Investigating the potential benefits of natural gestures and skeletal tracking in One-to-Many Presentation Environments

Shane Mulcair

August 2015

A Research Dissertation submitted in partial fulfilment for the Degree of Masters of Science in Software Engineering and Database Technologies at the National University of Ireland Galway.

College of Engineering and Informatics, Discipline of Information Technology

Head of Department: Dr. Michael Madden

Thesis Supervisor: Karen Young

Certificate of Authorship

Final Thesis Submission

MSc in Software Engineering and Database Technologies

Discipline of Information Technology

National University of Ireland, Galway

Student Name:	Shane Anthony Mulcair
Telephone:	0857069684
E-mail:	shanemulcair@outlook.com
Date of Submission:	20/08/15
Title of Submission:	Investigating the potential benefits of natural gestures and skeletal tracking in One-to-Many Presentation Environments
Supervisor Name:	Karen Young

Certification of Authorship:

I hereby certify that I am the author of this document and that any assistance I received in its preparation is fully acknowledged and disclosed in the document. I have also cited all sources from which I obtained data, ideas or words that are copied directly or paraphrased in the document. Sources are properly credited according to accepted standards for professional publications. I also certify that this paper was prepared by me for the purpose of partial fulfilment of requirements for the Degree Programme.

Signed: Shane Mulcair Date: 20/08/15

Acknowledgements

I'd like to take this opportunity to thank my supervisor, Karen Young, for all her help and feedback during this thesis programme.

To my parents, whose support has been continuous, and whose love and friendship means the world. They've supported every bit of my education so far, and will always encourage learning.

And to my girlfriend Emily, who has been there through every stressful piece of this work, and still believed I'd complete it.

Table of Contents

Certificate of Authorship 2

Acknowledgements 3

List of Figures 6

List of Tables 6

Abstract 8

Thesis Statement 9

Introduction 10

Introduction to Area of Study 10

Research Objectives 11

Overview of the Chapters 12

Literature Review 14

Introduction 14

Human-Computer Interaction 14

Skeletal and Gesture Tracking 17

Skeletal versus Glove-based tracking 19

Consistency and discoverability of gestures 21

Gesture tracking in other formats 23

Limitations of gesture tracking 25

Conclusion 28

Skeletal Tracking and the Microsoft Kinect 29

Introduction 29

Kinect Version One 29

Kinect Version Two 30

Competing sensor systems 31

Conclusion 32

Research Methodology 34

Introduction 34

Research Objective 34

Review of the Research Method 35

Quantitative Research 35

Qualitative Research 35

Choice of Research Method 36

Detailed Description of Research Method 36
Application choice and setup 36
Component One – Gesture Tracking 38
Component Two – Skeletal Tracking 41
Users 43
Data Collection 44
Limitations...45
Conclusions 46
Results and Findings...47
Introduction 47
Users 47
Survey Results 48
Have you worked with gesture control before? 48
Have you made presentations before? 49
Did you find the swipe gestures easy to perform? 50
Did you find the start/stop gestures easy to perform?..................................51
Did you find that gestures made it easier to present?..................................52
Did you find the registration (hand raised) gesture easy to perform? 53
Did you find it easy to move around the map using your hand? 54
Did you find the “push” gesture easy to perform? 55
User Feedback..55
Did you have trouble getting your gestures recognised? 56
Did you have trouble with the skeletal tracking test? 57
Any additional comments...57
Conclusion..58
Discussion...59
Introduction 59
Research Questions 59
Evaluation of Findings...59
Can gesture control replace a hand held control when making a presentation?59
Where is the optimum placing of a sensor, to track a user making a presentation? 60
Does the user feel comfortable with the controls, and are the gestures intuitive enough that they can be made instinctively, rather than pausing for consideration each time? 60

Overall comments 61
Conclusion..63
References 65

List of Figures

Figure 1 - Sega Activator, one of the first commercial gesture tracking systems. www.gamewalker.link 10
Figure 2 - Kinect volume measurement from Dutta, 2011.............................26
Figure 3 - Kinect tracking a face, with vertices joining the returned points....30
Figure 4 - Microsoft Kinect, Xbox One edition...30
Figure 5 - Initial Presentation setup including kinect 38
Figure 6 - Presentation setup for left-handed presenters..............................39
Figure 7 - Presentation setup for right handed users.....................................40
Figure 8 - Wireless Presentation Remote, www.kensington.com..................40
Figure 9 - Room setup for skeletal tracking...42
Figure 10 - Kinect tracked points explanation - www.microsoft.com............42

List of Tables

Table 1 - Have you worked with gesture control before? 48
Table 2 - Have you made presentations before?...49
Table 3 - Did you find the swipe gestures easy to perform? 50
Table 4 - Did you find the start/stop gesture easy to perform?......................51
Table 5 - Did you find that gestures made it easier to present? 52
Table 6 - Did you find the registration gesture easy to perform? 53
Table 7 - Did you find it easy to pan around the map? 54
Table 8 - Did you find the push gesture easy to perform? 55

Abstract

This thesis sought to discover whether simple gesture controls could help a presenter navigate a slide deck, or similar presentation material, when presenting in front of an audience. It begins with a review of current literature on the topic of natural gesture control, and relevant parts of Human Computer Interaction.

Once the review is complete, the chosen sensor is presented, with justification for its use in a prototype application for testing. The experimental setup is then covered, as the users were tasked with making presentations at the front of a room, using only gesture controls to navigate the presentation. The user feedback was gathered through the use of short surveys, with a combination of yes/no and open-ended questions. The feedback was collated and presented in charts, while the open ended feedback was discussed in points.

Finally, the results are compared and contrasted with the literature covered in the review, to point at areas where the results aligned or disagreed with the literature. One of the most significant was the reinforcing of the comments made by researchers, that the users should be able to define their own gestures for their tasks.

It was found that, by putting the users under a semblance of pressure (the presentations were 20 slides long, with a maximum of twenty seconds to cover each slide), over 65% of the users could navigate the slides easily with gestures.

The users themselves provided some valid feedback on the placement of the sensor – it was initially perpendicular to the projector screen, but users found it more beneficial to be at a 45 degree angle to the screen, facing the user. In this way, the presenter could look at the slide, and then turn to the audience, while still keeping their arms in view of the Kinect.

Thesis Statement

"Investigating the potential benefits of natural gestures and skeletal tracking in One-to-Many Presentation Environments".

Introduction

Introduction to Area of Study

Computing has slowly moved through various input methods, from individual switches, to punch cards, to keyboards, towards today's mix of touch, voice, gesture, pointing device and keyboards. Each of these methods has been found to have a particular scenario where it is best used. Touch works best for smaller devices. Voice, for individual users in quiet areas, or users wearing microphones. Pointing devices, such as mouse or pen input, are best combined with keyboards, for users working at the computer. This, then, leaves gesture input.

Gesture input has advanced considerably in the past decade, moving from using webcams, fitted gloves or expensive multi-camera setups, towards low-cost sensor and tracking systems. A far cry from the Sega Activator, which simply used light beams to decide where the user had moved, giving only 8 possible inputs, systems such as the Wiimote motion controller, or the Kinect, have made gesture tracking accessible to developers. This thesis shall look at the field of gesture tracking and input, and its application to a specific environment.

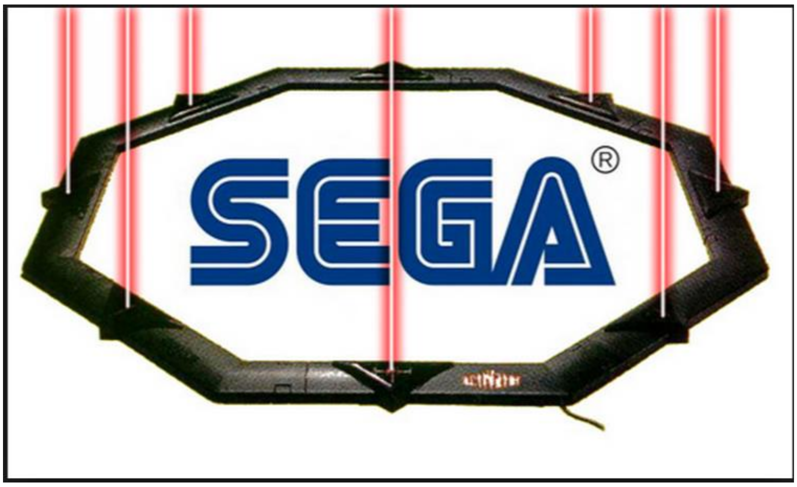

Figure 1 - Sega Activator, one of the first commercial gesture tracking systems. www.gamewalker.link

The field of gesture tracking has a considerable amount of prior and ongoing research, especially in potential uses. This thesis seeks to suggest one possible target scenario for gesture input; namely, in the one-to-many presentation scenario. The scenario is familiar to anyone- whether in an office, classroom or lecture hall, the speaker stands near the screen, discussing their slide deck, before an assembled group. The speaker generally uses either a handheld remote, or presses a key/mouse button on the computer to advance to the next slide. This is quite limiting, in that the speaker must either walk back and forth to the computer (or use an accomplice to control the slides), or else be in range of the remote at all times.

Gesture control, then, could offer a potential solution to the problem of navigating the presentation. Once gestures and skeletal tracking are accepted as a possible method, then deeper analysis is required. The decision needs to be made on what is tracked- the whole body, the head, or the hands? From watching keynotes, it is clear that people are most accepting of a speaker moving their hands while talking. Even when holding a remote, speakers will use that hand to gesture to emphasise points. It makes sense, then, to use the hands for tracking, while being sensitive to these other unintentional gestures.

Gestures are an important piece of human interaction. Easily understood by humans, it is necessary to carefully restrict and define their motions when tracking for computer input.

Research Objectives

This thesis shall define several gestures for control of applications in the given scenario. It will not be enough to simply look at slide control (for which a simple “swipe” gesture may be enough control). Instead, there should be analysis and exploration of other presentation methods- looking at map data through pan and zoom, for example. Moving an image around on a screen requires the use of both hands, to “grasp” the image, and then “pull” or “push” to zoom in and out. These gestures may seem simple, but need to be explored to discover if users can actually find them to be comfortable and a worthy replacement for moving back to the keyboard podia.

It next becomes necessary to discuss *how* this tracking shall be accomplished. The thesis shall, with feedback from users, discover where is most comfortable for a sensor to be placed with respect to the speaker, and in what way they will interact. The chosen sensor is a Microsoft Kinect, a system which incorporates visual and infra-red cameras to track a user's movements. The movements are modelled onto a tracked user skeleton, which provides points of reference when interpreting. Natural gestures use hands and arms to create a symbol- the raised hand is a familiar one from early childhood, to get attention in school. That same gesture could serve to identify which user wishes to control the system. This sensor shall observe the speaker, and, on detecting a defined gesture, should interpret the gesture into its defined action, and pass it to the target application.

Overview of the Chapters

- Chapter One gives a brief introduction to the area of study, as well as to the research objectives, as well as a brief summary of the subsequent chapters
- Chapter Two gives a critical review of the current field of research into the chosen thesis area. It begins with an introduction covering the development of HCI interfaces, and the argument for investigating the chosen method of input. The rest of the chapter covers the themes of HCI, skeletal tracking, and gesture input.
- Chapter Three covers the justification for using the Microsoft Kinect as a chosen approach for the prototype application. The previous and current sensors are compared, and contrasted with several competing sensors.
- Chapter Four presents the research method, and the rationale used in creating a prototype application. This chapter is where the research question is covered, and explained in terms of the questions that will be answered. Other research methods will be mentioned here, and it will explained as to why the specific technology and approach were taken. This chapter contains the focus of the research that was done, so it also includes a detailed description of the research, as well as the limitations encountered.

- Chapter Five presents the research findings, following the research themes identified in the literature review in chapter two. The chapter focuses on what was found to answer the research questions from chapter four.
- Chapter Six contains an evaluation of the results from chapter five, contrasting them with the research questions from chapter four, to discuss areas where the theory and practice either correlated, or conflicted.
- Chapter Seven concludes the paper, by presenting the main conclusions of the previous chapters, framed in terms of what the findings actually meant, in the context of the questions. Naturally, the chapter concludes with suggestions for areas of further study.

Literature Review

Introduction

The field of human-computer interaction has changed rapidly in the past decade, through the introduction of new interaction methods which are specific to certain areas. Touch has found applications in small and medium-sized devices, with the iPhone leading the way for all-touch smartphones in 2007, followed by the release of the iPad in 2010. These were by no means the first of their kind, but they were the first commercial successes which introduced the world to interaction with an operating system without a mouse or keyboard.

In the field of computer gaming, the Wii sold the equivalent of one console for every seventy people on the planet. For the first time, a gesture-based system was popular and widely available, and massively outsold competing games consoles.

It is into that world, the world of affordable, widely-available touch and gesture-based systems, that the question is raised; Outside of gaming and simple on-screen gestures, is there a use for natural user interaction? This paper shall look at natural user input in the specific realm of the presentation. In other words, a lecturer or presenter using gestures and skeletal tracking to aide in making presentations.

This chapter reviews the current literature on the topic of gesture tracking, with emphasis placed on hand and arm gestures. It begins with a brief introduction to HCI, before addressing the use of skeletal tracking for gesture input, and the applicable research. The aim of the overall paper is to build on this research, using a Microsoft Kinect v2 as a proxy for building a prototype application. This will be explored in later chapters.

