Development and Evaluation of Facial Gesture Recognition and Head Tracking for Assistive Technologies

This Thesis Submitted to The University of Kent

For The Degree of *Doctor of Philosophy*

In *Electronic Engineering*

By

Shivanand Prabhoolall Guness

August 2015

I would like to dedicate this thesis to my loving parents ...

Declaration

I certify that I have read and understood the entry in the Project Handbook on Plagiarism and Duplication of Material, and that all material in this Thesis is my own work, except where I have indicated with appropriate references.

Shivanand Prabhoolall Guness

August 2015

Acknowledgements

I would like to thank my supervisors Dr. Deravi, Dr. Sirlantzis, Dr. Pepper and Dr.Sakel for all their help, support and encouragement during the course of this research. I would also like to thank all my lecturer and staff at the School of Engineering and Digital Arts for their help and support during my PhD course. I would also like to thank Richard Bonner, Asad Ali, Chang Li and Scott Yang for all their help and friendship during this course. I would also like to thank all for participating in my experiments. And I would like to add a special thanks for Andy Farrall for helping me turn my head tracker closer to becoming an assistive device. And finally, I would like to thank my parents and my sister for their love and support. These couple of years would have been impossible without them.

Abstract

Globally, the World Health Organisation estimates that there are about 1 billion people suffering from disabilities and the UK has about 10 million people suffering from neurological disabilities in particular. In extreme cases these individuals with disabilities such as Motor Neuron Disease (MND), Cerebral Palsy (CP) and Multiple Sclerosis (MS) may only be able to perform limited head movement, move their eyes or make facial gestures. The aim of this research is to investigate low-cost and reliable assistive devices using automatic gesture recognition systems that will enable the most severely disabled user to access electronic assistive technologies and communication devices thus enabling them to communicate with friends and relative.

The research presented in this thesis is concerned with the detection of head movements, eye movements, and facial gestures, through the analysis of video and depth images. The proposed system, using web cameras or a RGB-D sensor coupled with computer vision and pattern recognition techniques, will have to be able to detect the movement of the user and calibrate it to facilitate communication. The system will also provide the user with the functionality of choosing the sensor to be used i.e. the web camera or the RGB-D sensor, and the interaction or switching mechanism i.e. eye blink or eyebrows movement to use. This ability to system to enable the user to select according to the user's needs would make it easier on the users as they would not have to learn how to operating the same system as their condition changes.

This research aims to explore in particular the use of depth data for head movement based assistive devices and the usability of different gesture modalities as switching mechanisms. The proposed framework consists of a facial feature detection module, a head tracking module and a gesture recognition module. Techniques such as Haar-Cascade and skin detection were used to detect facial features such as the face, eyes and nose. The depth data from the RGB-D sensor was used to segment the area nearest to the sensor. Both the head tracking module and the gesture recognition module rely on the facial feature module as it provided data such as the location of the facial features. The head tracking module uses the facial feature data to calculate the centroid of the face, the distance to the sensor, the location of the eyes and the nose to detect head motion and translate it into pointer movement. The gesture detection module uses features such as the location of the eyes, the location of the pupil, the size of the pupil and calculates the interocular distance for the detection of blink or eyebrows movement to perform a click action. The research resulted in the creation of four assistive devices based on the combination of the sensors (Web Camera and RGB-D sensor) and facial gestures (Blink and Eyebrows movement): Webcam-Blink, Webcam-Eyebrows, Kinect-Blink and Kinect-Eyebrows. Another outcome of this research has been the creation of an evaluation framework based on Fitts' Law with a modified multi-directional task including a central location and a dataset consisting of both colour images and depth data of people performing head movement towards different direction and performing gestures such as eye blink, eyebrows movement and mouth movements.

The devices have been tested with healthy participants. From the observed data, it was found that both Kinect-based devices have lower Movement Time and higher Index of Performance and Effective Throughput than the web camera-based devices thus showing that the introduction of the depth data has had a positive impact on the head tracking algorithm. The usability assessment survey, suggests that there is a significant difference in eye fatigue experienced by the participants; blink gesture was less tiring to the eye than eyebrows movement gesture. Also, the analysis of the gestures showed that the Index of Difficulty has a large effect on the error rates of the gesture detection and also that the smaller the Index of Difficulty the higher the error rate.

Contents

List of Figures

List of Tables

List of Abbreviations

ALS	Amyotrophic Lateral Sclerosis
AU	Action Unit
CP	Cerebral Palsy
CRPD	Convention on the Rights of Persons with Disabilities
CVS	Computer Vision Syndrome
DASH	Disabilities of the Arms,Shoulders and Hand
EMG	Electromyography
FAU	Facial Action Unit
fMRI	functional Magnetic Resonance Imaging
HCI	Human Computer Interface
ICF	International Classification of Functioning, Disability and Health
ISO	International Organization for Standardization
ITA	Instantaneous Task Axis
LDA	Linear Discriminant Analysis
MND	Motor Neuron Disease
MS	Multiple Sclerosis
NTSC	National Television System Committee
PIADS	Psychosocial Impact of Assistive DeviceS
QUEST	Quebec User Evaluation of Satisfaction with assistive Technology
ROC	Receiver Operating Characteristic
SDK	Software Development Kit
SIFT	Scale-Invariant Feature Transform
SVM	Support Vector Machines
TBI	Traumatic Brain Injury
WHO	World Health Organisation

List of Symbols

AUC	Area Under Curve
FPR	False Positive Rate
GER	Gesture Error Rate
ID	Index of Difficulty
ID_e	Effective Index of Difficulty
IP	Index of Performance
MT	Movement Time
TP	Throughput
TPR	True Positive Rate
TP_e	Effective Throughput

Chapter 1

Introduction

In 2011, a survey conducted by the World Health Organisation (WHO) estimated that 1 billion people around the world live with some form of disability [1]. Approximately 10 million people in UK have disabilities with a neurological diagnosis. For a multitude of reasons, the number of people with profound disability stemming from neurological disorders is increasing with a resulting impact on their quality of life as well as that of their caregivers. However, the value of assistive technologies in enabling patients to improve their quality of life and also reduce carer strain is emphasised in a Royal College of Physicians Report [2].The cost of caring for neuro-disabled persons in Europe has been estimated at 795 Billion Euro[3].

Assistive technology has been identified in international law as a necessary tool for the independence and inclusion of persons with disabilities in society. The United Nations (UN) has had a substantial role to play for the interests of persons with disabilities to be defended. Persons with disabilities were included as from the 1950s in resolutions of the General Assembly of the UN [4] with finally a Convention on the Rights of Persons with Disabilities (CRPD) being adopted in 2006. In 2008, the CRPD became a legally binding instrument of international law imposing obligations on state parties. The CRPD is the first human rights treaty to be ratified by a regional integration organisation which is the European Union. Individuals are rights holders rather than "objects" of services [5]. They are therefore entitled to the respect, protection and fulfilment of all of their human rights on an equal basis. The need for equality requires in some cases positive actions. Therefore, it is compelling to adopt an affirmative approach to include persons with disabilities in society. Education, expression and participation including exercising political rights, health, work and enjoyment of family life are all important components of an individual's life. Disability cannot be justification of invisibility.

The UK has ratified the CRPD in 2009 and has also accepted the individual complaints procedures as well as the inquiry procedure forming part of the mechanisms set up for persons with disabilities. Since then, in line with its reporting obligations, it has submitted its initial report as a state party in 2011. Assistive technology has been identified by the government as a useful tool to help persons with disabilities in line with its obligations under article 26 of the CRPD for the habilitation and rehabilitation of persons with disabilities [6, para 279 – 290]. Article 26 (1) requires state parties to take measures for persons with disabilities "to attain and maintain maximum independence, full physical, mental, social and vocational ability, and full inclusion and participation in all aspects of life" [7, Article 26(1)]. To reach these objectives, states are compelled by the Convention to "promote the availability, knowledge and use of assistive devices and technologies" [7, Article 26(3)].

For individuals with disabilities access to a computer and/or communication aid may help mitigate the effect of communication impairments. Often this can be achieved through the identification of suitable access sites e.g. hand, foot, arm or head. Some patients, however, suffer from such profound disability that they might be unable to talk but can only make small head movements and facial gestures such as eye blink or eyebrow movement. In some cases, there may not even be enough head movement to enable the use of an access device such as a head mouse like SmartNav [8] and so the only remaining access site may be small facial gestures. Although there are other options available - e.g. the use of eye gaze, but existing systems using eye gaze technology such as MyTobii [9] are complicated and expensive [10]. Additionally, their set-up/configuration places a significant burden on both the user and the caregiver [11].

1.1 Motivation

The aim of this research is to investigate and develop a low cost, reliable automatic gesture recognition system - HeadTracker - that will enable the patients with the most severe disability access to electronic assistive technologies e.g. communication devices. This system is designed for individuals suffering from neuro-disabilities such as Traumatic Brain Injury (TBI), Stroke, Motor Neuron Disease (MND), Cerebral Palsy (CP) and Multiple Sclerosis (MS). In extreme cases, these individuals may only be able to move their eyes or make small facial gestures.

The proposed system, using web cameras coupled with computer vision and pattern recognition techniques, is used to enable a user to move a cursor across a computer display and perform clicking actions. The movement of the user is calibrated by performing small head movements. Once the calibration stage is completed, a homography matrix is generated to map the limited movement on the screen. The next stage of development would involve enabling the system to adapt to any change in the condition of the individuals. For example, a patient with a progressive disease e.g. MND may begin by using head movements to control a mouse and eye blinks to emulate a mouse click action. As the condition progresses and head movement is no longer reliable, the system will then adapt to control the mouse via eye movements, with blinks being used to control mouse clicks. Eventually moving from eye movement to using facial gestures involving movement of eyebrow will provide a switch input to the assistive technology. This adaptability to the changing condition of the patients will potentially make it easier for patients and carers to continue using the device since they would not have to learn to operate different interfaces.

The finished Kinect based head tracker with eye blink as a facial switching mechanism was usedin the COALAS project [12]. The project focused on the development of an autonomous cognitive platform, combining an intelligent wheelchair coupled with a humanoid robot. The head tracker was mounted on a motorised wheelchair to enable users to control it by moving their head [13].

1.2 Contributions

The main contributions of this work include:

1. Development of an evaluation protocol based on Fitts' test with a central position and the randomised selection of target and the development of a low cost Head Tracking system using either a web camera or the Kinect [14]. The evaluation protocol is found in Chapter 3 and the development of the Head Tracker is found in Chapter 4.

2. Development and evaluation of an eye blink and an eyebrow movement switch both a web camera and an RGB-D sensor (Kinect [14]). The evaluation protocol is found in Chapter 3 and the development of the eye blink and eyebrows movement switch are discussed in Chapter 5.

3. Creating of a dataset for facial gestures with both colour facial and depth images. The participants were requested to point at different locations on the screen and perform timed gestures consisting of blinking, eyebrows movement and mouth closure gestures. The detail about the data captured is found in Section 3.6 and sample of the documents such as the participant information sheet, questionnaire and consent form can be found in Appendix A.

1.3 Publications

[1] Guness, S. P., Deravi, F., Sirlantzis, K., Pepper, M. G., and Sakel, M. (2012). Evaluation of vision-based head-trackers for assistive devices. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 4804–4807.

[2] Sakel, M., Deravi, F., Sirlantzis, K., Pepper, M., and Guness, S. P. (2012). Poster 134 Developing a Vision Based Gesture Recognition System to Control Assistive Technology in Neuro-Disability.

[3] Guness, S. P., Deravi, F., Sirlantzis, K., Pepper, M. G., and Sakel, M. (2013). A novel depth-based head tracking and gesture recognition system. In *Assistive Technology Research Series*, volume 33, pages 1021–1026.

[4] Catley, E., Guness, S. P., Sirlantzis, K., Howells, G., and Kelly, S. (2015). A novel depth-based head tracking system for proportional wheelchair control. In *TechAAL 2015*.

[5] Guness, S. P., Deravi, F., Sirlantzis, K., Pepper, M. G., and Sakel, M. Development and evaluation of a mouse emulator using a multimodal real time head tracking system with facial gesture recognition as a switching mechanism. In *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*.

1.4 Thesis Structure

Below is the outline of the thesis:

- Chapter 2 contains the Literature Review for the project. This chapter looks at the past work done and the state-of-the-art in the field of assistive technology. It also looks at the field of gesture recognition namely facial gestures such as eye blink and eyebrows movement.

- Chapter 3 deals with the Experimental Framework of the thesis. In this chapter, the framework to perform the different experiments is discussed. It includes details on the evaluation of devices and assessing the performance of the gesture recognition system.
- Chapter 4 deals with the evaluation of our proposed eye blink and eyebrows detection algorithm. The Blink and the eyebrows movement detector are tested against two facial gesture databases namely the Cohn-Kanade [15] database and the ZJU Blink Database [16]. The performance of the gesture recognition system is tested on the data captured during the data capture sessions to analyse the GER for all the target locations. This chapter also compares the three switching mechanisms - Dwell click, eye blink clicking and eyebrows movement clicking - using a web camera and the Kinect are evaluated and compared. The data obtained from the KentAssist headtracker system is evaluated and compared with existing systems such as SmartNav and CameraMouse.
- Chapter 5 deals with the experimentation carried out during our data capture session. The data obtained compares the impact of different switching facial gestures i.e., eye blink and eyebrows movement on different sensors i.e., a web camera and a Kinect for Windows sensor. It includes details on the data captured and the data capture sessions.
- Chapter 6 contains the conclusion and recommendations about future works that can be carried out.

Chapter 2

Literature Review

This chapter is divided into two parts. The first part introduces assistive devices and provides a glimpse at the difficulties faced by individuals suffering from MS,and MND. In the second part, we are looking at head tracking methods and also looking at the gaze tracking under different types of light sources i.e. infrared and ambient lighting source to compare the techniques used for both types. Techniques used for face detection and for facial feature extraction are also investigated. The facial features investigated are the eyes and face. Evaluation techniques and experimental protocols used in the different experiments involving evaluation of assistive devices and gesture recognition are also reviewed. The different datasets used in the evaluation of facial gestures are also reviewed.

2.1 Assistive Technology

An equipment, or product, that can be used to increase, maintain, or improve the functional capabilities of individuals with disabilities can be defined as an Assistive Device [1].

MND also known as Amyotrophic Lateral Sclerosis (ALS) is a neurodegenerative disorder characterised by loss of motor neurones in the cortex, brain stem, and spinal cord. MND affects the bulbar, limbs, and respiratory muscles. In clinical terms, MND is characterised by progressive weakness, atrophy, spasticity, dysarthria, dysphagia, and respiratory compromise, ultimately resulting in death or mechanical ventilation in the vast majority of patients [17].

In Holper et al. [18], MS is described as being a chronic, often progressive disease of the central nervous system. In 2004, a report by the WHO estimated that there were between 1.1 to 2.5 million cases of MS worldwide [19].

In Rosenbaum et al. [20], CP was described as a group of permanent disorders of the development of movement and posture, causing activity limitation. The motor disorders of CP are often accompanied by disturbances of sensation, perception, cognition, communication, and behaviour. It can also be accompanied by epilepsy, and by secondary musculoskeletal problems.

The Tables C.1, C.2, C.3 and C.4 in Appendix C provides us with an overview of the impairments associated with MS and MND based on the International Classification of Functioning, Disability and Health (ICF).

2.2 Device Development

2.2.1 Facial Feature Extraction

Facial feature extraction is the first step for the head/eye tracking, calibration and facial gesture recognition algorithms. A number of different facial features extraction techniques are discussed for facial features such as the face, the eyes and eyebrows.

2.2.1.1 Face Detection

In this survey, different types of methods used to detect the face are reviewed. Face detection is the first step in the applications such as head tracking system and facial gesture recognition system. In recent years a number of techniques have been developed for face detection. The most common techniques for face detection are based on classifiers such as [21, 22].

In Viola and Jones [21], a rapid and robust algorithm was developed for the detection of objects. The algorithms used shapes called Haar-like features. Viola and Jones used an integral table or a Summed-Area [23] table. The integral table is a representation of an image in terms of the sum of pixel intensity. It is used to greatly speed-up the processing of an image as the pixels do not have to be queried one by one. The Haar-like features perform lookups on the integral table to calculate the sum of pixel intensity of the area they cover. Using weak classifiers, they used the difference between the areas of the black and white regions of the Haar features to train classifiers to recognise objects. Classifiers were trained with frontal face images. A high detection rate is achieved by using the weak classifiers in a degenerate decision tree called Cascade. The cascade operates in a sub-window. The first classifier triggers the second classifier in the cascade if a positive result is obtained and if a negative result is obtained the sub-window is rejected. This algorithm was tested with face image database and proved to be both fast and very accurate [21]. In [24], the algorithm is tested with facial images and proved to be 15 times faster than previous algorithms. With the use of the Integral table, AdaBoost for feature selection and cascade of weak classifiers a very accurate and fast face detector was created by Viola and Jones [24].

Lienhart and Maydt [22] proposed an extended Haar-like features list to improve the performance of the face detector. The extended list of Haar-like features consists of four edge features of which two were rotated at 45 degrees, eight lines features of which four were rotated and two centre-surrounded features of which one was rotated. Two integral tables are required, one for the basic features and the second for the extended features. For the basic features, only one pass from left to right and from top to bottom was required to calculate the integral table. Whereas for the extended features, two passes - the first one, from left to right and top to bottom, and the second pass from right to left and from bottom to top - were required. They also proposed an optimisation stage for the classifier for a given hit rate. An iterative process using a gradient decent-like process to find the optimised parameters for the classifiers for a given hit rate was calculated. Overall the proposed method improved the performance by 23.8% and it was estimated that the rotated features contributed 10% and the post optimisation process contributed 12.5% to the observed improvement.

In Sigal et al. [25], an algorithm using the HSV [26] colour model is used. The algorithm is used to segment the image between background and foreground. It was tested with images from both indoors and outdoors, and also with varying illumination. The proposed algorithm performed about 24% better than static segmentation algorithms proposed by Jones et al. [27].

