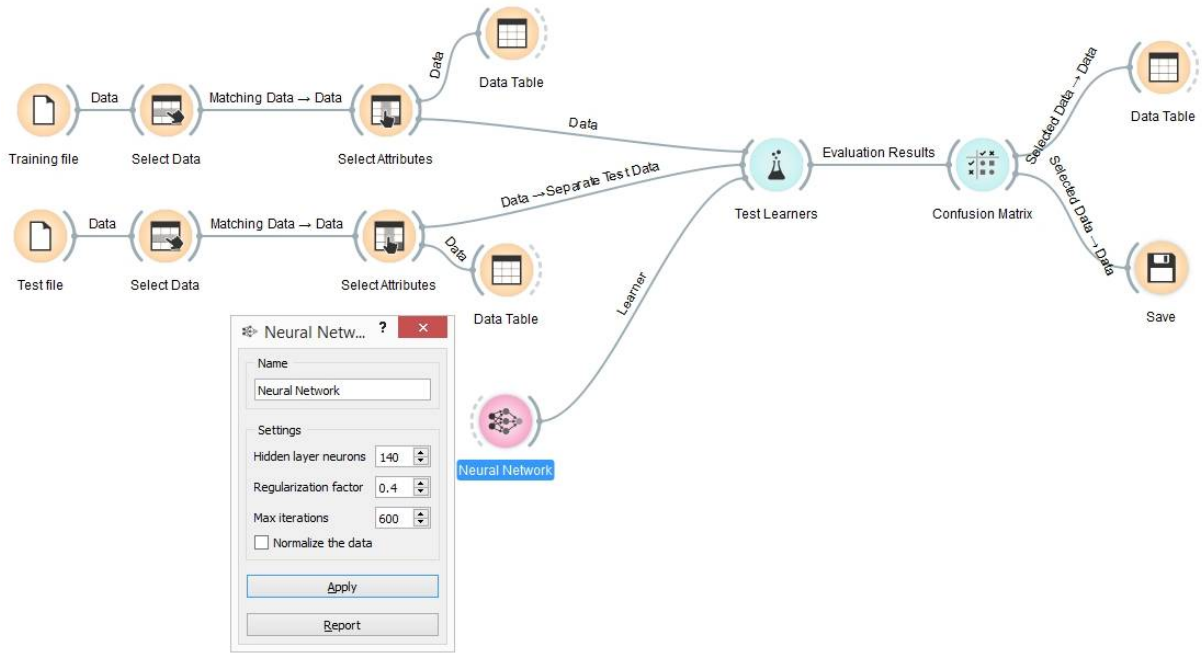


Figure 4.32. A canvas in Orange to perform neural network classifier learning using training data and testing using separate test data.

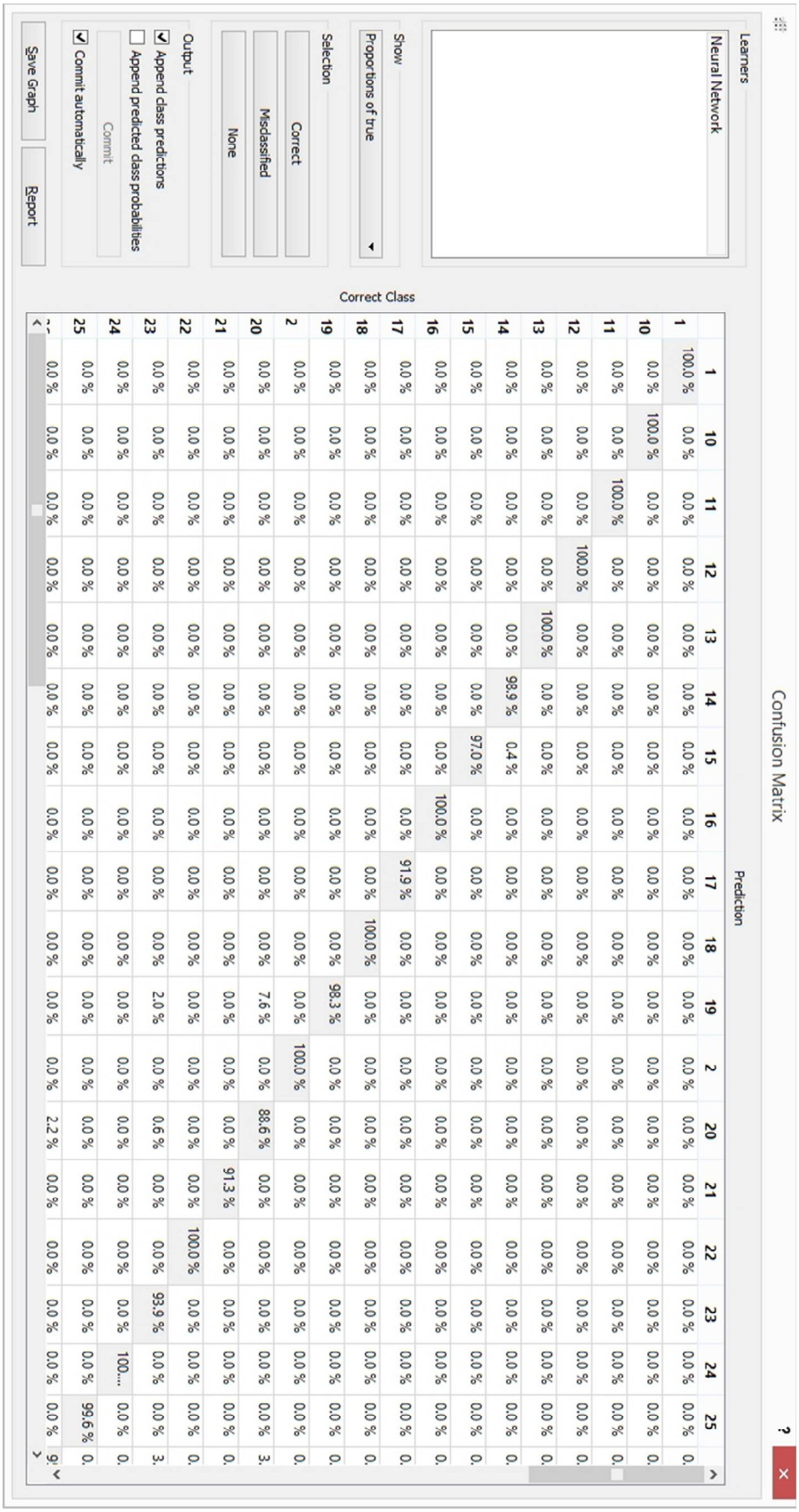


Figure 4.33 Confusion matrix from Orange's software showing classification results of neural network classifier on test data.

4.5 Matching Techniques

In general, there are two sets of human signatures in re-identification problem, the first is a training set A (also called gallery), and the second is a testing set B (also called probe). The signature P^B of each person in B needs to be matched with the corresponding signature P^A of each person in A. The matching mechanisms employed in this research are called Single-shot and Multi-shot re-identification which are described below.

4.5.1 Single-shot Re-identification

When each person in the training set is described by multiple images, and the testing set contains only a single image for each person, it is called an 'MvsS' scenario (Multiple training images vs Single test image).

4.5.2 Multi-shot Re-identification

To improve the robustness in the estimation, the output of multiple consecutive frames can be easily integrated, for example using a voting scheme. This is called multi-shot re-identification (MvsM - Multiple training images vs Multiple test images). The objective of MvsM re-identification is to re-identify a group of adjacent images, with the knowledge that images in each group are all of the same individual. Two simple decision methods have been tested, which are mean computation and median value, of the neural network outputs over several adjacent frames. CMC curves are calculated from these mean and median outputs for performance comparisons.

4.6 Experimental Results

The experimental results are presented in this section. The aim was to investigate the best combinations of features for viewpoint invariant person re-identification and to evaluate the performance of the ViMM feature descriptor.

To identify an individual from their feature vector, a one shot re-identification strategy was used, similar to that described by Munaro et al. (2014b), i.e., a feature vector of test data extracted from each RGB-D frame was matched to a person in the training set using a pre-trained classifier. For example, to test classification performance with "Turning 2", 16 048 feature vectors were inputted to the pre-trained classifier and the matching scores for each test person were averaged to calculate the classification ranks. The results were evaluated using the average Cumulative Matching Characteristics (CMC) curve and normalised Area Under the Curve (nAUC). The CMC curve is defined as the probability of finding the correct match within the first *n* ranks, with *n* ranging from 1 to the number of test subjects (Pala et al., 2015). For each frame, the classifier estimates a probability of identification to each person in the test dataset. If the highest estimated probability is assigned to the correct person, this is rank 1. If the second highest probability is assigned to the correct person, this is rank 2, and so on. An nAUC is the area under the entire CMC curve normalised over the total area of the CMC graph (Bazzani et al., 2014).

Twelve classifiers (C1 to C12), as listed in Table 4.6, are created using different feature sets extracted from training sets comprised of "Turning 1", "Free Walking 1", "Turning 1 + Free Walking 1", "Turning-bag 1", "Turning-bag 1 + Free Walking-bag 1", and "Turning 1 + Free Walking 1 + Turning-bag 1 + Free Walking-bag 1".

Table 4.6. Classifier Feature Sets

Classifier ID	Training Set	Feature Set
C1	Turning 1	ViMM (θ || **d** || **s** || **c**)
C2	Turning 1	Heights (**d**)
C3	Turning 1	Ellipses (θ || **s**)
C4	Turning 1	Colours (θ || **c**)
C5	Turning 1	Heights + Ellipses (θ || **d** || **s**)
C6	Turning 1	Heights + Colours (θ || **d** || **c**)
C7	Turning 1	Ellipses + Colours (θ || **s** || **c**)
C8	Free Walking 1	ViMM (θ || **d** || **s** || **c**)
C9	Turning 1 + Free Walking 1	ViMM (θ || **d** || **s** || **c**)
C10	Turning-bag 1	ViMM (θ || **d** || **s** || **c**)
C11	Turning-bag 1 + Free Walking-bag 1	ViMM (θ || **d** || **s** || **c**)
C12	Turning 1 + Free Walking 1 + Turning-bag 1 + Free Walking-bag 1	ViMM (θ || **d** || **s** || **c**)
C13	Turning 1 + Free Walking 1 + Turning-bag 1 + Free Walking-bag 1	ViMM v0 ex RGB (θ || **d** || **s** || **c** || **c'**)
C14	Turning 1 + Free Walking 1 + Turning-bag 1 + Free Walking-bag 1	ViMM v0 (θ || **d** || **s** || **c** || **c'** || **c''**)

A total of five main experiments (labelled Ex1 to Ex5) were designed to evaluate the performances of classifiers on different combinations of testing sets. A further two experiments (labelled Ex6 to Ex7) were designed to evaluate the performances of classifiers with expanded features with which the results are presented in Section 4.8.1. Table 4.7 summarises all the experiments in this chapter.

Table 4.7. List of Experiments

Experiment ID	Classifier ID	Testing Set	Feature Set	MvsS / MvsM
Ex1	C1	Turning 2	ViMM	MvsS
	C2	Turning 2	Heights	MvsS
	C3	Turning 2	Ellipses	MvsS
	C4	Turning 2	Colours	MvsS
	C5	Turning 2	Heights + Ellipses	MvsS
	C6	Turning 2	Heights + Colours	MvsS
	C7	Turning 2	Ellipses + Colours	MvsS
Ex2	C1	Free Walking 2	ViMM	MvsS
	C8	Free Walking 2	ViMM	MvsS
	C9	Free Walking 2	ViMM	MvsS
	C9	Free Walking 2	ViMM	MvsM M=1,5,10,15
Ex3	C10	Turning-bag 2	ViMM	MvsM M=1,5,10,15
	C9	Free Walking-bag 2	ViMM	MvsM M=1,5,10,15
	C11	Free Walking-bag 2	ViMM	MvsM M=1,5,10,15
Ex4	C12	Free Walking 2	ViMM	MvsM M=1,5,10,15
	C12	Free Walking-bag 2	ViMM	MvsM M=1,5,10,15
Ex5	C12	Around Tabletop	ViMM	MvsM M=1,5,10,15
	C12	Around Tabletop-bag	ViMM	MvsM M=1,5,10,15
Ex6	C13	Free Walking 2	ViMM v0 ex RGB	MvsM M=1,5,10,15
	C13	Free Walking-bag 2	ViMM v0 ex RGB	MvsM M=1,5,10,15
	C13	Around Tabletop	ViMM v0 ex RGB	MvsM M=1,5,10,15
Ex7	C14	Free Walking 2	ViMM v0	MvsM M=1,5,10,15
	C14	Free Walking-bag 2	ViMM v0	MvsM M=1,5,10,15
	C14	Around Tabletop	ViMM v0	MvsM M=1,5,10,15

The first, preliminary experiment (Ex1) used classifiers C1 to C7, with “Turning 1” for training and “Turning 2” for testing. As could be expected, because the training and test datasets are different recordings but of the same turning activity, the classification performance was very

good for ViMM. Figure 4.34 contrasts the performance of the complete ViMM feature vector with its subsets. Using the complete ViMM vector with 16,048 frames of test data, **96.87%** have rank-1 classification and nAUC = 99.91%. Comparing the ViMM subsets, Figure 4.34 shows that the subsets that include colour information consistently outperform those that do not. Classifier C4 ("Colours" only feature set) recorded **91.77%** rank-1 classification with 99.5% nAUC.

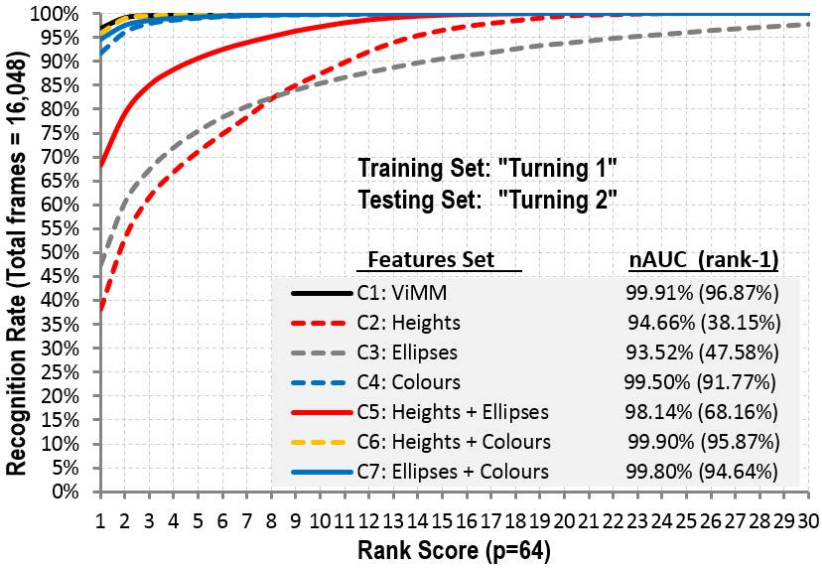


Figure 4.34. Ex1: CMC curves for ViMM and ViMM descriptor subsets using classifiers C1 to C7, as listed in Table 4.6. (Best viewed in colour)

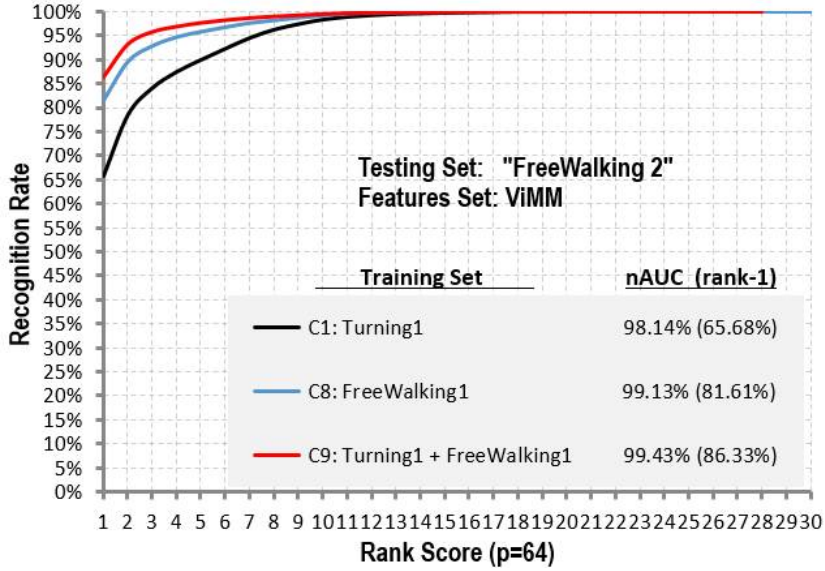


Figure 4.35. Ex2: CMC curves for the ViMM descriptor using classifiers C1, C8 and C9 as listed in Table 4.6. (Best viewed in colour)

Classifiers C6 (“Heights + Colours”), C7 (“Ellipses + Colours”) and C1 (ViMM) all achieved an nAUC greater than 99.8%. The best results of all were obtained with the complete ViMM feature vector.

The second experiment (Ex2) was performed to investigate the performances of classifiers C1, C8 and C9. This experiment used the more natural “Free Walking 2” dataset component for testing. The results are shown in Figure 4.35. The rank-1 performance of classifier C1 (“Turning 1” as a training set) is 65.68% and 98.14% nAUC. Classifiers C8 and C9 used the same ViMM feature vector but different training datasets: “Free Walking 1” for classifier C8 and a combination of “Turning 1” and “Free Walking 1” for classifier C9. The reason for these experiments was to evaluate the effects of using more diverse training datasets. It is evident in Figure 4.35 that the performance of classifiers C8 and C9 improves with the more diverse training dataset compared with classifier C1 which used only “Turning 1” (nAUC increases from 98.14% for C1 to 99.5% for C9 and rank 1 performance increases from 65.68% to 86.52%).

The re-identification methods and results reported thus far used a one-shot re-identification strategy where individuals are identified from a single frame in the test dataset. In the training dataset, each person is described by multiple images; this is an ‘MvsS’ scenario (Multiple training images vs Single test image). To improve the ‘MvsS’ results, a multi-shot re-identification technique (MvsM - Multiple training images vs Multiple test images (Farenzena et al., 2010)) was implemented by averaging the neural network outputs over several adjacent frames (using either the mean or the median decision method). This calculation example is illustrated in Figure 4.36.

	Neural Network output for each class					
Frame No.	**1**	**2**	**3**	**4**	**...**	**64**
1	0.50	0.30	0.10	0.10	...	...
2	0.40	0.20	0.18	0.10	...	...
3	0.30	0.10	0.40	0.10	...	...
4	0.60	0.30	0.10	0.10	...	...
5	0.50	0.10	0.30	0.10	...	...
Total Sum:	2.30	1.00	1.10	0.50	...	...
Mean:	0.46	0.20	0.22	0.10	...	...
Rank:	***1***	***3***	***2***	***4***	...	...
Median:	0.50	0.20	0.18	0.10	...	...
Rank:	***1***	***2***	***3***	***4***	...	...

Figure 4.36. MvsM example using “mean” and “median” decision methods.

CMC curves were calculated from these averaged outputs and are shown in Figure 4.37 and Figure 4.38, along with results for the single-shot method. The multi-shot (MvsM) method gives a rank-1 performance improvement of up to 6%.

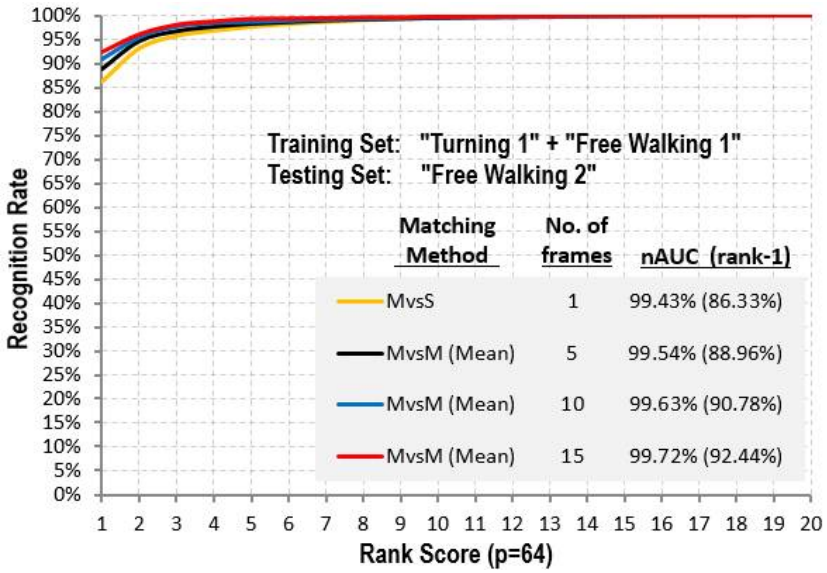


Figure 4.37. Ex2: CMC curves showing results for single-shot (MvsS) vs mean multi-frame (MvsM) method with different number of frames. (Best viewed in colour)

The third experiment (Ex3) was performed to investigate how classifier C9, being the best classifier so far, would perform on persons carrying backpacks. First, a preliminary test was carried out using classifier C10 (i.e. “Turning-bag 1” as training) and “Turning-bag 2” as testing

set, to see if ViMM feature descriptor was robust against the new appearance with bag. ViMM passed the test as illustrated in Figure 4.38.