Human-Computer Interaction

Human Computer Interaction is a field which has evolved gradually over the past 70 years, from using levers and switches on the first systems, through punch cards and keyboards, to the modern amalgamation of choices we have today. An average user today could, in a single day, use a touch screen to check their emails and messages, use voice control for setting a reminder, use a mouse and keyboard (or a laptop-style touch pad) for working, and even sign for a parcel using a keyboard and screen. The key point that can be seen from this, is that different input methods have their own appropriate areas. The mouse and keyboard paradigm has been effective for the longest time, but the keyboard was quickly replaced by a touch screen as smart phones evolved, followed by the introduction of tablet computing, where devices may have a power button, volume control buttons, and no other physical buttons.

This, then, is the field of HCI, Human Computer Interaction. In this paper we shall look at the input component of HCI. Input methods, for the purpose of this paper, are the commonly used methods of translating a human users actions into instructions understandable by a computer. These methods always comprise a hardware component (but may not necessarily have to be directly manipulated by the user).

Once keyboards were the established interface for computers, attention switched to more direct manipulation of items on the screen. Humans have a love for skeuomorphism, the desire to make virtual objects mirror their real ones. Folders still look like a file, even today. The save button looks like a floppy disk. This love of direct manipulation of objects was captured first in Sketchpad, by Sutherland in 1963. Using a light pen and a screen, basic interactions were possible- especially actions such as making the object bigger or smaller with a touch, or moving the object on screen. At the time, operating systems used text-only representations of files and folders. Indeed, the term “icon” was not coined until 1975 by David Smith.

However, despite the limitations of the display, research still focused on getting away from the keyboard. Obviously, the mouse was the most important step forward here- indeed, it has been credited with being the reason graphical user interfaces (GUIs) became popular. Despite this, researchers were still investigating other methods to map user inputs to computer input. Dataglove was the next stage in this- digital cameras were a decade away, so video tracking was impossible, so Sandin and Defanti proposed a wearable glove in 1977, connected via cable to the computer. This glove would track a user's hand movements, but, due to its use of light-based sensors, it was mainly used to manipulate sliders – the glove could detect the hand closing, as well as basic motion, so the "push" motion of a throttle could be interpreted.

The next breakthrough from the glove, became touch. Users could "click" with a mouse, certainly, but it is far more natural to point at the exact icon they want, and press it. Incredibly, the issues with touch input were outlined by Lee, Buxton and Smith in 1985. At the time, a computer mouse had only one button, which limited user input. By allowing multi-touch, the authors proposed, multiple different actions could be performed by varying the number of touches used (It was not until the iPad, in 2010, that these multi-touch inputs became a commercial reality). Unfortunately, as the authors noted, the largest problem with touch input, is in the feedback for users. Whether in resistive or capacitive touch, there is no "click", in the form of vibration or movement, to tell the users that their input was successful. Even when the Buxton followed this paper with another describing a multi-touch tablet, this same issue was mentioned.

The solution to the issue of feedback is still being worked on today – Kaaresoja, Brown and Linjama proposed using vibration as a method, identifying the users input, and vibrating the device in response. This method is used in most smartphones today for users with limited vision, but is impractical when scaling to larger devices – imagine a laptop or computer monitor with a strong enough vibration to provide feedback across the screen.

Other input methods include voice, or other specialised methods (light pens, remote controls for specific operations etc), however, these methods either require a specialised piece of hardware, or require extremely limited inputs in the case of voice. The difficulty is not in recognising the individual words (as far back as 1995, dictation software was capable of up to 100,000 words recognised, according to Kamm). The problem is understanding intent- the user must construct each instruction in such a way as to be unambiguous.

It should be clear, then, that the realm of HCI strives to map user intentions into valid and precise computer instructions. The keyboard gives exact to-the-letter input, at the cost of being awkward with GUI systems. Microsoft lists dozens of keyboard shortcuts to accomplish tasks (Control + C, Control + V would be familiar to any user of Copy and Paste), almost all of these tasks can also be completed using the mouse, and several clicks. Indeed, when the GUI was introduced to Windows, two games were added- Minesweeper and Solitaire. These games weren't designed as productivity sinks. Instead, Minesweeper teaches a user precise left and right clicks, Solitaire teaches clicking and dragging. Two simple games taught users the fundamentals of mouse control easily.

This all comes from the concept of What You See Is What You Get (WYSIWYG), a system of making the system act exactly as the user expects. When controlling a computer, the normal user does not wish to enter obscure commands- they wish to touch on an icon to select it, or press a button. The Sayre data glove mapped this in the way users could slide virtual handles. In the same way, the use of different HCI methods must mirror a user's desired input. Pen input may go in several directions- either simple handwriting entry, but also as a "pick" and "move" tool, dragging by moving the pen.

It is obvious though, that these methods- a wired glove, a wireless pen, a keyboard or mouse, are difficult to use when making a presentation. The glove will track every hand motion as the user speaks. The pen, mouse or keyboard have difficult analogies when attempting more than a click to progress a slide. It becomes possible then to either use a basic remote control with advance/retreat buttons, or to look at other potential options. This thesis shall explore the area of gesture tracking and input.

Skeletal and Gesture Tracking

One step forward, two steps back. That is how Norman and Nielsen (2010) described the current state of the art in natural gesture interactions. That is, the methods used to track and interpret gestures are a “usability disaster”, according to the authors. It is intimidating to see that opinion from Don Norman, the author of “The Design of Everyday Things”, and yet, there is ample positive research into the area of HCI, and specifically natural gesture input.

This thesis, then, aims to contradict the opinion of one of the most popular proponents of “good” design, as applied to a specific application. While Norman applies his criticism to touch input, the same observations are no less valid when looking at skeletal tracking-based gesture input.

It should be noted that, even within the field of skeletal tracking, there are subclasses of input types. Gestures are tracked in either Offline or Online modes. Offline records the motions, and applies analysis later. Ju, Black, Minneman and Kimber (1998) used video to record users who made presentations, in order to extract intent and context from the talks. The authors attempt to extract two different types of information from video- firstly, events where the speaker on screen starts talking, or switches to using their slides, and secondly, the gestures the speaker uses when discussing the slides. In this way, the authors attempt to determine the motions for “occlusion, motion, pointing, writing, and revealing” from offline analysis of the video. The issue with the applicability of their research today is their reliance on two things: firstly, a video of the presentation (so the gestures are not interpreted live), and secondly, many of their algorithms are based on the idea that the new information must enter the screen from a side (as they work with overhead projector slides, which are pushed on screen). However, despite these limitations, they did extract important information within their scenarios, recognising pointing gestures in their experiments.

The second type of gesture tracking is online tracking. First demonstrated in the oN Line System (NLS) in 1968, gestures are interpreted as they are made, for live computer interaction. The user interacts directly with the UI using gestures, so the tracking system acts as a form of middleware to translate the gestures into traditional computer inputs. Suma, Krum, Lange, Koenig, Rizzo and Bolas (2012) discussed a middleware for interpreting gestures for two disparate scenarios- video games, and accessibility for users with motor impairments. Simple gestures are identified that can be used across the operating system, rather than within a specific application, such as extending the arm towards the screen to press an on-screen button, as well as tracking a hand to move the on-screen mouse pointer. The authors make the case, supported by others such as Chaudhary, Raheja, Das and Raheja (2011), that the gestures do not simply track the hand; instead they track the shape and orientation of the hand, relative to other joints – in the examples given, the shoulders or elbows.

The paper covers much of the field of gesture tracking indirectly, by outlining methods to decompose and represent complex gestures. They work on the concepts of position, angle and body constraints, to track orientation of limbs and the whole body. Perhaps as importantly, the concept of a time constraint is identified – a gesture should have a start and end point in time, so that sequences can be defined. Using these constraints, both sequential and simultaneous gestures are explored, to be recorded as actions and given an interpretation for the computer.

These actions are important, as they must cover what Norman (2010) calls Consistency. The same gesture should always result in the same action from the computer, so the constraints identified must restrict the gesture to what is intended by the user. Both mouse and keyboard events are covered, which allows the FAAST framework to integrate gesture controls into many existing applications.

Skeletal versus Glove-based tracking

It need to be mentioned, that tracking of hand gestures has been performed before, using a glove-based approach. This ties the user to the computer with an input cable, but could potentially be made wireless. Pavolvic, Sharma and Huang (1997) make the case that a worn glove (at the time, using input and power cables) impacts on the “naturalness” of the gestures being made. At the time their paper was being published, they noted that most work was focussing on static gestures (postures, or simply holding the arm at a given angle and pose). But, as Chaudhary

(2011) correctly notes, *the motion of the hands conveys as much meaning as their posture does.* A glove based approach, while it has the ability to track individual fingers, loses the posture of the hand relative to the arm, which removes its utility in tracking sign language.

Pavolvic et al (1997) do make excellent points that are still valid today though – that the "common HCI still relies on simple mechanical devise – keyboards, mice and joysticks". In this, they only miss out on touch screens to encompass the majority – even today – of HCI. Look at a system such as the Oculus Rift, where entirely new hand-held controls are being created to attempt to keep the immersion of virtual reality. The authors make the case that many approaches to gesture tracking focuses on a subset of gestures, one of either hand tracking, pose classification, or hand posture classification. It was seen in Suma et al (2010) that the more modern systems take the opposite approach, of encompassing all three subsets of tracking, in order to properly grasp the intent of the user.

To move to a more modern paper, work done by Villaroman, Rowe and Swan (2011) takes a different perspective, completely overlooking the glove-based devices in their HCI overview, while mentioning the Nintendo Wii and Playstation Move. The reason for this is simple: tools attached to the Wii, Playstation or Xbox have been commercial successes, but a glove-based interaction has never been commercially successful. Sandin and Defanti, even as they created Dataglove in the seventies, could only find uses for the tool in manipulating virtual sliders and similar – it was a tool with a very narrow applicable focus. Of course, Villaroman et al (2011) focus on one specific technology in their work, the Microsoft Kinect.

While the authors don't discuss competing products (the Leap Motion, or equivalent sensors), they complete an excellent analysis of advantages of gesture and skeletal tracking in their work. Some limitations (such as a greater involvement being required on the part of the developer) are glossed over as advantages in the Kinects use as a teaching tool. In fact, despite the section being named "Advantages and Limitations", there are no real limitations mentioned in terms of the sensor, or the applicability of gesture tracking. Their included photos show the tracking of multiple users (2-4 users were supported under ideal conditions), but neglect to mention the problems reported by researchers such as Mankoff and Russo (2013), who identify issues with responsive tracking when the users move quickly towards or away from the sensor.

Consistency and discoverability of gestures

A key point that Norman and Nielsen (2010) make in their criticism of gesture interaction is that of consistency. Hespanhol et al actually did a comprehensive study on this connection, in their study in 2012. They looked at the interaction between gestures and a display, but at a distance. In their study, they identified the issue of gestures being misinterpreted by the vision system, and suggested that interactions are broken into two simple categories; namely, selection and rearrangement. To accomplish this, five gestures were identified: Pushing, Dwelling (holding the hand over one spot), lassoing ("drawing" a circle in the air around an icon), grabbing (closing the hand in an exaggerated gesture on an icon) and enclosing (using two hands to "frame" and area"). The participants in the study were not told the gestures they would need to use the system, so the study also serves as a study on how intuitive the gestures are. The only real issue, then, (besides the low participant numbers) was that the gestures went against those proposed in other studies above- namely, they used a single hand for 4 of the five gestures, and they only tracked the hand itself, not the arm or even the angle between hand and arm. This possibly limited the study, but as they note, even though the display was possibly too "busy" in the number of items displayed at once, the gestures were still easy to use according to the participants.

This study does raise valid concerns about the gestures identified. The "pushing" gesture would seem to be an intuitive one, as any user is familiar with buttons. However, its simplicity was also its weak spot, in that the change in distance on the z-axis (towards/away from the sensor) led to near-continuous false readings. The study reports that the game icons flickered as the users moved- the human shoulder is a pivot, so the hand will naturally move in an arc in 3 dimensions, not 2.

Perhaps the most important section of the study is in the findings, and this is a result duplicated in other studies. The gestures, once learned (whether by teaching or discovery) are easily reproduced by the user. These simple gestures, despite the trouble with asking users to figure them out without guidance, were then used repeatedly by the participants. The gesture that was the least intuitive, and most error-prone once discovered, was the “lasso” motion. This anti-clockwise circle drawn on screen is, according to the study, “physically exhausting”.

This gesture- an anticlockwise circle – is not intuitive, nor does it map to a familiar action from other paradigms. It was seen in the discussion of Dataglove that the glove input was used to map familiar “push” and “pull” movements to sliding a virtual control. In a similar vein, the original Windows games Solitaire and Minesweeper were created to introduce users to a mouse-based operating system, by mapping familiar gestures to those in the game (Nachimuthu and Vijayakumari, 2011). Minesweeper familiarised the users with the left and right click, in the paradigm of precision clicking on small tiles. Solitaire introduced users to general clicking, but also to drag-and-drop control. These gestures were consistent across the OS, so users could practice them unconsciously in the games while interacting with them in the wider operating system.

While the mouse has clear controls - movement in the plane of the desk, as well as two or three buttons – it is more difficult to formally define gestures as clear controls. Kammer and Wojdziak attempted this task in 2010 with their paper on formalisation of gestures. Although they focused on touch, in the form of multi-touch input, their work is no less valid for skeletal tracking-based gestures. This is easily argued, as they place emphasis on the gesture, rather than the input method. A key point, perhaps THE key point of their paper, is in the concept of an *atomic* gesture, i.e. one with a defined start and end point that is consistent. Their gestures centre on a compass rose on a surface, so the computer can relate interactions such as a line, arrow or circle easily. But these same gestures can easily be translated to 3D space – a line drawn by a hand in the air is an atomic gesture with a start and end, after all. A brief mention is made of fatigue again – a consistent point made throughout many of the papers discussed in this review. For the first time, specific gesture types are mentioned in the context of fatigue, namely those which require stopping, then moving again, or repeated fluid gestures.