An adaptive skin detection algorithm is proposed by Dadgostar et al. in [28]. In this approach, a global skin detector is trained and used to threshold on the hue channel of the image. The area of motion in the image is detected, and the pixels in the area are examined to detect the skin region. The detected skin region is then used to retrain the skin detector, and thus a new threshold is obtained.

In Pai et al. [29], a fast face detection algorithm was developed. The proposed algorithm was shown to work with numerous facial variations such as brightness, closed eyes, open mouth, half profile. The experiment results demonstrated that the algorithm could detect a face in about 111 milliseconds with a 92.3% detection rate [29].

In Li et al. [30], an analysis of the pixel-wise skin segmentation approach using colour pixel classification is carried out. Skin classification using the Bayesian classifier and the multilayer neural network classifier performed better than the piecewise linear classifier and the Gaussian classifier. It was also found that a histogram with 256 bins per channel performed better and that there were only small changes in performance when the bin sizes were reduced to 128 and 64 bins. Also, they showed that the RGB and HSV colour spaces are more robust to changes in histogram size.

Soetedjo et al. [31], proposed a two-stage face detection method using skin detection. In the first stage, the coarse region of face skin was detected using fixed boundaries. The region was further refined to identify the fine skin area using flexible boundaries. The fine skin region is the actual face area detected. This method was tested with images both from a number of frontal face databases and different illuminations. The proposed algorithms had a detection rate of about 87%.

In Sun et al. [32], proposed a method whereby a local skin model is detected in an image and used to train a global skin model. The aim of their method was to increase the accuracy and reduce the detection rate degradation caused by varying illumination conditions. The local skin colour model was generated by sampling skin pixels from the current image. It was demonstrated that the proposed method can achieve better accuracy compared to the method proposed by Jones et al. in [33] where the skin model was generated by training a classifier using a very large database of images collected from the web.

In Jain et al. [34], a benchmark framework was proposed for face detection. In Ban et al. [35], a face detection method was proposed, based on the skin colour likelihood using a boosting algorithm. Using both skin and non-skin colour with features from Haar-like features [21] and Local Binary Pattern [36, 37] were used to create a cascaded classifier. The proposed method using Haar-like features and using LBP were tested against Pointing'04 [38], the IMM [39] and the CMU-PIE [40] dataset to test the robustness of the face pose variation. The algorithms were also tested against the Caltech frontal face database [41], FDDB [34] and the Bao face database [42]. The method is shown to be robust and good against face pose.

For face detection [21], is a very popular algorithm with a very good performance. Haar-Like features are used in [22, 35]. It is also used in facial detection benchmark framework [34].

2.2.1.2 Eye Detection

Eye detection is important because one of the aims of this research is to create a blink detection module to be used with the head tracking system. The eyes are an important feature of the face; it can also be used to validate the presence of a human face. A number of different techniques have been developed during recent years [43–47]

In Iskander et al. [43], the method proposed detects both eyes by looking for two circular symmetric objects. A simple quadruple axis spatial-domain-based symmetry indicator was used to detect the symmetric objects. It has been shown that the proposed method performs better than the energy-based detector for moderate levels of noise. Also, the method appears to be more robust than its frequency domain counterpart and correctly detects the eye in a broad variety of images [43].

A novel approach for eye detection in grey-scale face images was proposed by Peng et al. [44]. The proposed algorithm consisted of both feature-based and template-based method to detect the location of the eyes. This method assumes that the face was already detected and the size of the face was known. The first step in this method uses the feature-based method to locate the eye pair in the face image. The second step involves using the template-based method to detect the centre of their iris of both eyes. This method was tested on the ORL face database [48, 49] which consists of 10 images of 40 individuals. The ten images consisting of five images frontal, left, right, top and bottom faces wearing spectacles and another five images in the same pose but without spectacles. The algorithm has a success rate of 95.6% with the 229 images of 29 people without glasses. The success rate for the images with glasses was very low and after analysis the low success rate was attributed to reflections on the glass which prevented the template from locating the iris[44].

In DeSantis et al. [45], an algorithm was proposed where firstly the area where the user is found is segmented and then a second segmentation is performed by using a boolean operation on the first four levels of segmentation to obtain the pupil as the darkest area on the face. This method is invariant to changes in orientation, occlusion and illumination.

A method of eye detection using morphological and colour image processing is proposed by Tanmay[46]. In this paper, the method proposed used the skin model to detect the face of the user and the eyes were detected by converting the image to National Television System Committee (NTSC) [50, 51] colour space and performing morphological operations. This method was found to be very efficient and during testing on frontal images, 90% rate of success was reported but it was found that this technique was not suitable for profile images [46].

In Han et al. [47], Haar-Cascade [24] was used to locate the eyes and a template of the eye region was generated. The template was used to track the eyes. A search region for the eyes was used to reduce the area and was generated using the algorithm in [52]. The application was installed on a tablet and, the accelerometer and gyroscope of the device were used to estimate the location of the eyes.

In this section, the focus has been to detect the facial features such as the face, the eyes and the eyebrows. Viola and Jones [24], can be used to find most of the features required for the development of the facial gesture detection system. As it can be seen from the survey a number of the systems reviewed depend on [24] to initialise their methods such as in [47].

2.2.2 Tracking

2.2.2.1 Head/Face Tracking

Head tracking is very important in this research as one of the main aims of this research is the development of a low-cost head tracking system.

Different technologies have been used for head tracking. Anson et al. [53], compared three head tracking systems using either ultrasonic, infrared or gyroscopic sensors. The first device, the HeadMaster Plus device was an ultrasonic-based system and used a light headset containing four microphones and a control box, which sends ultrasonic sound towards the user. The microphones are used to triangulate the source of the sonic sound. The Tracker 2000 was the second device. It used an infrared camera and required that the users wear a reflective dot on their forehead. The third device was Tracer. The Tracer system consisted of a solid-state gyroscope, which was contained within a headset. The gyro tracked the rotational movement of the head which was then converted to a cursor movement on the screen.

Recently, a number of vision-based assistive devices have been reported [54–58]. Betke [58] attributes this to the fact that personal computers have become more powerful and also camera system for computers are nowadays readily available, and this has contributed to advances in computer vision. CameraMouse [56, 59] is a vision based tracking system, which can track features on the user's body. The face or facial features are most commonly used for tracking and the users have to select the feature they want to track. In Cloud et al. [54], the performance of CameraMouse was evaluated for different users, features tracked (nose, lower lip, interior of left eye and the thumb) and with different applications such as BlinkLink [58]. It was observed that the nose, being the most prominent feature of the face, was the most effective feature for tracking.

Gorodnichy and Roth [60] presented a system for tracking the nose in both in 2D and 3D in low-resolution images. Their proposed system used a stereo camera system to track the nose. The nose was selected as the facial feature to track because it is the most prominent feature of the face; it is always visible even with different head orientations, and it has a unique convex shape. The proposed algorithm was used to build a hands free system.

In Jang et al. [61], a 3D head tracking system was proposed. A set of feature shapes found on the head were selected and Scale-Invariant Feature Transform (SIFT) [62] was used to match the shape in the consecutive frames. The feature shape was generated using SIFT features and stored in a head feature database along with the pose of the head. The head feature database was created from different views of the face i.e. while the user was turning his head. During this learning phase, the database captured numerous features from the user's head during rotation. Their method performed the on-line registration by trying to match the maximum number of SIFT features from the head database to the extracted SIFT features and thus the head movement and pose were calculated. This system was tested on the 3D Pose dataset and was found to have an average error of 2.1°, 3.7° and 4.6° for roll, pitch and yaw respectively. The system was also tested for occlusion and was found to be resistant to partial occlusion.

In Kjeldsen [63], a vision based head tracking algorithm was proposed to address the issue of change in illumination during face tracking and also to offer the user greater accuracy in pointing. The system used frame difference and templates to track the movement of the head. A Sigmoid function with the distance of the movement towards the target was used to reduce drift and filter out any erratic head movement of the user.

Pereira et al. [57], proposed a head tracking system that uses a low-cost infrared camera and the user must wear a reflective dot. The effectiveness of the head tracker was tested using Fitts' Law [64] as recommended by the International Organization for Standardization (ISO) 9431-9:2000 [65]. The mean TP was 0.75 ± 0.12 bits/second , the mean movement time for the tests at ID of 2 bits was 3.02 ± 0.44 and the mean movement time for tests at ID of 5 bits was 5.77 ± 1.12 seconds. The statistical analysis of the data showed that no statistically significant differences were detected between the different attempts based on the mean throughput ($p = 0.218$), movement time of tests with an ID of 2 bits ($p = 0.179$), and movement time with an ID of 5 bits ($p = 0.396$).

In a study conducted by Ashdown et al. [55], head movement is used to enable a user to switch between multiple monitors. The proposed method used three cameras - two on top of the monitors and one near the keyboard. The proposed algorithm generated a 3-D model of the user's head. It was found that the mean distance moved by the cursor using the head in pixels was reduced by 32% i.e. it requires less effort to move the screen cursor using the head tracker than a standard mouse. However, it was also seen that the mean time to complete a task increased by 24% and that it was more difficult to move the pointer between points that are close together.

In [66], a head tracker using the Lucas-Kanade optical flow algorithm [67] was proposed. Haar-cascade [24] was used to find the face region. The pyramidal implementation of the optical flow algorithm was used to determine the best features to track the face region. The best features to track were based on the spatial gradient matrix having large minimum eigenvalues. These features were tracked by the optical flow algorithm. The proposed system was found to be robust to changes in illumination and also complex/busy backgrounds.

Epstein et al. [68] proposed the Kernel-Subset-Tracker, which is an exemplar-based module for CameraMouse [59]. The purpose of the proposed module was to improve the tracking of CameraMouse when the feature being tracked is lost either because the optical flow algorithm [67] tracker lost the facial feature it was tracking or when the user made a rapid head movement with CameraMouse. For the training of the kernel, a training set consisting of images of the face having a size of 100 by 100 pixels at different positions. The image training set is used to generate templates for each position. Three different image-based kernels were created: Threshold kernel, Normalised threshold kernel and the Normalised radial intensity kernel. In the Threshold kernel, a threshold is used to binerise two sequential images. The binerised images are merged and the intersection of the white pixel was located. This kernel was not robust to changes in illumination and object scale. The normalised threshold kernel used the mean intensity of the respective images as a threshold and normalised the number of intersecting pixels by the total number of pixels. The normalised threshold kernel was found to be invariant to changes in the brightness and contrast.

2.2.2.2 Gaze/Eye Tracking

Eye tracking is an important part of the proposed system. In this section, the behaviour of the eyes is investigated along with the different types of eye tracking systems and the techniques used to keep track of the movement of the eyes. In our proposed system, we intend to use eye tracking techniques to track the location of the eyes and use this to both validate the face region and also to improve the performance of the blink gesture detection module.

One of the characteristics of the eyes is that they do not remain still. This is because only a small central area of the retina called the fovea is able to perceive with high acuity and thus when the user views a scene, the eyes have to move to capture the whole scene. This type of movement is called the saccades movement. Fixation occurs when the eyes are not moving and the gaze is held at a specific location. A blink occurs when the eyelids close to moisturise the cornea due to either environmental issues, physical activity or fatigue. During blinking, the eyes are not visible [69].

Adolphs [70] investigated the behaviour of the pupil to emotion. The study was looking at how the size of the pupil of individuals changes while looking at sad faces. Pupil size is well-known to be influenced by stimulus luminance, but it turns out also to be influenced by other factors, including salience and emotional meaning [70]. Experiments were carried out using two groups of individuals to investigate the effect of face stimuli on the size of the pupil. A functional Magnetic Resonance Imaging (fMRI) scanner was used to investigate the relationship between the size of the pupils and brain activations of the participants. The experiments were not conclusive as the subjects in the experiment did not judge emotion on the faces used as stimuli.

In Harrison et al. [71], the size of the pupil is investigated to see how it changes while viewing the expressions of another person. The study showed that the size of the pupil becomes smaller while viewing sad facial expressions and there is no visible change observed while the person is viewing expressions of happiness, neutral or anger.

Wang et al. [72] used eye tracking and pupil dilation to see if a person is telling the truth. They conducted experiments by sending messages within a sender-receiver game, where the subjects were expected to exaggerate the messages to other participants. It was observed that the pupils of the subject were dilated when deceptive messages were sent and that the dilation of the pupil was related to the magnitude of the deception.

In a study carried out by Privitera et al. [73], it was found that the dilation of the pupil was strongly influenced by the visual task being conducted. In both experiments, the participants were asked to fixate the central location of the screen and to press a button when the target was found. For their first experiment, the ten participants were requested to report the presence of an icon. The participants were shown a series of seventeen icons out of which two had targets on them and the other fifteen were neutral. The stimuli were presented to the viewers in a rapid serial visual presentation (RSVP). It was noted that the pupil dilation was significantly associated with target detection. In the second experiment 500 by 500 pixel grey-scale satellite images were shown to seven trained imagery analysts and the target image was a helipad. The second experiment was concluded in 120 seconds which was at a more rapid rate than the first experiment. Privitera stated that the amplitude of the dilation depended on the frequency of targets and the time of the detection. It was observed that larger dilations of the pupil occurred with trials having fewer targets and where targets were viewed earlier during the trial.

Castelhano et al. [74] investigated the influence of the gaze of another person on the direction of gaze of an observer. In their experiment, the participants were shown a sequence of scene photographs that told a story. Some scenes contained an actor fixating an object in the scene. It was observed that the first fixation points of the participants were the face of the actor, the eye then moved to focus on the object the actor was focusing on. Furthermore, it was observed that even in the presence of other objects in the scene, the participants would always focus their gaze on the object the actor was looking at [74]

Another study conducted in [75], the influence of task on the movement of the eye is being investigated. The experiments consisted of twenty participants who were asked to view colour photographs of natural scenes but under two different instruction sets. The first instruction set was to do a visual search of the image and the second task was to memorise the image. The results of the experiments show that the fixation points and the gaze duration of the different participants were influenced by the task they were performing. It was also seen that the areas of fixation were different for both tasks but the movement amplitude and the duration of the fixation were not affected [75].

In Agustin et al., a low-cost gaze tracking system was built. The system consisted of a low-cost Sandberg web camera with six infrared LED lights. The system was compared against two commercially available systems. The commercially available systems were SMI IViewX RED and Tobii 1750. The test conducted using the three eye trackers were target acquisition and eye typing [76].

In Chao-Ning et al. [77], a 3D eye gaze and head movement tracking system using four cameras that did not require any calibration was proposed. The system was designed to be used for real-time eye tracking in an open environment to get a wide-angle view and long range of the scene. Their system consisted of four modules. The first module was the face tracking module that was responsible for tracking the face of the user using the AdaBoost algorithm. The second module was used to detect the eyes by finding the centre of the pupil and using images from the different cameras to determine the location in 3D. The third module was the active control module. This module was used to integrate the face tracking module and the eye tracking module with the input control module. The input control module was used to control the mouse cursor on the screen. This module was used to move the cursor to follow the gaze of the user and to click on buttons on the screen by detecting blinks from the user [77].

In Asteriadis et al. [78], the paper proposed a method to monitor the state of the user as information was displayed to the user on a computer monitor. The proposed system does not require any calibration. It also allows the user free movements and works under ambient lighting conditions. The method used to estimate the head pose uses the point between the nostrils and two points on each eyebrow. For nostrils detection, an area around a segment of the perpendicular to the inter-ocular line was extended starting from the middle of the eyes. The darkest row of this area is considered as the vertical position of the nostrils and the middle point was assumed to be the middle of the nose. Similarly, two points on each eyebrow are extracted by assuming that the darkest points in a neighbourhood above the eye corners were the eyebrows. The features were extracted using a web camera and used to determine the level of engagement of the user with the content being shown on his screen. The features extracted the head pose, eye gaze, eyebrow movements, head horizontal and vertical speed components, mouth horizontal and vertical opening, backwards and forward movement of the user. Feature tracking was performed using a three-Pyramid Lucas-Kanade [67, 79] algorithm. The eye tracking was done using the centre of the eye and monitoring the distance from the eye corners and the eyelids. The resolution of the video generated was 720 by 576 pixels at 25 frames per second. The amount of data obtained per video was about 10,000 to 12,000 frames [78].

In Lin et al. [80], a mouse emulator using a web camera to track the face and the eyes of the users based on skin colour is proposed. The image is converted from the RGB colour space into the YCbCr space and the illumination component Y is removed. Skin models in the CbCr subspace were generated using the elliptical model discussed [81] and used for tracking. A K-Nearest Neighbour(KNN) classifier was trained to identify the type of illumination. The classifier was trained with ten images per illumination condition and the skin model was built for each type of lighting. The classifier was tested with 30 images per illumination conditions and it was found that the overall accuracy of the classifier was 92%. To detect the face image subtraction was carried out to identify moving objects and the skin model was used to identify the face region. For the identification of the eyes, the Y component of the YCbCr images is used. As the iris is the darkest part of the face as it has a low intensity of luminance, it is extracted from the image. The centre of the eyes and face regions were used to move the cursor on the screen. When the head moves the eye centres move but the centre of the face stays the same and hence by comparing the degree of movement of the eyes centre with reference to the centre of the face, the cursor location on the screen is estimated.

In Magee et al. [82], a real-time gaze detection system using a low-cost camera was proposed. In their first experiment, eight participants were asked to perform 20 eye movement to the right and the left of the monitor. They performed the task while moving between 2 and 5 feet from the camera. Overall the system correctly identified 87.5% of the attempts but to achieve this performance the participants had to perform 248 attempts instead of 160. In the second experiment, four participants were asked to play three games of BlockEscape - a game where the user had to move an object in a maze using their eyes in this case - with 3 devices, the EyeKeys [82], the CameraMouse [54] and a keyboard. EyeKeys and CameraMouse were found to have the same number of wins. It was noted that manual intervention was required to reset CameraMouse when the feature being tracked was lost whereas with EyeKeys, the user just had to look at the monitor to correct the error. As it can be seen both from Castelhano et al. [75] and Privitera et al. [73], the eyes are influenced by the task being conducted. The pupil and eyes are also sensitive to different stimuli such as images, light and emotion. There are a number of issues with using eyes for both navigation and also to trigger an action such as a blink gesture to emulate a mouse click as stated in [11].