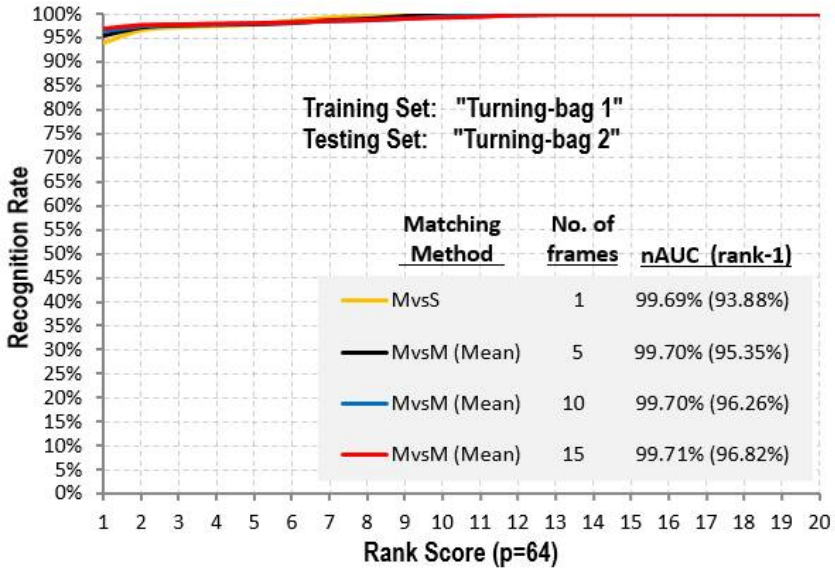


Figure 4.38. Ex3: CMC curves showing results for single-shot (MvsS) vs mean multi-frame (MvsM) method with different number of frames. (Best viewed in colour)

As expected, the performance of classifier C9 decreased significantly because of information missing, as a result of the bag appearance not being included in the classifier C9 (Figure 4.39). The experiment was repeated with classifier C11 on the same “Free Walking-bag 2”. The results shown in Figure 4.40 recorded significant improvement over the previous results from Figure 4.39. This was anticipated because the appearance features of persons with backpacks have now been included in the training set.

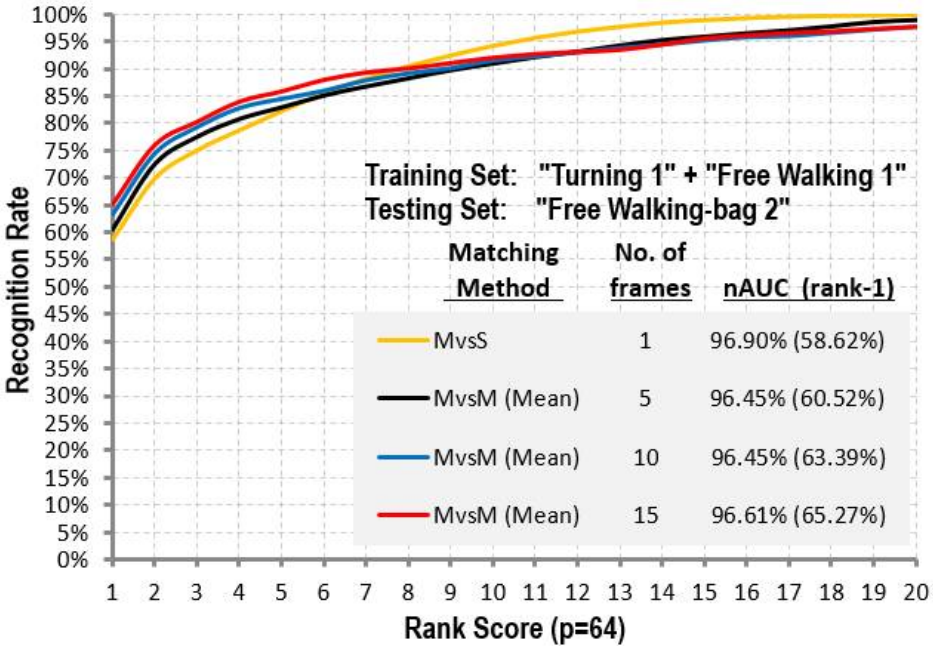


Figure 4.39. Ex3: CMC curves showing results for single-shot (MvsS) vs mean multi-frame (MvsM) method with different number of frames. (Best viewed in colour)

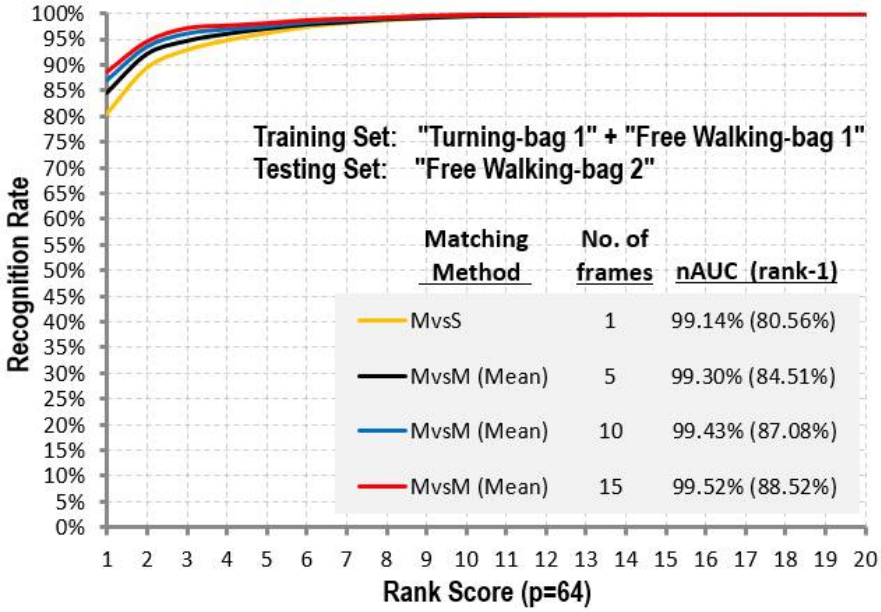


Figure 4.40. Ex3: CMC curves showing results for single-shot (MvsS) vs mean multi-frame (MvsM) method with different number of frames. (Best viewed in colour)

The fourth experiment (Ex4) was carried out to see if combining C9 and C11, now defined as C12, can handle scenarios of persons walking with and without backpacks. Classification performance (Figure 4.41) is found to be consistent with experiment result shown in Figure 4.37 although a very little drop in performance of nearly 1% was seen (from 92.44% to 91.49%).

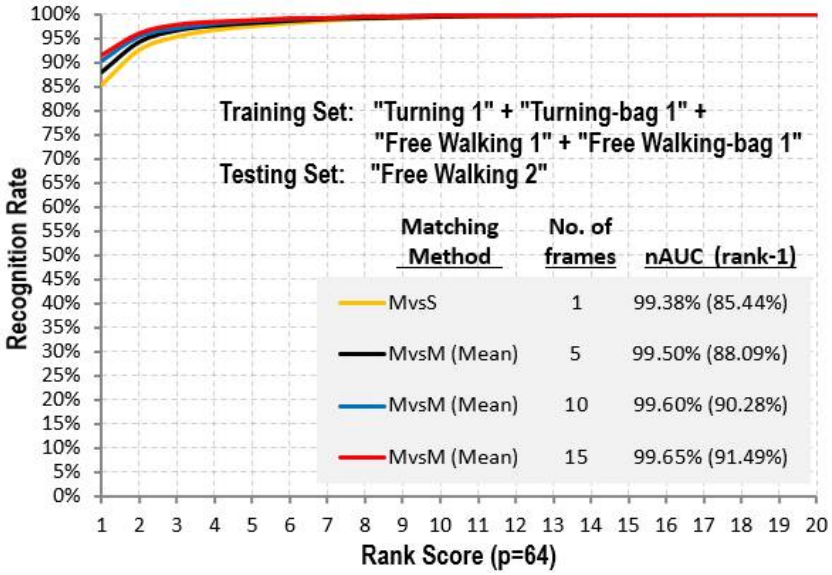


Figure 4.41. Ex4: CMC curves showing results for single-shot (MvsS) vs mean multi-frame (MvsM) method with different number of frames. (Best viewed in colour)

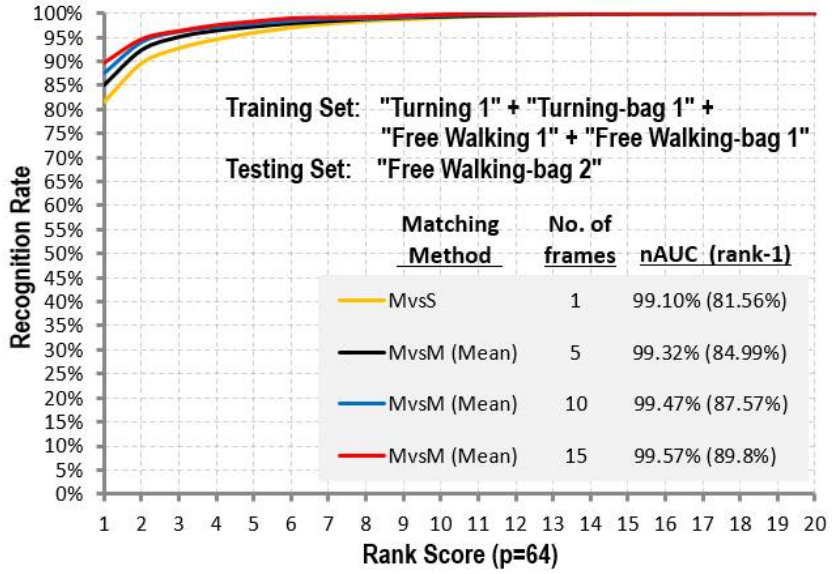


Figure 4.42. Ex4: CMC curves showing results for single-shot (MvsS) vs mean multi-frame (MvsM) method with different number of frames. (Best viewed in colour)

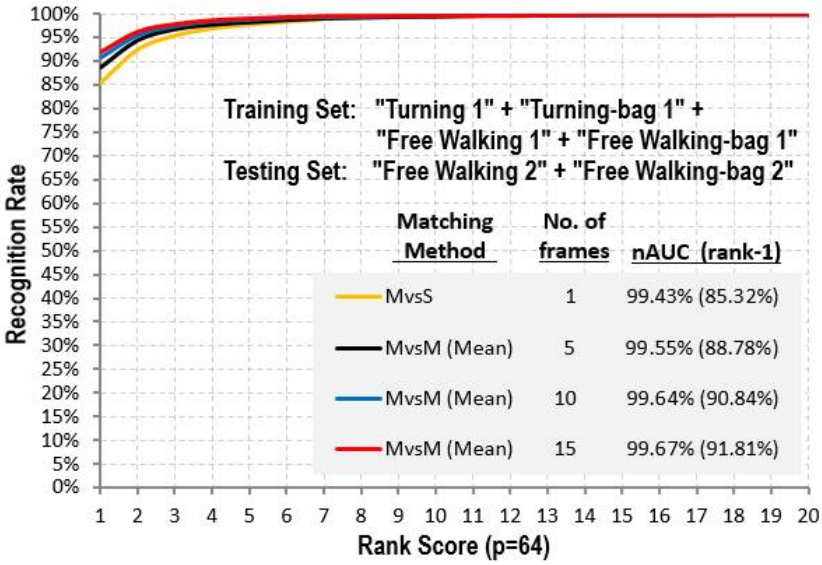


Figure 4.43. Ex4: CMC curves showing results for MvsS vs MvsMmethod. "Free Walking 2" contains testing data for person with ID=01 to 32, while "Free Walking-bag 2" contains person with ID=33 to 64. (Best viewed in colour)

Using the same C12 as the classifier, a test was carried out to find out if classifier C12's performance on test data "Free Walking-bag 2" was similar to that of "Free Walking 2". It can be observed from the results in Figure 4.41 and Figure 4.42 that performance of classifier C12 on "Free Walking-bag 2" was lower than that of classifier C12 on "Free Walking 2". This was believed to have been caused by the inaccuracy of ellipse fitting method for some styles of carrying backpacks such as shown by the last image in Figure 4.19.

Analysis of the results was done to see if there exists any range of body orientation that performs better or worse than any other ranges. One might assume that the frontal side of body might perform better than the other body orientations as have been reported by most work in the literature because it contains the most visual cue. The best performance (i.e. accurate classification) is indicated by rank-1 and the less accurate performance is indicated by rank-n with **n** > 1. This analysis was done to the experiment in Figure 4.37. It can be seen from Figure 4.44 that the distribution of single-shot results of 27,351 frames with rank-2 and above are fairly distributed amongst all body orientation angles. This shows that all body orientations performed

equally well. The next test was to validate an assumption that, for experiment in Figure 4.39, the reduced performance of the classifier on “Free Walking-bag 1” was contributed by the absence of appearance features of persons carrying bags in classifier C9 with the training set “Turning 1” + “Free Walking 1”. Specifically this “bag-carrying” appearance in the test set only resides in the body orientation range of about 140^{O} to 220^{O}. Therefore one could expect more rank-2 and lower classifications (rank-1 is the highest) in the 140^{O} to 220^{O} range. This is shown to be true as evidenced in Figure 4.45.

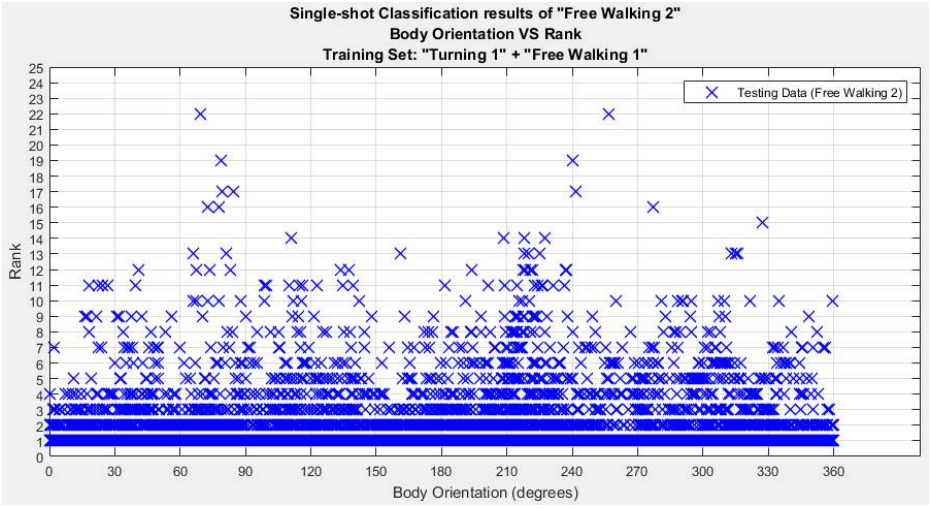


Figure 4.44. Fair distribution of ranks vs different body orientations. Number of testing data (frames) is 27,351.

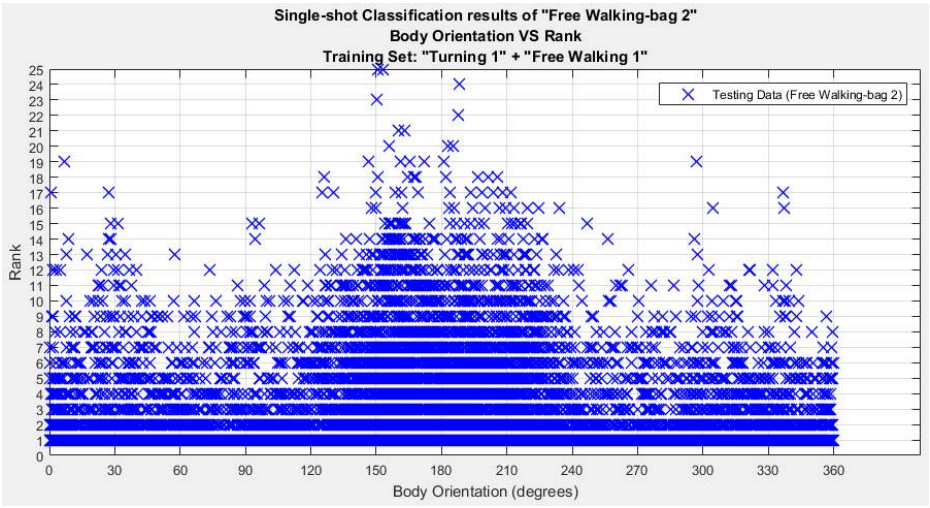


Figure 4.45. Distribution is concentrated more (high ranks) around 140^{O} to 220^{O}. This angle range represents the back view of persons carrying bags or backpacks.Number of testing data (frames) is 27,451.

The improvement of re-identification results shown in Figure 4.44 via multi-shot matching technique (M=15) is illustrated in Figure 4.46. Figure 4.47 represents multi-shot results for Figure 4.45.

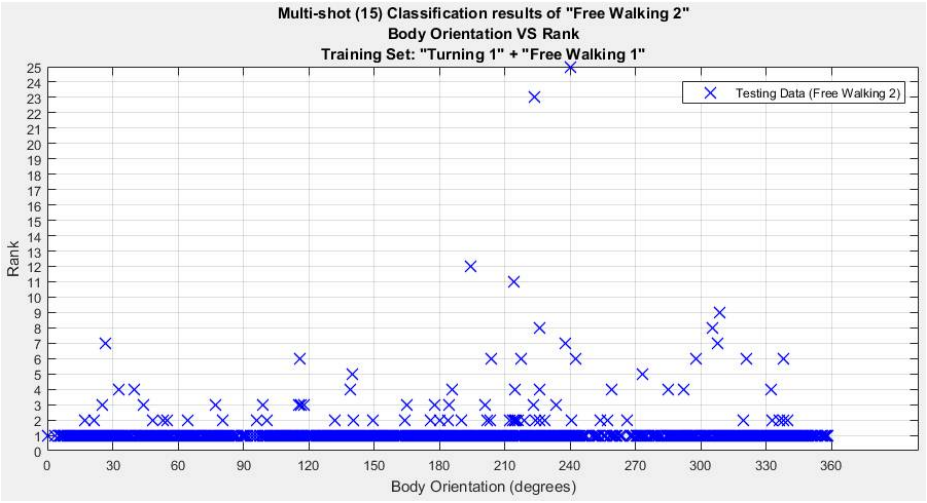


Figure 4.46 Distribution of ranks vs different body orientations looks cleaner than Figure 4.44, when multi-shot matching technique is applied. Number of testing data (frames) is 1,823.

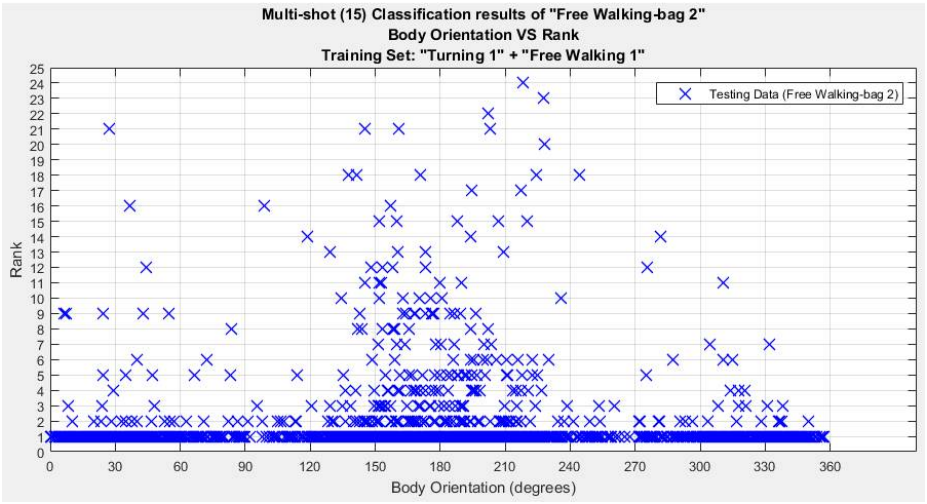


Figure 4.47. Distribution of ranks vs different body orientations looks cleaner than Figure 4.45 when multi-shot matching technique is applied. However distribution of high ranked classification is still concentrated more around 140^{O} to 220^{O}. Number of testing data (frames) is 1,830.