The only failure point of this paper is in the lack of exploration of complex gestures. Due to the complexity of the formal notation outlined in the paper, complex gestures are omitted for the sake of "simplicity and comprehensibility" – but, it must be pointed out that, if the formal notation veers from comprehensibility once the gestures progress from simple, then the notation may need to be reworked.

Gesture tracking in other formats

As mentioned above, Kammer et al (2010) explored gestures in the realm of multi-touch surfaces. Using a surface to perform the gestures on removes many of the difficulties of tracking a gesture across three dimensions, but it also, as Marquardt, Jota, Greenberg and Jorge (2011) note, only focusses on one mode of interaction, ignoring the potential of the "complete interaction space". Their work is similar to combining a Leap Motion (www.leapmotion.com) to track a specific space around the screen, while also allowing the user to interact directly with the physical touch surface.

Murquardt et al make an interesting point, in that research generally aims to expand either *on the surface* or *above the surface* gestures (i.e. touch or skeletal tracking), rather than in a way that combines the two paradigms. By combining the two, they argue that developers would create what they call a "continuous interaction space". The purpose of the gestures is the same as those mentioned by Hesphanol in 2012, but, with the addition of "lifting" gestures to reveal hidden items.

A lot of the work done by the authors is to work around the problem of occlusion. Because their system is using tracking dots monitored by infra-red cameras, the simple motion of the hands can hide the dots from the cameras. A system of eight cameras tracks the dots, and the combined touch/special gestures help in giving the system a start/end point to track. However, despite the effort into "practical" uses of gestures, there is also a far more science-fiction style look at less practical applications, such as using gestures while holding a tablet computer over a touch screen, or using real-world toy bricks or other objects to interact with the screen. Lenovo attempted product ranges like this with their Horizon table, which allowed interaction via touch or using hockey pucks, or other compatible equipment, but this was not a commercial success. By simply looking at the most popular commercial touch screen, the iPad, it's clear that the use of extra peripherals and real-world objects gets in the way of the interaction, rather than aiding in it.

Like several other papers mentioned here, however, Murquardt and his team do not cover the limitations or drawbacks of their suggested approach. As Norman (2012) noted, the problems faced by UI developers when working with gestures are similar to those faced by original UI researchers in Xerox PARC. In a very simple example of the problems inherent in gestures, Norman asks the question "What gesture signifies a copy rather than a move?" In mouse and keyboard computing, users know control+x versus control+c, almost from muscle memory. But how does that same simple differentiation get mapped to gestures? In their book on the subject of designing natural user interfaces, Wigdor and Wixon (2011) argue that, in order to feel "natural" to the user, the best approaches do NOT mimic another experience. So in our copy versus move example, the gestures to perform the actions, Wigdor and Wixon propose, should in no way attempt to mimic a virtual action.

The problem with this, as Malizia and Bellucci (2012) note, gestures have both cultural and human aspects. Some presenters "speak" using their hands, where only a broad gesture may be intended as computer control, and any other hand-waving may simply be part of their natural motion. Other users may have "quieter" body language, and much smaller gestures could accomplish the same task. Consider also the cultural impact of gestures. The thumbs-up gesture, which in the West implies "ok", and could be used to signal a confirmation to the gesture control system, is an insult in some cultures. This simple cultural difference supports Malizia and Belluccis' (2012) argument that, when designing gestural controls, it may be more beneficial to allow the users to personalise them, once the basic controls are mapped.

To support this in a different medium, consider the gesture controls on an Apple computer running OSX. The touchpads on Apple macbooks have long been held up as an example of responsive and accurate mousepads. But, beyond clicking, controlling a mouse cursor, and scrolling using two fingers, there are many less obvious gestures included that seem forced. A simple example is the "pinch to zoom" gesture, familiar to any user of touch screens. But that same control, to zoom in and out, has been replicated in a "double tap with two fingers" gesture that is not obvious. It could be argued that this follows the Wgdor and Wixons argument of not mimicking an action, but at the same time, this is moving away from an established gesture, which seems counter-intuitive.

The Apple guide to the included OS gestures is troubling, as it breaks the consistency argument given by Norman (2010) in multiple ways. Moving between one version of the OS to the next changes given gestures from single to double taps, for example, which is problematic for users- they may try the gesture once, then a second time, and that time-based component may allow or disallow the gesture. If a user retries the single tap quickly, the system may interpret the gesture as the required double-tap. But if the user misses the time allowed, then the gesture to them is inconsistent.

Limitations of gesture tracking

It is difficult to find research that directly criticises gesture tracking. Norman is obviously a critic, as mentioned throughout this review. However research papers are based on finding a positive outcome, so few will base an entire paper on a critical comparison.

To consider the downfalls, then, we must look at the conclusions of the papers with a critical eye. Take, for example, Nickel and Scemanns (2004) work on 3D tracking of head and hands. At their best, when tracking both head and hands to measure where a user was pointing, they reported a false positive rate of between 13% and 26% - from one in every eight, to one in every **four** inputs. While tracking devices have advanced significantly since then, even papers such as Villaroman et al in 2011 reported a large drop in precision as the user moves away from the sensor. Dutta (2011) made similar remarks when evaluating the Kinect, and noted that the field of vision is almost pyramid shaped:

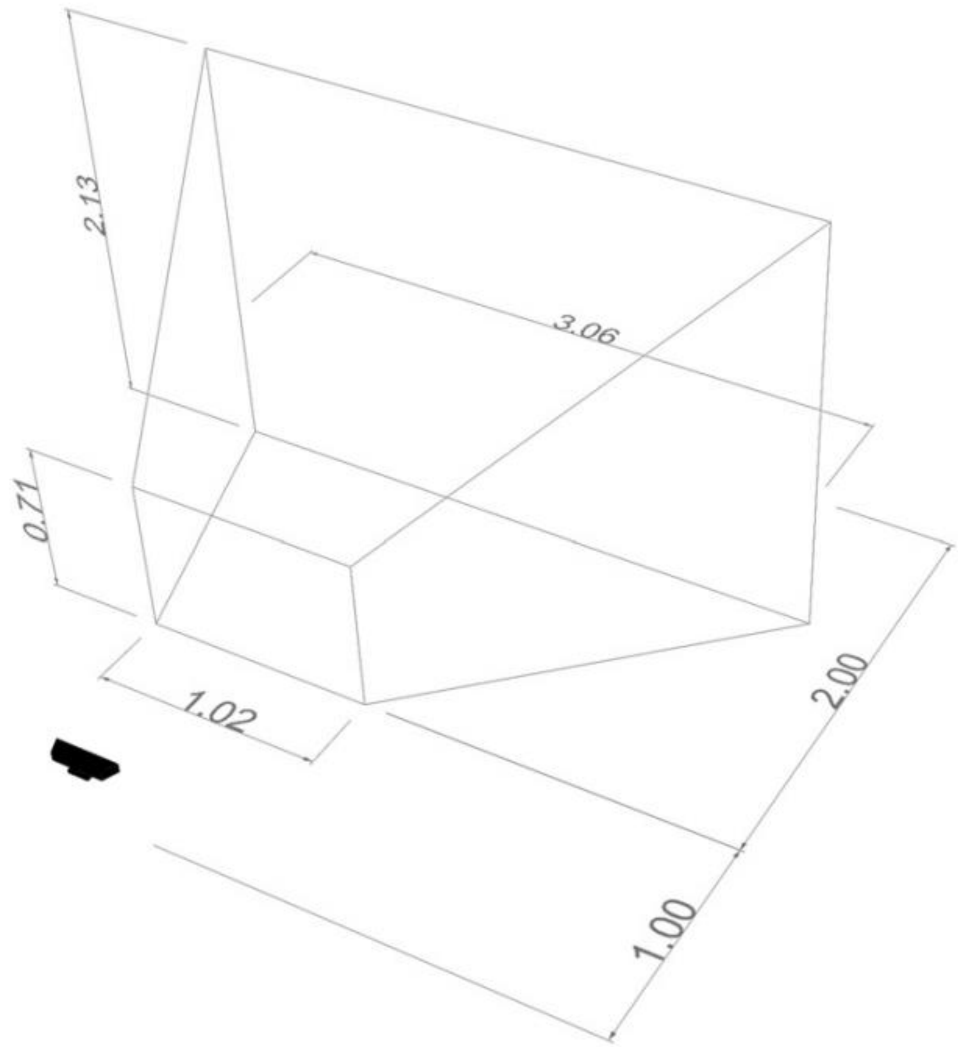

Figure 2 - Kinect volume measurement from Dutta, 2011

This shape is common to any single-camera system, and presents another obvious problem: that of occlusion. Occlusion is simply the blocking of the sensor from tracking a limb, due to the users own body. Marquardt and his team (2011) worked around this difficulty by using 8 cameras around the table, to continuously track both hands, even if one hand inadvertently blocks a sensor from seeing the other hand. With the Kinect, Leap Motion, and other comparable systems, this is not a viable option unless multiple bulky sensors are used. Ren, Liu and Lim (2014) outlined this option while tracking surgical instruments, by using two kinect sensors to view the entire work area, even with occlusion.

Garg, Aggarwal and Sofat outlined some other commonly-discussed issues with gesture tracking in 2009, mentioning the main problem of the limited subset of commands (as we have seen earlier, researchers such as Osunkoya (2013) and Pavlovic (1997) used around five gestures in their studies), as well as one not mentioned previously, that of low light conditions. At the time their paper was written, only 6 years ago, many computer vision systems used webcams or similar cameras to track the user. This has the obvious limitation of being dependent on reflected light to accurately track. The Kinect, and other systems such as the leap motion, use infrared light to track in low light, as well as normal lighting. The leap motion has an infrared emitter and receiver to track with high fidelity, while the Kinect is precise enough to register the users pulse while tracking them in infrared.

In a different approach, Kenety and Parker (2013) discuss the recent research by automobile manufacturers. The authors identify one of the key problems with current gesture tracking- what one manufacturer defines as a gesture may be different from another. Malizia and Belluccis’ (2012) paper worked around this by proposing that, outside of very basic gestures such as swiping left/right, or pointing, that any further gestures are personalised by the user. But in an application such as a car, it must be asked, are gesture controls even a safe option?

Conclusion

It can be seen from the existing research outlined above, that there is an extensive body of work into gesture controls, and gesture tracking. Norman and others have argued, correctly, that gestures have not found their niche, but it can be argued that the inputs need the correct technology behind them, before they can find widespread appeal. The failures of systems such as the Sega Activator would serve as a valid argument against motion control, if not for the widespread success of the Nintendo Wii. It takes the correct identification of possible uses, rather than releasing a product in search of a market.

This problem is no more apparent than the Leap Motion, a motion controller that, even with half a million units sold, failed to find a valid use case. Compare this with the original Kinect sensor, which gained a Guinness World Record for fastest selling consumer electronics device, with over 24 million sensors sold. Much of this was due, similarly to the Wii, to the use of the sensor with fitness and sports games – a huge gaming market, properly identified.

With the research outlined, multiple areas of use have been identified and analysed, such as teaching or controlling systems using gestures. In the following chapters, one particular use will be explored further, that of gesture control to control presentation software. It will be necessary to identify the gestures that comply with Normans (2010) concept of consistency, while also keeping in line with Hespanhol et als' (2012) recommendations on finding intuitive and effective gestures.

Skeletal Tracking and the Microsoft Kinect

Introduction

This chapter serves as a brief overview of the Microsoft Kinect sensor system, and the justification for using it as the chosen sensor when building the test application. As this is generally a technology overview, and a mention of competing products, it was not a good fit for the literature review, though it does serve as an introduction to the Methodology chapter, where the sensor shall be utilised.

The chapter gives an overview of the two Kinect versions, and then moves to competing systems for comparison.

Kinect Version One

Version one of the Kinect was originally created to run with the Xbox 360, a console first released in 2005. This console used (by modern standards) a low power processor, combined with only 512MB of system RAM. The Kinect sensor, released in 2010, was designed not only to be used by that console, but to use no more than 10% of the system resources while doing it.

To work around this, the sensor had a limited resolution of 640x480 pixels in both its colour camera, and infra-red depth camera. Software switches were introduced to toggle the tracking from normal to “near mode” tracking, which focuses the processing on tracking more points on a single skeleton. While normal tracking can monitor up to four users at one time (with 20 joints per user in terms of articulation), the near mode emphasises a single user, generally seated, at close range. This mode is optimised for head and shoulder tracking, returning over 100 points in a 3D array when tracking the face. Shown below is an example of the returned points when performing face tracking, with lines connecting the points to form a mask.

In early research approaches using the Kinect v1, there was no way to access the output of the sensor easily. For this reason, researchers used the raw image and depth coordinates returned, in order to use model based tracking to teach a system how to track a hand (Oikonomidis et al, 2011). This approach changed when an official software development kit was released, which enabled researchers to use the skeleton data returned, rather than the raw images.