2.2.3 Calibration

Camera calibration is an inevitable part of a marker-less motion capture system, a simple and useful calibration method can be conducive to reduce the complexity of the marker-less system.

2.2.3.1 Using Calibration

Bernet et al. [83], demonstrates a technique that can be used for a rapid calibration of an eye tracking system. In this paper, a quick calibration method is compared to a method using a grid consisting of calibration points. In the proposed method, the user is required to fix their gaze on a single point whereas in the grid calibration method the user is required to fix 9 points. In Agustin et al. [76], the gaze tracking system proposed uses multiple calibration points. The method suggested by Bernet et al. [83], was tested with both a fixed and a portable gaze tracking system. In the fixed system, the camera is directed at the pupil of the user. This system is very accurate to determine the gaze of the user on a flat surface in front of the user such as a screen. The portable system consists of two cameras mounted on a headgear. The first camera is directed at the pupil of the user and the second camera is directed at the scene the user is viewing. The mobile system gives the user a greater liberty of head movement. During their research, it was demonstrated that the quality of the calibration of the single point calibration was similar to the quality of the calibration for multi-point calibration method. The reference distance used for calibration is 1 metres [83].

Hansen et al. [84], proposes a method using homography with multiple calibration points to estimate the gaze. The method can be used when head movements are present but without any requirements for camera calibration or geometric calibration. This method also does not require any particular lighting source, but for experimentation purposes to extract the pupil and glint, an infrared camera is used. During calibration, the location of the pupil and the location of the calibration point being displayed is used to calculate the homography matrix. This method combines both geometry and interpolation based techniques and thus obtained both the robustness associated with geometry-based models and at the same time keeping the simplicity and flexibility of interpolation-based methods.

2.2.3.2 Calibration Free

In Asteriadis et al. [78] and Chao-Ning et al. [77] no calibration was required from the user. Nagamatsu et al. [85], proposed a 3D gaze tracking system without calibration and using two cameras and two light sources. Each camera is directed to an eye. The optical axes of both eyes and the visual path are calculated. The location of the gaze point is estimated to be at the midpoint of the intersection of the optical axes with the screen and the estimated visual paths intersect. To be able to estimate the optical axis, the centre of the pupil and the Purkinje - the reflection of the light source - are required.

In Kaminski [86], both the orientation of the head and the gaze is detected by creating a geometrical model of both the face and the eye. In this paper, three points i.e. the centre of the eyes and the midpoint between the nostrils. The points A, B and C form an equilateral triangle i.e. all the three sides of the triangle are equal. This relationship between the centre of the eyes and the midpoint between the nostrils can be used to track the face. The direction of movement can be estimated and used to estimate the direction of the gaze [86].

2.2.4 Gesture Detection

In this section, the interest is in processing video information to recognise blink and eyebrow movement gestures. This recognition can be to emulate a mouse click or a switch action to access and control a computer/communication aid.

2.2.4.1 Spontaneous and deliberate gestures

Gestures can be either spontaneous or deliberate. It has been reported that the frequency of spontaneous eye blinks can change from 20 to 30 blinks/min depending on the mental task the person is performing [87], and can decrease to about 11 blinks/minute during visually demanding tasks [88]. As for eyebrows a study by [89], the spontaneous characteristics of the raised eyebrows is compared to spontaneous smiles. In [90], it was found that deliberate smiles had a higher amplitude and were faster than spontaneous smiles. The duration of spontaneous smiles was found to be between 0.5 seconds to 0.75 seconds. In the experiments carried out in [89], the results obtained for spontaneous smiles were similar to the ones obtained in their previous study and also that raised eyebrows gesture followed a similar pattern to smiles.

In Bartlett et al. [91], a spontaneous facial expression system was proposed. The aim of the system was to detect facial expressions on frontal faces in a video stream. The proposed system was based on using Gabor filters. AdaBoost was used for selecting the best Gabor filters. The method was compared using two different types of classifiers - Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM).The implementation consisted of two stages. The implementation using SVM classifiers got the best result. The system obtained 93% correct classification on the Cohn-Kanade expression dataset [15]. The system can classify 17 Action Unit (AU) either occurring by itself or with other AU and has a mean accuracy of 94.8%.

It has been reported that spontaneous eye blink can change from 20 to 30 blinks/min depending on the mental task the person is performing [87], and can decrease to about 11 blinks per minute during visually demanding tasks [88]. In [92], the spontaneous eye-blink duration is determined to 250ms or less and voluntary eye-blink is considered to be less than 2 seconds i.e. 2000ms. This is also confirmed by [87] where the duration of a spontaneous blink due to mental stress to be greater than 2 seconds (30 blinks/second).

Therefore, a voluntary blink duration can be between 250ms to 2000ms. Unintentional facial gestures also could interfere with the detection of the voluntary blink detection. Thus, the threshold for blink duration should be set to a value greater than 750ms [88, 89]. The blink duration should be set to a duration between 750ms and 2000ms. In [93] , it is stated that the activation time for most eye typing system is between 450ms and 1000ms. In the case, the activation time for both the eye blink and the eyebrows movement gesture were set to a range between 1000ms and 2000ms.

2.2.4.2 Blink Detection

Grauman et al. [52] proposed two systems called BlinkLink and EyebrowsClicker. The BlinkLink software tracked both the motion within the eye region and the eye itself. The EyebrowsClicker tracked the eyebrows region and detected the rising and lowering of the eyebrows. To initialise the location of the eye and eyebrows regions, the user has to perform the gestures and by analysing the area of motion on the face, the respective regions are detected. A template of each region is generated. To detect a blink, the correlation score of the eye region and a template of both the closed eye and open eye are compared. For eyebrows gesture, the distance from the eyes and the eyebrows are monitored to detect the rise and fall motion of the eyebrows. Blink detection had an overall success rate of around 95.6% and was tested on 15 healthy individuals and one person suffering from TBI. EyebrowsClicker had an overall success rate or 89% and it was tested with six individuals, but the software had to be reinstated twice during the data capture session because the tracking of the eyebrows was lost.

The blink detector from [52] was used in [47]. Computer Vision Syndrome (CVS) detector uses the blink detector to monitor the blink rate of the user and every 20 minutes informed the user that their blink rate had decreased.

In [94], the proposed method was an enhancement of the work in [52]. They proposed an automatic initialisation process which is activated by a voluntary blink, and an on-line template of the eyes is created. The system auto initialises when the correlation falls below a threshold. Frame difference is used to locate regions of motion and is assumed to be eye region. The largest connected components are extracted and the templates for the open and closed eyes are generated. The eye template is used to track the eye. The normalised correlation coefficient is calculated to try to match the template. The correlation score from the matching with the closing and opening templates are used to detect a blink. For an open eye, the correlation range was from 0.85 to 1.0. During a blink, the score falls to a range between 0.50 and 0.55. If the score is below 0.45, it means the tracking failed and templates have to be re-initiated. The system was tested on 8 participants, a total of 2,288 true blinks were analysed, and the system was found to have an overall accuracy of 95.3% with 43 missed blinks and 64 false positives. With blink length being ten frames i.e. one-third of a second, there were 23 missed blinks and 33 false positives out of a total of 1,242 blinks. With blink length being 20 frames i.e. two-third of a second, it was found that in 504 blinks, the missed blinks increased from 23 to 58 and the false positives rose from 33 to 68.

Yunqi et al. [95] proposed an eye blink detection algorithm which was used in a drowsiness driver warning system. The proposed system used Haar-like [24] features and AdaBoost to detect the face of the user. Some pre-processing was performed on the image and an edge detection algorithm was used to find the eye corners, the iris and the upper eyelid for each eye. The curvature of the upper eyelid was compared with the line connecting the two eye corners and if most of the upper eyelid curvature was under this line, the eye was considered closed. The algorithm was tested on images captured during a real driving session and a 94% accuracy was obtained for the eye state detection.

Missimer et al. [96] proposed a blink detection algorithm based on the analysis of the differences in three consecutive images. Blobs are generated from the merging of two difference images produced. Three points are used for tracking, the centre of the upper lip and the upper part of both eyebrows. Also, optical flow is used to monitor these three points. The eye templates are generated based on the tracked points and used to train the system. The system is reported to having a success rate of 96.6% and was tested on 20 healthy individuals.

Val et al. [97], presented a robotic assistive aid. The system was controlled via eye blinks, for example a Right Eye Blink moves the robot forward, a Right Eye Blink followed by a Left Eye Blink instructs the robot to turn right and doing the Left and Right Eye blink causes the robot to come to a stop. An infrared emitter and an optic sensor were used to detect the eye blink.

An active contour model was employed in Krolak et al. to detect a voluntary blink i.e. a blink which is more than 2 seconds long [98]. The face is detected using the Haar-features [24] and the location of the eyes was found based on the known geometrical dependencies of the human face. Once the eye region is located, the image is converted from the RGB colour space to the YCrCb colour space and the active contour model is applied. The proposed system uses two active contour models, one for each eye. The system was tested on 30 participants and each participant was asked to perform 30 blinks consisting of 15 short blinks and 15 long blinks. The overall accuracy of the proposed system was 96% i.e. 98% of long eye blinks and 94% of short eye blinks were correctly detected.

In [99] and [100], a real-time, non-invasive eye blink detected based on image difference and template matching were proposed. The system relies on the difference between frames and template matching to track the eye and detect an eye blink.

In Mackenzie and Behrooz [101], a new text entry system is proposed using the API from EyeTech Digital system. The proposed system allowed a user to type using their blink. The system was implemented using a scanning ambiguous keyboard [102–104], a new form of scanning keyboard that allows English text to be entered in less than two scanning intervals per character. The proposed system was tested with 12 participants. The participants had to input text using the system at three different scanning intervals i.e. 1000ms, 850ms, and 700 ms. It was found that the average text entry rate was 4.37 wpm and the accuracy of the system was greater than 97%. The highest average text entry rate was 4.58 wpm and the interval was 850 ms. The average text entry rate of 700ms and 1000ms scanning intervals were 4.37 and 4.19 wpm respectively.

In Lee et al. [105], a method to detect an eye blink in different facial poses was proposed. The method used AdaBoost [22] for both detecting the face and the eyes. Moreover, to overcome the limitation of Adaboost with rotated faces, the Lucas-Kanade Tracker[79] was used. To use the Lucas-Kanade tracker, points to track are selected from within the eye regions using the points to track algorithm [106]. The tracked eye region is normalised using the algorithm in [107] and binerising the image. To detect an eye blink, two features are extracted from the tracked eye region, the first feature is the cumulative difference in the number of black pixels. The second feature is the ratio of the height to width of the eye. These two features are used with the SVM classifier to determine the state of the eye.

2.2.4.3 Eyebrows Movement Detection

In Tuisku et al. [108], the functionality of a new wireless prototype called Face Interface was investigated. The system combined the use of voluntary gaze direction and facial muscle activation, for pointing and selecting objects on a computer screen, respectively. The subjective and objective functionality of the prototype was evaluated with a series of pointing tasks using either frowning (i.e., frowning technique) or raising the eyebrows (i.e., raising technique) as the selection technique. Pointing task times and accuracy were measured using three target diameters (i.e. 25, 30, 40mm), seven pointing distances (i.e., 60, 120, 180, 240, 260, 450, and 520mm), and eight pointing angles (0^0, 45^0, 90^0, 135^0, 180^0, 225^0, 270^0, and 315^0). The results showed that the raising eyebrow selection technique was faster than the frowning technique for the objects that were presented in the pointing distances from 60mm to 260mm. For those pointing distances, the overall pointing task times were 2.4s for the frowning technique and 1.6s for the raising technique. Fitts' law computations showed that the correlations for the Fitts' law model were r =0.77 for the frowning technique and r =0.51 for the raising technique. Furthermore, the index of performance (IP) value was 1.9 bits/s for the frowning technique and 5.4 bits/s for raising the eyebrow technique. Based on the results, the prototype functioned well and was adjustable so that two different facial activation techniques can be used in combination with gaze direction for pointing and selecting objects on a computer screen.

In Lombardi at al. [109], an eyebrow clicker was proposed. The eyebrows were detected by first locating the eyes. The location of the eyebrows was assumed to be just above the region of the eyes. During the initial stages of the proposed system, the eyes were found by comparing frames between eye closing and opening movement. The eyebrows were located above the eye region. Once the eye and eyebrows are located, their relative locations are recorded. A raised eyebrow movement is detected when the distance between the eyes and the eyebrows is increased and this causes a mouse clicking action to be initiated. The system was tested and was found to achieve an accuracy of 93%.

In Lancioni et al. [110], replaced micro-switch devices used by two persons with CP using camera-based micro-switches. The proposed camera-based micro-switching system relied on the use of opaque colour spots to be placed on the face and a USB web camera. In the first case, a photocell used as a mouse switch to detect mouth opening and closing and, an infrared emitter and detector to detect blink were replaced. For the mouse switch, colour spots were placed on the nose and on the left corner of the mouth under the lower lip. In the second case, a colour spot was placed on the left eyebrow. The computer monitored the spots and when the spots moved for a predetermined time, it was considered as a switch activation. The camera-based micro-switch does not require any support frame and thus it is easier to use and more acceptable by their users. In [111], a further study with three participants with profound multiple disabilities was carried out. Eyelid closure, mouth closing and eyebrows lifting were used as switching mechanisms. The increased response of the participants during the experiments show the potential of this method for people with minimal movement.

The depth data from an RGB-D sensor would be used to segment, detect and track the movement of the head. The system is completely non-invasive i.e. no marker or reflective surface would be used. The depth data with the RGB data are used to capture facial gestures such as eye blink and eyebrows movement and used a switching mechanism. For facial feature detection, the method proposed by Viola and Jones [24] could be used. For the detection of the eyebrow, the method proposed by Lombardi at al. [109] could be adapted to measure the distance of the corners of the eyebrows along as the top most location of the eyebrows and monitoring the movement of the eyebrows against features such as the nose or the interocular midpoint between both eyes. Skin detection methods could be used as a backup in case Haar-Cascade [24] fails for a certain pose. The system has to be robust and be able to work with the head in a number of different poses.

2.2.5 Discussion

For face detection and facial feature extraction required for head tracking and gesture recognition, the Haar-Cascade [21], is a very popular [22, 35, 47, 66, 98, 105] algorithm with a very good performance. It is also used in facial detection benchmark framework [34]. Haar-Cascade can be used to extract all the different features such as face and eye required for head tracking and facial gesture detection. Methods using skin colour [29–33, 46, 80, 81] might be too computationally expensive and thus would either require expensive hardware to handle the processing or would not work in real-time situations.

A number of the systems are invasive [57, 76, 110] i.e. they require the user to wear an artefact such as in Pereira et al. [57] where the user had to wear a reflective dot or had led light shown on their face such as in Agustin et al. [76]. The proposed system is aimed to provide the users with a non-invasive system to increase comfort and the acceptability of the system by the user.

During the survey, no depth based algorithms or systems could be found before 2011. The depth data became readily available to the mainstream public at a low cost by the advent by the introduction of the Kinect [14] in the Xbox360 gaming console. Using the depth data to reduce the area the algorithm needs to process would both boost the performance of the device and keep computational requirement of the system low. Haar-Cascade [21] works best with a full-frontal face and has had issues with changes in the pose of the face which was addressed in [22]. The depth data could be used to reduce the issue with the pose of the face instead of using skin detection as a backup to locate the face. For a head tracking system, it is assumed that the head of the user would be the nearest to the sensor in this case, the depth data could be used to locate the object nearest to the RGB-D sensor.

The proposed system is a low cost system. The use of mainstream consumer grade products such as the Kinect and using robust algorithms with low computational requirements such as the Haar-Cascade [21] helps to keep the cost of the system low.

2.3 Device Evaluation

2.3.1 Objective Evaluation

Betke et al. [58], describe the advances made in the development of assistive software and the use of emerging technology to create intelligent interfaces using both assistive technology and Human Computer Interface (HCI). The example of the CameraMouse [54] is used as an interface system with different assistive devices and software such as Midas Touch [59], Dasher [112], etc. are included to highlight the use of HCI and assistive devices.

Abascal et al. [113], highlighted some opportunities and challenges that designing human-computer interfaces suitable for the disabled can pose. For people suffering from disabilities, HCI can be used to design better interfaces which could be accessible to individuals with disabilities and thus improving socialisation, providing better access to communication facilities and having a greater control over their environment.

Fitts' test [64] was developed in 1954 to model human movement. The result of the experiments showed that the rate of performance of the human motor system is approximately constant over a broad range of movement amplitudes. Mackenzie et al. [114], adapted the Fitts' Law for assessing HCI. This work was later embedded in an International Standard for HCI, ISO 9431-9:2000 [65] providing guidelines for measuring the users' performance, comfort and effort. The performance of the device was measured by making the user perform tasks using the device. There are six types of tasks - one-direction, multi-directional, dragging, freehand drawing and, hand input, grasp and park (homing/device switching). ISO 9431-9:2000 [65] requires that the input device be tested for at least two different ID. ID is a measure of the difficulty of the task [115] . In this paper, Douglas et al. [116], investigated the validity and practicality of the ISO framework using both multi-directional and the one-direction Fitts' Tests for two devices namely a touch-pad and a joystick.

In Hwang et al. [117], proposed the sub-movement model. In the sub-movement model, it is assumed that a rapid aiming movement is made up of a sequence of smaller, discrete sub-movements. A rapid aimed movement is considered to start with an initial sub-movement. A sub-movement ends either when the target is reached, speed is zero, acceleration changes from positive to negative while the speed is still less than 75% of the peak speed or a relative maximum occurs in the acceleration while the acceleration is negative. If the target is not reached a new sub-movement is started. The Instantaneous Task Axis (ITA) is an imaginary line connecting the start point and the target's centre -

it represents the ideal path. It is used to measure the error between the ideal path and the path taken by the user and thus it can be used to denote the accuracy of the sub-movement. The proposed model has two types of errors, Coincident error and Perpendicular error. The Coincident error is the difference between the sub-movement's component along the ITA axis and the target centre. There are three types of Coincident errors: an overshoot, an undershoot and a counter-productive sub-movement. The overshoot occurs when the Coincident error is greater than zero i.e. the cursor has moved beyond the target centre. The undershoot occurs when the Coincident error is less than zero i.e. the cursor stopped before the target is reached. The counter-productive sub-movements represent the sub-movements which move the cursor away from the target. The Perpendicular error is the distance between the sub-movement endpoint and the ITA. It is measured in a direction normal to the ITA. The error can be positive or negative depending on the position of the endpoint to the ITA. Thus, if the sub-movement endpoint is counter clockwise to the ITA, it is negative. And if it is counter clockwise to the ITA, it is positive.