The fifth experiment (Ex5) was carried out to evaluate the classifier C12's performance on a scenario such as in public space where an interactive tabletop display is located. This scenario demonstrated some challenging situations for person re-identification such as non-upright standing posture causing inaccurate height measurement, hand gestures on the display obscuring front body's appearance cue, and occlusion problem when a person is behind the

tabletop. These situations were believed to have caused the decrease in performance of classifier C12 on “Around Tabletop” test set as shown in Figure 4.48.

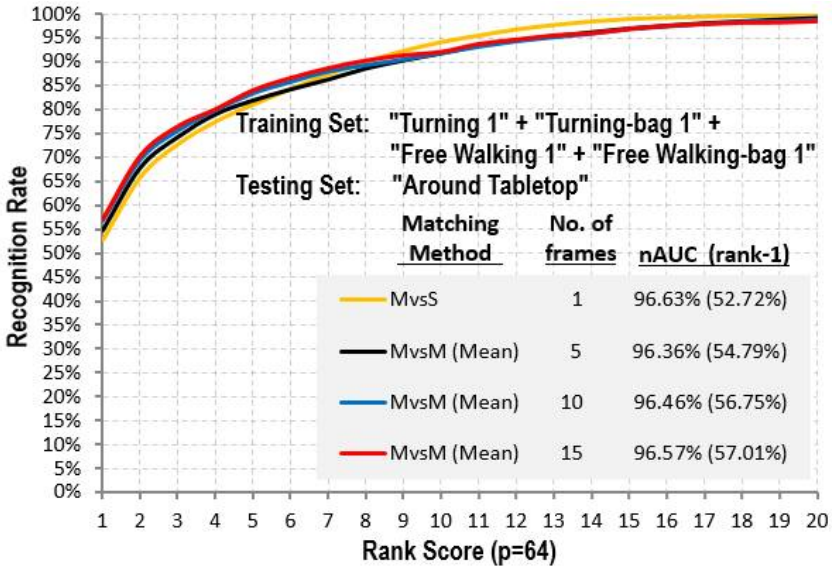


Figure 4.48. Ex5: CMC curves for classification results of single-shot (MvsS) and multi-shot (MvsM) method. (Best viewed in colour)

The occlusion effect of the tabletop display on the classification performance was investigated. It was discovered that occlusion might have contributed to the drop of performance as shown by rank-vs-distance plot in Figure 4.49. A cluster of plots with rank-2 and lower for distance range of more than 3.5m (d > 3.5m) indicates incorrect classification occurred more at this distance range. The furthest two sides of the tabletop display where occlusion happened, are in the same distance range, illustrated in Figure 4.49. Improvement of classification results was achieved when test data with d > 3.5m was removed as illustrated in Figure 4.50. This suggests that occlusion from the tabletop on the body indeed caused the little drop on the re-identification performance.

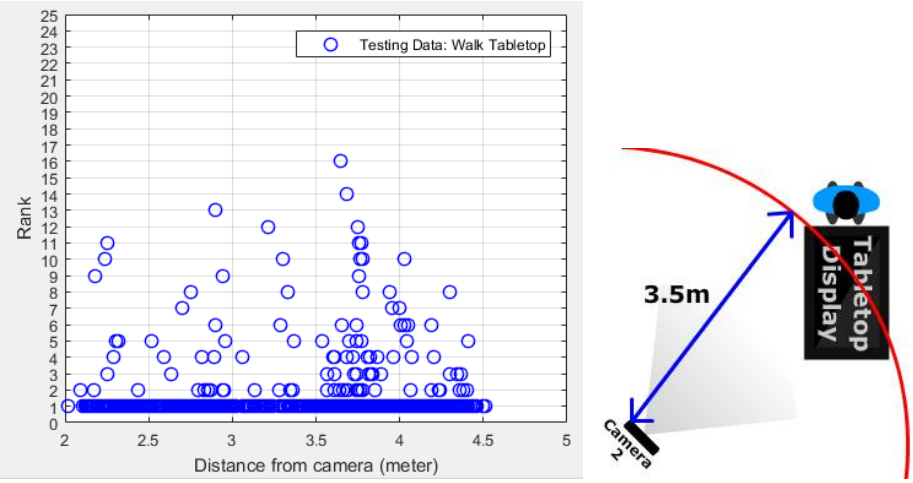


Figure 4.49. A cluster of incorrect classification (rank-2 and lower) for distance range more than 3.5m (left). Distance of tabletop display from the camera (right).

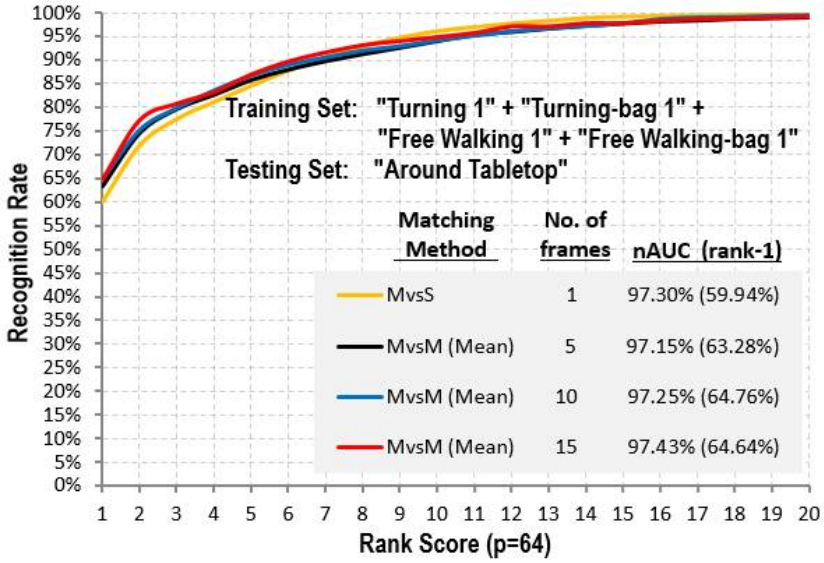


Figure 4.50. Ex5: CMC curves for classification results of single-shot (MvsS) and multi-shot (MvsM) method with maximum distance of 3.5m. (Best viewed in colour)

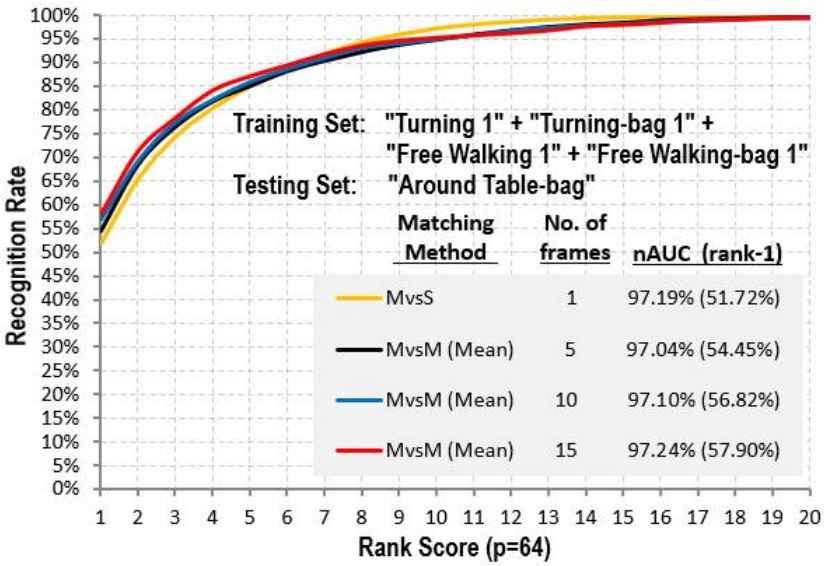


Figure 4.51. Ex5: CMC curves for classification results of single-shot (MvsS) and multi-shot (MvsM) method with no maximum distance. (Best viewed in colour)

Table 4.8 lists rank-1, rank-5 and rank-10 re-identification performances of single-shot (MvsS) and multi-shot (MvsM) matching methods (where available) reported in the literature (Munaro et al., 2014b) and (Pala et al., 2015). In the absence of compatible public datasets, it is, of course, not possible to directly compare ViMM with other re-identification methods. However, they are included in Table 4.8 to contrast with ViMM performance. The ViMM results selected for comparison here were achieved using the new KinectV2 RGBD-ID dataset with classifier C9: “Turning 1” + “Free Walking 1” as training and “Free Walking 2” as testing.

Table 4.8 Classification results of selected methods from literature.

Dataset: BIWI RGBD-ID (50 people, unconstrained viewpoints)

	Rank-1		Rank-5		Rank-10	
Method from (Munaro et al., 2014b) **\ Frames**	Single	Multi*	Single	Multi	Single	Multi
Skeleton descriptor	27%	32%	60%	-	81%	-
Point Cloud matching	33%	43%	60%	-	78%	-
Face + SURF	44%	57%	72%	-	85%	-
Face + SURF + Skeleton	52%	68%	84%	-	90%	-

Dataset: KinectREID (71 people, three viewpoints)

	Rank-1		Rank-5		Rank-10	
Method from (Pala et al., 2015) **\ Frames**	Single	Multi 10 frames	Single	Multi 10 frames	Single	Multi 10 frames
SDALF multimodal	-	42%	-	70%	-	83%
eBiCov multimodal	-	45%	-	70%	-	83%
MCMimpl multimodal	-	52%	-	78%	-	87%

Dataset: RGBD-ID (79 people, two viewpoints)

	Rank-1		Rank-5		Rank-10	
Method from (Pala et al., 2015) **\ Frames**	Single	Multi 5 frames	Single	Multi 5 frames	Single	Multi 5 frames
SDALF multimodal	-	58%	-	95%	-	99%
eBiCov multimodal	-	55%	-	92%	-	99%
MCMimpl multimodal	-	88%	-	97%	-	99%

Dataset: KinectV2 RGBD-ID (64 people, unconstrained viewpoints)

	Rank-1		Rank-5		Rank-10	
Proposed Method \ Frames	Single	Multi	Single	Multi	Single	Multi
ViMM (Classifier C9)	86.3%	5 frames 89.0%	98%	5 frames 98.3%	99.8%	5 frames 99.5%
		10 frames 90.8%		10 frames 98.6%		10 frames 99.6%
		15 frames 92.4%		15 frames 99.3%		15 frames 99.8%

Rank-1, rank-5 and rank-10 classification results of methods reported in (Munaro et al., 2014b) using the BIWI RGBD-ID dataset, (Pala et al., 2015) using the KinectREID and RGBD-ID datasets, and for the ViMM method using the KinectV2 RGBD-ID dataset. *The number of frames used for the BIWI RGBD-ID MvsM results are not reported in (Munaro et al., 2014b).

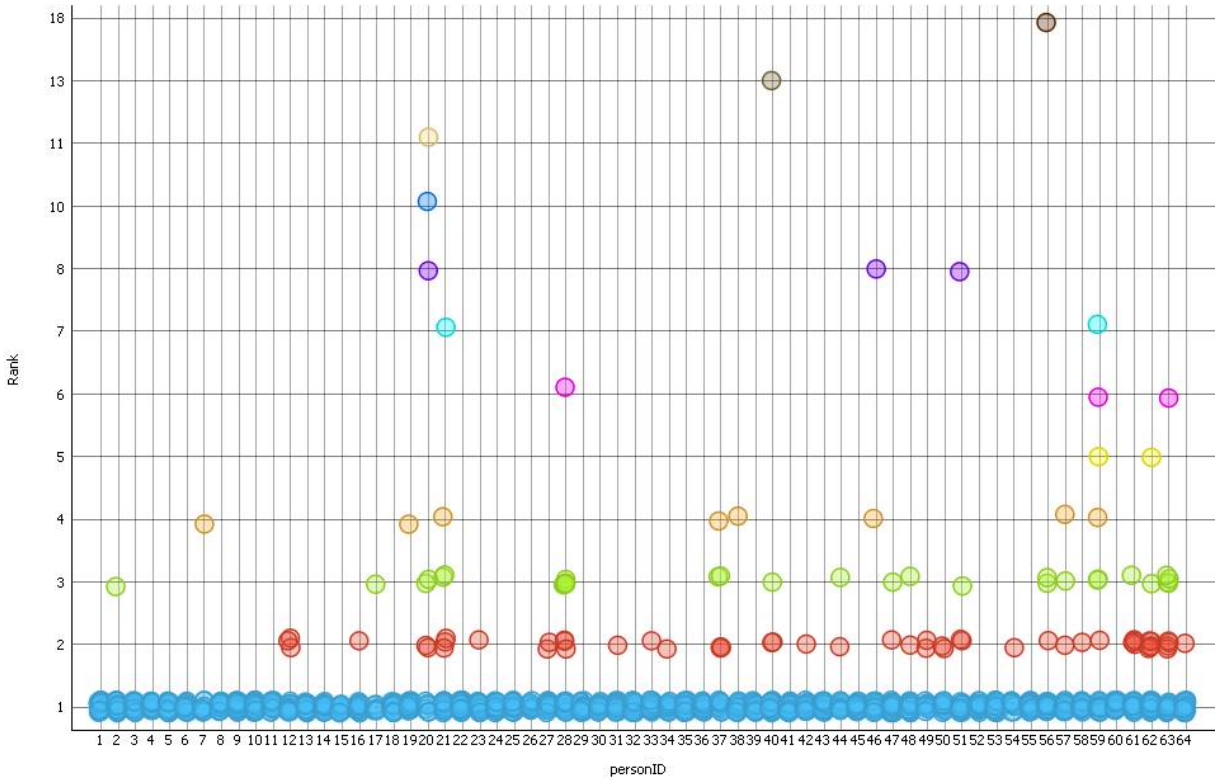


Figure 4.52. This figure represents the rank-1 result of ViMM in Table 4.8. The scatter plot shows "MvsM (M=15)" rank-1 classification results of classifier C9 on "Free Walking 2" test data, based on rank, for each person with ID=1 to 64.

It can be observed from Figure 4.52 that 30 people have perfect rank-1 classification results while very few have above rank-10.

4.7 Prototype Application

A prototype application was developed to validate the effectiveness of ViMM on actual video feeds from RGB-D recordings in the dataset. The application also features a walking direction computed from the body orientation estimation. This classifier was trained using "Turning 1" + "Free Walking 1" + "Turning-bag 1" + "Free Walking-bag 1" and the test video was played by a player in Kinect Studio v2.0. The prototype application written in C#, captures the colour and depth frames from the Kinect Studio, performs features extraction and sends re-identification query to the pre-trained classifier for every frame received. The whole re-

identification process is performed at 15 fps and decision is performed at 1 fps for the multi-shot decision technique with M=15.

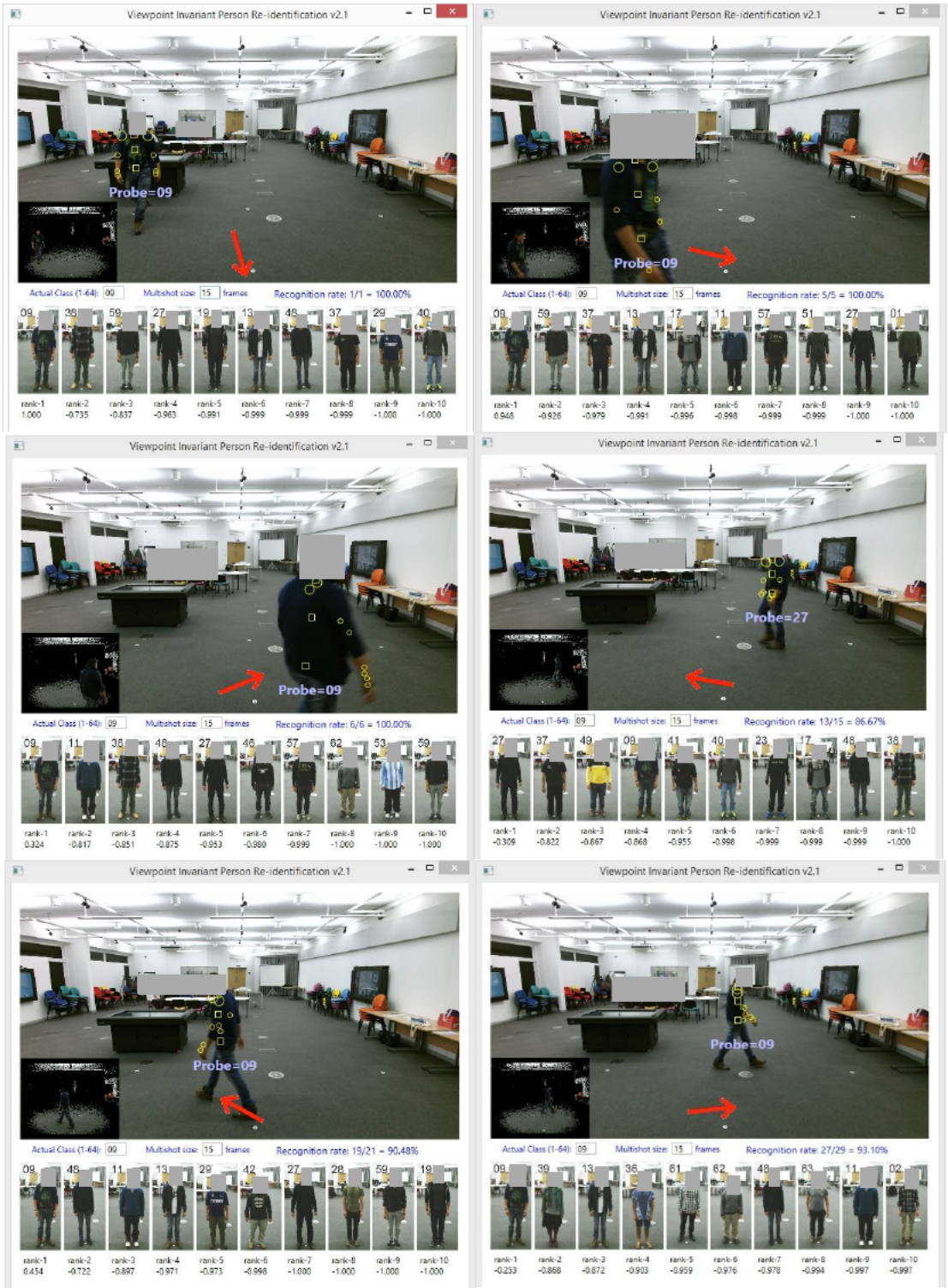


Figure 4.53 Screen grabs from prototype application showing a sequence of frames extracted from a dataset component "Walking 2" with arrows indicating the directions of body orientations.

4.8 Expanded Features Set

The ViMM feature descriptor was designed from a combination of 2-D and 3-D appearance features of the upper half of a human body because of their high chance of continuous visibility compared to the lower body parts. However for further exploration, the features from lower body parts (i.e. colours at knee and foot level) were included into ViMM and was named as ViMM v0 (version 0) as described by Listing 1 in Section 4.4.2. There were two variations of ViMM v0, one, called ViMM v0 ex RGB, only includes the *rg*-chromaticity values for images at knee and foot level, and the other, called ViMM v0, includes all R, G and B values for all the five body parts, i.e. shoulder, mid-spine, hips, knee and foot level.

4.8.1 Experimental Results

Results of the experiments are presented below and the summary of results when using the expanded features set is presented in Table 4.9.

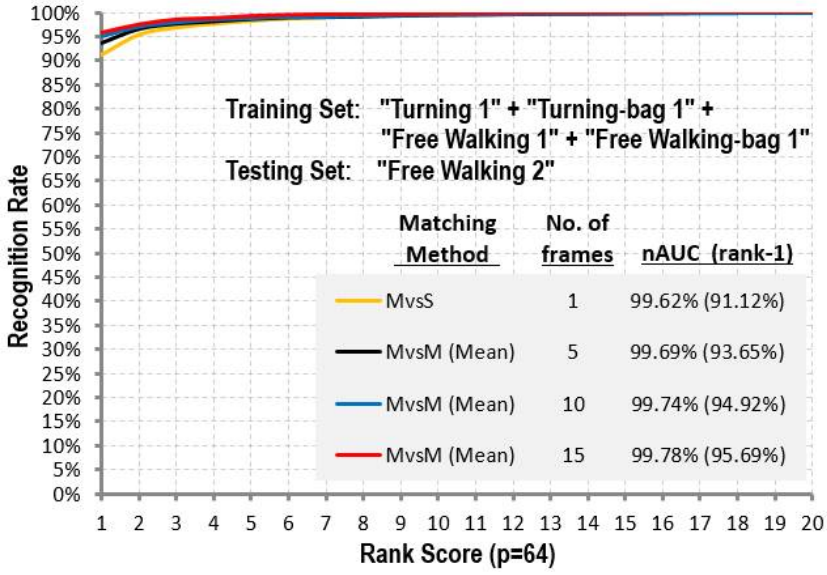


Figure 4.54. Ex6: Classification results for ViMM v0 ex RGB (without RGB colour information).