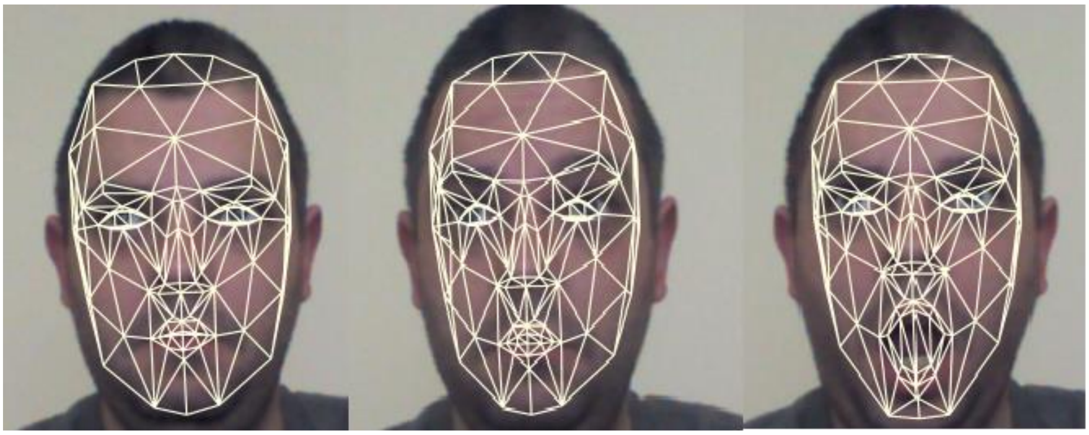

Figure 3 - Kinect tracking a face, with vertices joining the returned points

Kinect Version Two

Figure 4 - Microsoft Kinect, Xbox One edition

The Kinect v2 sensor, shown above, is a self-contained infra-red camera, visible light camera, time-of-flight processor, and microphone array. The first three items enable accurate depth perception and motion tracking at short and medium ranges, the fourth item enables the use of voice control, with an emphasis on direction. The four microphones in the base are capable of returning a direction vector, to where the sound is coming from. It uses this system to identify which visible user is speaking.

The Kinect was chosen as the sensor, rather than using another camera tool such as a webcam, because it returns three dimensional point data for everything in its field of vision. Many of the older systems used a single camera with multiple light sources, to enable the system to process shadows and skin tone, to track the hand (Raheja et al, 2010). This requires processing every 1/3 of a second, a predefined skin tone, and a specific light setup. The Kinect removes these requirements, by using two cameras (visible and infra-red light) in one system, with a known and unchanging separation distance. The sensor was designed to be used in living rooms, so it is more resistant to lighting requirements.

The newer sensor, thanks to a higher-resolution camera, returns 25 “joint” objects for each skeleton object it tracks. Wrist, hand tip, and thumb, are three of the joints, which allow hand position and orientation to be tracked easily.

This was the chosen sensor for the work done in the following chapter, due to its ready availability, research completed using its predecessor, and the authors’ familiarity with the .net programming language family.

Competing sensor systems

Mention should also be made of the competing skeletal and gesture tracking systems available, which make use of similar or different technologies. The Leap Motion (www.leapmotion.com) is far smaller than the Kinect, is powered from its host computer, and has higher fidelity for finger tracking. Indeed, Weickert, Bachmann, Rudak and Fisseler (2013) noted an accuracy to 1.2mm when tracking movement. Their study presents the high precision in all planes of tracking, especially when the comparison is made to the Kinect, but fails to mention the "significant drop in accuracy" when objects move more than 25cm from the sensor (Guna, Jakus, Pogacnik, Tomazic and Sodnik, 2014). This would, despite its low cost, remove the leap motion from contention when tracking a moving user, some distance from the sensor – in the case of this paper, the user could be a metre or more from the sensor.

Other sensors mentioned by papers mentioned above include the Wii remote or Playstation Eye (Villaroman et al, 2011), or the Xtion PRO Live mentioned in the literature review for a paper by Armin, Mehrana and Fatemah (2013) In Villaromans' case, the two choices mentioned are the opposite of the Kinect, as they require the user to hold a device (either a Wiimote, or a specialised Playstation controller with a coloured ball on top). Armin et als choice, and the Primesense Carmine, both offer equivalent specifications to the Kinect version one, with the same 640x480 depth resolution. Litomisky did a comparison of these cameras in 2012, which concluded that the Kinect was the recommended choice, based on price and resolution. It should be noted that he is talking about the Kinect version one, the second version offered the considerable improvements outlined above.

The Nintendo Wii should be mentioned, for its ability to track objects with a high degree of fidelity, as long as reflective tape is employed. While this moves closer to glove-based tracking than skeletal tracking, the Wii is notable for its controllers having a 1024x768 camera with built-in "blob" tracking, i.e. the systems tracks the object the reflective piece is attached to, as a complete object. Johnny Lee (johnnylee.net/projects/wii) completed several novel approaches to tracking, using the Wii, before he moved to work on the Kinect at Microsoft.

Conclusion

This chapter was written to give justification for the Kinect as an appropriate research tool. Many of the authors referenced in the literature review completed their research using a Kinect, but, rather than simply copying them, it was felt that a proper investigation should be made into *why* the various authors chose the sensor. In the “competing sensor systems” section, the million unit selling Leap Motion was shown to have been noted to lose resolution quickly as the tracked body moved away from the unit.

The competing systems noted by researchers, such as the Primesense Carmine, were all at least equivalent to the Kinect version one. Some updated systems offer equivalent resolution to the Kinect v2, but at a financial cost in the thousands of dollars. For this reason, the Kinect offers a viable sensor that will allow comparisons to be made against other, much more expensive systems.

Research Methodology

Introduction

This chapter will outline the process in which the “own research” section of the thesis shall be run. The chapter shall define the research objectives, and the methods employed, as well as the limitations of the research methods, before ending with conclusions. The results of this research shall be explored in a later chapter.

Research Objective

The aim of this research was to examine the potential of using skeletal and gesture tracking while making a presentation in front of a group. The research attempts to discover whether a hand-held remote, or standing within reach of the computer, can be replaced by gesture control.

In carrying out the research, the following questions were asked:

- Can gesture control replace a hand held control when making a presentation?
- Where is the optimum placing of a sensor, to track a user making a presentation?
- Does the user feel comfortable with the controls, and are the gestures intuitive enough that they can be made instinctively, rather than pausing for consideration each time?

Review of the Research Method

Quantitative Research

Quantitative research assumes that facts have an objective reality, and variables can always be defined and measured (Siegle, 2008). Its' purpose is to generalise and make predictions across a much wider scale than that used in the study.

This research approach centres on statistical techniques, forming hypotheses and supporting them with the gathered data. The hypotheses formulated can then be applied to the entire population of interest, rather than just the subset used in the study. In this method, large numbers of subjects are generally used. A study using this method would use, for example, surveys with number-based feedback, or true/false categories.

Qualitative Research

Qualitative research focuses less on the sharply-defined variables, and more on the "soft science" areas, of opinions and feelings. A study using this method would, in contrast to quantitative methods, use free-form feedback to gather subject opinions. Of course, the opinions in this free-form method can then be gathered and analysed to form a pseudo-quantitative outcome, but the key point is that the hypotheses are not formed before the data is gathered, but rather the data is used to form the results later.

In this method, the aim is to form a complete, detailed description, generally using detailed interviews or observation.

Choice of Research Method

Due to the nature of the questions posed above, it was best to use qualitative research to explore the research area. The chosen method allows understanding into human opinions and behaviour, which is highly necessary when considering the thesis topic, namely that gesture control can improve presentations. This measure of "improving" the presentation method needs to take into account user opinions and feedback, based on their interactions with the application.

The word "improve" in this case is subjective, as it is opinion based. For this reason, there is a need to consult subject-matter experts when performing the research, and to have these people, who are familiar with presenting in front of groups in different contexts (business, formal, informal, lecture) test the actual application and deliver their feedback

Detailed Description of Research Method

This section will outline the actual approach taken, in building an application and having users test it.

Application choice and setup

The decision had already been made above on the choice of sensor to program against. From this, it naturally followed to use the Software Development Kit (SDK) built specifically for it. The chosen SDK was version 2.0, available directly from Microsoft. The sensor was a Kinect for Windows, version 2, as outlined in a previous chapter. The application was developed in c#, on a Windows 8.1 computer. For the experimental setup components, the computer was connected to a projector, as well as the Kinect sensor. The following sections include diagrams of the room setups for the experiments.

The Microsoft SDK was chosen, in contrast to the OpenNI open source SDK, as the Microsoft offering has been specifically designed for the Kinect sensor, while OpenNI targets Structure sensors, which are a broad range of tools which are PrimeSense compatible. The specific targeting of the sensor by Microsoft makes it easier to develop for, especially when using the object-orientated programming language c#.

The application runs as a command-line service, with no specific UI, as the purpose is to aide in input to other applications – on its' own, the application simply interprets input. It is the interpretation, and passing the interpreted commands to the currently selected application (PowerPoint, for example) which forms the complete setup.

In defining gestures and tracking motions, the decision was taken to pass the gestures to other applications as standard commands. It will be seen in the first section that "swipe" gestures are used- these map to Page Up and Page Down commands. In the second section, the user directly controls the mouse pointer – and their "press" is passed as a left button click.

The application has two key features. In the first section, it acts as a service to interpret user gestures while they make a presentation (Gesture tracking). The second section allows the user to directly control the mouse cursor, interacting with the computer using their hand (Skeletal Tracking). The two sections will be outlined below.

Component One – Gesture Tracking

The first component tracks a user as they make a presentation to a room. The user faces the audience, with a Kinect between the presenter and the audience. The setup was initially as outlined in the diagram below:

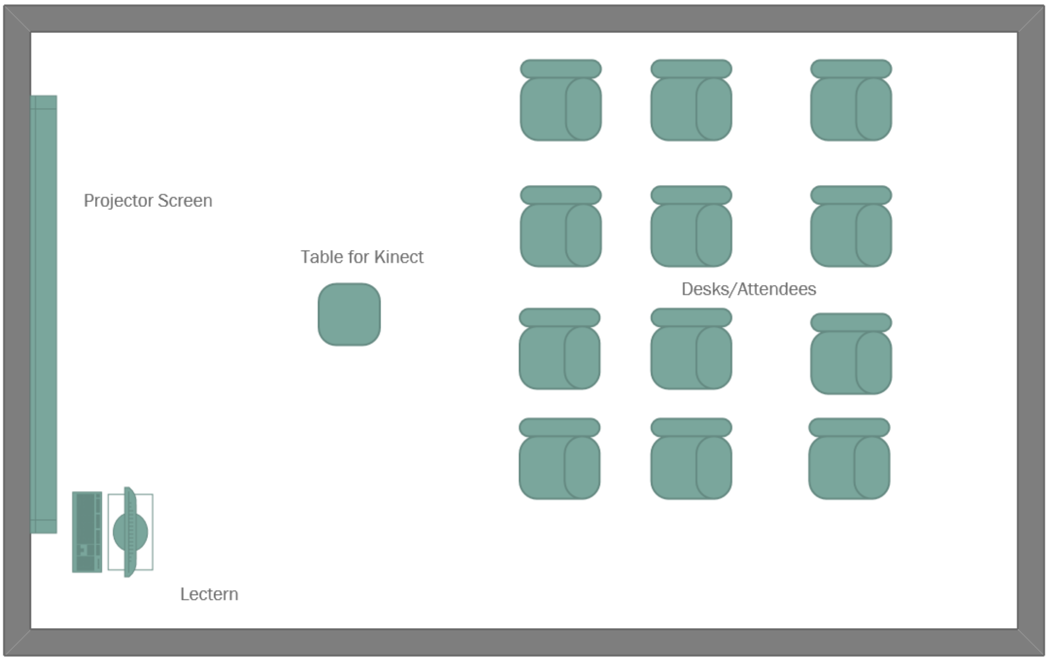

Figure 5 - Initial Presentation setup including kinect

This worked, but had difficulties registering some swipe motions, due to occlusion and wrong angles. The user was expected to stand in the centre, but this was not a natural place to stand for the lecturers and presenters. So, for the user setups after this, they were changed to accommodate either left- or right- handed people:

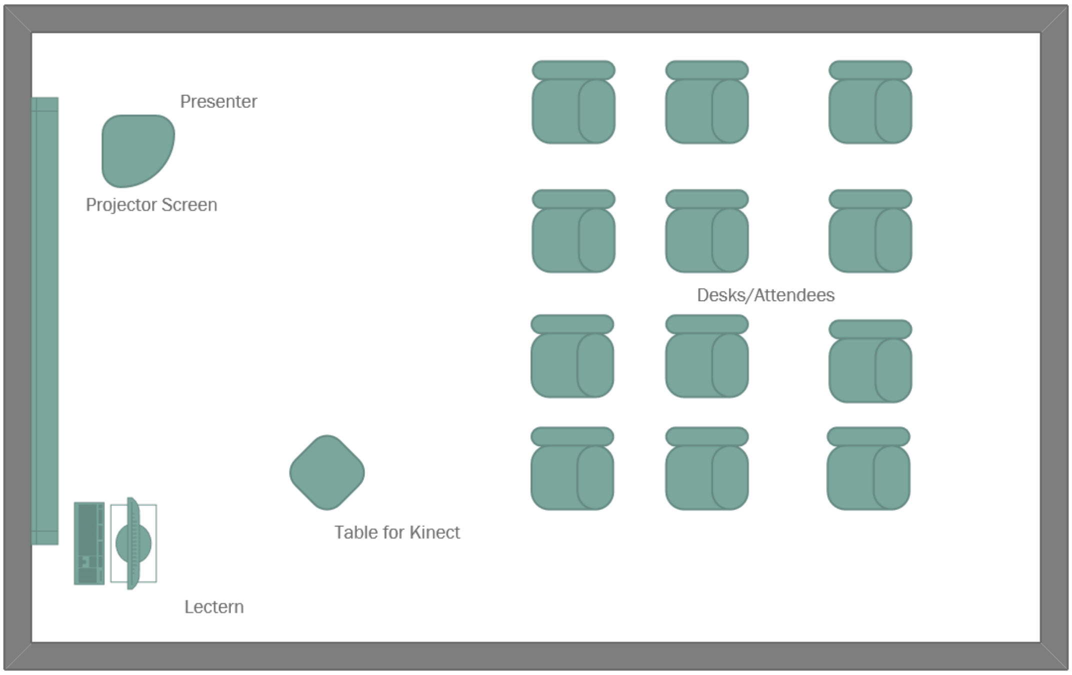

Figure 6 - Presentation setup for left-handed presenters.