2.3.2 Subjective Evaluation

2.3.2.1 Independent Questionnaire for assessment of comfort

The user-based approach is the assessment of the effect of the system on the user. The measurement is obtained from feedback or surveys obtained from the end users or their carers. This questionnaire has been used in a number of studies [116, 118–121]. In Douglas et al. [116], when evaluating devices, the comfort of using the devices was assessed using the IS0 Independent Questionnaire for Assessment of Comfort [65]. The participants were asked to answer the questionnaire after having used the devices.

2.3.2.2 Psychosocial Impact of Assistive DeviceS(PIADS)

Psychosocial Impact of Assistive DeviceS (PIADS) is a 26-item questionnaire, which is used to measure the quality of life of people using assistive devices. The quality of life measurement is from the point of view of the disabled person. PIADS can be divided into three sections: competency, adaptability and self-esteem. The competency section consists of 12 items, the adaptability section consists of six items and self-esteem includes of 8 items. Competency related to the same functional capabilities, independence and performance. The adaptability section associated with the motivation of the participant in social interaction and risk taking. The self-esteem section relates to the self-confidence, self-esteem and emotional well-being [122].

2.3.2.3 Disabilities of the Arms, Shoulders and Hand(DASH)

In the study conducted by Gummesson et al. [123], the Disabilities of the Arms,Shoulders and Hand (DASH) outcome measurement was used to evaluate the effect of surgery on 109 patients. The DASH outcome measurement is obtained from a self-report questionnaire consisting of 30-item and each of the items is rated on a five-point Likert [124] scale. The raw score obtained from the questionnaire is then transformed to a 0 to 100 scale, where 0 stands for no disability and 100 for maximum impairment. The score for the disability/symptom scale is called the DASH score. The information obtained from the questionnaire corresponds to the status of the patient's health. The survey was conducted with the patients before and after the surgery. It was found that DASH can be used to detect and differentiate between small and large changes in disability over time after surgery in patients with upper-extremity musculoskeletal disorders.

2.3.2.4 Quebec User Evaluation of Satisfaction with assistive Technology(QUEST)

Quebec User Evaluation of Satisfaction with assistive Technology (QUEST) is an assessment tool to evaluate the user assessment of assistive devices [125]. The evaluation of the assistive device is based on the satisfaction of the users with the device. QUEST consists of 27 variables, and for each the user has to specify the importance the user associates with the variable and indicate the level of satisfaction. The objective of QUEST is to provide a reliable and valid means of assessing the outcome of assistive technology from the user's perspective. QUEST has been updated to QUEST 2.0 [126] which contains only 12 questions with a five-point Likert [124] scale for satisfaction. The 12 questions selected for QUEST 2.0 where the most important questions i.e. the highest ranking questions in terms of performance measurement, and the importance scale was removed. QUEST 2.0 also looks at both the satisfaction with the device and the service. It can be applied to a wider context.

2.3.3 Discussion

The proposed device evaluation framework consists of two parts, the objective and subjective evaluation. The objective evaluation is the measurement of the performance of the participant and the device being evaluated. The subjective evaluation is the perceived performance of the device from the participant's perspective.

For the objective evaluation for HCI, the method proposed by MacKenzie et al. [114], based on Fitts' Law [64] has been used in a number of studies and it is also the guideline set forward by the ISO for evaluating the performance, comfort and effort required to use a non-keyboard input device was followed [65].

The subjective evaluation of assistive devices is important to understand the relationship between the users and their devices. The IS0 Independent Questionnaire for Assessment of Comfort [65] can be used to evaluate any input device and not only assistive devices. Thus, it can be used to compare the perception of both abled and disabled participants taking part in the evaluation. It can also be used for device evaluation across a range of different disabilities. Both QUEST [125, 126] and PIADS [122] are an evaluation tool based on the satisfaction of the person using an assistive device. This evaluation is aimed exclusively at people suffering with disabilities. It can also be used in longitudinal studies where it could be administered periodically to the user to get an idea of the user's satisfaction over time. DASH [123] is used only before and after an intervention such as surgery and is specific to people suffering from upper-extremity musculoskeletal disorders. The ISO Independent Questionnaire for Assessment of Comfort [65] seems best suited for the evaluation of the device because initially the devices would be tested with volunteers both healthy and those suffering from disabilities. The QUEST [125, 126] or PIADS [122] could be administered only with the volunteers suffering from disabilities.

Using both an objective and a subjective evaluation could be interesting as it would offer both the performance measured during the task but also the perspective of the user on the different devices. This approach could provide a better understanding of the relationship between the user and their devices and also which aspects of the device are most important to the user. It could also help to find the best suited both based on the performance during the test but also based on what the user thinks is the best for him. This would make the user more inclusive in the decision process and thus contribute to decreasing device abandonment issues.

2.4 Experiment Design

Experiments based on Fitts' Law [64] and adapted by [114] for HCI forms part of ISO standards for non-keyboard input device [65]. The experiments are used to either evaluate devices which are in development such as [57, 127–129], or to evaluate the performance against existing and established devices such as [53, 119, 130].

Anson et al. [53], conducted an investigation to compare the efficiency of three commercially available head tracking devices. To test the efficiency of these devices, each participant was asked to produce a series of drawings using each device. The drawings were presented to each participant in a predetermined order. To prevent the participants from memorising the drawings, the order of using the different devices was balanced to control for learning effect. For each participant, the time taken to draw and the number of errors were recorded. It was found that the HeadMaster Plus produced the most consistency. According to the feedback provided by the participants the two fastest devices i.e. the HeadMaster Plus and Tracer were the most uncomfortable to use and the Tracker 2000 system was the one preferred by most of the participants.

A comparative study was conducted in Man and Wong [119]. The aim of the study was to evaluate the individual performance of two participants suffering from CP on four different devices namely the CameraMouse [59], The ASL Head Array mouse emulator, the CrossScanner and the Quick Glance Eye Tracking System. It was found that the CrossScanner had the highest accuracy rate across both participants and also between the different devices. Followed by the CameraMouse, which was the most comfortable. The ASL Head Array Mouse emulator was too difficult to use and was too uncomfortable for both users. Both participants were not able to use the Quick Glance Eye Tracking System because of their condition.

In a study conducted by Williams et al. [130], the efficiency of three different alternate pointing devices was compared. The devices used in the study were a standard mouse as a baseline, a head orientation system and an Electromyography (EMG) based system. Fitts' Test was used to measure the performance and the throughput of the devices. In the study, it was found that with the head orientation system, the movement both to a single direction targets (i.e. targets at the cardinal points - North, South, East and West directions) and in diagonal movements were much better than with the EMG system. In the case where an EMG based system was used the diagonal movements followed a square shape path indicating that the participants were moving the cursor first in one direction and then the other, rather than simultaneously controlling the two cursor directions.

In Pereira et al. [57], a head tracker was developed, and evaluated with ten individuals with cervical spinal cord injury. The device developed used an infrared camera and a reflective dot. The evaluation involved a Multi-Directional Fitts' Test with 16 different orientations. The participants had to undergo 12 sequences of tests for two ID – 2 bits and 5 bits. The results of the 12 attempts showed that the mean throughput movement time for ID 2 bits and the average movement time for ID 5 bits of the proposed system were 0.75 ± 0.12 bits/second, 3.02 ± 0.44 bits/seconds and 5.77 ± 1.12 bits/seconds respectively. There was also no significant difference in the average movement time of the first and last attempts. From the results, it was seen that the device was easy to use and the movement time showed that the device could adequately emulate the movement of a mouse.

Salivia and Hourcade [127] conducted an experiment with two persons with Parkinson's Disease(PD) to demonstrate the difficulties they encountered while performing the pointing task. The pointing task involved pointing at targets of different size and distances, in 8 different orientations. The data obtained in completing the tasks were compared with the data from previous studies with children [128] and older adults [131]. The difficulties of the participants in completing the tasks were analysed by using the duration and accuracy of the tasks and the characteristics of the sub-movement [117] of the participants while performing the tasks. It was found that the performances of the two participants were different. In the first session, the participant referred to as "Bob" had a higher MT than the participant "Dave". Also, Dave had double the number of re-entry than Bob whereas Bob had numerous sub-movement away from the target i.e. more than 60 pixels from the target, and Dave could complete a task in less than 6 sub-movements. In the second session, Bob decreased his MT and increased his number of re-entry, whereas Dave decreased his number of re-entry but his MT was similar. Bob required only four sub-movements to complete the task and the number of sub-movements required by Dave was six, similar to session 1. Also, the accuracy of Dave decreased by 6%.

In another study Salivia et al. [129], conducted an experiment with sixteen participants with different impairments - Parkinson's Disease(3), CP(3), Carpel Tunnel Syndrome(3), Stroke(1), 1 individual with Development deficiencies, 1 person with damage to the

Central Nervous System, MS(1), Spina Bifida(1) and 2 individuals who did not want to disclose their disabilities. The participants were asked to perform two sets of pointing tasks. In the first session, PointAssist was disabled and in the second session it was enabled. In session 2, the parameters such as the length and speed of the sub-movement were personalised for each user. Half of the participants participated in session 2. It was found that there was a significant difference in the click accuracy between the sessions with and without PointAssist. For the press and release accuracy, it was found that the use of PointAssist had a significant effect on the pointing accuracy.

Most of the experiments surveyed use both healthy volunteers and individuals with impairments such as [57, 119, 127, 129, 131]. Using people with disabilities and involving them in the development of assistive devices could help to develop better devices which would match what they are looking for and thus could contribute to a decrease in the number of people abandoning their devices.

2.5 Database

In [15], the CMU-PITTSBURGH AU-Coded Face Expression Image database was presented. The database consists of facial images obtained from 210 adults from different ethnic background between the age of 18 and 50. For the data capture - two cameras were used, one directly in front of the participant and the other positioned 30^o to the participant's right. For one-third of the participants, the lighting of the room was ambient with the addition of a high-intensity lamp and for the others, two high-intensity lamps were used with reflective umbrellas to provide uniform lighting. The resolutions of the image captured are 640 by 490 for grey-scale and 640 by 480 for colour. The participants were asked to perform 23 facial displays consisting of both single action units and combinations of action units, and each display started and ended with a neutral face. Sixty participants were asked to perform a 30^o degree head rotation with facial expression. The data obtained were coded by two certified Facial Action Unit (FAU) coders.

Table 2.1 FAU from [15]

AU	Facial Muscles	Description of Muscle Movement
1	Frontalis, pars medialis	Inner corner of eyebrow raised
2	Frontalis, pars lateralis	Outer corner of eyebrow raised
4	Corrugator supercilii, Depressor supercilii	Eyebrows drawn medially and down
5	Levator palpebrae superioris	Eyes widened
6	Orbicularis oculi, pars orbitalis	Cheeks raised; eyes narrowed
7	Orbicularis oculi, pars palpebralis	Lower eyelid raised and drawn medially
9	Levator labii superioris alaeque nasi	Upper lip raised and inverted; superior part of the nasolabial furrow deepened; nostril dilated by the medial slip of the muscle
10	Levator labii superioris	Upper lip raised; nasolabial furrow deepened producing square-like furrows around nostrils
11	Levator anguli oris (a.k.a. Caninus)	Lower to medial part of the nasolabial furrow deepened
12	Zygomaticus major	Lip corners pulled up and laterally
13	Zygomaticus minor	Angle of the mouth elevated; only muscle in the deep layer of muscles that opens the lips
14	Buccinator	Lip corners tightened. Cheeks compressed against teeth
15	Depressor anguli oris (a.k.a. Triangularis)	Corner of the mouth pulled downward and inward
16	Depressor labii inferioris	Lower lip pulled down and laterally
17	Mentalis	Skin of chin elevated
18	Incisivii labii superioris andIncisivii labii inferioris	Lips pursed
20	Risorius w/ platysma	Lip corners pulled laterally
22	Orbicularis oris	Lips everted (funnelled)
23	Orbicularis oris	Lips tightened
24	Orbicularis oris	Lips pressed together
25	Depressor labii inferioris, or relaxation of mentalis, or orbicularis oris	Lips parted
26	Masseter; relaxed temporal and internal pterygoid	Jaw dropped
27	Pterygoids and digastric	Mouth stretched open
28	Orbicularis oris	Lips sucked
41	Relaxation of levator palpebrae superioris	Upper eyelid droop
42	Orbicularis oculi	Eyelid slitaa

Continued on next page

Table 2.1 – continued from previous page

AU	Facial Muscles	Description of Muscle Movement
43	Relaxation of levator palpebrae superioris;orbicularis oculi, pars palpebralis	Eyes closed
44	Orbicularis oculi, pars palpebralis	Eyes squinted
45	Relaxation of levator palpebrae superioris;orbicularis oculi, pars palpebralis	Blink
46	Relaxation of levator palpebrae superioris;orbicularis oculi, pars palpebralis	Wink

An extension to the Cohn-Kanade [15] was created by Lucey et al. called the extended Cohn-Kanade (CK+) database in [132]. Following a similar protocol setup in [15], an additional 107 sequences from 26 participants with posed facial expression having a duration between 10 to 60 frames were created. Non-posed expressions were extracted from the original database from around 84 subjects. The criteria for the non-posed expression were to start and end with a neutral face, perform the gesture without having been instructed and, not to have any occlusion or motion. In total 122 smiles were extracted from 66 participants.

Emotion	Criteria
Angry	AU23 and AU24 must be present in the AU combination
Disgust	Either AU9 or AU10 must be present
Fear	AU combination of AU1+2+4 must be present, unless AU5 is of intensity E then AU4 can be absent
Happy	AU12 must be present
Sadness	Either AU1+4+15 or 11 must be present. An exception is AU6+15
Surprise	Either AU1+2 or 5 must be present and the intensity of AU5 must not be stronger than B
Contempt	AU14 must be present (either unilateral or bilateral)

In [133], the MMI Face Database was presented. The database consists of more than 1500 of both static images and image sequences obtained from 19 participants from different ethnic backgrounds between the age of 19 and 62. 44% of the participants were female. During the data capture, the participants were instructed to perform 79 series of expressions consisting of either single action unit or multiple action units. The sequence started and ended with a neutral face. The resolution of the static images was 720 by 576 pixels and were coloured. The images consisted of 600 frontal images and 140 dual-view images i.e. both the frontal and the profile view of the face was visible with the use of a mirror. From the images sequence, there were 60 profile view and 750 dual view video sequence captured at 24 frames per second and having a length between 40 and 520 frames. Also, two FACS coders were used to code the images.

The is no database available with the head pointing in different directions, facial gestures and depth data. The advantage database [15, 132, 133] is that the data was collected from a large number of participants with different ethnic background and the images have been annotated with AU which makes it easy to find images with a specific gesture. The MMI Face database [133] had only frontal or profile images and also it contained only one full blink gesture (tested in 2013). In the case of head tracking, the profile images are not usable as the face of the user is not pointing at the screen. The ORL face database [48, 49], also contains frontals faces with the head pointing in different directions but it does not have depth data or facial expressions. The ZJU Blink Database [16] consists of 80 videos each with a number of blinking gestures being performed by participant. In a number of the videos, the eyes were obstructed by glasses and the database does not have depth data. The Cohn-Kanade [15] database consists of frontal images of people performing various expressions including eyebrows movement whereas [132] also include the gestures being performed at different facial poses but both database lack depth data.

With RGB-D sensors such as the Kinect [14] now readily available, the creating of an RGB-D facial expression database similar to [15, 132, 133] consisting of both colour(RGB) and depth data, with the participants performing timed gestures with different head pose is needed.

Chapter 3

Experimental Framework

In this chapter, the experimental framework for the evaluation of assistive devices is discussed. The contribution discussed in this chapter is the novel evaluation framework proposed based on the Fitts' Law [64, 134]. It is used to perform the evaluation of assistive devices (both existing and the ones being developed). And the experimental setup is also discussed. It includes a detailed discussion on the evaluation of the assistive devices used in Chapters 4 and 5. The hardware and software used both for development and evaluation are also discussed in sections 3.1 and 3.2 respectively. In section 3.1, the different devices such as the SmartNav [8], the Logitech Web camera [135] and the Kinect for Windows [14] are discussed. In section 3.2, the software based assistive device - CameraMouse [59] and the software used for the development are discussed. In section 3.3, the proposed evaluation framework based on the modified Fitts' [64, 114, 134] Test is discussed. Section 3.4 contains the set up of the different devices and the procedure for the experiment. The description of the stimulus presented to the participants and both the objective and subjective evaluation carried out during the experimentation is described in section 3.5. Section 3.6, contains information on the different data capture scenarios proposed. The discussion and conclusion of this chapter are presented in sections 3.7 and 3.8 respectively.

3.1 Hardware

In this section, the hardware such as the different sensors and devices used is discussed.

3.1.1 Sensors

3.1.1.1 Web camera

Figure 3.1 Logitech Pro 9000 web camera [135]

The sensor used for this project was a web camera from Logitech. The web camera used was a Logitech Pro 9000 [135]. It was selected as it can be used to create high-definition video (up to 1600 by 1200 pixels) and can capture video up to 30 frames per second. This device is low cost and easy to set up. It can be used with any Windows machine and can be configured to zoom in when capturing images.

3.1.1.2 SmartNav

SmartNav [8] uses infrared light and requires the use of an infrared reflective dot. The LED on top of the sensor emits infrared light. The reflective dot reflects the infrared light back to the CMOS sensor. SmartNav pre-processes the data and converts it into a video signal. The video signal is pre-processed by the SmartNav software on the desktop to threshold the image using either a pre-set value or a value set by the user. The blob obtained is tracked by the software and used to move the cursor on the screen.