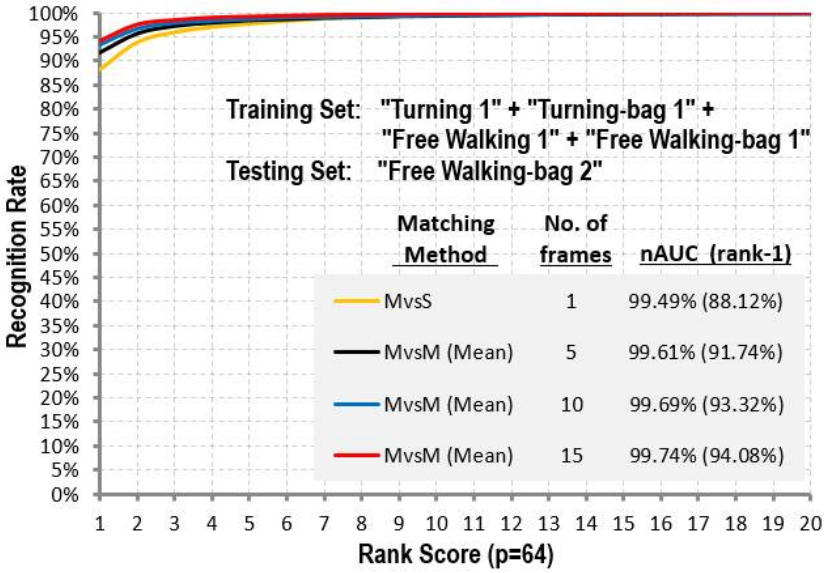


Figure 4.55. Ex6: Classification results for ViMM v0 ex RGB (without RGB colour information).

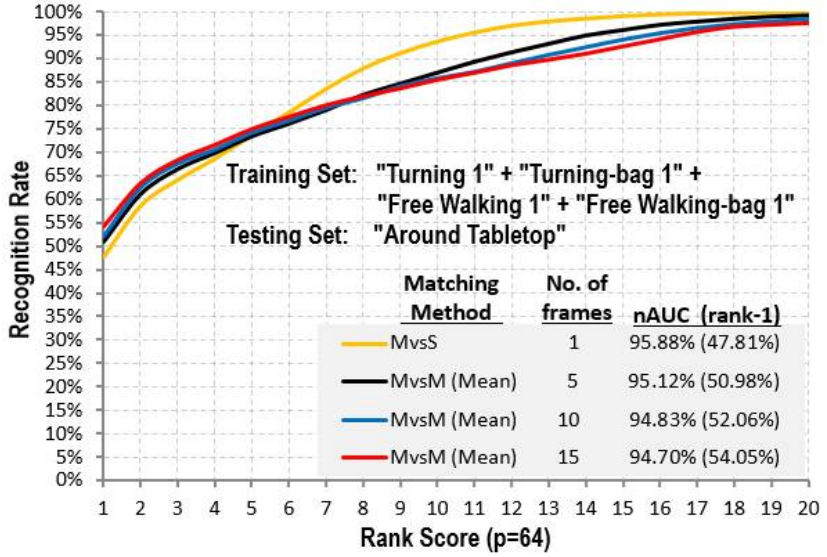


Figure 4.56. Ex6: Classification results for ViMM v0 ex RGB (without RGB colour information).

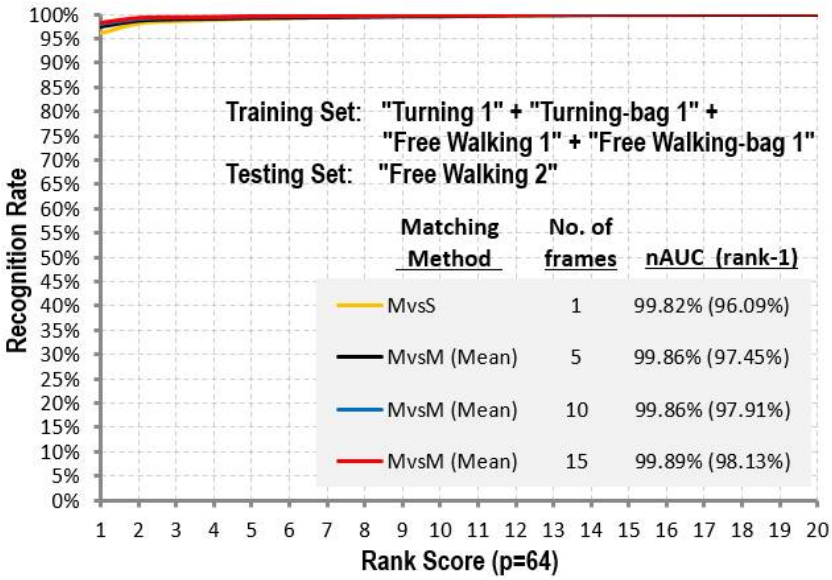


Figure 4.57. Ex7: Classification results for ViMM v0 (with RGB colour information).

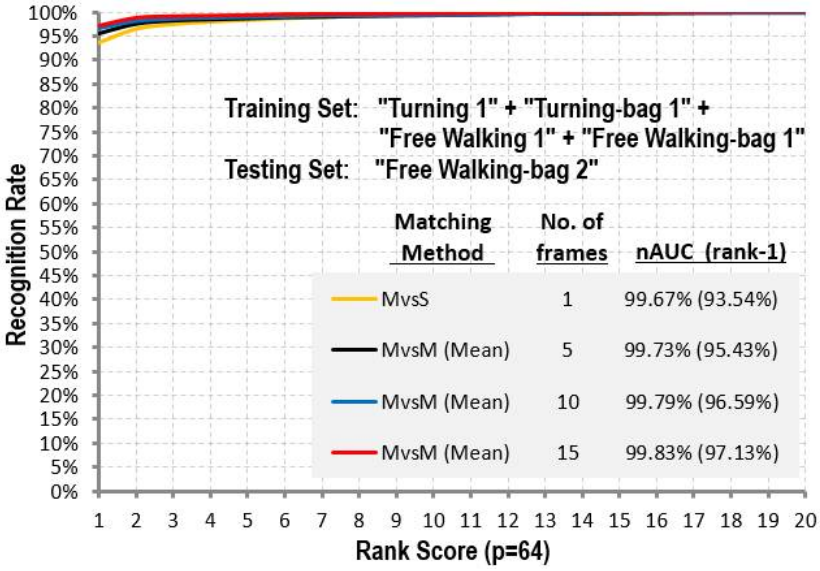


Figure 4.58. Ex7: Classification results for ViMM v0 (with RGB colour information).

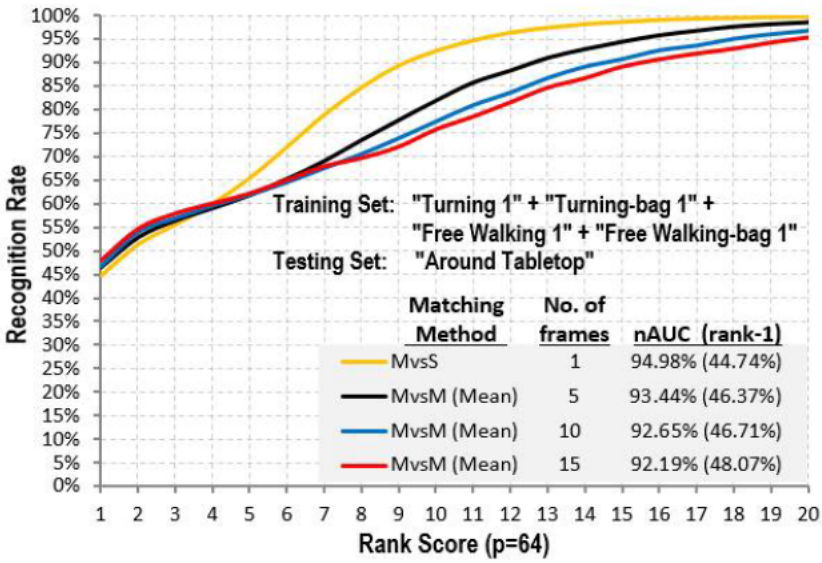


Figure 4.59. Ex7: Classification results for ViMM v0 (with RGB colour information).

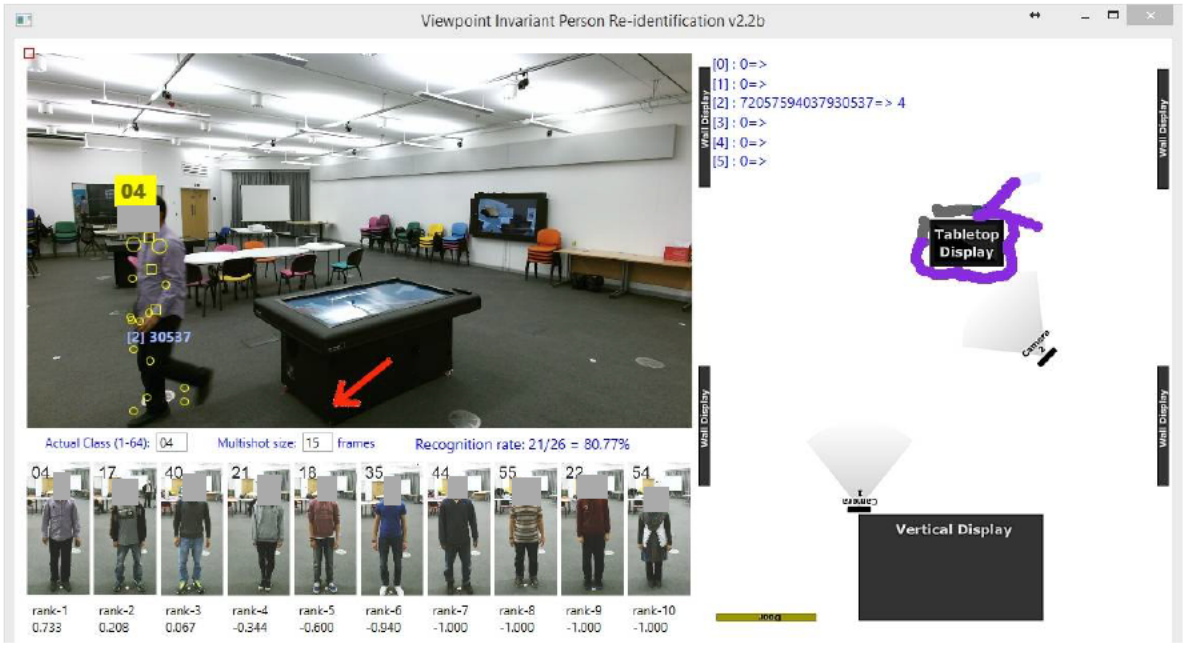


Figure 4.60. Example scenario where using "ViMM v0 ex RGB" causing incorrect classification when a person is behind the tabletop display, hence the lower body parts was occluded. Correct classification (80.77%) is indicated visually by blue tracking plot on the right side of the figure.

Figure 4.61. Incorrect classification problem from Figure 4.60 was solved with the use of ViMM. Correct classification (100%) is indicated visually by blue tracking plot.

Table 4.9. Summary of results of re-identification for ViMM, ViMM v0, and ViMM v0 ex RGB.

Experiment ID	Classifier ID	Testing Set	Feature Set	Rank-1 (%)		
				MvsS	MvsM M=5	MvsM M=15
Ex4	C12	Free Walking 2	ViMM	85.44	88.09	**91.49**
Ex4	C12	Free Walking-bag 2	ViMM	81.56	84.99	**89.80**
Ex5	C12	Around Tabletop	ViMM	52.72	54.79	**57.01**
Ex6	C13	Free Walking 2	ViMM v0 ex RGB	91.12	93.65	**95.69**
Ex6	C13	Free Walking-bag 2	ViMM v0 ex RGB	88.12	91.74	**94.08**
Ex6	C13	Around Tabletop	ViMM v0 ex RGB	47.81	50.98	**54.05**
Ex7	C14	Free Walking 2	ViMM v0	96.09	97.91	**98.13**
Ex7	C14	Free Walking-bag 2	ViMM v0	93.54	95.43	**97.13**
Ex7	C14	Around Tabletop	ViMM v0	44.74	46.37	**48.07**

It was expected that classification results would improve when the feature representation from the lower body parts are included as evidenced by Figure 4.54 and Figure 4.55. A significant improvement of rank-1 classification from 91.49% to 95.69% for "Free Walking 2" test set was observed when colour information for knee and foot were included (see Figure 4.54). Similar

improvement of rank-1 classification was observed, from 89.80% to 94.08% for “Free Walking-bag 2” test set. Experiments were also performed on the similar training and testing sets using ViMM v0 feature descriptor but with the RGB colour values added on top of the *rg*-chromaticity values (see Figure 4.55). It is worth emphasising that this might not be suitable for real scenario as RGB colour information is not illumination invariant. The rank-1 classification results for “Free Walking 2” test set was 98.13% (see Figure 4.57), and 97.13% (see Figure 4.58) for “Free Walking-bag 2” test set. Further tests of the expanded features set on “Around Tabletop” test set revealed rank-1 performance drop by nearly 3% when using classifier C13, and nearly 9% when using classifier C14, compared to performance of classifier C12. This was expected due to occlusion of lower body parts caused by the tabletop (e.g. when a person moved behind the tabletop).

The performance recorded by the expanded features set “ViMM v0” even though very good, it may not be useful for practical usage because of its non-illumination-invariant property. It however can indicate that the ViMM possesses strong discriminant quality for person re-identification and is robust when used as function of body orientation angle.

.

4.9 Conclusions

A novel viewpoint invariant feature descriptor for person re-identification was presented. An experimental dataset comprising sixty-four people turning and walking freely was acquired using Kinect V2 cameras and was used to test the classification performance of the ViMM feature descriptor. The results showed that the performance of the ViMM descriptor vector compares favourably with other methods and that performance improves further if observations from multiple frames are combined. The method was previously tested using Kinect V1 cameras but ellipse fitting estimation proved inaccurate as a result of noisy depth data. The improved specification of the Kinect V2 camera (Lachat et al., 2015) provides the accuracy and resolution required for reliable performance resulting in robust ellipse fitting estimation which is necessary in obtaining accurate body orientation estimation. The nearly identical performances of the person re-identification when classifiers were trained using different machine learning algorithms suggest that the ViMM feature descriptor possesses strong discriminant properties as the descriptor does not depend on specific machine learning algorithm to achieve the best performance.

The limitation of the proposed method is that it requires observed persons to stand or walk with normal pose – that is, with both arms down (small movements when walking are not problematic). Raising hands to the shoulder level or above will disrupt the ellipse estimation hence affecting extraction process of features that are based on the ellipse estimation. This limitation, however, could be obviated by using joint information to detect instances of non-compliant pose.

CHAPTER 5:

Multi-person Re-identification

5.1 Introduction

It is envisaged that when people walk around an exhibition space, each of them can be tracked and labelled with a unique identifier. It is not uncommon for people in such spaces to be with or amongst other people when moving, standing, or interacting with exhibits. For this reason, identifying individuals in a multi-person scene will be desirable. The focus of this chapter is multi-person re-identification. It continues on from person re-identification presented in Chapter 4, with the purpose of investigating the suitability and effectiveness of the C12 classifier in multi-person scenarios. C12 was selected because it covers all conditions of people walking with and without bags, pre-trained using "Turning 1" + "Free Walking 1" + "Turning-bag 1" + "Free Walking-bag 1" training sets and ViMM feature descriptor.

5.2 Multi-person Re-identification

The free-walking and walking to tabletop activity components 15, 16, 17 and 18 listed in Table 5.1 (and originally listed in Table 4.4) were used for experiments. However, unlike in Chapter 4 where these activities were performed individually, here these activities were performed concurrently by groups of 3 and 4 people. The C12 classifier and multi-shot matching method (MvsM) with M=15 was used for classification.

Table 5.1. Dataset components for multi-person re-identification.

Dataset Component ID	Description of activity	Carry bag?	Individual/ Group
15	**Free walking** in front of a Kinect camera		Group
16	**Free walking** in front of a Kinect camera	Yes	Group
17	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.		Group
18	**Walking towards** a **tabletop display,** and go around the table while interacting with the table briefly from each side of the table.	Yes	Group

5.2.1 Experimental Design and Datasets Creation

In activity 15, people were asked to walk freely in a space within the field of view of a camera, in random directions, for approximately 15 seconds. Activity 16 is the same as 15 except that people were required to carry backpacks, handbags or briefcases. The recordings produced by these activities do not truly represent normal scenarios, for example, it was observed that in this activity people tended to walk and turn quite quickly and occlusions happened frequently as they moved between each other and crossed paths with each other a number of times. In more normal scenarios, people's movement might be expected to have more consistency in pace or direction. For example, rush hour commuters may move very quickly, but their directions are more consistent. Visitors to a museum may move more slowly but their direction may be very varied. The combination of fast pace and varied direction in these activities adds to the challenge implicit in multi-person re-identification.

In activity 17, the same people were asked to approach a tabletop display from different directions. This activity was designed in such a way to represent people coming from any possible direction to any side of the tabletop display. The camera was placed at an angle facing two diagonally opposite corners of the tabletop on a straight line (as can be seen in Figure 4.25

in the previous chapter). This position was derived empirically as the position for best observation of arriving individuals though, of course, it was not possible to avoid occlusions altogether.

5.2.2 Experimental Results

The aim of the experiments was to evaluate the performance of the ViMM feature descriptor in the multi-person scenario using the classifier C12 trained using “Turning 1” + “Free Walking 1” + “Turning-bag 1” + “Free Walking-bag 1”. Dataset components 15, 16, 17 and 18 were used for testing. The method of testing is described below.

The tracking outputs of selected participants from the ViMM prototype application are shown in Figure 5.1 for the “Free Walking 2” activity. Unfortunately, the occlusions were too frequent to enable the acquisition of 15 consecutive frames per person needed for the MvsM re-identification algorithm and without these it was not possible to compile cumulative classification statistics. As expected, the number of occlusions increased with the number of people participating in the activity. Recognition for individuals for whom valid frames were acquired was achieved, for example, in Figure 5.3 the person with ID=05 recorded a rank-1 classification of 96.77 %.

The classification results for “Around Tabletop” and “Around tabletop-bag” activities are shown in Figure 5.3 and Figure 5.4, respectively.

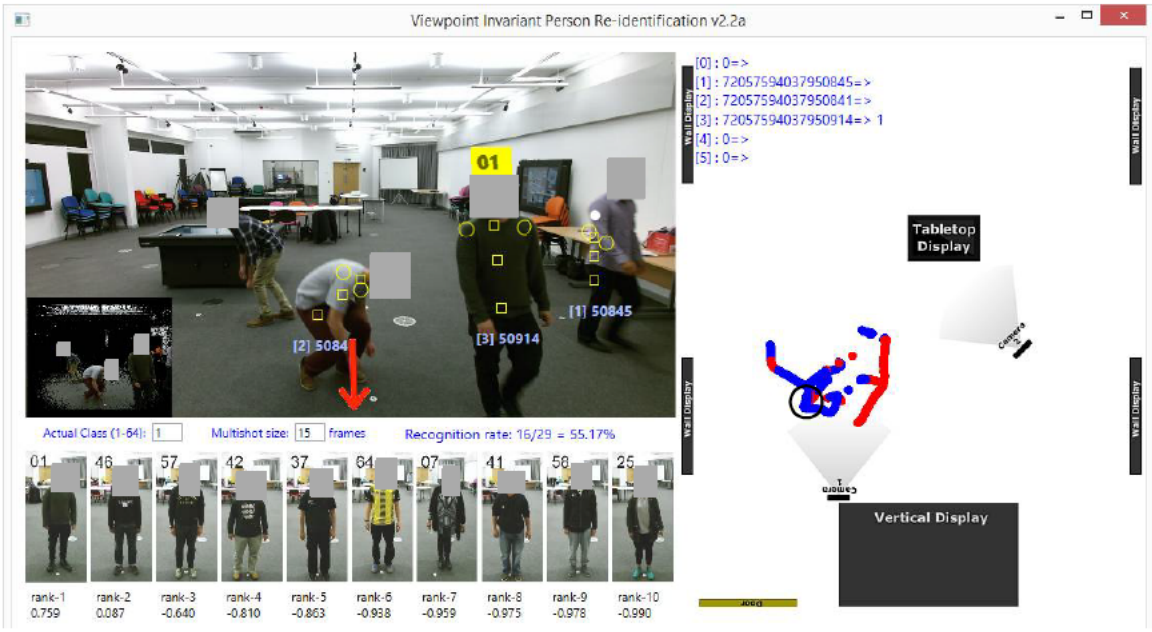


Figure 5.1. The red plot indicates misclassifications happened earlier in the video. The current position is marked by the black circle on the map. The red arrow takes the estimate of body orientation angle directly from the ViMM's ellipse fitting method.