In this configuration, the Kinect is looking at a presenter who can turn their head to reference their slides, which was reported as a more natural way to speak. The above picture shows the left handed approach, so the user moves their left hand to signify a "swipe" motion. In the below diagram, the same setup is flipped, for right handed users. This simple change from the first diagram, with the Kinect perpendicular to the projector screen, allows the user to naturally face the sensor and the audience, while also being able to keep track of the slides without drastic turning motions.

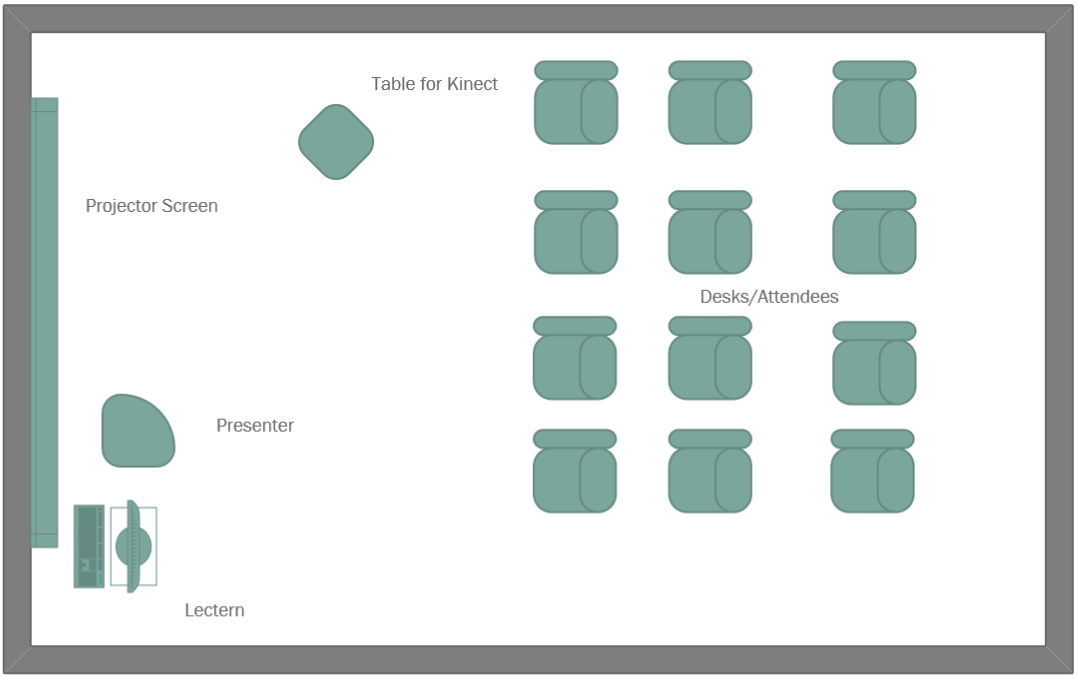

Figure 7 - Presentation setup for right handed users

The application itself focused on the most basic of presentation motions- advancing or reversing the slides, using a “swipe” motion, with the user moving their dominant hand across their body in a sweeping motion. The simplicity of the motion was chosen to replicate the most common peripheral used by presenters – the wireless presentation remote.

Figure 8 - Wireless Presentation Remote, www.kensington.com

Notice that, despite other manufacturers' attempts to add extra controls, this device simply offers forward/backwards buttons, a "stop" button to cease the presentation, and as an option, a laser pointer. It is the first three that are focussed on in the application.

Forwards and backwards controls are modelled, as mentioned, using the sweep motion. There then needed to be included a gesture to begin/end the presentation. After consulting with the users, this was created as a gesture with both hands raised in front of the body, palms towards the sensor. Users made this motion to begin the presentation (if they had not already started it while at the lectern), and made it again at the end to exit from the presentation.

Component Two – Skeletal Tracking

The second component of the application is more involved, as it continuously tracks the position of the users' hand, rather than waiting for a gesture. As shown in the diagram below, the configuration of the room changes slightly too, as the presenter is facing the projector screen directly, rather than facing the attendees. To accommodate this, the Kinect is placed on a table underneath the screen, facing the user. The user indicates their dominant hand by raising it above their head, and the application gives a pop-up bubble in the bottom right corner to indicate that it has registered the hand selection.

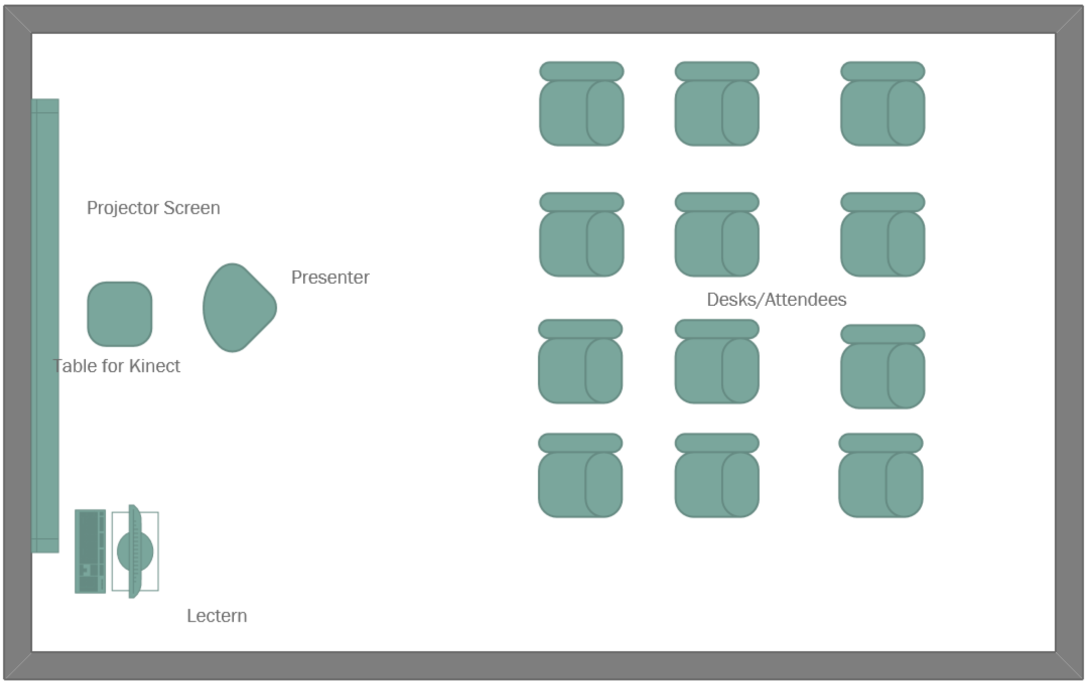

Figure 9 - Room setup for skeletal tracking

Once the user has completed this step, the mouse cursor then follows the users hand motion. The points tracked by the Kinect are returned as 3D- points for the hand:

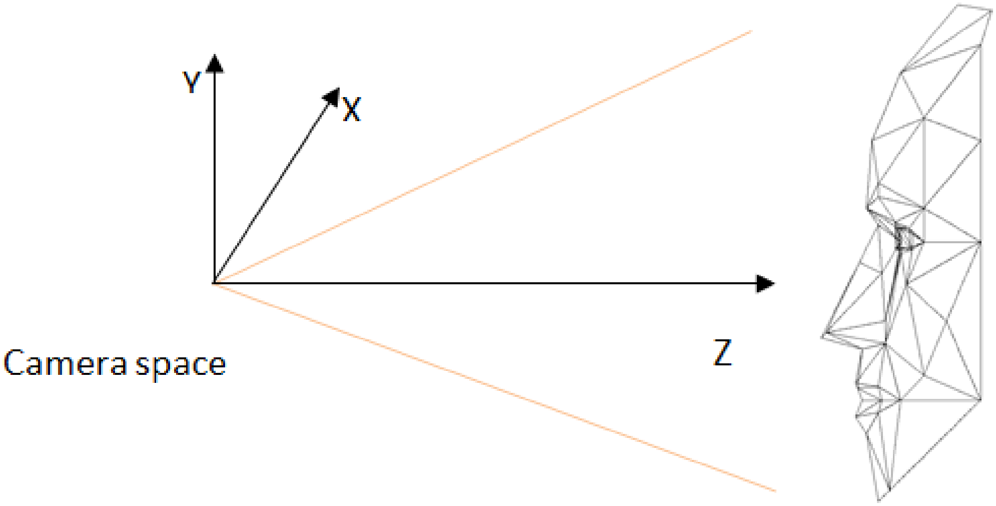

Figure 10 - Kinect tracked points explanation - www.microsoft.com

These points were then translated using the Kinect Mouse Cursor project files, an interpretative layer provided by Microsoft (2012). The user could “push” towards the screen as a button press, which, in an application such as a game or using a full-screen map (for example, in a classroom environment, with the student as user), allows the user to “grab” the screen and move it. This gesture acts as a left mouse button move. The press was interpreted as a change in the z-axis shown in the diagram above, while keeping the x and y values relatively constant. The word “relatively” is used as the users arm will naturally shake, especially if the arm is raised over the head. Once the user brings their hand back along the z axis again, the “press” is completed, and the click released.

Users were asked to use Google Maps in full-screen on the projector, and use skeletal tracking to move the map onscreen. Differing from some of the studies in the literature review, the users had each gesture demonstrated first, before they were asked to duplicate it.

Users

For the tests run on this experimental setup, it was felt that using a technique similar to that followed by Hespanhol et al (2012) should be followed – a narrow group of users who could each spend time getting to know the system, before making a five-minute presentation of their choice, using the PechaKucha method (www.pechakucha.org) of 20 slides, with 20 seconds of speaking for each slide.

For these tasks, the narrow group consisted of five university lecturers, five users from the authors’ workplace who had experience with presentations, and five users who did not have presentation or lecture experience.

These users were each asked to read through a supplied sample presentation, and then to present it in front of an empty room. This was due to initial user feedback of being wary of making mistakes when learning the system.

Data Collection

The users were asked to fill out a short survey once they had finished their presentation. The target time for the 20 slide presentation is 6 minutes, 40 seconds, so the users were also timed, to measure whether or not the gesture controls had an impact on the time taken to complete the task. The survey questions are given here:

1. User type: Lecturer / Presenter / Neither
2. Have you ever used gesture controls before? (Yes/No)
3. Have you made presentations before? (Yes/No)
4. For the first application, did you find the “swipe” gesture easy to perform? (Yes/No)
5. In that same application, did you find the start/stop gesture to be easy to perform? (Yes/No)
6. Did you have trouble with the gestures? (Yes/No , with space for freeform comment)
7. Any other comments? (Freeform comments)

The users were then asked to use the skeletal tracking component of the application, to pan a google map view along a defined route (a section of Irish geography laid out as a road map). The users were then given a second short survey:

1. User type: Lecturer / Presenter / Neither
2. Have you worked with gestures before? (Yes/No)
3. Have you used any kind of skeletal tracking before? (Yes/No)
4. For the second application, was the “push” gesture easy to perform? (Yes/No)
5. Did you find it easy to move around the map? (Yes/No)
6. Did you have any trouble with the gestures? (Yes/No with space for freeform comment)
7. Any other comments (Freeform comments)

The results and charts from these surveys shall be outlined in the next chapter.

Limitations

There are several limitations in the research methods outlined above. For completeness, they need to be outlined here, before any results are presented. This really serves as an initial look at results, as it gives some of the learnings raised before and after the research was conducted.

- The first, and most obvious limitation, is the small number of users for the tests. This is due to the nature of the research method. The experiment needed to be configured in a room with limited availability, and each user needed time to become familiar with the system. Due to a third (five) of the users not having made presentations before, the decision was made to have them present only to the author, rather than attempt the presentations to a classroom or group of their peers.
- The second limitation is in not using another sensor as a comparison against the Kinect. Since this is the same limitation of studies such as Osunkoya (2013), Guna (2014), and Villaroman (2011), it was felt that the results can be presented in the same atmosphere. Dutta (2011) evaluated the first version of the Kinect against far more expensive systems, with favourable results. For this reason, the Kinect should serve as a suitable proxy for other sensors.
- The number of gestures is restricted in this application. The first component uses swipe gestures (left/right), and a hands raised pose for beginning/ending the presentation. The second uses only one gesture – the “push” to simulate a button press, for click and drag and click controls. While this may seem like a small subset of gestures, consider that they map the controls needed to navigate most applications (with the exception of those that require right-click), as well as the controls for either presentations, or navigating image

galleries (left/right swipes). Hespanhol (2012) used only five gestures in their study, with the caveat that they did not teach the gestures to their test subjects, and instead asked them to be discovered.

Conclusions

Building the application itself was probably the easiest part of this project. As noted with the office layout diagrams, there was an extra learning curve for the author, in finding the correct way to keep a user tracked correctly. The software is unremarkable, as already noted, in that it's a console application. In later versions it will be reduced further to a service running in the taskbar – similar to the Synaptics mouse drivers many Windows users are already familiar with. The application needs to be abstracted to such an extent that it forms an invisible interface layer, in the same way that a touch screen, mouse or keyboard are now – the user simply instinctively reaches for the control.

This chapter outlined the application, and the experimental setup. It is in subsequent chapters that we discuss the findings, and then present them in the context of the literature review.

Results and Findings

Introduction

This chapter outlines the results of the surveys gathered from the users who participated in testing both application components. The results are presented as charts, organised by question.