(a) (b)

Figure 3.2 (a) SmartNav sensor (b) Infrared Reflective dots used with SmartNav sensor. [8]

3.1.1.3 RGB-D sensor

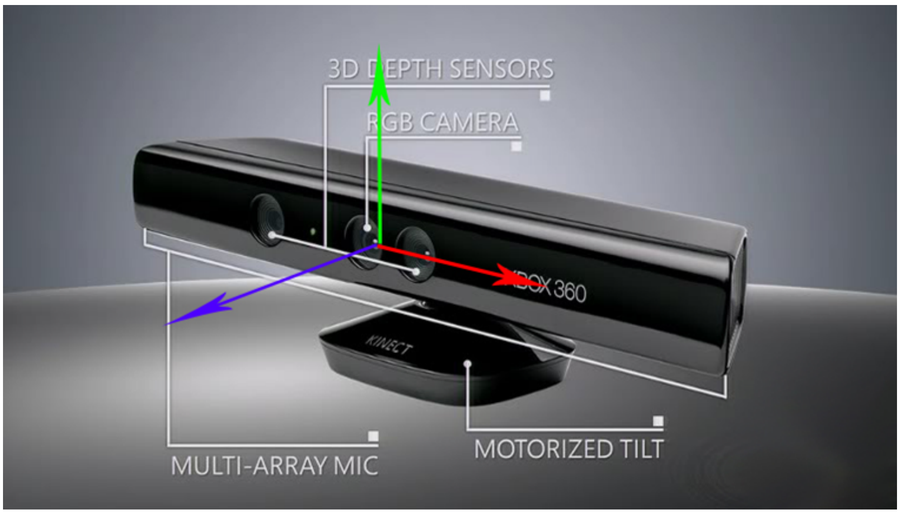


Figure 3.3 Kinect for Windows (Version 1)

The Kinect [14] consists of both a depth camera and an RGB camera. The Kinect for Windows version 1 is used in this research. The depth camera contains a near infrared projector which projects a known pattern (sparkles) in near infrared light, and an infrared receptor which captures the reflected projected lights. The depth map is calculated by using the calibration data of the sensor. During the calibration process, a check board is held at different known distances and the reflection pattern recorded. The depth is calculated based on the disparity of reflected patterns with the reference patterns obtained during calibration. The RGB image is 640 by 480 pixels and the depth image also has a resolution of 640 by 480 pixels. The Kinect sensor operates in 2 different modes - the near mode and the default mode. In the near mode, the range of the sensor is 0.4m to 3m and in the default mode the range of the sensor is 0.8 meter to 4 meters.

3.1.2 Computer

The computer used for the development and testing of the application was a 2.33 GHz Core 2 Duo laptop with 3 GB of RAM. The monitor used had a resolution of 1680 by 998 pixels. The computer specification was the standard specification of the computers in the lab.

3.2 Software

3.2.1 CameraMouse

CameraMouse [54, 59]is a vision based tracking system. The system can track features on the user's body and the feature is selected by the user. To track the movement of the head, facial features such as the nose is selected to be tracked. Once a facial feature has been selected, optical flow is used to track the movement of the feature. CameraMouse is used as an interface system with different assistive devices and software such as Midas Touch [59], Dasher [112] etc. are included to highlight the use of HCI and assistive devices

3.2.2 Development tool

For the implementation of this project Visual Studio 2010/2011 was used as the IDE [136]. The language used was C#. The EMGU CV library [137] which is a C# wrapper for OpenCV [138, 139] was also used. To be able to use the EMGU CV library, the source code for EMGU CV had to be built using the .NET 4.0/4.5 Framework and the HeadTracker was built using the .NET 4.0 Framework. The decision of using these .NET and the Microsoft Windows platform was based on the fact that the Kinect [14] device and Software Development Kit (SDK) are also primarily aimed at the Microsoft Window platform. Matlab [140] was also used to analyse the generated data. The 2011 version of Matlab was used.

3.3 Evaluation Methodology

The main goal of the proposed evaluation framework is to evaluate the performance of pointing devices. Fitts' originally proposed a method to model the human hand movement in order to improve human-machine interactions [64, 134]. Each task has an Index of Difficulty (ID) which is based on the size of the target and the distance of the target from the starting point. The evaluation framework used is based on Fitts' Test. In [114], Mackenzie suggested seven recommendations on the use of Fitts' Law models for the comparison of pointing devices. The seven recommendations are:

1. Use the Shannon [141] formulation of the ID. In Shannon's Theorem 17, the effective information capacity C (in bits/second) of a communications channel of bandwidth B (in s^{-1} or Hz) is expressed as a ratio of the signal power (S) and the noise (N) as shown in Equation (3.1):

$$\begin{aligned} C &= B \times log_2(\frac{S+N}{N}) \\ &= B \times log_2(\frac{S}{N}+1) \end{aligned} \tag{3.1}$$

The original formulation proposed Fitts'[64] is shown in Equation (3.2).

$$ID = log_2(\frac{2A}{W}) \tag{3.2}$$

Equation (3.2) represents the ID of a task where ID is equivalent to the effective information capacity of the channel C in Shannon's Theorem 17. The ID is represented as the ratio of the amplitude (A) of the movement and the width (W) of the target is used. The amplitude replaced the power of the signal (S) and the width replaced the noise (N) in Equation (3.1). The amplitude was multiplied by 2 and the +1 was removed. When Mackenzie [114] adapted Fitts' Law to be used with HCI, ID was redefined as shown in Equation (3.3). Similarly, to Equation (3.2), the ratio of the amplitude of movement (A) and width of the target was used but also the +1 was added again.

$$ID = log_2(\frac{A}{W} + 1) \tag{3.3}$$

2. The range of amplitude of the movement and width of the target should be chosen so that the resulting ID is between 2 to 8 bits.

3. A measure of subjects' movement end-points must be gathered, where obvious outliers - such as double-clicks - may be removed from the data.

4. The end-points i.e. the location where the user performed the selection task, should be used to adjust for accuracy. The adjustment is done by using the effective target width (W_e) and the effective index of difficulty (ID_e).

5. Use least-squares linear regression to fit the data and to find the intercept and the slope of the Fitts' Law equation. This is used to find the measure of the goodness of the fit.

6. Calculate the predicted movement time using the regression model and the ID.

7. Calculate the throughput (TP).

The parameters added to those proposed by MacKenzie [114] in this experiment are in Table 3.1.

Table 3.1 Variables captured during Fitts' Test using formula from [142]

Variable	Description
UserID	Identification number of the users performing the test
SessionID	Identification number of the session
DeviceID	Identification of the device being used
DeviceRank	The order in which the devices were tested
Orientation	The orientation of the task
ID	Index of difficulty of the task
TimeToTarget	The amount of time it took to reach the target
TimeToCompleteTask	The amount of time it took to complete the task and complete the task

3.4 Experimental Set-up

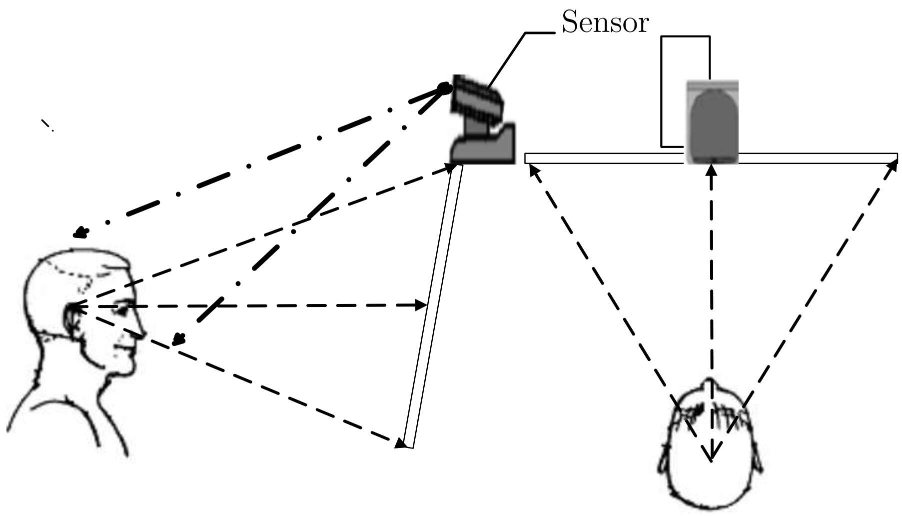


Figure 3.4 Experimental set-up.

The participant was asked to perform a series of Fitts' Tests [64, 142]. The Fitts' Test was used to evaluate two devices: a 2D vision-based head tracker using the Logitech web camera and a 3D head tracking system using the Kinect device. The experiment was performed using the two facial gestures (blink and eyebrows movement) as a switching mechanism.

Figure 3.5 Participant performing experiment

Figure 3.6 Equipment setup in lab

3.5 Device Evaluation

The devices evaluated in this work incorporate a camera and an algorithm for tracking the head movement and detection of the eye blink or eyebrow movement facial gestures. The camera is either the Microsoft Kinect for Windows [14] sensor which can provide 3D (RGB and depth) data or a Logitech web camera which can only provide 2D (RGB) data. Raw data is extracted in the form of images and depth maps. The efficacy of head tracking and gesture recognition is compared for the 3D and 2D vision-based systems using a modified Fitts' test.

3.5.1 Stimuli

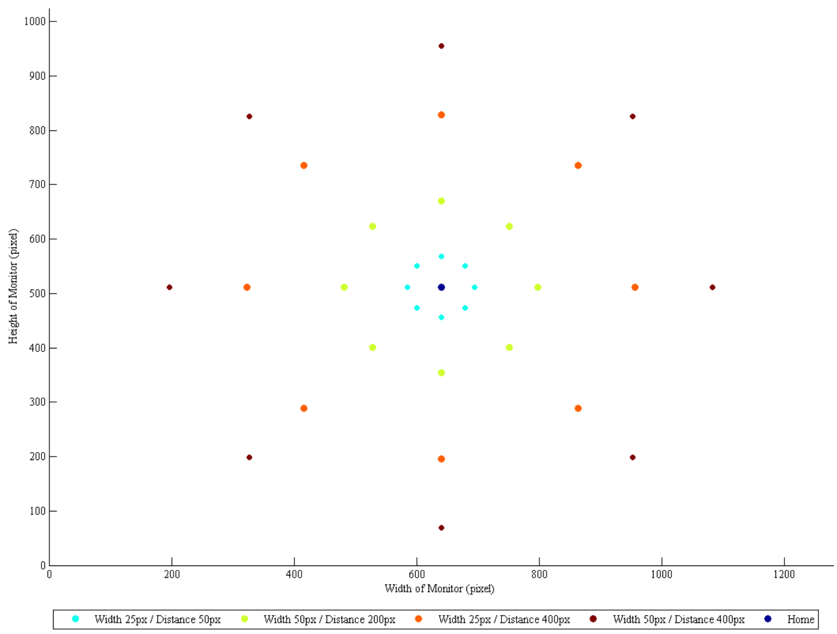
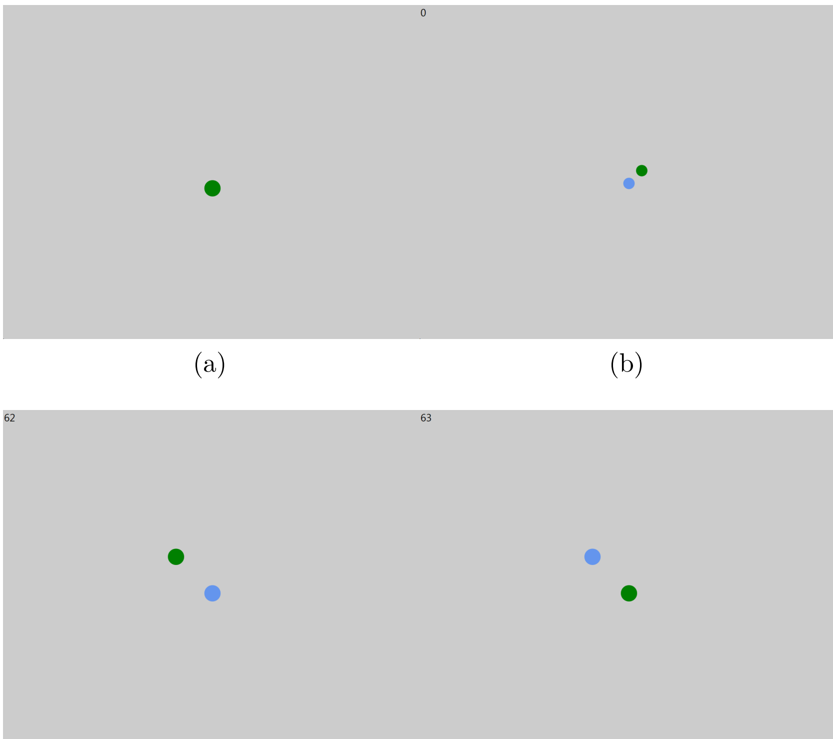


Figure 3.7 Constellations of target locations

The stimuli used are based on the multi-directional Fitts' test specified in [65]. The Home location is placed centrally as it is assumed to be the default location where users would position their head pointer. Using the Home location approach in the Fitts' Test would double the number of data collected i.e. 64 (16x4) data points as opposed to 32(8x4) data points with an eight point Multi-Directional task and thus increase the accuracy of the data being captured. The targets are selected randomly and shown to the participants by our application. The random selection of the targets helps to simulate a real world scenario where a user would be doing different actions with different ID.

The participants have to move the cursor to the central home location, and once a click is detected, a target location is displayed to the participant. The participant has to move the cursor to the location of the target once again and perform a Dwell click - clicking action performed by holding the cursor still over a target for a predefined length of time. Eight different targets, each at a different orientation for each target size (W) and distance (D) are displayed to the participant. In Figure 3.8, four of the different locations selected during a session are shown (see Appendix B for all the locations)



(a) (b) (c) (d)

Figure 3.8 Example of the location of the Stimuli

3.5.2 Objective Performance Measure

3.5.2.1 Fitts' Test

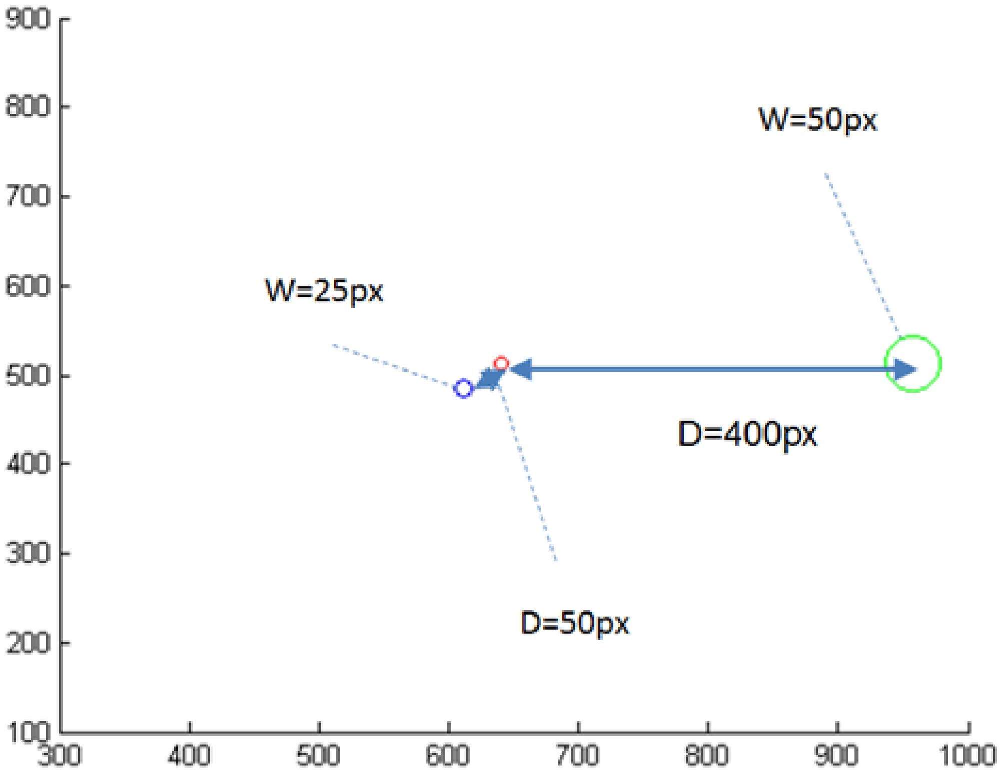


Figure 3.9 Example of target and starting point

Fitts' originally proposed a method to model the human hand movement in order to improve human-machine interactions [134]. Each task has an ID which is based on the size of the target and its distance from the starting point. The ID represents the cognitive-motor challenge imposed on the human to accomplish the task and is measured in bits as shown in Equation (3.4).

$$ID = log_2(\frac{D}{W} + 1) \tag{3.4}$$

where D represents the distance from the starting point to the target and W is the width of the target. Figure 3.9, shows an example of the target and the starting point.

$$MT = a + b \times ID \tag{3.5}$$

The relationship between MT and ID shown in Equation (3.5) is a linear one where a is the y-intercept i.e. where the line cuts the y-axis and b is the gradient or slope of the line.

The Index of Performance (IP) in bits/second of a device is given in Equation (3.6).

$$IP = \frac{1}{b} \tag{3.6}$$

where b is the gradient of the line described in Equation (3.5). A positive value of IP indicates that the device gets more difficult to use as the interaction becomes more challenging.

$$ID_e = log_2(\frac{D_e}{W_e} + 1) \tag{3.7}$$

ID_e,is the effective index of difficulty, in bits, and is calculated from the distance (D_e) from the start location to the target and W_e, the effective width of the target. W_e,is the effective width of the target and it is calculated from the observed distribution of the target selection coordinates [116].

$$W_e = 4.133 \times SD \tag{3.8}$$

where SD is the standard deviation of the selection coordinates. As the test are multi-directional, SD is calculated by using the Euclidean distance between the point where the click action occurred and the mean point $(\overline{x}, \overline{y})$. Equation (3.9) shows the calculation for SD.

$$SD = \sqrt{\frac{\sum_{i=1}^{n}\left(\left(x-\overline{x}\right)^2+\left(y-\overline{y}\right)^2\right)}{n-1}} \tag{3.9}$$

Equation (3.10) represents the TP_e in terms of ID and MT and Equation (3.11) is used to calculate the TP_e in bits/second.

$$TP_e = \frac{ID_e}{MT} \tag{3.10}$$

calculated as follows in our application.