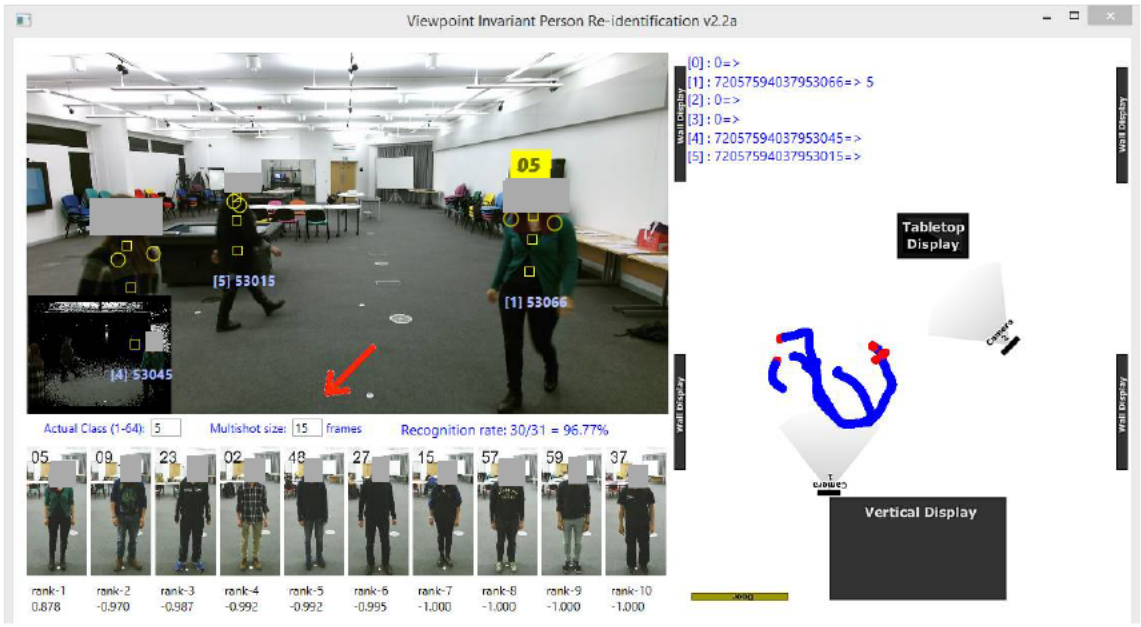


Figure 5.2. The scenario above has three persons walking freely within the field-of-view of the camera. Less occlusions occurred with three people when compared to four. The person with ID=05 recorded rank-1 classification of 96.77%.

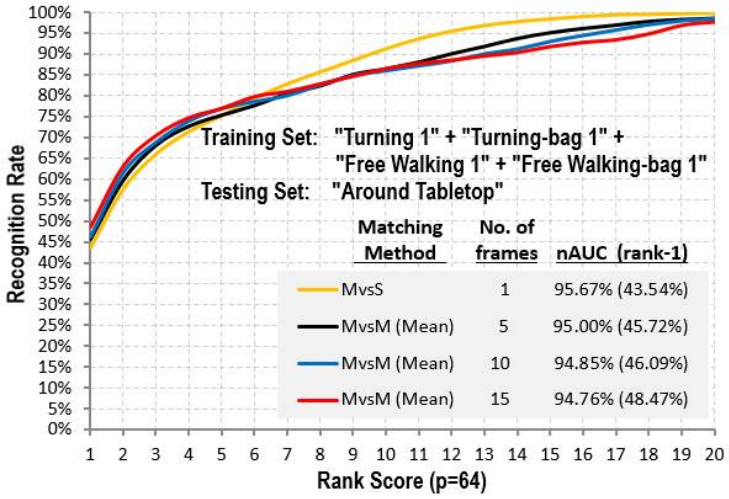


Figure 5.3.Classification results for classifier C12 on testing set “Around Tabletop”.

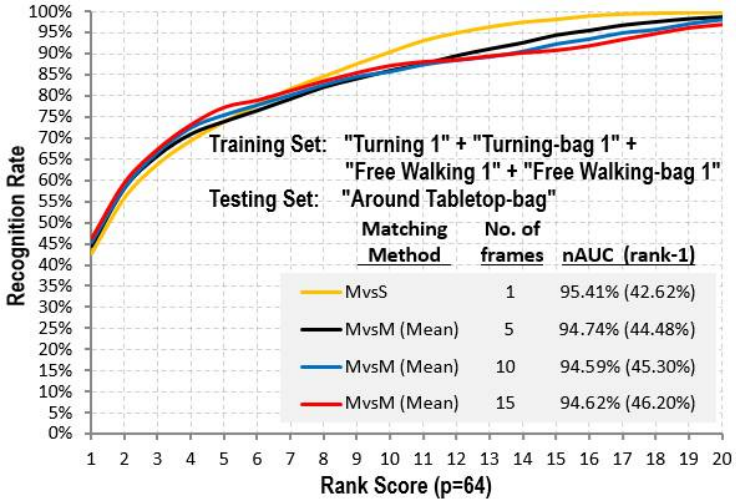


Figure 5.4. Classification results for classifier C12 on testing set “Around Tabletop-bag”.

5.2.3 Discussion

The results achieved in this multi-person re-identification testing highlighted the problems of occlusions when spaces become more populated and using only one camera. Where valid numbers of frames could be acquired, the results suggest that ViMM re-identification has the potential to perform satisfactorily. The multi-person “Around Tabletop” activity achieved only rank 1 classifications in the region of 40-50%. From observation this reduced performance was, again, due to occlusions, in particular, occlusions caused by individuals at the tabletop (i.e.

close-by) blocking the view of those approaching. The occlusions problem can be solved if multiple cameras are used by placing cameras around the tabletop.

It is hypothesised that people approaching the tabletop display could be identified whilst walking towards the tabletop and, once they are about to reach the table, the human-sensing tabletop system could detect their presence and connect to their identities from the re-identification system using their sensed position at the table. The human-sensing tabletop system could then provide personalised services to those around the tabletop. By using the body orientation information from the re-identification system, the tabletop system will also know when people are leaving the table and will perform the necessary clean-up of personalised services such as closing opened folders or objects.

5.2.4 Conclusions

Integration between human-sensing tabletop display (or any other types of display) and multi-person re-identification system was achievable with satisfactory results. Simultaneous re-identification of multiple people can be performed by running six instances of the re-identification application, each instance taking one person and performing re-identification. This method was tested on a laptop computer with a configuration of 2.4 GHz Intel i7 processor and 16 GB RAM. With four instances running at the same time, the overall system performance slows with the re-identification process performed at 2 fps. System performance is improved to 10 fps when three instances running at the same time, and with two instances, the performance is greatly increased to 20 fps. It is expected that this limitation can be minimised by code optimisation, and eliminated when a faster machine is used or a general-purpose GPU (GPGPU) parallel programming implementation is employed.

The future recommendations from this chapter would be to place extra units of the Kinect cameras around the space or tabletop so that occlusions can be minimised especially for the “tabletop” scenario.

CHAPTER 6:

Context-Aware System with Person Re-identification

The aim of this chapter is to demonstrate how the integration of the person re-identification system with a context-aware system can be realised. The human-sensing tabletop system (presented in Chapter 3) will be used as an example, but generally any context aware system may benefit from the identity and location information. This chapter presents examples of selected individuals and groups from the ViMM dataset components performing activities, simulating scenarios that can be found in an interactive environment such as a digital exhibition at a museum.

6.1 System Design and Architecture

Context aware systems typically have their own sensing module or layer comprising of sensors, raw data retrieval, pre-processing and data storage management (Kohli and Jetawat, 2012). The last layer is an application layer acting as the consumer of the sensing module. The ViMM person re-identification module (labelled and illustrated in Figure 6.1) acts as a sensing module and context provider for context aware systems, in which the output of the context provider is simply a world coordinate location of an identified person that is within the field of view of the camera. A context aware system would continuously check the output from ViMM module if there is any person detected within its space. If a person is detected, ViMM would provide the identity and location information of the person to the context aware system.

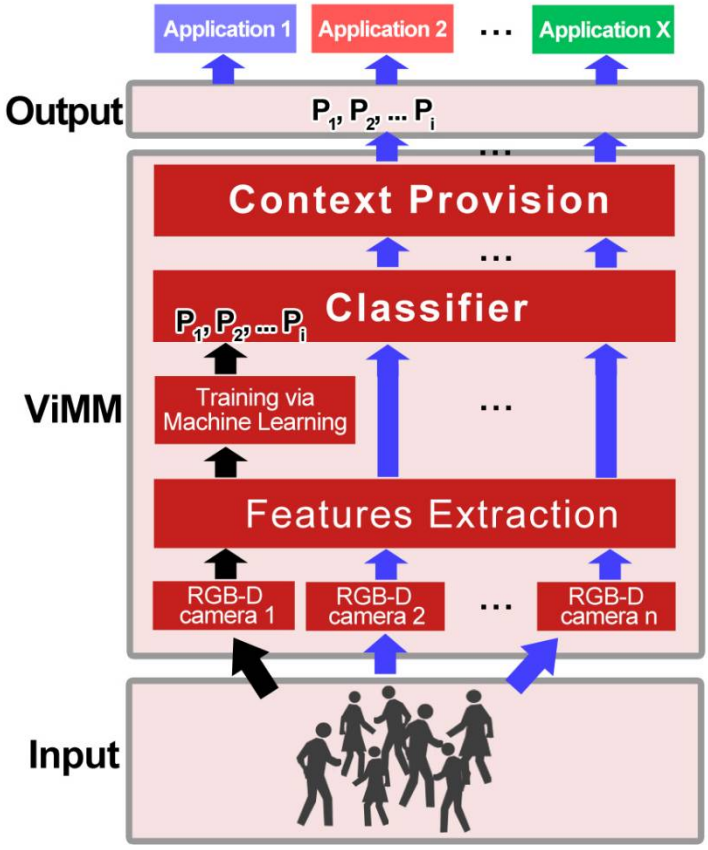


Figure 6.1. Architecture of Person Re-identification with Context-Awareness. Application layer at the top represents any context aware systems.

6.2 Simulated Scenarios: Digital Exhibition at a Museum

To demonstrate how an intelligent human-sensing tabletop could work, let us assume that there were a number of tabletop displays in the museum, and the tabletop being used in the example in this chapter was running an image gallery application containing high resolution images of heritage objects. In a normal condition, a visitor would approach the tabletop display and start interacting with the gallery application by performing multi-touch gestures on the images. Examples of such gestures include touch to select, drag to move and pinch to zoom.

It is worth mentioning that the ViMM module maintains a look-up table containing the identity, body orientation and real world location of detected persons either locally or at a central location accessible via network. There can be other tables in a database storing history of every person's locations, and contents of virtual folder belonging to every person.

6.2.1 Scenario 1: Person working on a tabletop display

The scenario is described below by a series of illustrations in Figure 6.2 to Figure 6.5. It should be mentioned that the colour images from the Kinect v2 camera, presented in figures in this thesis are mirrored, keeping the original data streams from the Kinect. There is no built-in way for the SDK to "un-mirror" the image, and no attempt was made to perform this as it does not affect the quality of this research.

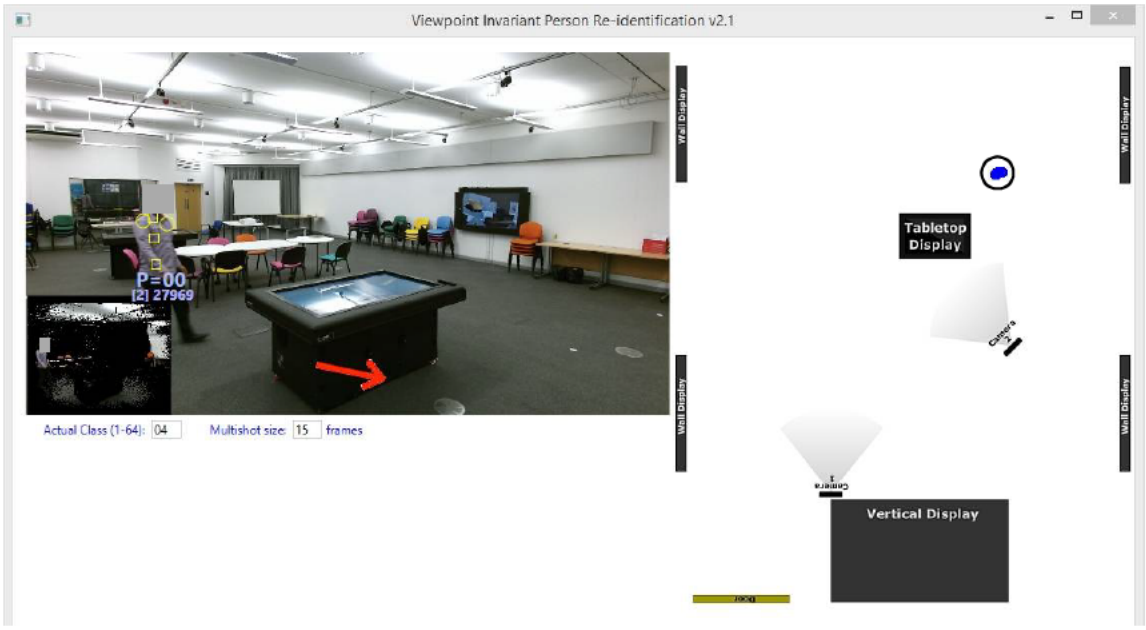


Figure 6.2. A person is seen entering the field of view of the camera. The blue trail on the map shows the position tracking of the person. The current position is marked by the black circle on the map. The red arrow takes the estimate of body orientation angle directly from the ViMM's ellipse fitting method.

A person was walking towards the tabletop display. It took around a second after the person was first detected, before ViMM could produce re-identification output. This was because ViMM used MvsM matching technique with M=15, that is 15 frames are needed before decision could be made.

Figure 6.3. A person arrives at the tabletop display. ViMM identifies this person as having an ID=04. Location tracking is marked on the map by the blue trail.

The identity of the person was known with a simple query on the look-up table. If the person had previously created a virtual folder containing images he collected before, the tabletop would retrieve the images and present the folder to the person automatically. Otherwise, a new virtual folder would be created. The person may reorganise his virtual folder by deleting or adding new images. The body orientation information could indicate that the person was currently interacting with the content on the table. The tabletop would be able to use this information to re-orient interface elements such as buttons and texts, and other images on the tabletop to be in a correct orientation.

Figure 6.4. The person is seen interacting with the content on the tabletop display.

The person moved to the other side of the table as can be seen from the blue trail in Figure 6.4. The tabletop would move his virtual folder to a new position by following his current location. The orientation of interface elements and other images were dynamically changed to align with the person's orientation.

Figure 6.5. The person is seen leaving the tabletop display. ViMM achieves 100% rank-1 classification in this scenario.

The tabletop could deduce from the location distance and the body orientation angle, that the person was leaving the table. The tabletop then could perform a clean-up i.e. save, close and remove the virtual folder from the display, and reset the application for new users.

6.2.2 Scenario 2: Person walking around a space

Figure 6.6. The person starts walking in an area within the field of view of the camera. Location tracking is marked on the map by a blue trace. The black circle on the map indicates the person's current position.

In this scenario, no context aware system was demonstrated. The illustrations in Figure 6.6 to Figure 6.8 demonstrates the locations of person with ID=25 being tracked by ViMM module. The position and body orientation of the person could be exploited by the nearby wall display, such as the one on the left, shown in the map image. The wall display would only know that a person is coming to it if the position of the person is within the proximity of the display and the body orientation is facing the display. However re-identification in front of the display is not possible since the area is not within the field of view of the camera.

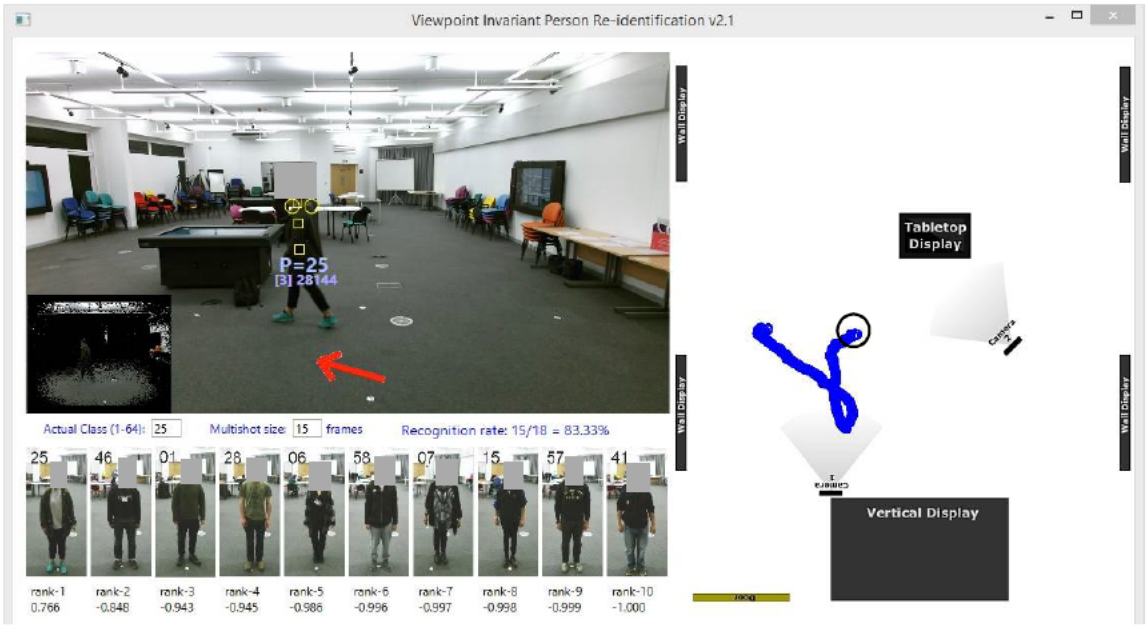


Figure 6.7. The person is seen to continue walking while ViMM is performing re-identification.

Figure 6.8. ViMM records 88.46% rank-1 classification for the person with ID=25 in this scenario.

The Figure 6.8 shows the identity and location history information of the person that could be stored in the central database. Other context aware system could make use of this information for example, to send a recommendation of next exhibition area to visit, to a mobile phone that had been paired (earlier) with the identity of the person.

6.2.3 Scenario 3: A group of people working on a tabletop display

In this scenario groups of four and three people walked towards a tabletop display, performed simple gestures on the display and moved to new locations around the display.

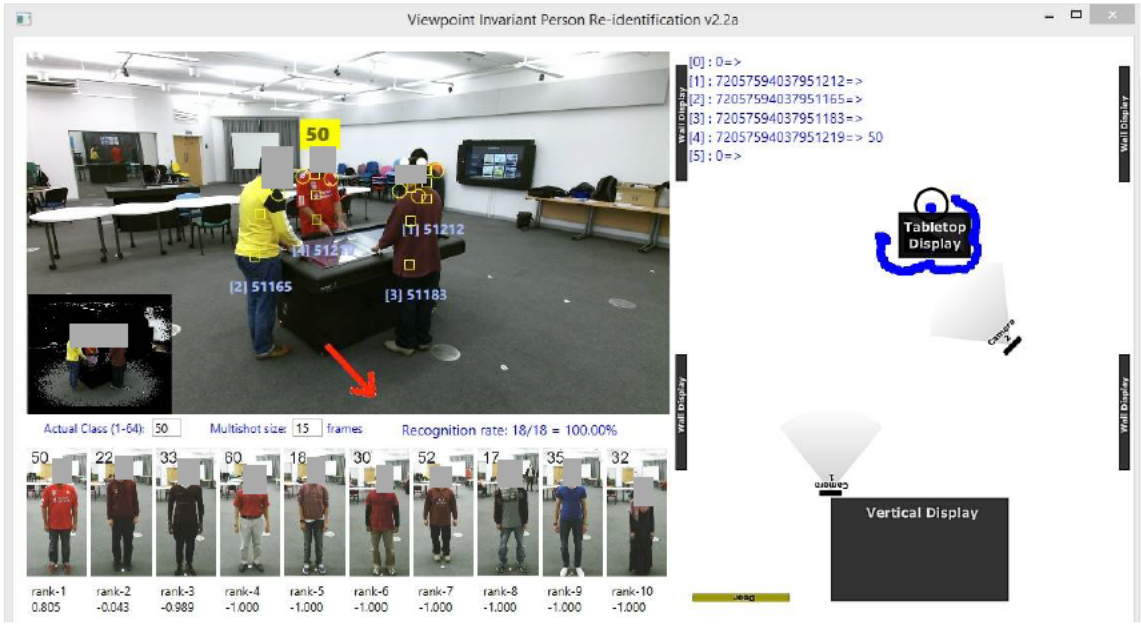


Figure 6.9. The blue plot on the map on the right shows the location tracking of person with ID=50. The gap plot is caused by a temporary occlusion before moving to a new location. The black circle marks the current location of person with ID=50.