Users

The users consisted of 5 University lecturers, 5 Project Managers, and 5 users from other disciplines. The first two user groups all have substantial experience with making presentations and lectures. The final user group were chosen as they had no prior experience with making presentations.

While the user group is small, the ten who were experienced in presentations were essentially subject matter experts – all had extensive presentation experience in front of small and large groups, and all were comfortable with technology. An original question on the survey was "Do you feel comfortable being tracked by a sensor". These users either laughed or pointed to the webcams on their laptops, explaining that there was little difference. The final group, who did not have presentation experience, generally had the same reaction. One user gave feedback in the freeform comment area, which will be mentioned there.

Survey Results

Have you worked with gesture control before?

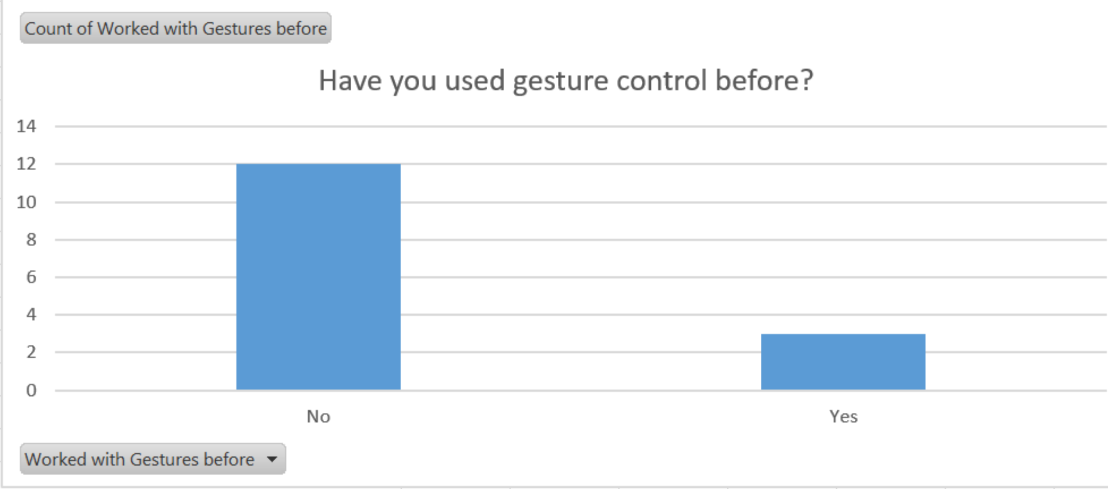

Table 1 - Have you worked with gesture control before?

Of the 15 users, only 3 had even passing experience with gesture controls. Two were from using the Kinect for gaming on the Xbox console, only one (a university lecturer) had used gesture controls outside of games. This is significant in two ways.

Firstly, even within the lecturer group, composed of five *Computer Science* lecturers, only one had used gesture controls. This shows how specialised the research area is even now. The lecturer who had used the controls had used them for a student's dissertation, rather than working with them directly himself.

Secondly, the two users who had used the controls for gaming reported that (in spite of the Kinect, Wii, and PlayStation Eye) the games were easier to play with a physical controller, unless specifically built for the gesture input. The games they reported to have used the Kinect well, were sports and fitness games – the same game type that made the Wii so popular.

Have you made presentations before?

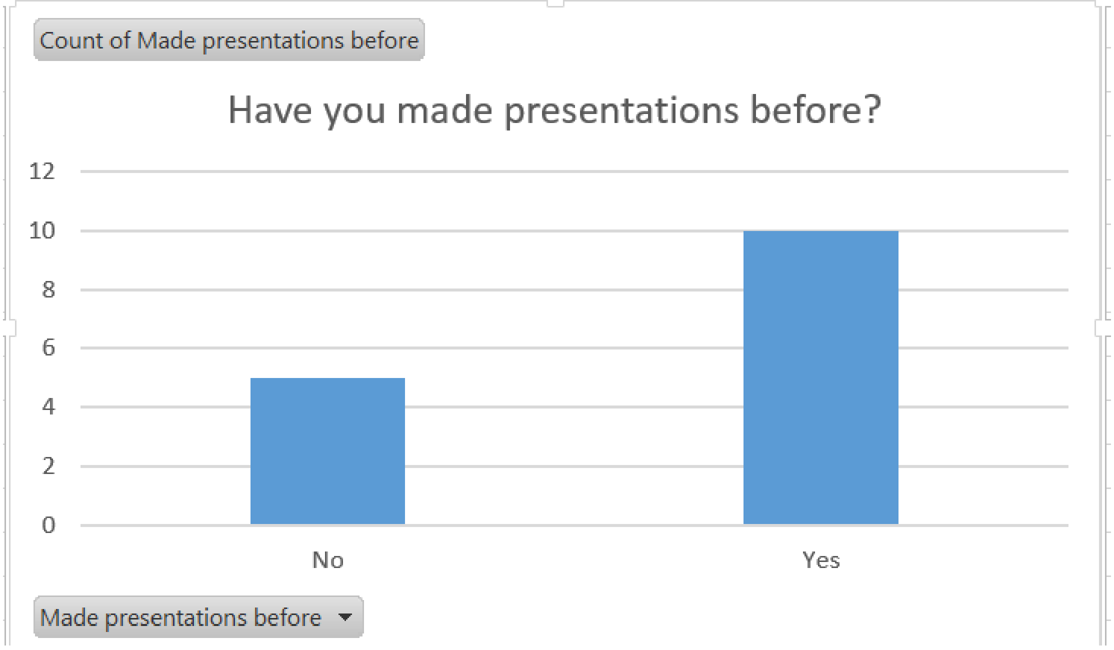

Table 2 - Have you made presentations before?

As specified earlier, of the fifteen users, ten had prior experience with one-to-many presentations, and five did not. In the ten who had, five were university lecturers who made presentations via PowerPoint several times per day, and the other five made frequent presentations in office environments.

Did you find the swipe gestures easy to perform?

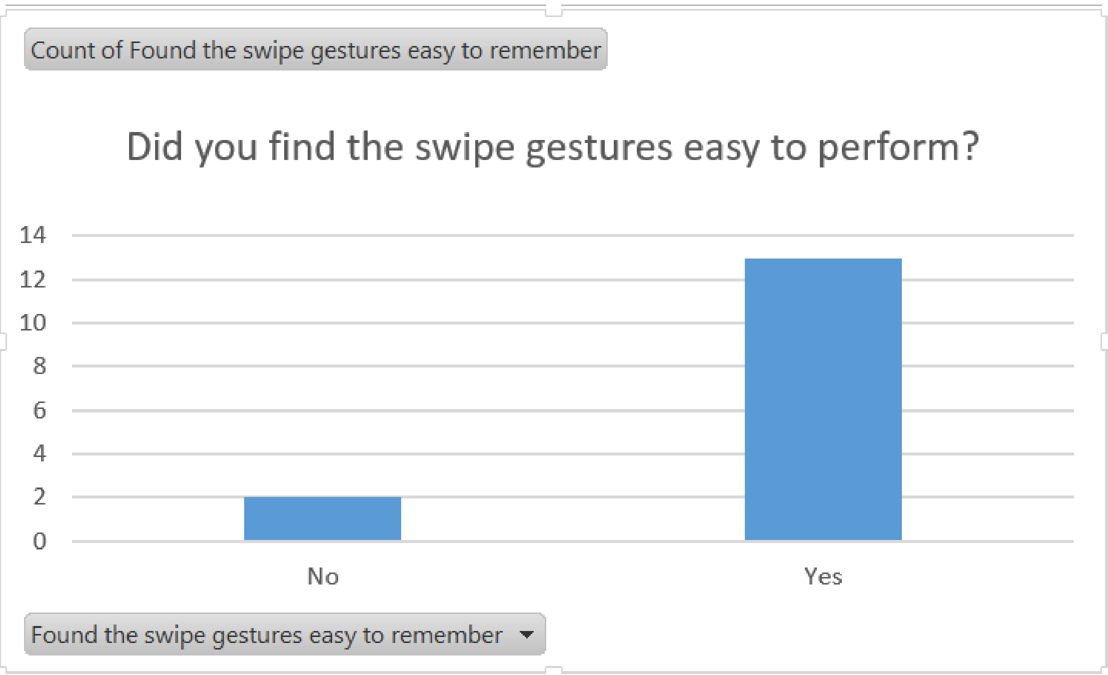

Table 3 - Did you find the swipe gestures easy to perform?

Only two users reported large difficulties with the swipe gesture – but, as can be seen later, several other users had valid comments and criticisms. The two users who had significant trouble (generally needing two or three attempts for most valid inputs) were making the swipe gesture diagonally, rather than horizontally. This should be taken into account when defining a gesture- it's not just the horizontal component that may change, but the vertical. Obviously there are limits to how much error can be accounted for, but this error rate would suggest a need for more flexibility.

Did you find the start/stop gestures easy to perform?

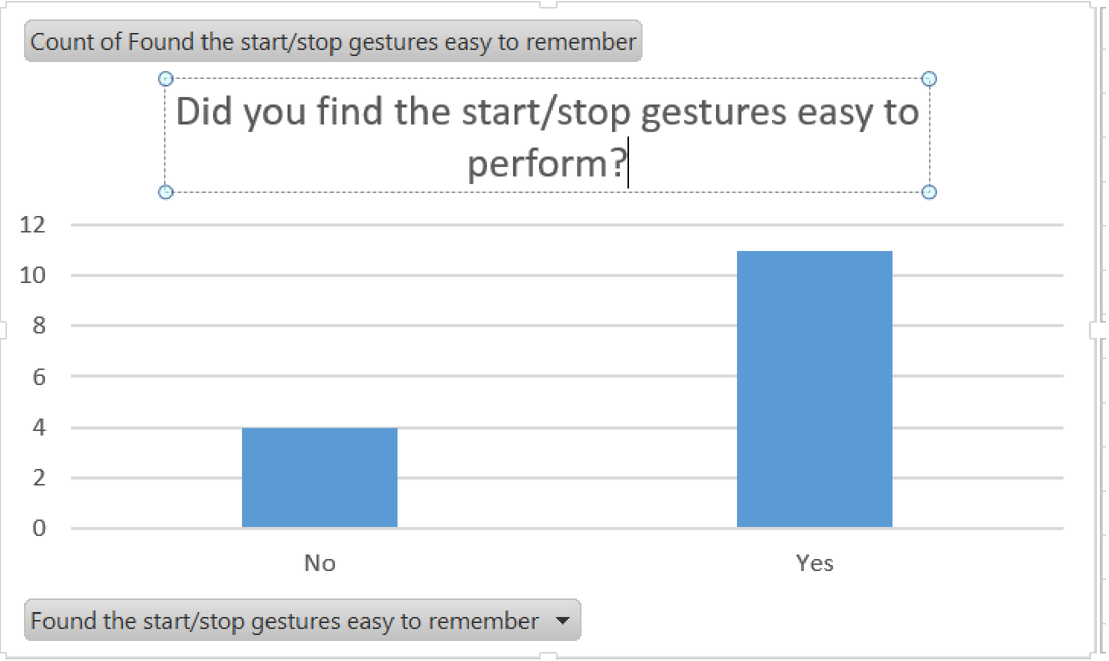

Table 4 - Did you find the start/stop gesture easy to perform?

Four of the users reported real difficulty with the start/stop gesture. As outlined earlier, this was a pose with both hands raised to chest height, palms towards the Kinect. While 11 of the users found it easy to perform, several others questioned the necessity of the motion. Their comments will be outlined later in the feedback section, but the gesture (or at least its intent) was generally seen as superfluous, even by the users who found it easy.

Did you find that gestures made it easier to present?

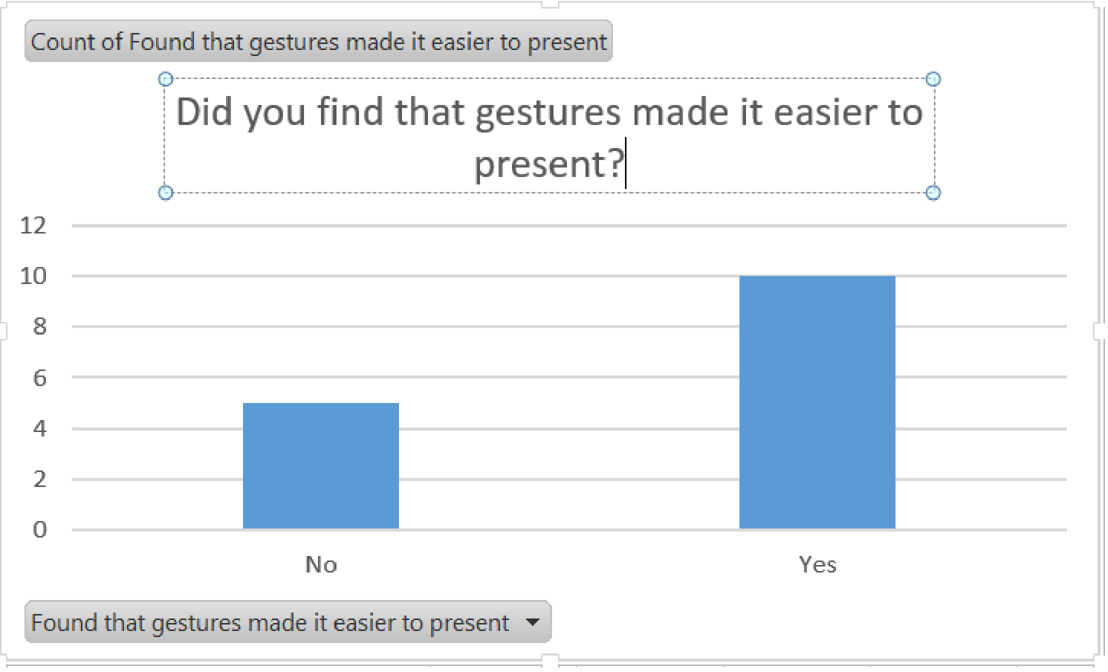

Table 5 - Did you find that gestures made it easier to present?