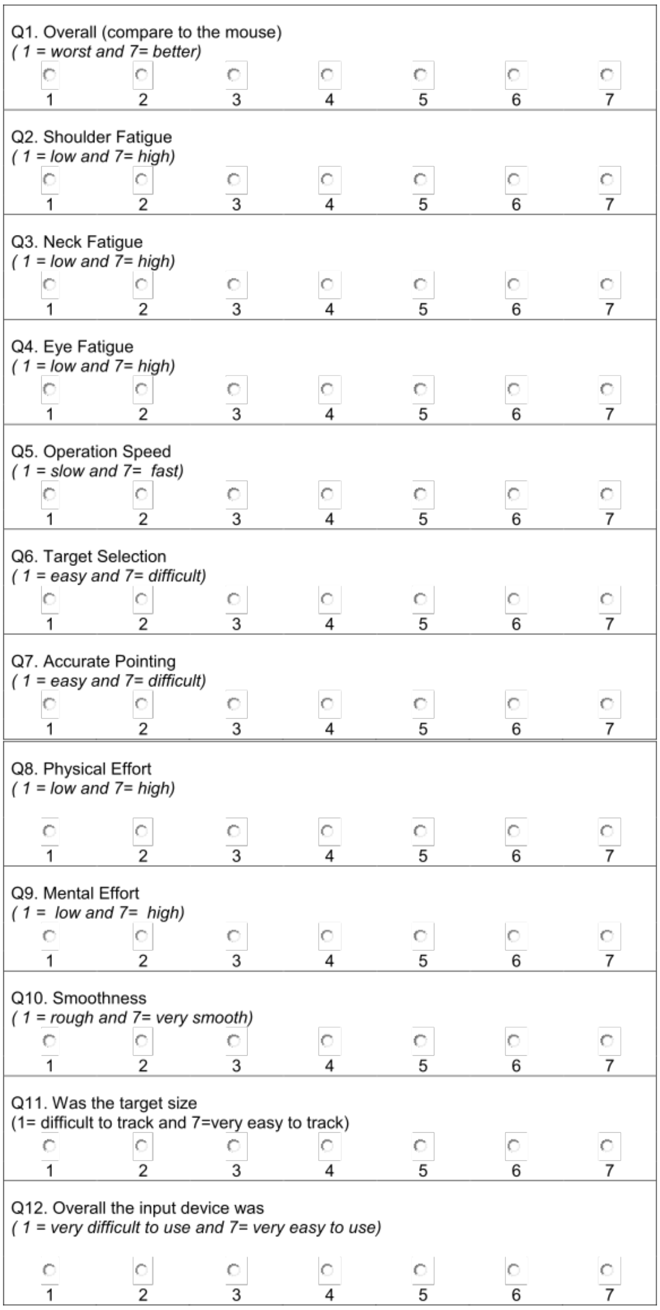
$$TP_e = \frac{1}{y}\sum_{i=1}^{y}\left(\frac{1}{x}\sum_{j=1}^{x}\frac{ID_{eij}}{MT_{ij}}\right) \tag{3.11}$$

where MT is the mean movement time, in seconds, for all trials (j) within the same condition (i). The condition can be for type of device or user depending on the outcome required. For example in our case, we want to calculate the TP_e of each device for each ID. It represents the overall efficiency of the device in facilitating interactions.

The experiments showed that the rate of performance of the human motor system is approximately constant over a wide range of movement amplitudes. Fitts' Law [64, 114] states that MT should increase with an increase in the ID i.e. as the difficulty of the task increases, the time to complete the task also increases. Fitts' Law was adapted in Mackenzie et al., to assess HCI devices [114]. Therefore, it is thought Fitts' test is an appropriate tool for assessing the performance of the head tracking and gesture recognition system.

3.5.3 Subjective Performance Measures

The performance measurements were supplemented with a user survey to establish the usability of the different devices. The questionnaire used in the survey is based on the Device Assessment used by Douglas et al. [116]. Participants answered the survey after they had completed the experiment with all four devices. The purpose of the questionnaire is to obtain the perceived performance of the different systems from the participants' perspective. The questionnaire consisted of twelve questions. Question 1, is looking at the comparability of the devices with a standard mouse. Operational speed, target selection and accuracy are addressed by Questions 5 to 7. Issues of fatigue and comfort are addressed in Questions 2 to 4 and 8 to 9. Moreover, the usability is addressed in questions 10 to 12.



Q1. Overall (compare to the mouse)
(1 = worst and 7= better)
1 2 3 4 5 6 7

Q2. Shoulder Fatigue
(1 = low and 7= high)
1 2 3 4 5 6 7

Q3. Neck Fatigue
(1 = low and 7= high)
1 2 3 4 5 6 7

Q4. Eye Fatigue
(1 = low and 7= high)
1 2 3 4 5 6 7

Q5. Operation Speed
(1 = slow and 7= fast)
1 2 3 4 5 6 7

Q6. Target Selection
(1 = easy and 7= difficult)
1 2 3 4 5 6 7

Q7. Accurate Pointing
(1 = easy and 7= difficult)
1 2 3 4 5 6 7

Q8. Physical Effort
(1 = low and 7= high)
1 2 3 4 5 6 7

Q9. Mental Effort
(1 = low and 7= high)
1 2 3 4 5 6 7

Q10. Smoothness
(1 = rough and 7= very smooth)
1 2 3 4 5 6 7

Q11. Was the target size
(1= difficult to track and 7=very easy to track)
1 2 3 4 5 6 7

Q12. Overall the input device was
(1 = very difficult to use and 7= very easy to use)
1 2 3 4 5 6 7

Figure 3.10 Questionnaire to record experience of user while using the different devices [65, 116]

3.6 Data Capture

3.6.1 Ambient/Normal Environment

In this scenario, the data capture session was conducted in the ambient environment of our lab. During the session people were walking by and trying to look over the shoulder of the participant, thus adding background noise. The objective of this scenario is to test the effectiveness of the use of depth data to reduce noise by removing the background and to provide simulation of a workplace environment as close as possible to highlight issues that such devices and their users might experience in their daily life.

The motivation behind the creating the dataset is because this database captures the gesture being performed in different orientation. The Kinect which is a commercially available and low cost sensor was used to capture the data in both 2D and 3D. The gestures were timed, so it could be possible to generate simulated videos of gestures of variable duration using the images. The data was captured in a normal working environment.

3.6.2 Target Population

The targeted population was adults from 18 to 65 years old. The experiment was carried out on a population size of approximately 22 healthy participants and one person suffering from MS. Participation to this experiment was voluntary.

3.6.3 Scenario

The first step in the data collection process, the participants were given both the Participant Information Sheet and the consent form to read. They were encouraged to ask the investigator any questions they might have or clarifications they require. Once the participants were comfortable with the data capture and understood both the consent form and the Participant Information Sheet, they were asked to sign the consent form, the data capture process could begin.

If the participant had opted to provide image or video data in the consent form, the images of the participants performing the gestures at different orientations were captured as specified in Section 3.6.3.1. The second data capture was to capture the participant using the proposed assistive devices. The gesture image capture was performed first in order to get the participants used to the system. If the participants did not opt out of performing the gesture image data capture, then they were given at least three practice runs on the devices, so as to enable then to get familiar with the evaluation system.

3.6.3.1 Gesture Image Data

For each participant, each gesture was collected five times at 1000ms time intervals with ambient lighting or variable lighting. The gestures were collected for the frontal face where the stimulus of the gestures appeared in the centre of the screen. The stimulus or target was moved along different orientations around the central location and the gesture was captured at each location. For facial gestures such as eye blink, an audio cue in the form of a computer beep was generated to inform the users to open their eyes after the eye blink specified duration of 1000ms had passed.

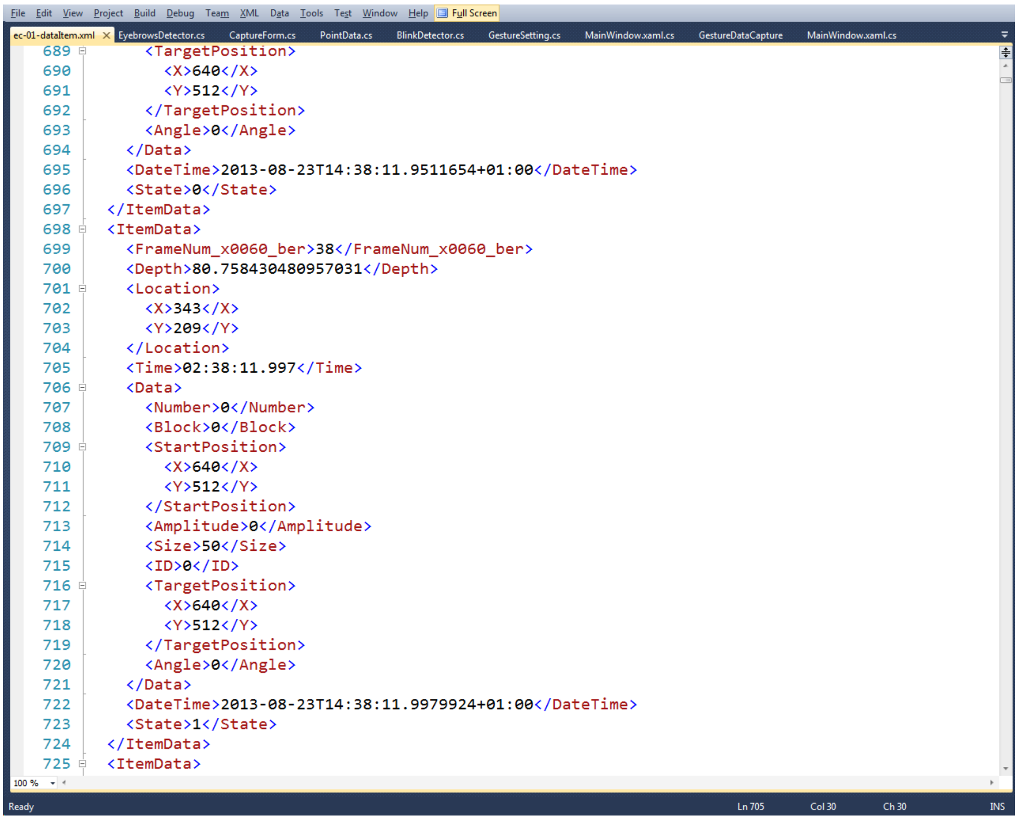


Figure 3.11 Example of XML file

For each gesture capture session an RGB image, a depth map data, and in an XML data file was captured at regular intervals. A data file in the format of an XML file was also stored for each gesture captured. The file contained the id of the user, the gesture being captured, the sequence of the gesture, the time interval, the coordinates of the maximum depth, and the minimum depth.

Kinect sensor generates an RGB image and depth data at a rate of 30 frames/second. Both the RGB and Depth data had a 640 by 480 pixel size. The raw data was captured with features such as skeletal tracking disabled. The RGB and Depth data were aligned i.e. the RGB image provided the pixel value of a specific pixel and the Depth data provided the corresponding depth value of the specific pixel. A queue was used to store the data files to disk and a Solid State Disk was used so as to be able to store the data with minimum time required.

The images data collected along with the XML file are used to train the SVM classifier to detect the different gestures. The data in the XML file represents the ground truth of when the gesture was performed.

A total of 855,391(205.4GB) images consisting of both RGB and Depth data files were collected from 22 participants performing eye blink, eyebrows movement, tongue or mouth opening/closing. Only eight participants performed all four gestures, the others were either embarrassed or unwilling to perform the tongue or the mouth opening/closing gestures. But all participants performed the eye blink and eyebrows movement gestures.

3.6.3.2 Device Evaluation Data

For the evaluation of the different assistive devices, the participants were asked to take their time and try to select as many targets as possible. In order to prevent the participants from getting frustrated when they were unable to capture a target location, they were asked to try at least three times before abandoning the target and moving on to the next one.

During the data capture, no image or video data was collected. The only data collected were the start location of the task, the target location, the type of device used, the amount of time spent in a location, whether a click action was performed and whether the cursor was on the target. The data was used to calculate the time taken to move from the start position to the target location and the time taken to perform the selection action by clicking.

Once the evaluation was completed each participant was given a questionnaire to fill on his assessment of the devices he had just used. The participant was given a short 5 minute break before performing the next evaluation. In some cases, the participants opted to perform the data capture on a different day. This was allowed as long as the device assessment questionnaire was complete.

A total of 5184 Fitts' Test were performed by 22 participants. Three participants either did not complete or withdrew their consent; their data was deleted from the data capture.

3.7 Discussion

The proposed evaluation method consists both of an objective and a subjective evaluation. For the objective evaluation, the performance measurement is carried out by using a modified Fitts' Test. The Fitts' Test used is based on the Multi-Directional task with circular targets and a central location called the Home location. The Home location is assumed to be the default location where users would position their head pointer. Using the Home location approach in the Fitts' Test would double the number of data collected i.e. 80 (16x5) data points as opposed to 40 (8x5) data points, with an eight point Multi-Directional task as the user would have to move from the target to the Home and thus increase the accuracy of the data being captured. The targets are also presented randomly to the participants by our application so as to prevent the user from anticipating the next target and to decrease the influence of a previous movement on the current one. The participants have to move the cursor to the central home location and once the click or selection action is performed, the next target location is displayed to the participant. The participant has to once again move the cursor to the location of the target and perform the clicking action. Eight different targets, each at a different orientation for each target size (W) and distance (D) are displayed to the participant. This approach also has the added advantage of mimicking or simulating a normal interaction on a computer.

The subjective evaluation is based on [65]. The questionnaire has been adapted to clicking using gestures by adding question 2 to 4. As the devices are head trackers and the users would be performing clicking action using facial gestures, it would be interesting to see if the use of these gestures would affect the user. Question 2 and 3 are aimed at the use of the head tracking device and Question 4 is targeted to the users using the blink gesture.

3.8 Conclusion

The contribution of this chapter is the evaluation framework based on a modified Fitts' Test. The framework consists both of an objective and a subjective evaluation of the device. The modified Fitts' Test is the objective evaluation. The proposed test has a central location and the user has to move between the central location and the targets. The targets are also presented to the user at random to prevent the user from anticipating the next target location. The test captures the time taken by the user to perform the task but also how long it took the user to first reach the target. The time taken by the user to reach the target provides us with a better understanding of the difficulty in performing the clicking action and as a result the difficulty in performing the gestures.

Chapter 4

Head Tracking and Gesture Detection

In this chapter, the development of the headtracker and the gesture detection algorithm are discussed. The contribution discussed in the chapter is the development of the head trackers (both 2D and 3D) and the gesture detection system used as switching mechnism for the assistive devices. The devices are evaluated based on the experiments discussed in Chapter 3. This chapter provides the details on the development of the head tracker and the experiments carried out to evaluate the effectiveness of the head tracker. The performance of the head tracker is also compared against the existing devices discussed in Chapter 3. For comparison against SmartNav [8] and CameraMouse [59], a Dwell clicking switch is used. Both eye-blink and eyebrows movements switches are used to compare against the 2D head-tracker. It was found that the two commercial devices performed better than our head-trackers. By using the depth data, it was observed that the TP_e of dwell clicking increased by a third (from 0.21 to 0.30 bits per second) and that of blink clicking doubled (from 0.15 to 0.28 bits per second). The eye blink detection algorithm performance was similar to that of the Dwell click switch using depth. The eyebrows detection algorithm had the worst performance. The performance of the gesture detection algorithms used in Chapter 5 is investigated. It was found that there was no sufficient evidence at a 95% confidence interval level to point to a relationship between the orientation in which the gesture is performed and the GER of the algorithms. Thus, this might indicate that the detection algorithms are robust to change in orientation. It was also found that the gesture being performed has a moderate effect on the error rates but no significant effect was found based on the sensor used to capture the gesture. It was also found that there might be significant evidence of a relationship between the error rate of the gesture and the *ID*.

In Section 4.1 contains information on the development of the head tracking and the gesture detection system. The different technologies used to track the head and detect the gestures - eye blink and eyebrows movement - are presented. In Section 4.2, the results are analysed and discussed. Finally, Section 4.4 contains the conclusions of the investigation.

4.1 Design

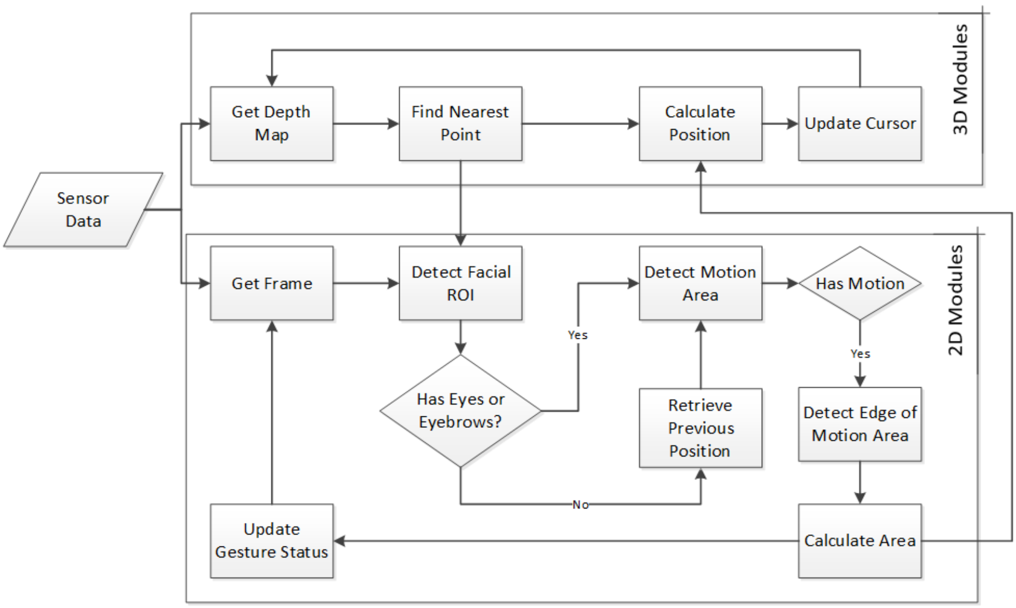


Figure 4.1 Algorithm to detect Blink and Eyebrows Movements.