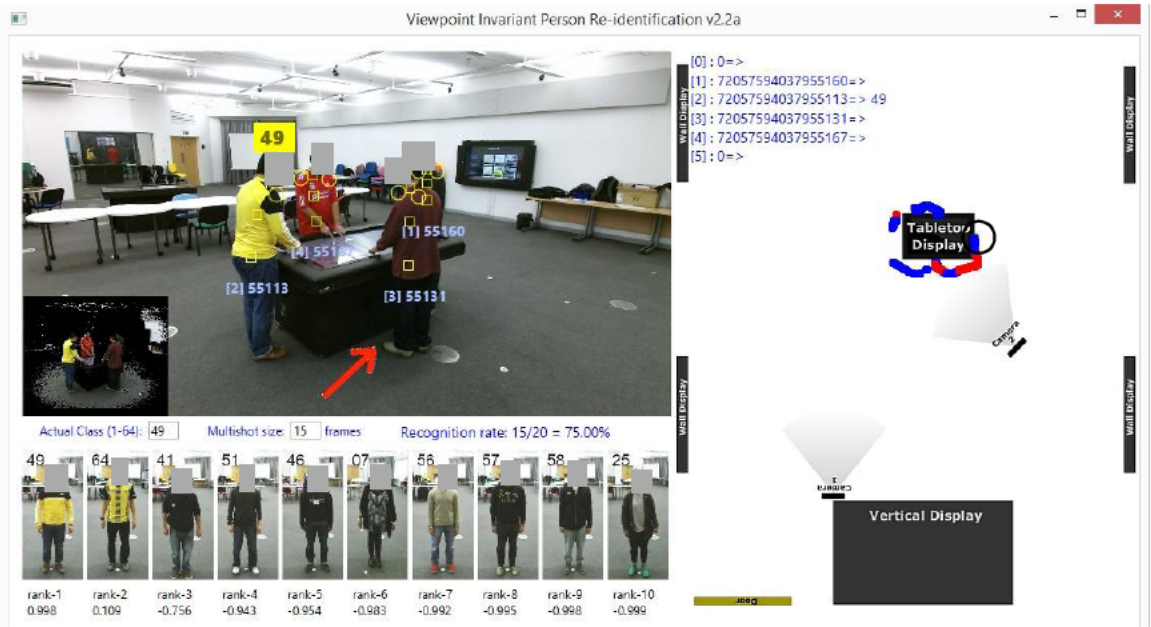


Figure 6.10. The blue trail on the map marks the location tracking of person with ID=49. The red plot indicates location tracking with incorrect classification that occurs temporarily especially when a person moves to a new location, while being occluded in the process.

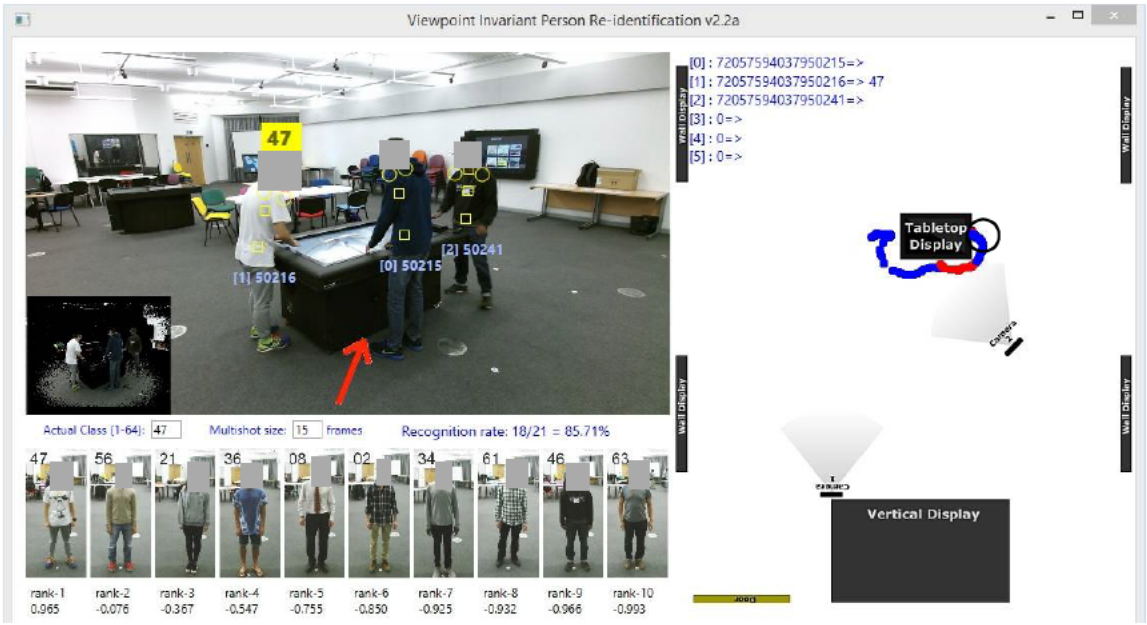


Figure 6.11. The blue plot on the map marks the location tracking of person with ID=47. The black circle indicates the current location of the person.

It can be observed from illustrations in Figure 6.9 to Figure 6.11 that a number of misclassifications (indicated by a red plot) had occurred in the simulated scenario. This is believed to have been contributed by rapid occurrences of occlusions causing difficulties for ViMM to perform proper feature extraction.

6.2.4 Scenario 4: A group of people walking around a space

In this scenario, a group of four people walked freely around a space. Some bending down actions were performed in between walk as illustrated in Figure 6.12.

Figure 6.12. The free walking actions include some bending down actions. The red plot in the map is contributed by the person with ID=08 being occluded by other persons, and also from a bending down action, causing a few misclassifications. The black circle marks the current location of person with ID=08.

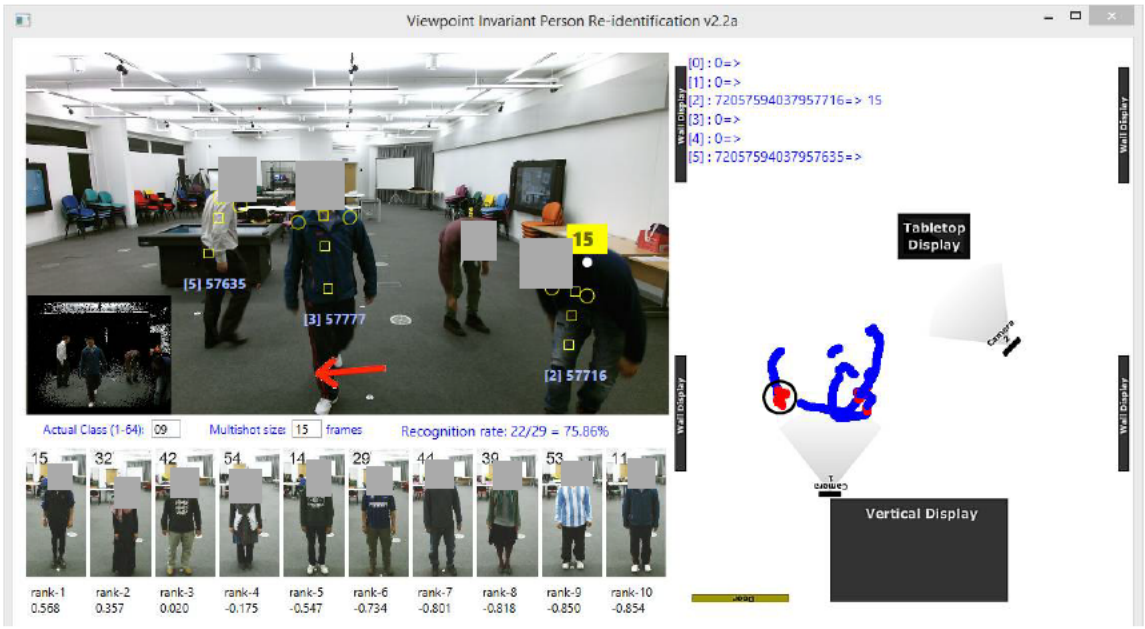


Figure 6.13. The person with ID=09 is being misclassified as person with ID=15 when performing the bending down action.

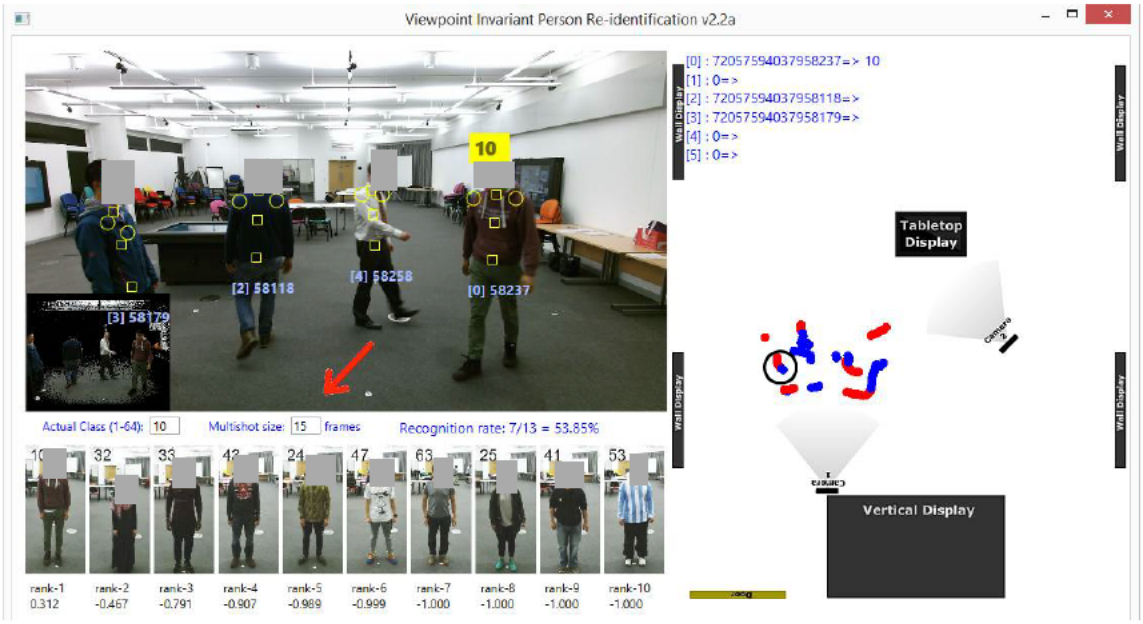


Figure 6.14. The red plot on the map is a result of misclassification caused by a high number of occlusions happened to person with ID=10, in addition to misclassification from some bending down actions.

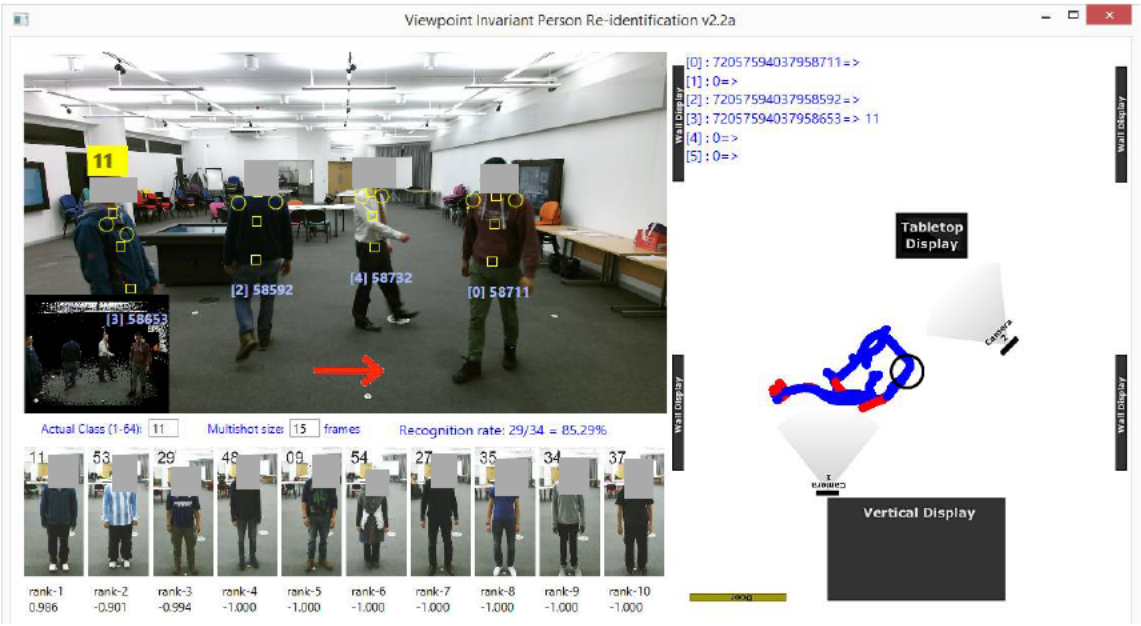


Figure 6.15. ViMM records 85.29% rank-1 classification for the person with ID=11 in this scenario. Bending down action and occlusions again have contributed to the misclassification indicated by a red plot.

It can be observed from illustrations in Figure 6.12 to Figure 6.15 that a number of misclassifications (indicated by a red plot) had occurred in the simulated scenario when the person being tracked performed bending down action for a number of times. Occlusions persistently contributed to the misclassification as similarly happened to other scenarios in this

chapter. This misclassifications can be obviated by using joint information to detect instances of partial occlusion, and non-compliant pose such as bending down action.

6.3 Conclusions

In this chapter, ViMM person re-identification for single person and multi-person scenarios were tested in simulated scenarios of a digital exhibition at a museum. ViMM person re-identification (will be called with just ViMM hereafter), has been demonstrated to be capable of being a context provider by sharing its output (i.e. person identity, current location and body orientation angle) to other context aware systems. The sharing mechanism can be in the form of a look-up table containing ViMM's output, either stored locally or at a central location accessible via network. A context-aware system wishing to consume this information would just have to look at the look-up table for person's location within its proximity.

CHAPTER 7:

Discussion, Conclusions and Future Work

7.1 Discussion

This section discusses the results of experimental chapters 3, 4 and 5 as a whole.

The human-sensing tabletop display presented in Chapter 3 is an example of a context-aware system. It has the capability of sensing up to a maximum of six people around it with highly accurate user position estimation. The proposed configuration uses a very minimum number of twelve infrared sensors. It can sense people approaching the table within 10 cm range, but it cannot tell the identity of each person. As demonstrated by the personal image gallery application in Figure 3.10 of Section 3.3.2, users had to enter name or login information to retrieve their saved galleries. The proposed ViMM person re-identification system presented in Chapter 4, uses multi-modal soft-biometric features to perform re-identification of a person. The ViMM feature descriptor is suitable for use in interactive environments such as museums, libraries, etc. ViMM is envisioned to compliment the human-sensing tabletop mentioned earlier, to give an enhanced level of personalisation. With reference to the image gallery above as an example, ViMM re-identification allows the tabletop to perform automatic login, hence automatic retrieval of a personal gallery as soon as a person approaches the tabletop. An effective integration between person re-identification and context-aware systems will not only eliminate a layer of interaction (i.e. login as identification step), but it will amaze users by letting the interaction scheme disappear to users' attention, which allows them to remain focused on content. The successful implementation of these two concepts will also create new

aesthetics of interaction, where people will have an illusion that they are dealing with something alive and experience a kind of "magic" under their control.

It is demonstrated from the results presented in Chapter 4 that in Experiment 2 ("Ex4"), ViMM achieved a very good rank-1 classification result of 91.49% on testing set "Free Walking 2" (without bags) using classifier C12. The matching method used was multi-shot (MvsM) with M=15 frames and a "mean" decision technique was used on the 15 frames. The ViMM feature descriptor is shown to be robust having maintained a performance of 89.8% in "Ex4", albeit with a small drop in rank-1 classification, for a scenario where all participants carried items such as backpacks, handbags, and briefcases. The drop is believed to have been caused by the inaccuracy of ellipse fitting method for some styles of backpacks such as the example shown in the last image in Figure 4.19. It is observed that handbags with long straps and briefcases did not affect how ViMM extracted its features since the handbags and briefcases positions were below the hip level. The experiment however does not represent a realistic scenario since it is uncommon to see all people carry backpacks at the same time at places like museums. So a new set of testing data was created containing 32 people with bags and the other 32 without. A very good rank-1 classification of 91.81% was achieved and it is 0.3% better than that of "Free Walking 2" (without bags) when used as the test set.

Although ViMM gives good results in single person experimental scenarios, some issues arose when ViMM was tested with multiple people in the camera's field of view. Partial occlusions affect ViMM feature extraction as a result of incomplete sets of depth points causing parameters to be wrongly estimated and the ellipse fitting algorithm to return incorrect estimates of body orientation angle. A simple strategy to deal with partial occlusions is to make a system to give a null result if partial occlusion is detected although this could result in unacceptably large numbers of 'missed' frames in practice. More sophisticated strategies could attempt to estimate

missing depth samples by extrapolation from previous frames or to allow classification to be attempted from a partial ViMM vector. Such strategies to handle partial occlusions could be the subject for future work in this area.

7.2 Conclusions

Future intelligent environments benefiting from robust and integrated person re-identification have the potential to offer enhanced levels of personalised services via natural interaction. A fundamental obstacle to the realisation of these kinds of intelligent environments lies is the limitation of person re-identification accuracy. A survey of literature has shown that no system with an accuracy above 90% of rank-1 was previously reported. Most previously published work in re-identification has been performed with the use of 2-D cameras, and with the aim of reducing the search time of a target from the hundreds of people in a gallery set. Only recent work has benefitted from the use of depth data from RGB-D cameras, and only very recently have higher resolution depth-sensing cameras, such as the Kinect v2, become available.

7.2.1 The Research Questions

The research questions have been answered. The first question asked in this thesis was:

Can the performance of previously proposed human aware multi-touch tabletop systems be achieved at much lower cost (i.e. low in construction price and computational power) using a reduced number of sensors?

The human aware multi-touch tabletop system was successfully demonstrated in Chapter 3 with demo applications developed to demonstrate its effectiveness. The configuration of the proposed system is very simple, and very cost effective both in construction price and computational power (i.e. resource usage) when compared to existing systems (Annett et al.,

2011, Klinkhammer et al., 2011) reviewed in the literature. For example the use of twelve infrared distance sensor and two I/O boards costs around USD 450 compared to Klinkhammer et al. (2011)'s system that uses 96 infrared sensors which would cost approximately 8 times higher at around USD 3,600. The computational power required to process inputs from twelve infrared distance sensors is obviously much lower than that of inputs from 96 infrared distance sensors.

This second research question was:

Can person re-identification accuracy be improved by supplementing clothing appearance descriptors with 3-D anthropometric parameters extracted from depth data, using RGB-D cameras, in unconstrained settings?

This research has shown that the depth information collected from Kinect v2 of a detected person can be used to collect depth points at various heights of the body. This in turn, allows depth points collected at shoulder level to be used to determine the body orientation by fitting an ellipse to the partial set of points. Tests showed that accurate estimation of body orientation can be obtained using the ellipse fitting method.

The 3-D body parameters of a human body cross section at shoulder, mid-spine, and hip levels, combined with appearance based features at shoulder, torso and hip levels, and, body orientation form the complete ViMM feature descriptor. In other words, the colour and combination of 2-D and 3-D anthropometric properties are logged as a function of body orientation. Training a neural network classifier with features from two activities from the Kinect V2 RGBD-ID dataset components "Turning 1" and "Free Walking 1" is sufficient to achieve over 92% rank-1 classification performance. Testing the classifier on the "Free Walking 2" dataset component reported a single-frame re-identification performance of 86.3%

rank-1 classification and nAUC of 99.4%, and performance is increased to 92.4% rank-1 classification and nAUC of 99.7% when multiple frames were combined using the MvsM method. However it is not uncommon for people to be seen carrying bags in spaces such as museums. Hence using the classifier that has been re-trained with additional dataset components "Turning-bag 1" and "Free Walking-bag 1", ViMM was tested with participants carrying bags and backpacks and it achieved similarly good performance of 85.32% rank-1 classification with nAUC of 99.43%, and 91.81% rank-1 classification with nAUC of 99.67% when MvsM (M=15) was used.

As a concluding statement, it can be said that the ViMM feature descriptor achieved viewpoint invariant re-identification as evidenced from the results of the experiments, and ViMM person re-identification could aid context-aware systems deliver targeted personalised services by sharing its output which includes person identity, current location and body orientation.