A full third of the users reported that the gestures did not make it significantly easier to present. Two had difficulties with the swipe gestures from the beginning, swiping on a diagonal rather than horizontal plane, and taking 3-4 minutes longer to finish the presentation than the other users.The other two-thirds of the users gave more positive feedback. Some gave contrastive criticism of the gestures, and many rated the experience as fun or enjoyable.

This is the most significant result of the study – two thirds of the (admittedly small, though experienced) group of users reported an improvement, whether in terms of technical speed or simply not having to move back and forth to the computer. Of the ten users who had experience with presentations, three were users of wireless presentation remotes, and these were the same users who gave a negative answer to this question. These same users, and the two inexperienced presenters, were also the same users with "noisy" body language – they gestured frequently while speaking, and moved around the stage while presenting. This introduced errors to the inputs, and also meant that they occasionally were facing away from the sensor while speaking.

Did you find the registration (hand raised) gesture easy to perform?

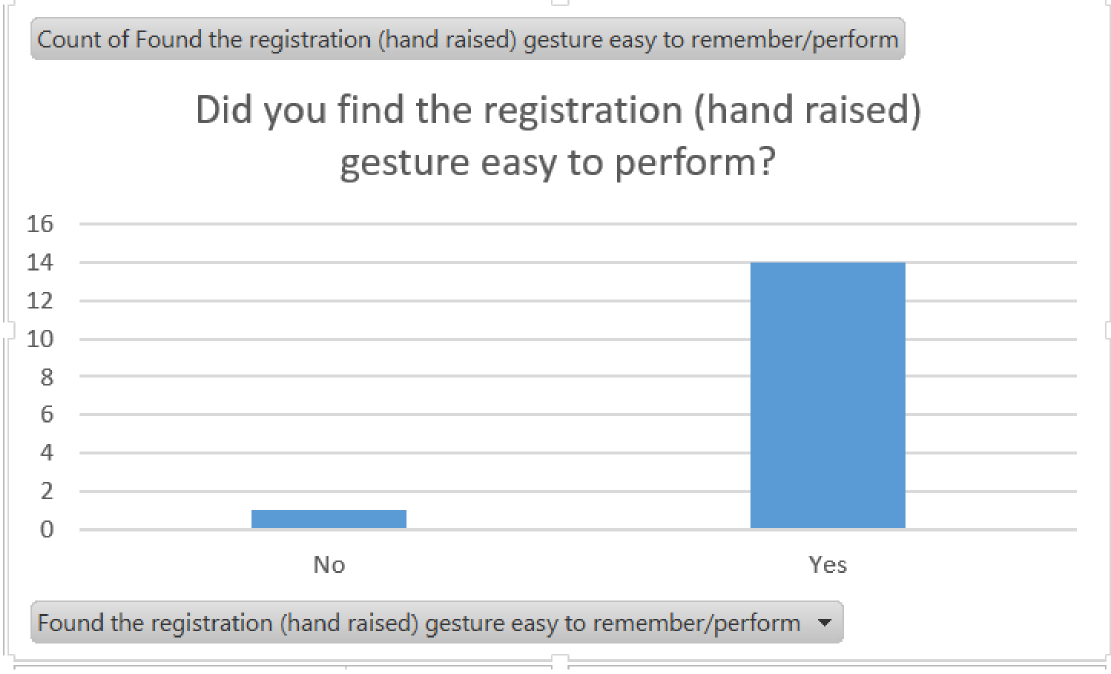

Table 6 - Did you find the registration gesture easy to perform?

Only one user found this gesture difficult. As noted in the Research Methodology chapter, users were shown the gestures and allowed to practice for the presentation segment. For this section, they were shown the relevant gestures and asked to follow the map. This user forgot the registration gesture (left or right hand raised above shoulder level), so they took two minutes longer to get the map moving, before the system essentially accidentally recognised which hand they wanted to use, when they kept that hand above shoulder level waving in frustration.

This gesture was the easiest of those defined in the application – a pose with one hand raised. Unlike the stop/start pose defined for the presentation component, users could see the purpose of the gesture, in telling the system which hand they intended to use.

Did you find it easy to move around the map using your hand?

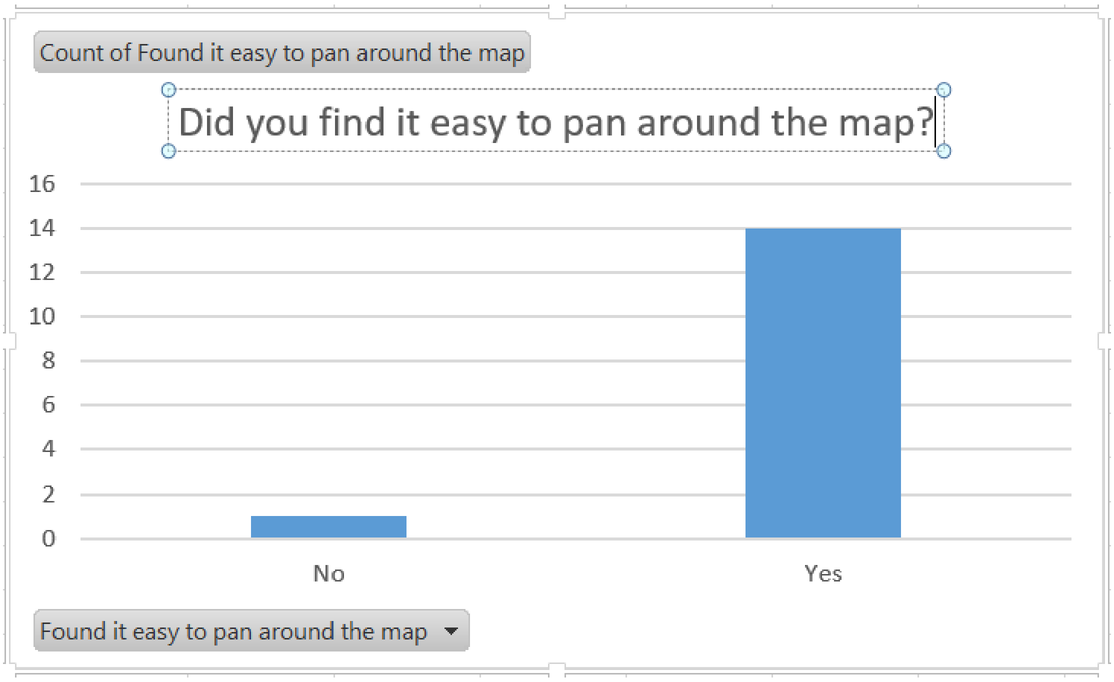

Table 7 - Did you find it easy to pan around the map?

Only one user found it difficult to use the mouse cursor with skeletal tracking. This was from the same user who had difficulties with the registration gesture, and the user simply found the experience to be frustrating, rather than involved or enjoyable. However, the majority (14 of 15) users found this method to be intuitive and easier than the “swipe” used earlier. The reason reported for this opinion was that they could see the mouse cursor moving on screen, following their hands, which gave them instant feedback.

Did you find the “push” gesture easy to perform?

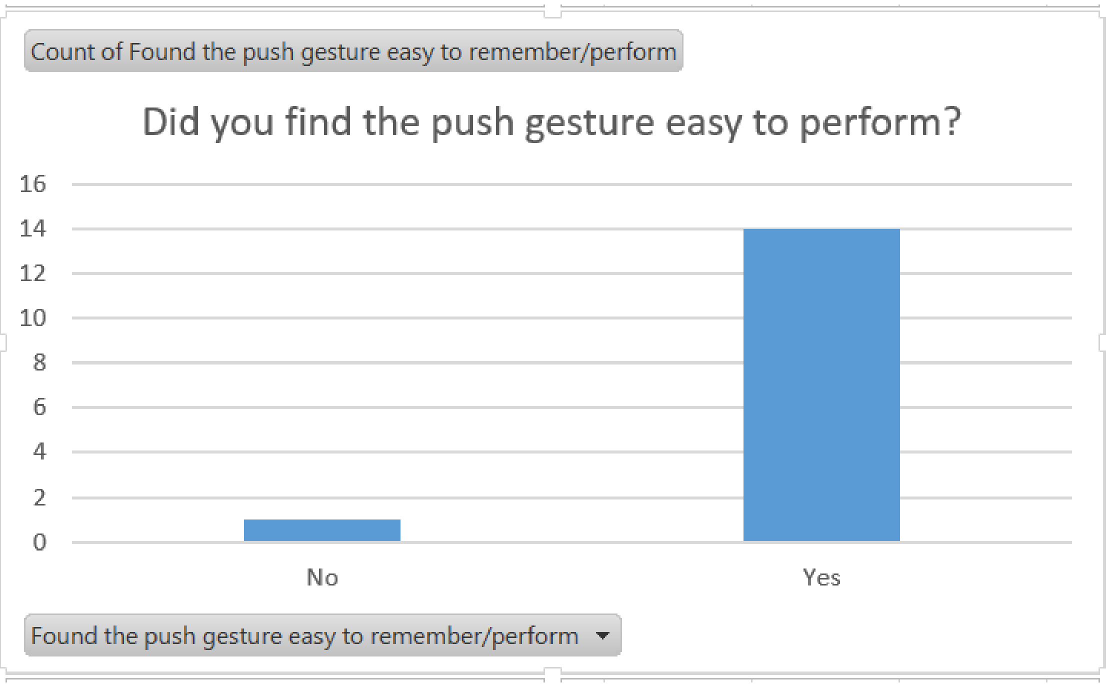

Table 8 - Did you find the push gesture easy to perform?

Again, the user who had issues with registration reported issues with this gesture. Other users left comments in the freeform section, which will be mentioned lately, but this was overall a recognisable gesture to the users, and was quickly reproduced when they attempted it. For users who did not have difficulty with the registration gesture, there was a variance of only 20 seconds (about 11%) between the fastest and slowest user, when completing the task on the map.

User Feedback

The first piece of user feedback influenced the entire testing process, and directly answered one of the research questions; the placement of the sensor. As seen in the diagrams in the Research chapter, the Kinect was originally placed perpendicular to the projector screen, facing the screen. This meant that the Kinect would best track the user only when they were standing in front of their slide presentation, rather than off to one side, where they could turn their head to see the slides or their audience. The users who initially saw the experimental setup suggested a change of Kinect position, which directly influenced the results.

In addition to the binary yes/no questions on the survey, users were also given two areas with freeform answers allowed, for each section of the application.

Did you have trouble getting your gestures recognised?

Seven of the fifteen users answered "yes" to this, with comments ranging from "twice, near the start", through "a few times", up to "almost every slide". The most troubling comment was that using natural body language while speaking caused this: "moving my hands made the slides move unintentionally". The worst outcome for gesture tracking is the unintentional gesture, and these will need to be worked on, to ensure that only deliberate gestures are tracked.

The "almost every slide" comment was from the same user who reported difficulty with the stop/start gesture. It identifies an area for further research, in discovering why some users have far more difficulty than others. One question asked in the survey as a "soft" question was the notion of being uncomfortable while tracked. This user was the only one in the group to answer affirmatively. It may be that, if a user is uncomfortable with the tracking, that there is a hesitation to their gestures not present in other users. Smaller, less obtrusive cameras may help here.

The surveys are anonymous, but the pair of surveys were both linked by user – so it can be related that the same user who had concerns on being tracked, felt far more comfortable in the second test, where they faced the screen. In that configuration, the Kinect was under the projection screen, rather than in their direct line of vision.

Did you have trouble with the skeletal tracking test?

In contrast to the above comments, five users reported negatively on the skeletal tracking piece. Of these, one only had difficulties at the top of the screen, and another commented on how quickly the mouse cursor moved. Since this is directly related to the speed at which the user moved, it suggests that practice may be the best solution in this instance.

In two of the users who simply answered “yes” to this question, with no further feedback, it was noticed that, when they were trying to “press” to click and drag, they were leaning their entire body forward, while keeping their hand in almost the same place. This interfered with the measure of the z-value change (towards/away from the sensor), and can be solved with user training.

Any additional comments

Thanks to the users being essentially subject-matter experts, there were some extremely helpful and relevant comments left in the surveys:

- One of the most important comments was that of transitioning more than one slide at a time. The user in question suggested a similar kind of “inertial scrolling” to that offered by trackpads or touch screen, wherein a faster gesture would jump more than one slide.
- There were several comments on the usability of the “swipe” gesture. Some users would have preferred both hands to have been used. In that case, their dominant hand would swipe to advance the slides, and their non-dominant hand would swipe to backtrack. This is a valid option to introduce.
- One issue that was mentioned was one fulfilled by the wireless remote- the laser pointer. As the users are standing to the side of their slides, they have no way to point to a specific area on the slide with a gesture. This will need to be investigated further, to find a suitable replacement.

- For the second task, in moving the mouse cursor using skeletal tracking, users found it tiring to control the pointer when it was at the top of the screen. This is due to the scaling performed by the code, and can be adjusted so that the hand does not to be raised as far when moving to the top of the screen.
- Two users commented on the mouse moving if they turn away from the screen. Until the point where the hand is occluded by the body, the Kinect is still tracking the user. The comment raises the possible need of a "stop" gesture – perhaps by raising the other hand in a "stop" motion – to pause the tracking if the user requires it.
- The "push" gesture was seen by three users as taking too long. They raised the possibility of a "grab" motion (by closing the hand) instead to click and drag the screen, rather than having to push forward, which they felt signified a button press.