Figure 4.1 shows an overview of the facial gesture recognition system, which is the same for both the 2D vision and the 3D Kinect system. Depth data is used only to filter the region of interest when processing the facial image - only objects within a meter of the 3D sensor were included in the region of interest, and all other background was removed before further processing. The facial gesture recognition system uses only the RGB data from the sensors. Thus, these algorithms works with both the Logitech [135] web camera and the Kinect [14]. Facial areas of interest such as the head, eyes region, left eye and right eye are detected using a Haar-Cascade [24].

4.1.1 Face Detection and Facial Feature Extraction

In this section, the face detection and facial feature extraction techniques used are discussed. The facial features extracted are used for both head tracking, eye tracking, and facial gesture recognition.

Viola Jones [24] contains classifiers for all the features required by the application. A cascade is configured to detect the face. In the case where there are multiple faces detected, the largest face is selected. This is because the head tracker is currently tracking the head of only one user at a time, and if there are more than one faces detected, then the nearest one being the largest is automatically selected.

Once a face has been detected, the Viola Jones [24] algorithm is used to search for an eye pair and the nose. When an eye pair region is detected, the region is divided into two equal sections. The left section is the search region for the left eye and the right section is the search region for the right eye.

If the location of the eye region is not found, the image is searched for the location of both the right and left eye.

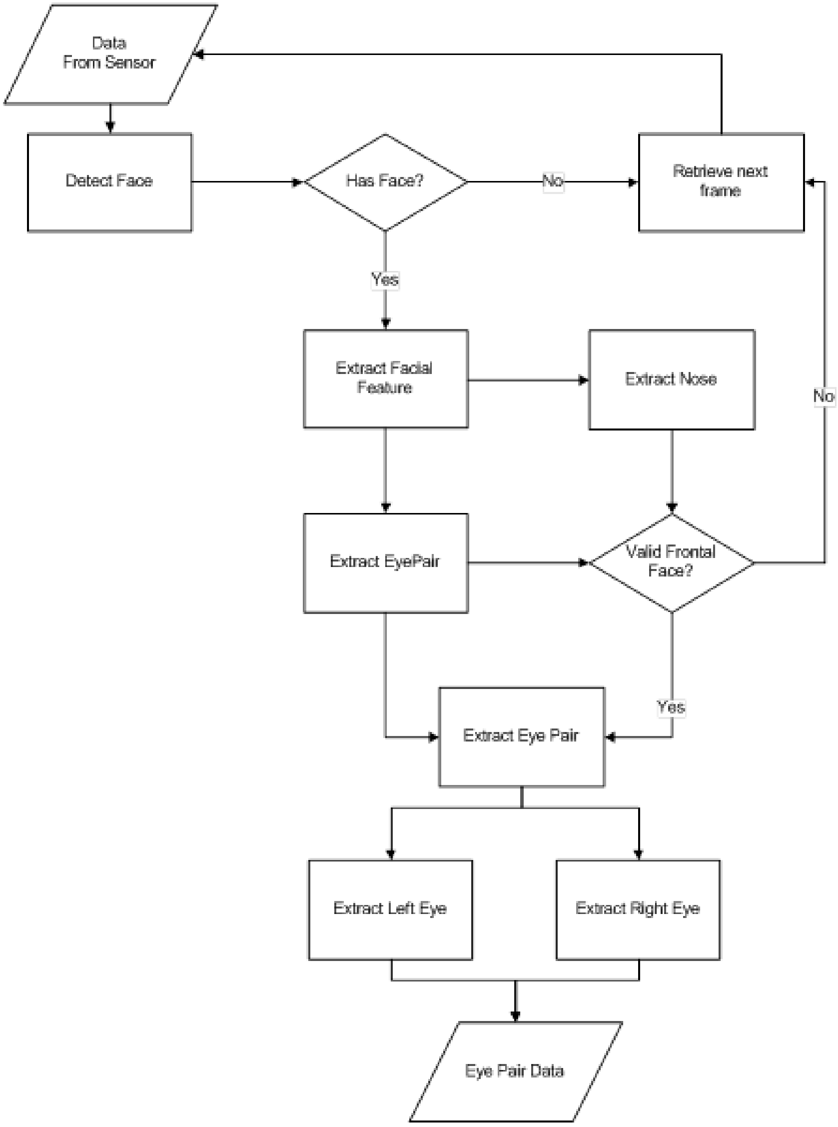
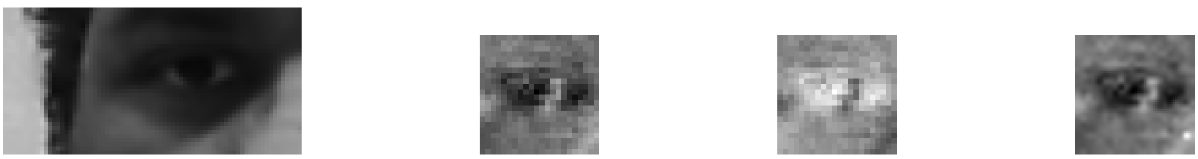


Figure 4.2 Facial Feature Extraction flow diagram.

4.1.2 Eye Tracking



(a) (b) (c) (d)

Figure 4.3 (a) Grey-scale Right eye Image (b) Saturation of eye in HSV (c) Area to segment eye region (d) Dark Pupil location.

Figure 4.2 (a) Grey-scale Image (b) Detected facial features (c) Face region (d) Eye pair region (e) Left eye (f) Right eye.

The eye tracking module is made up of different components as it can be seen in Figure 4.4. The aim of the eye tracking module is to track the eyes of the user and to calculate the interocular distance. The first component of the eye tracking module is face detection.

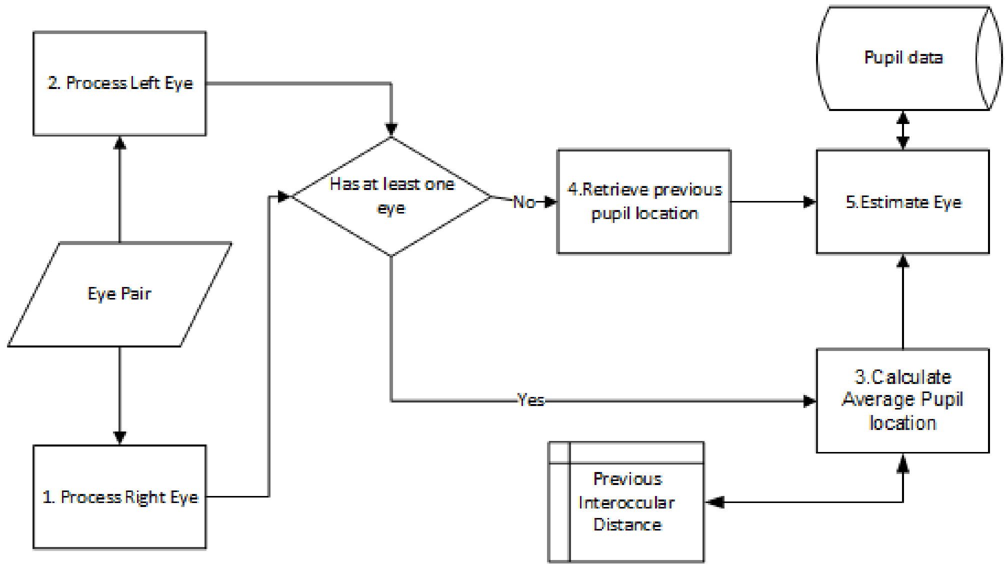


Figure 4.4 Eyetracking

The face detection component is used to verify that the face of the user is in the frame of the camera. If the face is present, then the second component, the facial feature extraction component can be used. The facial features extraction component is used to find the exact location of the facial features such as the eye, nose and eye region on the face. The positions of the iris and pupil are used to estimate the direction the user is looking towards. Both eyes are used in a view to create a more robust and accurate system. The location of the nose is used to determine if the face has moved between the time the last frame was processed and the current frame. If no head movement has been detected, then when processing the next frame, the locations of the eyes are assumed to be the same. This is used in order to reduce the processing time required, and thus produce a highly responsive system. Table 4.1 represents the data collected during eye tracking.

Both for eye tracking and also to verify the presence of an eye, the location of the pupil has to be located. The image of the extracted eye is obtained from the Detect Facial ROI module. The region retrieved is converted into the HSV colour space. The HSV image has 3 channels. The first channel is the Hue, the second is the Saturation channel and the third is the Value channel. To detect the pupil, the second channel is used. The saturation image is negated so as to make the pupil the darkest region of the image. Noise is removed from the image using morphological closing operation. The image is binarised using the minimum intensity of pixel in the processed image. Once this is done, contour of the darkest region is located on the image. Contour was used instead of using Hough transform to identify the pupil because the shape of the pupil is not a perfect circle, and also due to the location of the sensor relative to the face, the pupil would not always be a circle.

Table 4.1 Data collected while eye tracking

#	Field Name	Description
1	Frame Number	Unique incremental number
2	Time	The time the eye location was captured
3	Interval	The interval time used by the sensor
4	X coordinate of left pupil	The X coordinate of the centre of the left pupil
5	Y coordinate of left pupil	The Y coordinate of the centre of the left pupil
6	Tracking status of left pupil	The tracking status of the left pupil (0 -not tracked,1 - tracked)
7	X coordinate of right pupil	The X coordinate of the centre of the right pupil
8	Y coordinate of right pupil	The Y coordinate of the centre of the right pupil
9	Tracking status of right pupil	The tracking status of the right pupil (0 -not tracked,1 - tracked)
10	Type of data	The type of the data
11	Size of left pupil	The size of the left pupil

Continues on next page

Table 4.1 – Data collected while eye tracking(*Continued from previous page*)

12	Size of right pupil	The size of the right pupil
13	Interocular distance	The distance between the left and right pupil of the user
14	Number of white pixel in the right eye	The number of white pixel in the right eye
15	Number of white pixel in the right eye	The number of white pixel in the left eye
16	Area of right eye	The area in pixel of the right eye region
17	Area of left eye	The area in pixel of the left eye region

4.1.3 Blink Detection

As defined in Section 4.1.7, a blink is defined as the transition from open eye to close eye and to open eye again. The eyes close for a period of 1 second or more and then revert to the *Open* state. If the closure of only one eye is detected when two open eyes have been detected, the system assumes there is no blink. If only one eye has been detected due to the pose of the head, then a blink event is detected if only the detected eye is closed for a period of 1 second or more and then reopened.

The data captured during eye tracking in Table 4.1 is used to track the state of the eye. When the method is initialised, it is assumed that the state of the eye is set to *Open*. The Detect Edge of Motion Area and Calculate Area modules are used to detect the contour of the largest object in the region of the eyes and to calculate the area of the object. So if a motion is detected in the area of an eye or the eye pair area, the width, height and area of these regions are re-calculated. If the tracking status of any of the facial features is 0 i.e. the facial feature has not been detected, then the previous location of the feature is used but the interocular distance is not changed. These features are used with the SVM classifier. The SVM classifier was trained with about 20% of the initial data captured during our data capture sessions. The images with no blink gesture i.e. an open eye, were labelled as the negative class and images with the blink gesture i.e. a closed eye were labelled as the positive class. From preliminary experiments, it was found that the best threshold for the classifier was 90% and 60% of the training data was required for training purposes. The RGB and Depth data collected during the data capture in section 3.6.3.1 were used to train the classifier.

4.1.4 Eyebrows Movement Detection

In the case of the eyebrows detection, the states of the eyebrows are monitored by measuring the distance of the eyebrows points to the mid-point of the interocular distance, the centroid of the nose region and the centroid point of the face region or if the face is not detected the centroid of the area closer to the sensor (4.9) is used. As defined in Section 4.1.7, only the transition from down to raise and then to the down state again is recognised as a valid eyebrows movement. The eyebrows region is detected using the location above the eye region. The state of the eyebrows is initially set to *down*. To recognise eyebrows movement, both eyebrows have to be raised for a period of 1 second or more and subsequently return to the down state. In cases where only one eyebrow is detected due to the pose of the user, the raised and downwards movement of the detected eyebrows is sufficient to generate a click action.

4.1.5 Gesture Algorithm Evaluation

Ground truth data is required to generate a Confusion Matrix (4.11). The ground truth data is collected during the experiment in the form of *ExpectedOutcome* and *ObservedOutcome*. When the cursor is within the target region, a gesture is expected and thus the *ExpectedOutcome* is set to 1. When the cursor is outside the target area, the *ExpectedOutcome* is set to 0 i.e. no gesture is expected. When the user executes a gesture, the *ObservedOutcome* is set to 1. The default value for the *ObservedOutcome* is 0 i.e. no gesture is expected. By using the *ExpectedOutcome* and the *ObservedOutcome*, the Confusion Matrix is generated for each gesture, for each task and the *TruePositive*,*FalsePositive* and *Accuracy* rates are calculated based on [143].

4.1.6 Head Tracking

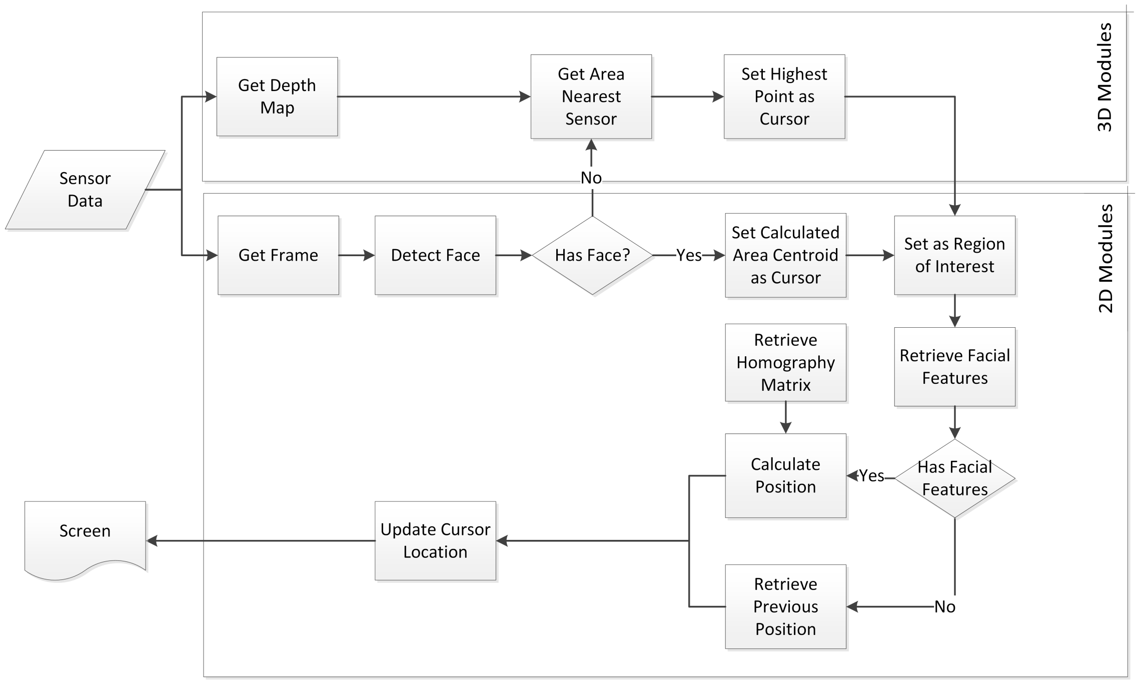


Figure 4.5 Algorithm to head tracking in both 2D and 3D.

Figure 4.5 shows an overview of the head tracking both the 2D vision system and the 3D Kinect system. The head is detected and tracked in subsequent frames using Haar-Cascade [24]. Depth data is used only to filter the region of interest i.e. nearest to the sensor when processing the facial image - only objects within a meter of the 3D sensor were included in the region of interest and all other background was removed before further processing. Using the depth data to find the region of interest reduces the area the Haar-Cascade [24] has to be applied to and thus reduces the amount of processing to be carried out. In cases where the head is not detected using the cascade, it is assumed that the area contains the head of the user but the location of the cursor does not change and the area is used for the detection of other facial features such as the eyes, nose etc.

The proposed system is non-invasive i.e. the user is not required to wear any additional markers such as a reflective dot [57] or have led light such as in [76]. The depth data is used to reduce the area the Haar-Cascade [24] has to process to find the facial features thus both boosting the performance of the head tracking and keeping computation to a minimum.

The calibration process used to map the location of the user's head to the screen location is discussed below.

4.1.6.1 Calibration

The calibration process is required to map the movement of the user to locations on the screen. This process is also used to adapt the system to the head movement of the user. If the user can only perform small head movements, these movements are mapped to the screen and thus enable the user to access the whole area of the screen. During the calibration process, the locations of the tracked features are stored along with the calibration point locations. The feature location and calibration point pairs are used to calculate the homography [144, 145] matrix. If more than three of the locations are collinear, an error message is displayed to the user and the user must restart the calibration process.

In the adaptive calibration process, the participants are asked to look at imaginary locations on the screen including the screen corners for a period of 10 seconds. This enables the participants to move their heads to the extent of their range of movement. The purpose of the head tracker is to enable navigation on the whole screen. Thus, it is assumed that the area with the highest positive feature detection is the best area to use in order to track the user.

The homography matrix is generated by using the average head location of the user during the calibration process. The homography matrix is used to calculate the head point of the user on the scene.

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} \\ h_{2,1} & h_{2,2} & h_{2,3} \\ h_{3,1} & h_{3,2} & h_{3,3} \end{bmatrix} * \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4.1}$$

where the x' and y' represent the coordinates of the projected point on a screen of the computer. x and y is the coordinate of the point on the head being tracked. And $h_{1,1}$, $h_{1,2}$,$h_{1,3}$, $h_{2,1}$, $h_{2,2}$, $h_{2,3}$, $h_{3,1}$, $h_{3,2}$ and $h_{3,3}$ represent the coefficient of the homography matrix.