7.2.2 Summary of Contributions and Findings

In summary, the outcome of the research has benefited the field of person re-identification and context awareness in the following areas:

i. Demonstrating a case for the integration of person re-identification and context awareness such as human sensing system for use in a real-time intelligent environment setting.
ii. Designing a novel robust viewpoint invariant multi-modal (ViMM) feature descriptor for person re-identification, based on 2-D and 3-D appearances, allowing a person to be re-identified from unconstrained viewpoints.
iii. Demonstrating viewpoint invariant re-identification in real-time.
iv. Devising a robust method to estimate body orientation angle by combining ellipse fitting to the partial depth data from RGB-D camera, at shoulder level with face detection.

v. Optimising neural network parameters for use with ViMM feature descriptor with body orientation, allowing a person to be re-identified from unconstrained viewpoints.

vi. Creating a new RGB-D based dataset of sixty-four people with sixteen activities for each person, acquired using Kinect version 2 camera. The higher resolution and more accurate depth sensing of the version 2 camera is important for the advancement of more robust re-identification systems.

vii. The design of an alternative sensor configuration for human-sensing multi-touch tabletop display (using a total of twelve infrared distance sensors), that is reliable and very simple to build at much reduced cost in terms of construction price, computing power, and setup time. It is also very mobile, transferring the setup to other tabletop display takes approximately only 15 minutes.

7.3 Future Work

There are many opportunities for further research in the area of person re-identification. Making interactive environments work robustly and reliably in the real world still poses many interesting challenges.

This section lists various possible directions for further work.

i. To incorporate the approach with other modes of identification e.g. with facial recognition and gait analysis, etc.

ii. To widen the scope of people-sensing to include people in wheelchairs, children, parents with pushchairs, etc.

iii. To improve the handling of partial occlusions in crowded areas by employing multiple sensors.

iv. To improve the processing of people in “non-compliant” poses. For example,

- manipulation of the point-cloud to correct for people bending over at the waist or leaning to the side, etc.
- separation and extraction of the arm profiles to make the chest and waist ellipse estimation more accurate.

In terms of use case enhancements, it is proposed that a mobile interface be developed to allow delivery of personalised services to mobile phones such as personal narration in a place where displays are not available. A live direction guidance can also be provided on the mobile since the location of the person is known to the system.

Publications

This section contains the copies of publications as listed below:

1) Yusof, H., Collins, T., Ch'ng, E., and Woolley, S. (2016), "Viewpoint Invariant Multi-modal Person Re-identification Using RGB-D Cameras". Submitted to the journal of the IEEE Transactions on Consumer Electronics on 3rd January 2016, second revised submission on 9th May 2016.
2) Yusof, H., Eugene Ch'ng, E., and Baber, C. (2014), "Human Sensing for Tabletop Entertainment System." *Context-Aware Systems and Applications* (pp. 283-292). Springer International Publishing, 2014.

Appendices

This section contains the code listing 5 mentioned in Section 4.4.4 (C# implementation of Ellipse fitting), listing 6 mentioned in Section 4.4.9 (AutoIt Windows automation script to perform batch processing of feature extraction), and informed consent form for people re-identification test subjects.

Listing 5. Ellipse fitting algorithm in C#

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using Meta.Numerics.Matrices;
using Meta.Numerics;
using System.Diagnostics;

namespace KinectCoordinateMapping
{
    public class EllipseFit
    {
        public Matrix Fit(PointCollection points)
        {
            int numPoints = points.Count;

            Matrix D1 = new Matrix(numPoints, 3);
            Matrix D2 = new Matrix(numPoints, 3);
            SquareMatrix S1 = new SquareMatrix(3);
            SquareMatrix S2 = new SquareMatrix(3);
            SquareMatrix S3 = new SquareMatrix(3);
            SquareMatrix T = new SquareMatrix(3);
            SquareMatrix M = new SquareMatrix(3);
            SquareMatrix C1 = new SquareMatrix(3);
            Matrix a1 = new Matrix(3, 1);
            Matrix a2 = new Matrix(3, 1);
            Matrix result = new Matrix(13, 1);
            Matrix temp;
            SquareMatrix R = new SquareMatrix(2);

            double mean_x = 0, mean_y = 0;

            C1[0, 0] = 0;
            C1[0, 1] = 0;
            C1[0, 2] = 0.5;
            C1[1, 0] = 0;
            C1[1, 1] = -1;
            C1[1, 2] = 0;
            C1[2, 0] = 0.5;
            C1[2, 1] = 0;
            C1[2, 2] = 0;

            //2 D1 = [x .^ 2, x .* y, y .^ 2]; % quadratic part of the design matrix
            //3 D2 = [x, y, ones(size(x))]; % linear part of the design matrix
            for (int xx = 0; xx < points.Count; xx++)
            {
                Point p = points[xx];

                mean x += p.X;
                mean_y += p.Y;
            }

            mean_x /= points.Count;
            mean_y /= points.Count;

            for (int xx = 0; xx < points.Count; xx++)
            {
                Point p = points[xx];
                p.X -= mean x;
                p.Y -= mean y;
                D1[xx, 0] = p.X * p.X;
                D1[xx, 1] = p.X * p.Y;
                D1[xx, 2] = p.Y * p.Y;

                D2[xx, 0] = p.X;
                D2[xx, 1] = p.Y;
                D2[xx, 2] = 1;
            }
```

```
//4 S1 = D1' * D1; % quadratic part of the scatter matrix
temp = D1.Transpose() * D1;
for (int xx = 0; xx < 3; xx++)
    for (int yy = 0; yy < 3; yy++)
        S1[xx, yy] = temp[xx, yy];

//5 S2 = D1' * D2; % combined part of the scatter matrix
temp = D1.Transpose() * D2;
for (int xx = 0; xx < 3; xx++)
    for (int yy = 0; yy < 3; yy++)
        S2[xx, yy] = temp[xx, yy];

//6 S3 = D2' * D2; % linear part of the scatter matrix
temp = D2.Transpose() * D2;
for (int xx = 0; xx < 3; xx++)
    for (int yy = 0; yy < 3; yy++)
        S3[xx, yy] = temp[xx, yy];

//7 T = - inv(S3) * S2'; % for getting a2 from a1
T = -1 * S3.Inverse() * S2.Transpose();

//8 M = S1 + S2 * T; % reduced scatter matrix
M = S1 + S2 * T;

//9 M = [M(3, :) ./ 2; - M(2, :); M(1, :) ./ 2]; % premultiply by inv(C1)
M = C1 * M;

if (!Double.IsNaN(M[0,0]))
{

    //10 [evec, eval] = eig(M); % solve eigensystem
    ComplexEigensystem eigenSystem = M.Eigensystem();

    //11 cond = 4 * evec(1, :) .* evec(3, :) - evec(2, :) .^ 2; % evaluate a'Ca
    //12 a1 = evec(:, find(cond > 0)); % eigenvector for min. pos. eigenvalue
    for (int xx = 0; xx < eigenSystem.Dimension; xx++)
    {
        Vector<Complex> vector = eigenSystem.Eigenvector(xx);
        Complex condition = 4 * vector[0] * vector[2] - vector[1] * vector[1];
        if (condition.Im == 0 && condition.Re > 0)
        {
            // Solution is found
            for (int yy = 0; yy < vector.Count(); yy++)
            {
                a1[yy, 0] = vector[yy].Re;

            }
        }
    }
    //13 a2 = T * a1; % ellipse coefficients
    a2 = T * a1;

    //14 a = [a1; a2]; % ellipse coefficients
    result[0, 0] = a1[0, 0];
    result[1, 0] = a1[1, 0];
    result[2, 0] = a1[2, 0];
    result[3, 0] = a2[0, 0];
    result[4, 0] = a2[1, 0];
    result[5, 0] = a2[2, 0];


    //added by hafiz
    //calculating major and minor axes
    double orientation_tolerance = 1e-3;
    double orientation_rad, cos_phi, sin_phi;
    double mxnew, mynew;


    double a, b, c, d, e, f;
    f = result[5, 0];
    e = result[4, 0];
    d = result[3, 0];
```

```
c = result[2, 0];
b = result[1, 0];
a = result[0, 0];

// remove the orientation from the ellipse
if (Math.Min(Math.Abs(b / a), Math.Abs(b / c)) > orientation_tolerance)
{
    double tt = Math.Atan(b / (c - a));

    orientation_rad = 0.5 * tt;
    cos phi = Math.Cos(orientation rad);
    sin_phi = Math.Sin(orientation_rad);

    //[a,b,c,d,e] = deal(...
    double a0 = a, b0 = b, c0 = c, d0 = d, e0 = e;

    a = a0 * Math.Pow(cos_phi, 2) - b0 * cos_phi * sin_phi + c0 * Math.Pow(sin_phi, 2);
    b = 0;
    c = a0 * Math.Pow(sin_phi, 2) + b0 * cos_phi * sin_phi + c0 * Math.Pow(cos_phi, 2);
    d = d0 * cos phi - e0 * sin phi;
    e = d0 * sin_phi + e0 * cos_phi;

    //[mean_x,mean_y] = deal( ...
    mxnew = cos_phi * mean_x - sin_phi * mean_y;
    mynew = sin phi * mean_x + cos_phi * mean_y;
    mean_x = mxnew;
    mean_y = mynew;
}
else
{
    orientation rad = 0;
    cos_phi = Math.Cos(orientation_rad);
    sin_phi = Math.Sin(orientation_rad);
}

// check if conic equation represents an ellipse
double test = a * c;
//switch (test)
//case (test>0),  status = '';
//case (test==0), status = 'Parabola found';
//case (test<0),  status = 'Hyperbola found';

// if we found an ellipse return its data
if (test > 0)
{
    // make sure coefficients are positive as required
    if (a < 0)
    {
        a *= -1;
        c *= -1;
        d *= -1;
        e *= -1;
        //[a,c,d,e] = deal( -a,-c,-d,-e );
    }

    // final ellipse parameters

    double d2a = d / 2.0 / a;
    double e2a = e / 2.0 / a;
    double X0 = mean x - d2a;
    double Y0 = mean y - e2a;
    double X0_in, Y0_in;

    double A = result[0, 0];
    double B = result[1, 0] / 2.0;
    double C = result[2, 0];
    double D = result[3, 0] / 2.0;
    double E = result[4, 0] / 2.0;
    double F = result[5, 0];

    double T1 = 2 * (A * E * E + C * D * D + F * B * B - 2 * B * D * E - A * C * F);
    double T2 = Math.Sqrt((A - C) * (A - C) + 4 * B * B) - (A + C);
```

```
                    double T3 = -Math.Sqrt((A - C) * (A - C) + 4 * B * B) - (A + C);
                    double T4 = (B * B - A * C) * T2;
                    double T5 = (B * B - A * C) * T3;

                    double T6 = (A - C) / (2 * B);
                    double T7 = Math.Atan(1 / T6);

                    double aa = Math.Sqrt(T1 / T4);
                    double bb = Math.Sqrt(T1 / T5);

                    double long axis = Math.Max(aa, bb);
                    double short axis = Math.Min(aa, bb);
                    double orientation_rad2 = 0;

                    if (Math.Abs(A) < Math.Abs(C))
                    {
                        orientation_rad2 = 0.5 * T7;
                    }
                    else if (Math.Abs(A) > Math.Abs(C))
                    {
                        orientation_rad2 = Math.PI / 2 + 0.5 * T7;
                    }

                    X0_in = (C * D - B * E) / (B * B - A * C);
                    Y0_in = (A * E - B * D) / (B * B - A * C);

                    result[6, 0] = X0;
                    result[7, 0] = Y0;
                    result[8, 0] = long_axis;
                    result[9, 0] = short axis;
                    result[10, 0] = orientation_rad2;
                    result[11, 0] = X0_in;
                    result[12, 0] = Y0_in;
                }
            }
            return result;
        }
    }
}
```

Listing 6. AutoIt windows automation script to perform batch processing of features extraction

```
#include <MsgBoxConstants.au3>

Local $app = FileGetShortName("C:\Program Files\Microsoft
SDKs\Kinect\v2.0 1409\Samples\Managed\kinect2-reid-feature-extraction-
vimm2\KinectCoordinateMapping\KinectCoordinateMapping\bin\x64\Debug\KinectReidentification.exe")
Local $app1 = FileGetShortName("C:\Program Files\Microsoft
SDKs\Kinect\v2.0_1409\Tools\KinectStudio\KSutil.exe")
Local $i = 1
Local $j = 1
Local $ij
Local $fileName
Local $dataFile

While $i <65
          While $j < 13

                    $ij = StringFormat("%02i", $i) & StringFormat("%02i.xef", $j)
                    $fileName = "E:\" & $ij

                    $dataFile = FileGetShortName($fileName)

                    StartProcess()
                    $j += 1
          WEnd
          $i += 1
          $j = 1
WEnd
```

```
Func StartProcess()
	Local $iPID1 = Run(@ComSpec & " /c " & $app1 & " -play " & $dataFile, "", @SW_SHOWNOACTIVATE)
	While ProcessExists($iPID1)
		Sleep(1000)
	WEnd

	Local $iPID = Run($app, "", @SW_SHOWDEFAULT)
	Sleep(1500)

	Local $iPID2 = Run(@ComSpec & " /c " & $app1 & " -play " & $dataFile, "", @SW_MINIMIZE)

	While ProcessExists($iPID2)
		Sleep(2000)
	WEnd

	Sleep(5000)
	ProcessClose($iPID)