Conclusion

Overall, despite some issues that users had, the feedback was positive, with a 60-70% positive feedback result from the two application components. The presentation gesture system was regarded favourably, though some key points were raised by users around the strictness of the gestures, which shall be looked at further in future work. #

In the next chapter, the results will be discussed in the context of the existing work covered in the literature review, to find parallels with other research, and show where this research has made advances.

Discussion

Introduction

In this chapter, the results and findings presented above, are discussed in the context of the existing literature. Once more, the research questions are visited and compared with the results, to confirm that they have indeed been answered.

Research Questions

- Can gesture control replace a hand held control when making a presentation?
- Where is the optimum placing of a sensor, to track a user making a presentation?
- Does the user feel comfortable with the controls, and are the gestures intuitive enough that they can be made instinctively, rather than pausing for consideration each time?

Evaluation of Findings

This section compares the three research questions to the results, to assert whether or not they have been answered in the research component. For clarity, the questions shall be answered in order:

Can gesture control replace a hand held control when making a presentation?

With a positive response of 65 percent, the gesture system can definitely replace a hand held control. The issue came down to one user wanting a laser pointer, which could be simulated by using the skeletal tracking component to move the mouse pointer to the required location onscreen.

The second issue users had was with the question of gesture consistency. Norman (2010) raised this same complaint when discussing gesture controls. In this instance, the horizontal “swipe” gesture needs to be more forgiving of gestures which are within a few degrees of horizontal. Since a 45 degree gesture could be an entirely different input, a certain amount of leeway is allowed, but within reason.

Where is the optimum placing of a sensor, to track a user making a presentation?

This question was answered in the first instance by the users. Osunkoya (2013) did work on Kinect with PowerPoint controls, but failed to answer this question adequately, only covering the vision parameters of the Kinect. From the users, it was found to be more comfortable to place the Kinect at an angle to the projection screen, allowing the users to keep their attention between the audience and the slides, rather than worrying about keeping in view of the sensor.

The users who have what could be considered “noisy” body language had the most difficulty with gestures, as some of their more emphatic gestures were registered as inputs. Chaudhary (2011) made this point when he commented that a gesture is a natural part of language. The Kinect needs to be forgiving, so the optimum placement is key to avoiding occluding.

In the second scenario, the users faced the projection screen, and the Kinect was under the screen, facing the user. This was found to be the best location to track the user, as they were essentially forced to face the screen to accomplish their task, so they were continuously facing the Kinect in the best aspect.

Does the user feel comfortable with the controls, and are the gestures intuitive enough that they can be made instinctively, rather than pausing for consideration each time?

This is a more difficult question to answer, despite the research. Users felt comfortable, as noted by the 65 percent approval rate for their initial experience with the system. However, it was noted by the author, and in some of the surveys, that some of the gestures need to be altered.

Hespanhol (2012) came to the same conclusion with one of their five gestures. Anything that causes a user to stop and think, can be a barrier to adoption of gesture input. In their case, it was the anticlockwise circle drawn with a hand. In this papers case, it was the “start/stop” gesture with both hands raised. Users did not see the need for this gesture when they could simply start the presentation while they were at the podium, and then use the swipe gestures afterwards. This “start/stop” gesture was the one which faced the most resistance, and indeed caused the most trouble when users were beginning or ending their presentation.

The other gesture which needs to be adjusted was the “press”. Making the same mistake as Sturman (1994) with their glove-based input, the wrong gesture was mapped to the idea of “click and drag”. The author made the assumption that the “click” and dragging of a mouse, should be the same when using a gesture; namely, that the user would “press” and drag. User feedback corrected this assumption with the comment that a “grasp” motion was much preferable for click-and-drag, that the motion felt more appropriate.

Overall comments

The above research outcomes directly tie into comments given by several existing authors, most notably Wigdor (2011) and Villaroman (2011). Both of those authors worked with users and gesture control, and came to the conclusion that there needs to be more flexibility in the design of the individual gestures.

A component of the user feedback, supported by Armin (2013) is the concept of giving feedback to the user. In the skeletal tracking component of the research, the users were asked to navigate around a map using press and drag gestures. Users could not tell when they had successfully “clicked” before they could drag – which led to some extremely drastic movements on screen, rather than the fluid motions they expected. A tone to signify the click, or even an icon similar to that used by the xbox system, could help with this. In the xbox approach, the icon on screen changes to a closed hand when users are selecting an option. When shown videos of this, users responded favourably to the potential change.

Once more in this area, Hespanhol (2012) has valid research, in terms of gestures needing to be fluid. Norman (2012) made the same comment when criticising natural gestures, arguing that a gesture that was ephemeral, rather than the atomic gestures required for accurate gesture control. Norman makes a second comment which supports the need for adjustable, user-configurable gestures; namely, the localisation problem. A gesture in one country can have complete different intent in another, so gestures will need to be updated on a localisation basis, similarly to languages localisation.

In summary, the research is backed by the reviewed literature, and presents a valid case for using gestures with presentations. There are key questions which will need to be asked before this can be considered a full success though, and these will be outlined in the final chapter.

Conclusion

The field of HCI is rich with examples of alternative input methods to the traditional mouse and keyboard. Touch input made advances once the appropriate input devices were created and gesture input has the potential to do the same, once the use cases are identified and validated. This paper outlined and validated the potential for gesture controls in one to many presentations. However, there are some questions and areas of further research which need to be explored:

- The Kinect sensor also includes a voice recognition component. Would it have been preferable to replace the “stop/start” gesture with a voice command? In an office-style presentation, this may work, but it may be more difficult in a busy lecture theatre.
- The method to configure gestures per user needs to be explored. If height, body shape, or voice/face recognition could be employed, then it remains to develop a user-friendly method to record and define gestures.
- Several users still reported being uncomfortable when being tracked by the sensor. This complaint was not repeated when the Kinect was under the projector screen, out of the direct field of vision, so it could be possible to shrink the sensor to make it less obvious and intrusive. This is more of a social problem than a technical one, as users quickly become accustomed to sensors facing them – witness the proliferation of webcams in laptops.
- There needs to be a method of feedback on screen, or through audio, for the system to acknowledge that a button has been pressed, or the screen has been grabbed. Much work has been done in providing feedback to touch-screen users, in the form of vibration or tone, even to the extent of raising items on the screen to create tactile feedback. No such method was used in this paper, and its absence was noted by the users.

Another point that was not explored in the paper, was the notion of fatigue. Two users mentioned it verbally, but not on the surveys, that in the skeletal tracking component, moving the mouse around with their arms over their heads was tiring. While it was simply humorous complaining, it is a part of gesture control that has some validity. After all, the first uses of the Microsoft Kinect were for exercise games on the Xbox, and, even though the games were far more energetic than we would expect users to be with a normal input method, there remains the problem of creating gestures which are *comfortable*. After all, it is impossible to *improve* our presentations if the user is uncomfortable.

References

- "Xbox-360-Kinect-Standalone" by Evan-Amos - Own work. Licensed under Public Domain via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Xbox-360-Kinect-Standalone.png#/media/File:Xbox-360-Kinect-Standalone.png
- "Xbox-One-Kinect" by Evan-Amos - Own work. Licensed under Public Domain via Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Xbox-One-Kinect.jpg#/media/File:Xbox-One-Kinect.jpg
- Armin, K., Mehrana, Z., & Fatemeh, D. (2013, February). Using Kinect in teaching children with hearing and visual impairment. In E-Learning and E-Teaching (ICELET), 2013 Fourth International Conference on (pp. 86-90). IEEE.
- Chaudhary, A., Raheja, J. L., Das, K., & Raheja, S. (2011). A survey on hand gesture recognition in context of soft computing. In Advanced Computing (pp. 46-55). Springer Berlin Heidelberg.
- Engelbart, D. C., & English, W. K. (1968, December). A research center for augmenting human intellect. In Proceedings of the December 9-11, 1968, fall joint computer conference, part I (pp. 395-410). ACM.
- Epps, J., Lichman, S., & Wu, M. (2006, April). A study of hand shape use in tabletop gesture interaction. In CHI'06 extended abstracts on human factors in computing systems (pp. 748-753). ACM.
- Garg, P., Aggarwal, N., & Sofat, S. (2009). Vision based hand gesture recognition. World Academy of Science, Engineering and Technology, 49(1), 972-977.
- Guna, J., Jakus, G., Pogačnik, M., Tomažič, S., & Sodnik, J. (2014). An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking. Sensors, 14(2), 3702-3720.

- Hespanhol, L., Tomitsch, M., Grace, K., Collins, A., & Kay, J. (2012, June).Investigating intuitiveness and effectiveness of gestures for free spatial interaction with large displays. In Proceedings of the 2012 International Symposium on Pervasive Displays (p. 6). ACM.
- Jacob, R. J. (1991). The use of eye movements in human-computer interaction techniques: what you look at is what you get. ACM Transactions on Information Systems (TOIS), 9(2), 152-169.
- Ju, S. X., Black, M. J., Minneman, S., & Kimber, D. (1998). Summarization of videotaped presentations: automatic analysis of motion and gesture. Circuits and Systems for Video Technology, IEEE Transactions on, 8(5), 686-696.
- Kammer, D., Wojdziak, J., Keck, M., Groh, R., & Taranko, S. (2010, November). Towards a formalization of multi-touch gestures. In ACM International Conference on Interactive Tabletops and Surfaces (pp. 49-58). ACM.
- Kaaresoja, T., Brown, L. M., & Linjama, J. (2006, July). Snap-Crackle-Pop: Tactile feedback for mobile touch screens. In Proceedings of Eurohaptics (Vol. 2006, pp. 565-566).
- Kenety, B., & Parker, J. (2013). Interpret this! Gesture recognition and HMI design.
- Kinect Mouse Cursor (2012). https://kinectmouse.codeplex.com/
- Leap Motion. http://www.leapmotion.com
- Lee, J. (2009). 3D vision using a wiimote
- Lee, S. K., Buxton, W., & Smith, K. C. (1985, April). A multi-touch three dimensional touch-sensitive tablet. In ACM SIGCHI Bulletin (Vol. 16, No. 4, pp. 21-25). ACM.
- Litomisky, K. (2012). Consumer RGB-D Cameras and their Applications.
- Malizia, A., & Bellucci, A. (2012). The artificiality of natural user interfaces. Communications of the ACM, 55(3), 36-38.

- Mankoff, K. D., & Russo, T. A. (2013). The Kinect: A low-cost, high-resolution, short-range 3D camera. Earth Surface Processes and Landforms, 38(9), 926-936.
- McArthur, V., Castellucci, S. J., & MacKenzie, I. S. (2009, July). An empirical comparison of Wiimote gun attachments for pointing tasks. In Proceedings of the 1st ACM SIGCHI symposium on Engineering interactive computing systems (pp. 203-208). ACM.
- Myers, B. A. (1998). A brief history of human-computer interaction technology. interactions, 5(2), 44-54.
- Nachimuthu, K., & Vijayakumari, G. (2011). Role of Educational Games improves meaningful learning. i-Manager's Journal of Educational Technology, 8(2), 25.
- Norman, D. A. (2010). Natural user interfaces are not natural. interactions, 17(3), 6-10.
- Norman, D. A., & Nielsen, J. (2010). Gestural interfaces: a step backward in usability. interactions, 17(5), 46-49.
- Oikonomidis, I., Kyriazis, N., & Argyros, A. A. (2011, August). Efficient model-based 3D tracking of hand articulations using Kinect. In BMVC (Vol. 1, No. 2, p. 3).
- Osunkoya, T., & Chern, J. C. (2013). s" Gesture-Based Human-Computer-Interaction using kinect for Windows Mouse Control and Power Point Presentation". Department of Mathematics and Computer Science, Chicago State University, Chicago, IL, 60628.
- Pavlovic, V., Sharma, R., & Huang, T. S. (1997). Visual interpretation of hand gestures for human-computer interaction: A review. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19(7), 677-695.
- Raheja, J. L., Shyam, R., Kumar, U., & Prasad, P. B. (2010, February). Real-time robotic hand control using hand gestures. In Machine Learning and Computing (ICMLC), 2010 Second International Conference on (pp. 12-16). IEEE.

- Ren, H., Liu, W., & Lim, A. (2014). Marker-based surgical instrument tracking using dual kinect sensors. Automation Science and Engineering, IEEE Transactions on, 11(3), 921-924.
- Shin, M.C., Tsap, L.V., Goldgof, D.B.: Gesture recognition using bezier curves for visualization navigation from registered 3-d data. Pattern Recognition 37(5), 1011–1024 (2004)
- Sturman, D. J., & Zeltzer, D. (1994). A survey of glove-based input. Computer Graphics and Applications, IEEE, 14(1), 30-39.
- Suma, E. A., Krum, D. M., Lange, B., Koenig, S., Rizzo, A., & Bolas, M. (2013). Adapting user interfaces for gestural interaction with the flexible action and articulated skeleton toolkit. Computers & Graphics, 37(3), 193-201.
- Villaroman, N., Rowe, D., & Swan, B. (2011, October). Teaching natural user interaction using OpenNI and the Microsoft Kinect sensor. In Proceedings of the 2011 conference on Information technology education (pp. 227-232). ACM.
- Weichert, F., Bachmann, D., Rudak, B., & Fisseler, D. (2013). Analysis of the accuracy and robustness of the leap motion controller. Sensors, 13(5), 6380-6393.
- Wigdor, D., & Wixon, D. (2011). Brave NUI world: designing natural user interfaces for touch and gesture. Elsevier.