EndFunc
```

References

Wikipedia.com (n.d.) Atari Breakout Game [online]. Available from: https://en.wikipedia.org/wiki/Breakout_(video_game) [Accessed 1 March 2016]

Wikipedia (n.d.) Atari Pong Game [online]. Available from: https://en.wikipedia.org/wiki/Pong [Accessed 1 March 2016]

AutoIt (n.d.) AutoIt Windows Automation Scripting [online]. Available from: https://www.autoitscript.com/site/autoit/ [Accessed 28 January 2016]

Hotmath.com (n.d.) Conic Sections and Standard Forms of Equations [online]. Available from: http://hotmath.com/hotmath_help/topics/conic-sections-and-standard-forms-of-equations.html [Accessed 23 January 2016]

Ohad Gal (n.d.) Ellipse Fitting by Ohad Gal [online]. Available from: http://uk.mathworks.com/matlabcentral/fileexchange/3215-fitellipse/content/fit_ellipse.m [Accessed 23 January 2016]

Srikanth Kotagiri (n.d.) Ellipse Fitting C# Implementation [online]. Available from: https://skotagiri.wordpress.com/2010/06/19/c-implementation-for-fitting-an-ellipse-for-a-set-of-points/ [Accessed 23 January 2016]

Microsoft (n.d.) Kinect for Windows SDK 1.8 Body Joints [online]. Available from: https://msdn.microsoft.com/en-us/library/hh855347.aspx [Accessed 5 November 2015a]

Microsoft (n.d.) Kinect for Windows SDK v2.0 Body Joints [online]. Available from: https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx [Accessed 22 April 2016b]

Microsoft (n.d.) Kinect v1 Skeletal Tracking [online]. Available from: https://msdn.microsoft.com/en-us/library/hh973074.aspx [Accessed 18 February 2016c]

Microsoft (n.d.) Kinect v2 Hardware Features [online]. Available from: https://dev.windows.com/en-us/kinect/hardware [Accessed 18 February 2016d]

Microsoft (n.d.) Microsoft Kinect for Windows SDK 1.8 [online]. Available from: https://msdn.microsoft.com/en-us/library/jj131025.aspx?f=255&MSPPError=-2147217396 [Accessed 5 November 2015e]

NUI-Group (n.d.) Natural User Interface [NUI Group] [online]. Available from: http://wiki.nuigroup.com/Natural_User_Interface [Accessed 3 February 2016]

Tech Terms (n.d.) NUI Definition [Tech Terms] [online]. Available from: http://techterms.com/definition/nui [Accessed 3 February 2016]

Ljubljana, U. of (n.d.) Orange Data Mining Software [online]. Available from: http://orange.biolab.si/ [Accessed 22 February 2016]

Atari.com (n.d.) Pong Online Game [online]. Available from: https://atari.com/arcade#!/arcade/pong/play [Accessed 1 March 2016a]

Atari.com (n.d.) Super Breakout Online Game [online]. Available from: https://atari.com/arcade#!/arcade/superbreakout/play [Accessed 1 March 2016b]

Rhodes, H.T.F. (1956) Alphonse Bertillon: Father of Scientific Detection. Abelard-Schuman, New York

Pentland, A., Moghaddam, B. and Starner, T. (1994) View-based and modular eigenspaces for face recognition. *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, (245): 2–8

Fitzgibbon, A.W., Pilu, M. and Fisher, R.B. (1996) Direct least squares fitting of ellipses. *Proceedings of 13th International Conference on Pattern recognition* [online], 1: 253–257. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=546029

Mitchell, T.M. (1997) Machine Learning [online]. Available from: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0070428077

Addlesee, M.D., Jones, A.H., Livesey, F., et al. (1997) The ORL Active Floor Floor technology. *IEEE Personal Communication*

Burges, C.J.C.J.C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* [online], 2 (2): 121–167. Available from: http://www.springerlink.com/index/Q87856173126771Q.pdf

Halir, R. and Flusser, J. (1998) Numerically stable direct least squares fitting of ellipses. *Proc. 6th International Conference in Central Europe on Computer Graphics and Visualization*, 98: 125–132

Smith, S.W. (1999) The Scientist and Engineer's Guide to Digital Signal Processing, Second Edition. 2nd ed. California Technical Publishing, San Diego, CA

Orr, R.J. and Abowd, G.D. (2000) The Smart Floor : A Mechanism for Natural User Identification and Tracking. *Human Factors in Computing Systems*. 2000. pp. 275–276

Geusebroek, J.M., Van Den Boomgaard, R., Smeulders, A.W.M., et al. (2001) Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (12): 1338–1350

Dietz, P. and Leigh, D. (2001) DiamondTouch: A Multi-User Touch Technology. *Proceedings of the 14th annual ACM symposium on User interface software and technology*, 3 (2): 219–226

Dey, A.K. (2001) Understanding and Using Context. *Personal and Ubiquitous Computing* [online], 5 (1): 4–7. Available from: http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s007790170019

Shutler, J., Grant, M., Nixon, M.S., et al. (2002) On a Large Sequence-Based Human Gait Database. *Proceedings Fourth International Conference Recent Advances in Soft Computing*, [online], pp. 66–72. Available from: http://eprints.soton.ac.uk/257901/

Chien, Y., Huag, Y., Jeng, S., et al. (2003) Real-Time Surveillance System by Use of the Face. *Techniques*, (1): 10–12

Jain, A.K., Dass, S.C. and Nandakumar, K. (2004a) Can soft biometric traits assist user recognition? *Proceedings of SPIE* [online], 5404: 561–572. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.73.2055

Song, K. and Chen, W. (2004) Face recognition and tracking for human-robot interaction*. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)* [online], 3: 2877–2882. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1400769

Vogel, D. and Balakrishnan, R. (2004) Interactive Public Ambient Displays : Transitioning from Implicit to Explicit , Public to Personal , Interaction with Multiple Users. *UIST 2004*. 2004. pp. 137–146

Rakotonirainy, A. and Tay, R.S. (2004) In-vehicle ambient intelligence transport systems: Towards an integrated research. *7th International IEEE Conference on Intelligent Transportation Systems*. 2004. pp. 648–651

Zewail, R., Elsafi, a., Saeb, M., et al. (2004) Soft and hard biometrics fusion for improved identity verification. *The 2004 47th Midwest Symposium on Circuits and Systems, 2004. MWSCAS '04.*, 1: 225–228

Jain, A.K., Dass, S.C. and Nandakumar, K. (2004b) Soft Biometric Traits for Personal Recognition Systems. *Proc. of International Conference on Biometric Authentication (ICBA)* [online], (July): 731–738. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.5414

Zhang, H. (2004) The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference FLAIRS 2004* [online], 1 (2): 1 – 6. Available from: http://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf

Wakkary, R., Hatala, M., Lovell, R., et al. (2005) An ambient intelligence platform for physical play. *Proc. MULTIMEDIA 2005* [online], p. 764. Available from: http://portal.acm.org/citation.cfm?doid=1101149.1101313

Stephen Pheasant and Christine M. Haslegrave (2005) Bodyspace: Anthropometry, Ergonomics and the Design of Work, Third Edition. 3rd ed. London: Taylor and Francis

Rokach, L. and Maimon, O. (2005) Decision Tree. *Data Mining and Knowledge Discovery Handbook* [online], p. pp 165–192. Available from: http://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

Sivic, J., Zitnick, C.L. and Szeliski, R. (2006) Finding people in repeated shots of the same scene. *Procedings of the British Machine Vision Conference 2006.* **2006** [online]. pp. 93.1–93.10. Available from: http://eprints.pascal-network.org/archive/00002195/

Han, J. and Bhanu, B. (2006) Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28 (2): 316–322

Arandjelović, O., Hammoud, R. and Cipolla, R. (2006) On person authentication by fusing visual and thermal face biometrics. *Proceedings - IEEE International Conference on Video and Signal Based Surveillance 2006, AVSS 2006*, (2): 1–6

Gandhi, T. and Trivedi, M.M. (2006) Panoramic Appearance Map (PAM) for multi-camera based person re-identification. *Proceedings - IEEE International Conference on Video and Signal Based Surveillance 2006, AVSS 2006*. November 2006. pp. 1–5

Gheissari, N., Sebastian, T.B., Tu, P.H., et al. (2006) Person Reidentification Using Spatiotemporal Appearance. *Computer Vision and Pattern Recognition CVPR 2006*. 2006

Baldauf, M. (2007) A survey on context-aware systems Schahram Dustdar * and Florian Rosenberg., 2 (4)

Gray, D., Brennan, S. and Tao, H. (2007) Evaluating appearance models for recognition, reacquisition, and tracking. *Performance Evaluation for Tracking and Surveillance (PETS), 10th International Workshop on* [online], 3: 41–47. Available from: http://www.soe.ucsc.edu/~dgray/dgray-pets2007.pdf\nhttp://pets2007.net/

Balkenius, C. and Johansson, B. (2007) Finding Colored Objects in a Scene. *Lund University Cognitive Science (LUCS) Minor* [online], 12. Available from: http://www.lucs.lu.se/LUCS/M012/Minor12.pdf

Glas, D.F., Miyashita, T., Ishiguro, H., et al. (2007) Laser tracking of human body motion using adaptive shape modeling. *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems* [online], pp. 602–608. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4399383

Stokman, H. and Gevers, T. (2007) Selection and Fusion of Color Models for Feature Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29 (3): 371 – 381

Wang, X., Doretto, G., Sebastian, T., et al. (2007) Shape and Appearance Context Modeling. *2007 IEEE 11th International Conference on Computer Vision*. 2007 [online]. IEEE. pp. 1–8. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4409019 [Accessed 22 July 2014]

Gabriel, C., Salvador, N., Hervás, R., et al. (2007) Spontaneous Interaction on Context-Aware Public Display : An NFC and Infrared Sensor approach. *Proceedings of the 1st International Conference on Immersive Telecommunications (ImmersCom'07)*. 2007

Kotsiantis, S.B. (2007) Supervised machine learning: A review of classification techniques.

Informatica [online], 31: 249–268. Available from: http://books.google.com/books?hl=en&lr=&id=vLiTXDHr_sYC&oi=fnd&pg=PA3&dq=survey+machine+learning&ots=CVsyuwYHjo&sig=A6wYWvywU8XTc7Dzp8ZdKJaW7rc\npapers://5e3e5e59-48a2-47c1-b6b1-a778137d3ec1/Paper/p800\nhttp://www.informatica.si/PDF/31-3/11_Kotsiantis - S

Apostoloff, N. and Zisserman, A. (2007) Who are you ? – real-time person identification. *British Machine Vision Conference* [online], pp. 509–518. Available from: http://www.robots.ox.ac.uk:5000/~vgg/publications/papers/apostoloff07.ps.gz

An, A. (2008) Classification Methods. In Wang, J. (ed.) *Encyclopedia of Data Warehousing and Mining*. 2nd ed. Idea Group Inc. pp. 196–201

Tănase, C.A., Vatavu, R., Pentiuc, Ş., et al. (2008) Detecting and Tracking Multiple Users in the Proximity of Interactive Tabletops. *Advances in Electrical and Computer Engineering*, 8 (2)

Lin, Z. and Davis, L.S. (2008) Learning pairwise dissimilarity profiles for appearance recognition in visual surveillance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5358 LNCS (PART 1): 23–34

Valli, A. and Linari, L. (2008) Notes on Natural interaction. *Notes* [online]. Available from: http://portal.acm.org/citation.cfm?doid=1358628.1358676

Hamdoun, O., Moutarde, F., Stanciulescu, B., et al. (2008) Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. *2nd ACM/IEEE International Conference on Distributed Smart Cameras, ICDSC*. 2008

Bay, H., Tuytelaars, T. and Gool, L. Van (2008) SURF : Speeded Up Robust Features. *Computer Vision and Image Understanding (CVIU)*, 110 (3): 346–359

Schulman, D., Sharma, M., Bickmore, T., et al. (2008) The Identification of Users by Relational Agents. *Proc. AAMAS*. 2008. pp. 105–111

Ahn, Y., Park, Y., Choi, K., et al. (2008) Ubi-touch: designing an interactive home control system. *12th WSEAS International Conference on SYSTEMS*. 2008 [online]. pp. 126–130. Available from: http://www.wseas.us/e-library/conferences/2008/crete/Systems/sys1-17.pdf

Walther-franks, B., Teichert, J., Krause, M., et al. (2008) User Detection for a Multi-touch Table via Proximity Sensors. *Proc. ITS 2008*

Gray, D. and Tao, H. (2008) Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. *10th European Conference on Computer Vision*. 2008

Zheng, W.-S., Gong, S. and Xiang, T. (2009) Associating Groups of People. *British Machine Vision Conference* [online], 5 (1): 6. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.160.3228

Schwartz, W.R. and Davis, L.S. (2009) Learning Discriminative Appearance-Based Models Using Partial Least Squares. *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. October 2009 [online]. IEEE. pp. 322–329. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5395183 [Accessed 17 July 2014]

Toninelli, A., Pantsar-Syväniemi, S., Bellavista, P., et al. (2009) Supporting context awareness in smart environments. *Proc. M-PAC 2009*. 2009 [online]. ACM Press. p. 1. Available from: http://dl.acm.org/citation.cfm?id=1657127.1657134 [Accessed 16 November 2012]

Baltieri, D., Vezzani, R. and Cucchiara, R. (2010) 3D Body Model Construction and Matching for Real Time People Re-Identification. *Eurographics Italian Chapter Conference*. 2010

Swartout, W., Traum, D., Artstein, R., et al. (2010) Ada and Grace : Toward Realistic and Engaging Virtual Museum Guides. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2010. pp. 286–300

Rusu, R.B., Bradski, G., Thibaux, R., et al. (2010) Fast 3D Recognition And Pose Using the Viewpoint Feature Histogram. *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on* [online], pp. 2155–2162. Available from: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5651280

Schmidt, D., Chong, M.K. and Gellersen, H. (2010) HandsDown Hand-contour-based User Identification for Interactive Surfaces. *Proceedings of the 6th Nordic Conference on Human-Computer Interaction Extending Boundaries - NordiCHI '10* [online], p. 432. Available from: http://portal.acm.org/citation.cfm?id=1868964\nhttp://portal.acm.org/citation.cfm?doid=1868914.1868964

Prosser, B., Zheng, W.-S., Gong, S., et al. (2010) Person Re-Identification by Support Vector Ranking. *Procedings of the British Machine Vision Conference 2010* [online], pp. 21.1–21.11. Available from: http://www.bmva.org/bmvc/2010/conference/paper21/index.html [Accessed 27 February 2014]

Farenzena, M., Bazzani, L., Perina, A., et al. (2010) Person re-identification by symmetry-driven accumulation of local features. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010 [online]. Ieee. pp. 2360–2367. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5539926

Bak, S., Corvee, E., Thonnat, M., et al. (2010a) Person Re-identification Using Haar-based and DCD-based Signature. *2ndWorkshop on Activity Monitoring by Multi- Camera Surveillance Systems, AMMCSS 2010*. 2010

Bak, S., Corvee, E., Bremond, F., et al. (2010b) Person Re-identification Using Spatial Covariance Regions of Human Body Parts. *International Conference on Advanced Video and Signal Based Surveillance* [online], pp. 435–440. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5597114

Ballendat, T., Marquardt, N. and Greenberg, S. (2010) Proxemic Interaction : Designing for a Proximity and Orientation-Aware Environment. *Proc. ITS 2010*

Goffredo, M., Bouchrika, I., Carter, J.N., et al. (2010) Self-calibrating view-invariant gait biometrics. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 40 (4): 997–1008

Chen, C., Heili, A. and Odobez, J.M. (2011) A joint estimation of head and body orientation cues in surveillance video. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 860–867

Satta, R., Fumera, G., Roli, F., et al. (2011) A multiple component matching framework for person re-identification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6979 LNCS (PART 2): 140–149

Klinkhammer, D., Nitsche, M., Specht, M., et al. (2011) Adaptive personal territories for co-located tabletop interaction in a museum setting. *Proc. ITS 2011* [online], p. 107. Available from: http://dl.acm.org/citation.cfm?doid=2076354.2076375

Dantcheva, A., Velardo, C., Angelo, A.D., et al. (2011) Bag of Soft Biometrics for Person Identification New trends and challenges . *Multimedia Tools and Applications*, 51 (2): 739–777

Dong Seon Cheng, Marco Cristani, Michele Stoppa, L.B. and V.M. (2011) Custom Pictorial Structures for Re-identification. *BMVC* [online], pp. 68.1–68.11. Available from: http://www.bmva.org/bmvc/2011/proceedings/paper68/index.html

Ozturk, O., Yamasaki, T. and Aizawa, K. (2011) Estimating human body and head orientation change to detect visual attention direction. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6468 LNCS (PART1): 410–419

Che, Z.-G., Chiang, T.-A. and Che, Z.-H. (2011) Feed-forward neural networks training: A comparison between genetic algorithm and back-propagation learning algorithm. *International Journal of Innovative Computing, Information and Control*, 7 (10): 5839–5850

Wiethoff, A., Kowalski, R. and Butz, A. (2011) inTUIt – Simple Identification on Tangible User Interfaces. *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction (TEI '11)*. 2011. pp. 201–204

Annett, M., Grossman, T., Wigdor, D., et al. (2011) Medusa: A Proximity-Aware Multi-touch Tabletop. *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*. 2011. pp. 337–346

Baltieri, D., Vezzani, R., Cucchiara, R., et al. (2011) Multi-view people surveillance using 3D information. *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1817–1824

Jungling, K. and Arens, M. (2011) View-invariant person re-identification with an implicit shape model. *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2011*, pp. 197–202

Oliver, J., Albiol, A. and Albiol, A. (2012) 3D descriptor for people re-identification. *International Conference on Pattern Recognition (ICPR)*. 2012. pp. 1395–1398

Nakatani, R., Kouno, D., Shimada, K., et al. (2012) A Person Identification Method Using a Top-View Head Image from an Overhead Camera. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 16 (6): 696–703

Olivera, R.P. (2012) A Stereo Vision System based on Human Robot Interaction

Satta, R., Fumera, G. and Roli, F. (2012) Appearance-based people recognition by local dissimilarity representations. *Proceedings of the on Multimedia and security - MM&Sec '12* [online], p. 151. Available from: http://dl.acm.org/citation.cfm?doid=2361407.2361433

Ma, B., Su, Y. and Jurie, F. (2012) BiCov: a novel image representation for person re-identification and face verification. *Proceedings of the British Machine Vision Conference*, p. 11

Richter, S.R., Holz, C. and Baudisch, P. (2012) Bootstrapper : Recognizing Tabletop Users by their Shoes. *Proc. SIGCHI 2012*, (c): 2–5

Ramakers, R., Vanacken, D., Luyten, K., et al. (2012) Carpus A Non-Intrusive User Identification for Interactive Surfaces. *Proc. User Interface Software and Technology 2012*

Kohli, M. and Jetawat, A. (2012) Context Awareness and Natural Interaction in Ubiquitous Computing., 3 (6): 5486–5491

Vu, T., Ashok, A., Baid, A., et al. (2012) Demo : User Identification and Authentication with Capacitive Touch Communication., p. 4503

Han, J., Pauwels, E.J., De Zeeuw, P.M., et al. (2012) Employing a RGB-D sensor for real-time tracking of humans across multiple re-entries in a smart environment. *IEEE Transactions on Consumer Electronics*, 58 (2): 255–263

Ewerling, P., Kulik, A. and Froehlich, B. (2012) Finger and hand detection for multi-touch interfaces based on maximally stable extremal regions. *Proc. ITS 2012* [online], p. 173. Available from: http://dl.acm.org/citation.cfm?doid=2396636.2396663

Chen, L., Panin, G. and Knoll, A. (2012) Human Body Orientation Estimation in Multiview Scenarios. *Advances in Visual Computing* [online], 7432: 499–508. Available from: http://dx.doi.org/10.1007/978-3-642-33191-6_49

Barbosa, I.B., Cristani, M., Bue, A. Del, et al. (2012) Re-identification with RGB-D sensors. *Computer Vision–ECCV 2012. Workshops and Demonstrations*. 2012. pp. 1–10

Martinel, N. and Micheloni, C. (2012) Re-identify people in wide area camera network. *IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 31–36

Zhang, H., Yang, X. and Ens, B. (2012) See me, see you: a lightweight method for discriminating user touches on tabletop displays. *Proc. CHI 2012* [online], pp. 2327–2336. Available from: http://dl.acm.org/citation.cfm?id=2208392 [Accessed 16 November 2012]

Ch'ng, E. (2012) The Mirror between Two Worlds : 3D Surface Computing for Objects and Environments. In *Digital Media and Technologies for Virtual Artistic Spaces*. pp. 166–185

Taylor, J., Shotton, J., Sharp, T., et al. (2012) The Vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 103–110

Eisenbach, M., Kolarow, A., Schenk, K., et al. (2012) View invariant appearance-based person reidentification using fast online feature selection and score level fusion. *Proceedings - 2012 IEEE 9th International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2012*, (Avss): 184–190

Albiol, A., Oliver, J. and Mossi, J.M. (2012) Who is who at different cameras: people re-identification using depth cameras. *IET Computer Vision*, 6 (5): 378–387

Rofouei, M., Wilson, A.D., Bernheim Brush, A.J., et al. (2012) Your Phone or Mine? Fusing Body, Touch and Device Sensing for Multi-User Device-Display Interaction. *Proceedings of the Annual Conference on Human Factors in Computing Systems (CHI'12)* [online], pp. 1915–1918. Available from: http://dl.acm.org/citation.cfm?id=2208332

Liu, W., Zhang, Y., Tang, S., et al. (2013) Accurate estimation of human body orientation from RGB-D sensors. *IEEE transactions on cybernetics* [online], 43 (5): 1442–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23893759

Kviatkovsky, I., Adam, A. and Rivlin, E. (2013a) Color invariants for person reidentification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (7): 1622–1634

Kviatkovsky, I., Adam, A. and Rivlin, E. (2013b) Color invariants for person reidentification. *IEEE transactions on pattern analysis and machine intelligence* [online], 35 (7): 1622–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23681991

Yusof, H. (2013) Continous Human Sensing and Tracking for Tabletop Display [online]. Available from: http://www.youtube.com/watch?v=cViJ-y1xfpE

Southwell, B.J. and Fang, G. (2013) Human Object Recognition Using Colour and Depth Information from an RGB-D Kinect Sensor. *International Journal of Advanced Robotic Systems*, 10

Motta, T. and Nedel, L. (2013) Interaction with Public Displays Using a Natural User Interface Based on an Extended Version of Kinect SDK. *ucsp.edu.pe* [online]. Available from: http://www.ucsp.edu.pe/sibgrapi2013/eproceedings/wtd/114693.pdf

Vezzani, R., Baltieri, D. and Cucchiara, R. (2013) People reidentification in surveillance and forensics. *ACM Computing Surveys* [online], 46 (2): 1–37. Available from: http://dl.acm.org/citation.cfm?doid=2543581.2543596

Tao, D., Jin, L., Wang, Y., et al. (2013) Person re-identification by regularized smoothing kiss metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 23 (10): 1675–1685

John, V., Englebienne, G. and Krose, B. (2013) Person Re-identification using Height-based Gait in Colour Depth Camera. *Image Processing (ICIP)*. 2013. pp. 3345–3349

Shotton, J., Fitzgibbon, A., Cook, M., et al. (2013) Real-time human pose recognition in parts from single depth images. *Studies in Computational Intelligence*, 411: 119–135

Wei-Shi, Z., Shaogang, G. and Tao, X. (2013) Re-identification by Relative Distance Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (3): 653–668

Berger, K. (2013) The role of RGB-D benchmark datasets : an overview. *ArXiv: Computer Vision and Pattern Recognition (cs.CV)* [online]. Available from: http://arxiv.org/abs/1310.2053

Munaro, M., Basso, A., Fossati, A., et al. (2014a) 3D reconstruction of freely moving persons for re-identification with a depth sensor. *2014 IEEE International Conference on Robotics and Automation (ICRA)* [online], pp. 4512–4519. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6907518

Bedagkar-Gala, A. and Shah, S.K. (2014) A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32 (4): 270–286

Wang, Y., Hu, R., Liang, C., et al. (2014) Camera compensation using a feature projection matrix for person reidentification. *IEEE Transactions on Circuits and Systems for Video Technology*, 24 (8): 1350–1361

Ma, B., Su, Y. and Juri, F. (2014a) Covariance descriptor based on bio-inspired features for person re-identification and face verification. *Image and Vision Computing*, 32 (6): 379–390

Ma, B., Su, Y. and Jurie, F. (2014b) Discriminative Image Descriptors for Person Re-identification. In *Person Re-Identification, ser. Advances in Computer Vision and Pattern Recognition*. Eds. Springer London. pp. 23–42

Hu, G., Reilly, D., Alnusayri, M., et al. (2014) Dt-Dt. *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces - ITS '14* [online], pp. 167–176. Available from: http://dl.acm.org/citation.cfm?id=2669485.2669501

Yusof, H., Ch'ng, E. and Baber, C. (2014) Human Sensing for Tabletop Entertainment System. In *Context-Aware Systems and Applications*. pp. 283–292

Munaro, M., Fossati, A., Basso, A., et al. (2014b) One-shot person re-identification with a

consumer depth camera. In *Person Re-Identification, ser. Advances in Computer Vision and Pattern Recognition*. Eds. Springer London. pp. 161–181

Gong, S., Cristani, M., Loy, C.C., et al. (2014a) Person Re-Identification. Springer London

Bazzani, L., Cristani, M. and Murino, V. (2014) SDALF : Modeling Human Appearance with Symmetry-Driven Accumulation of Local Features. In *Person Re-Identification, ser. Advances in Computer Vision and Pattern Recognition*. Springer London. pp. 43–69

Reid, D.A., Nixon, M.S. and Stevenage, S. V. (2014) Soft biometrics; Human identification using comparative descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36 (6): 1216–1228

Gong, S., Cristani, M., Loy, C.C., et al. (2014b) The Re-identification Challenge. In *Person Re-Identification* [online]. London: Springer London. pp. 1–20. Available from: http://link.springer.com/10.1007/978-1-4471-6296-4_1

Rusňák, V., Ručka, L. and Holub, P. (2014) Toward natural multi-user interaction in advanced collaborative display environments. *Future Generation Computer Systems* [online], 54: 313–325. Available from: http://dx.doi.org/10.1016/j.future.2015.03.019

Wired.com (2015) Facebook's M Virtual Assistant [online]. Available from: http://www.wired.com/2015/08/facebook-launches-m-new-kind-virtual-assistant/ [Accessed 1 March 2016]

Lachat, E., Macher, H., Mittet, M. -a., et al. (2015) First Experiences With Kinect V2 Sensor for Close Range 3D Modelling. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS)*. 2015 [online]. pp. 93–100. Available from: http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-5-W4/93/2015/

Microsoft (2015) Kinect for Windows SDK v2.0 Face Tracking [online]. Available from: https://msdn.microsoft.com/en-gb/library/dn782034.aspx [Accessed 31 December 2015]

Pala, F., Satta, R., Fumera, G., et al. (2015) Multi-modal Person Re-Identification Using RGB-D Cameras. *IEEE Transactions on Circuits and Systems for Video Technology* [online], 8215. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7088601

An, L., Kafai, M., Yang, S., et al. (2015) Person Re-Identification with Reference Descriptor. *IEEE Transactions on Circuits and Systems for Video Technology* [online], (To be published). Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7067434

Martinel, N., Das, A., Micheloni, C., et al. (2015) Re-Identification in the Function Space of Feature Warps. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [online], 8828 (c): 1–1. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6975169

Clark, R. a, Pua, Y.-H., Oliveira, C.C., et al. (2015) Reliability and concurrent validity of the Microsoft Kinect V2 for assessment of standing balance and postural control. *Gait & Posture*

[online], pp. 3–6. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0966636215000740

Zheng, W.-S., Gong, S. and Xiang, T. (2015) Towards Open-World Person Re-Identification by One-Shot Group-based Verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* [online], 8828 (2): 1–1. Available from: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7152932

Wu, Z., Li, Y. and Radke, R.J. (2015) Viewpoint Invariant Human Re-identification in Camera Networks Using Pose Priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37 (5): 1095–110

Jaha, E.S. and Nixon, M.S. (2015) Viewpoint invariant subject retrieval via soft clothing biometrics. *Proceedings of 2015 International Conference on Biometrics, ICB 2015*, pp. 73–78

Zdnet.com (2015) Zdnet Report [online]. Available from: http://www.zdnet.com/article/amazon-to-flex-internet-of-things-artificial-intelligence-muscle-in-2016/ [Accessed 13 February 2016]