Once the homography matrix H has been solved with head location and calibration point pair using Equation (4.1), the matrix can be used to estimate the screen point of the user. During the calibration process, more than one head location is captured. The mean of all the points captured is used to solve the homography matrix. Also, for the homography matrix to be solved all nine calibration points are not required, only four points are sufficient. In our case, we use the four corners of a rectangle delimiting the region where the feature is detected and the central point of the rectangle to generate the homography matrix.

Figure 4.6 shows an example of the calibration area using the web camera. The green rectangle represents the calibration area; the blue rectangle represents the area where motion was detected, and the red area is the region detected by Haar-Cascade [24] as containing the head of the user.

4.1.6.2 Depth Data

The depth data obtained from the Kinect sensor is used to reduce the search area for the different Haar-Cascade features. This reduces the computational load and avoids

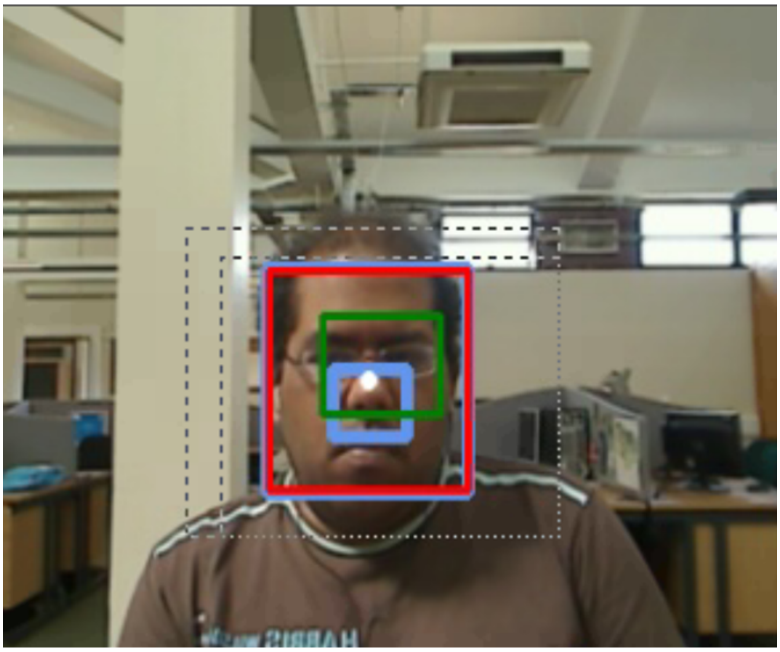


Figure 4.6 Snapshot of a user within the green calibration area.

background distractions, such as people, movements and changes in lighting and therefore increase the performance. A mask is created from the depth data and the object within 1000mm of the sensor is selected. The mask is used on the colour image to remove all the objects which are more than 1000mm from the sensor. An example is shown in Figure 4.7.

4.1.6.3 Depth Segmentation

In this approach it is assumed the user is within 1000mm of the sensor as stated in our experimental procedure. The first step is to capture both the RGB image and the Depth image from the sensor. Figures 4.7a and 4.7b show an example of images captured using the sensor.

For segmenting by depth, the mask is created with the pixels which are within 1000mm of the sensor. An example of the mask is shown in Figure 4.7c. The mask is applied to the RGB image and the result can be seen in Figure 4.7d.

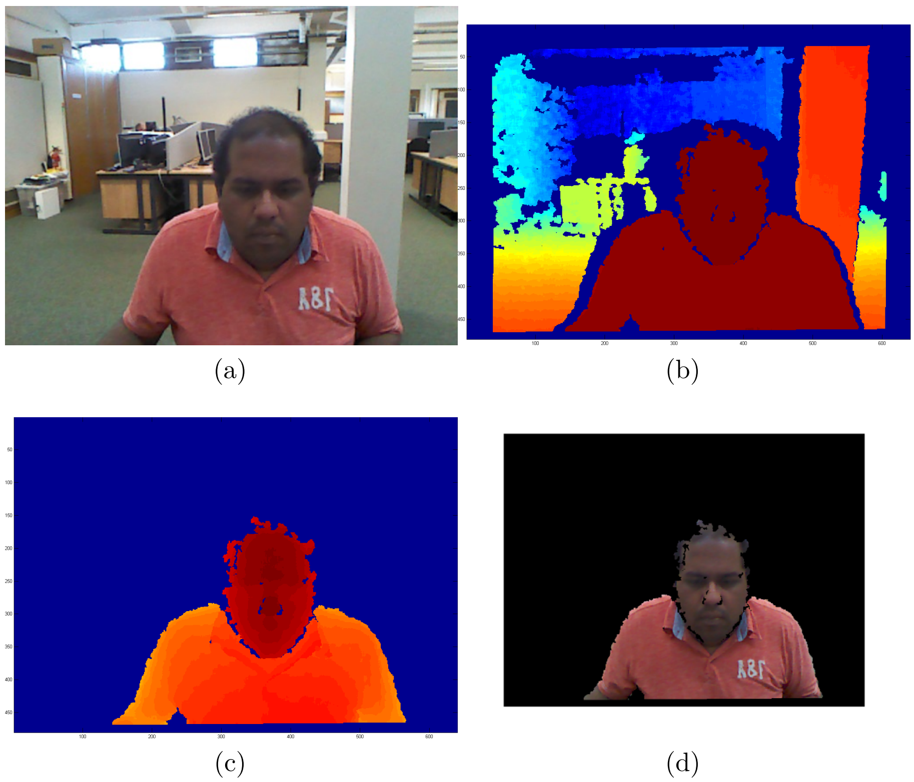
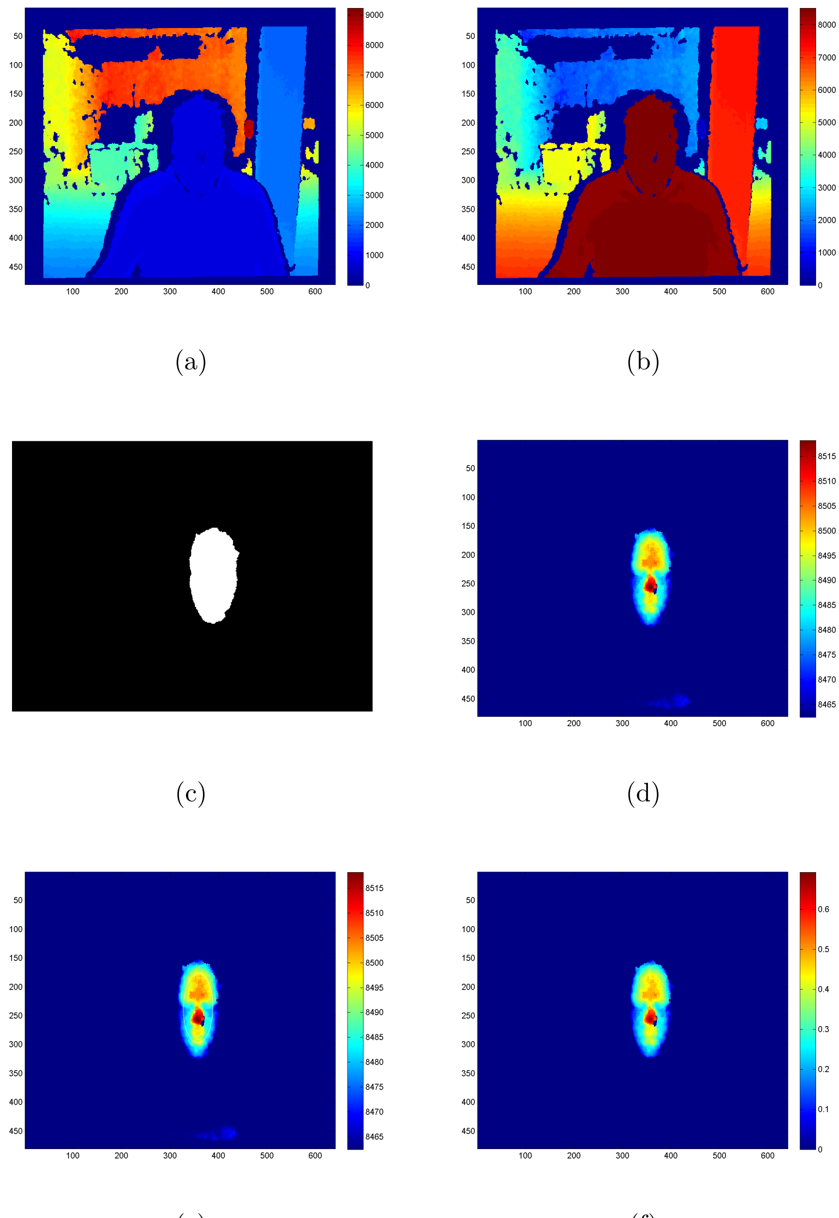


Figure 4.7 (a) Original colour Image (b) Original Depth data (c) Mask of 1000mm from the sensor (d) Output image.

The limitation of the fixed distance segmentation is that the torso area is also selected. An adaptive depth segmentation was implemented in order to overcome the limitation of the fixed distance segmentation. For this approach, the depth image has to be inverted. Inverting the image makes the background to zero and the points on the face get the maximum value. A Region of Interest is created in the depth image by selecting the area nearest to the sensor. The Region of Interest consists of the pixels which form part of the biggest object nearest to the sensor i.e. the head of the user. By applying contour on the Region of Interest, a mask is generated. An example of the mask is shown in Figure 4.9c. The mask can be applied to either the depth image or the RGB image. In Figure 4.9f, the depth image is normalised to scale the depth in the range 0 to 100.

procedure ADAPTIVE SEGMENTATION($depthdata$)
    $invDepthdata \leftarrow GetMaxDepth(depthdata) - depthdata$
    $maxDepth \leftarrow FindMaxinumDepth(invDepthdata)$ ▷ the point nearest to the depth sensor
    $contours \leftarrow ListOfContours(invDepthdata)$ ▷ returns an array of contours
    $mask \leftarrow zeros(sizeof(depthData))$
    $count \leftarrow Count(contours)$ ▷ get the total number of contours detected
    while $count > 0$ **do**
        if $contours \neq 0$ **then**
            $largestContour \leftarrow FindLargestContour(r)$ ▷ this is usually the firstcontour in the array. The array is sorted in decending order of area
            $b \leftarrow largestContour \prod depthData$ ▷ All the pixels outside the contour is set to zero
            if $GetMaxDepth(b) = maxDepth$ **then**
                $minDepth \leftarrow FindMinimumNonZeroDepth(b)$
                $mask \leftarrow Binerise(depthData \geq minDepth)$
                $count \leftarrow 0$
            else
                $contours \leftarrow contours - largestContour$ ▷ remove the largest array
                $count \leftarrow Count(contours)$
            end if
        end if
    end while
    return $mask$
end procedure

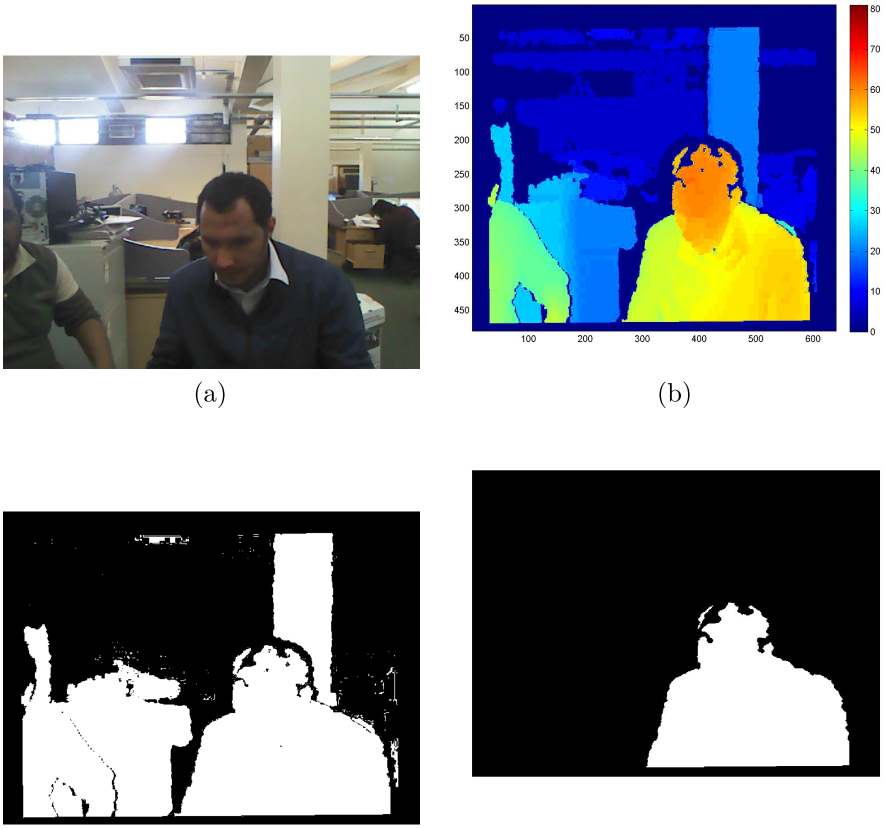
Figure 4.8 Adaptive segmentation algorithm



(a) (b) (c) (d) (e) (f)

Figure 4.9 Adaptive Segmentation (a) Original Depth data (b) Inverted Depth data (c) Mask of nearest object to sensor (d) Region of interest (e) Ellipse denoting face area within the ROI (f) Normalised ROI.

In Figure 4.10, an example is shown of how the Adaptive Segmentation can be used to remove potential distractions or imposter who might try to high-jack the session.



(a) (b)

(c) (d)

Figure 4.10 Imposter Example with Fixed and Adaptive Segmentation (a) Original Image (b) Inverted Depth data (c) Mask using fixed segmentation (d) Mask using adaptive segmentation.

4.1.7 Gesture Detection

To detect a blink, closure of both eyes has to be detected for a period of 1 second or more and then return to the open state. If closure of only one eye is detected, the system assumes there is no blink. Only the transition from an open eye to close eye and back to an open eye is recognised as a blink.

In the case of the eyebrow detection, the states of the eyebrows are monitored. The eyebrow region is detected using the location above the eye region. The state of the eyebrow is initially set to down. To recognise eyebrow movement both eyebrows have to be raised for a period of 1 second or more and subsequently return to the down state.

Only the transition from down to raised and then to the down state again is recognised as a valid eyebrow movement.

		Actual Outcome		
		p	**n**	**total**
Prediction Outcome	**p′**	TruePositive	FalsePositive	P′
	n′	FalseNegative	TrueNegative	N′
	total	P	N	

Figure 4.11 Confusion Matrix

To calculate the efficiency of the gesture detection algorithms a confusion matrix is generated for each gesture and the $TruePositive$, $FalsePositive$ and $accuracy$ rates are calculated [143]. The ground truth data is generated on-line, i.e. when the user is moving the cursor. If the cursor is within the target area, a gesture is expected by the system. If no gesture is detected, this is considered a $FalseNegative$ i.e. the gesture was considered a non-gesture. If no gestures are expected and a gesture occurs, this is considered a $FalsePositive$.

Figure 4.11 represents a confusion matrix that reflects the performance of the gesture detection system. The $TrueClass$ represents the classes expected i.e. in the case of blink detection; the two classes are Blink and No-Blink and the hypothesised class are the classes the algorithm detected.

$$TruePositiveRate = \frac{TruePositive}{P} \tag{4.2}$$

In Equation (4.2), the $TruePositiveRate$ is the ratio of the number of positive correctly classified gestures detected over the total number of the positive (P) gestures observed. The $TruePositive$ corresponds to the number of times the correct gesture was detected i.e. the number of times a blink was detected as a blink gesture by the algorithm, and the $FalseNegative$ corresponds to the number of times a blink gesture was detected as a no-blink gesture.

$$FalseNegativeRate = \frac{FalseNegative}{N} \tag{4.3}$$

In Equation (4.3), the $FalseNegativeRate$ is the ratio of the number of incorrectly classified negative gestures detected over the total number of the negative (N) gestures observed. The $FalsePositive$ corresponds to the number of times that a Non-Blink is incorrectly classified as a Blink gesture by the algorithm and the True Negative corresponds to the number of times a non-blink gesture was detected as a non-blink gesture.

$$Accuracy = \frac{TruePositive + TrueNegative}{P + N} \tag{4.4}$$

The accuracy of the gesture detection algorithm defined in Equation (4.4) is the ratio of the correctly classified gesture over the total number of gestures.

4.2 Result

4.2.1 Fitts' Test

Table 4.2 Parameters for Fitts' Test

Width (W)	Distance (D)	Index of Difficulty(ID)
25	50	1.585
50	200	2.322
50	400	3.170
25	400	4.088
13	400	4.990

Fitts' Test is conducted in blocks of 4 sequences with 8 different orientations for each device. Both the orientation and difficulty index of the target within each block are randomly selected. The approach is used to prevent the users from anticipating the location of the next target and thus ensure the independence between subsequent target selections.

There is a rest period of 3 seconds (3000 milliseconds) between each block of test. The rest period is to reduce the fatigue a user might experience during a test and to prevent the current test from influencing the following one.

4.2.2 Device Evaluation

Table 4.3 Mean Movement Time

Index of Difficulty	1.585	2.322	3.170	4.088	4.990
Mouse	0.980	1.085	1.337	1.457	1.525
SmartNav	1.933	1.566	2.002	2.714	3.584
CameraMouse	1.517	1.677	2.238	2.699	4.271
HeadTracker(with Dwell switch)	*4.596*	**3.302**	*4.681*	6.880	12.418

Table 4.3, shows the result of the preliminary evaluation of the headtracker against existing systems. In this case the Dwell switch was used because both CameraMouse [59] and SmartNav [8] have Dwell based clicking mechanisms. In this initial test the developed head tracker did not perform as well as the other devices. It can also be noted that for ID 2.322 bits, the mean MT decreases from 4.596 seconds to 3.302 seconds and increases again to 4.681 seconds for the next ID. This behaviour is not present in the other devices being investigated in this experiment. This could be due to the fact that the developed head tracker was not very stable i.e. it was difficult to maintain the head tracker still.

This also accounts for the fact that the mean MT for the different ID was higher than those of the other devices.

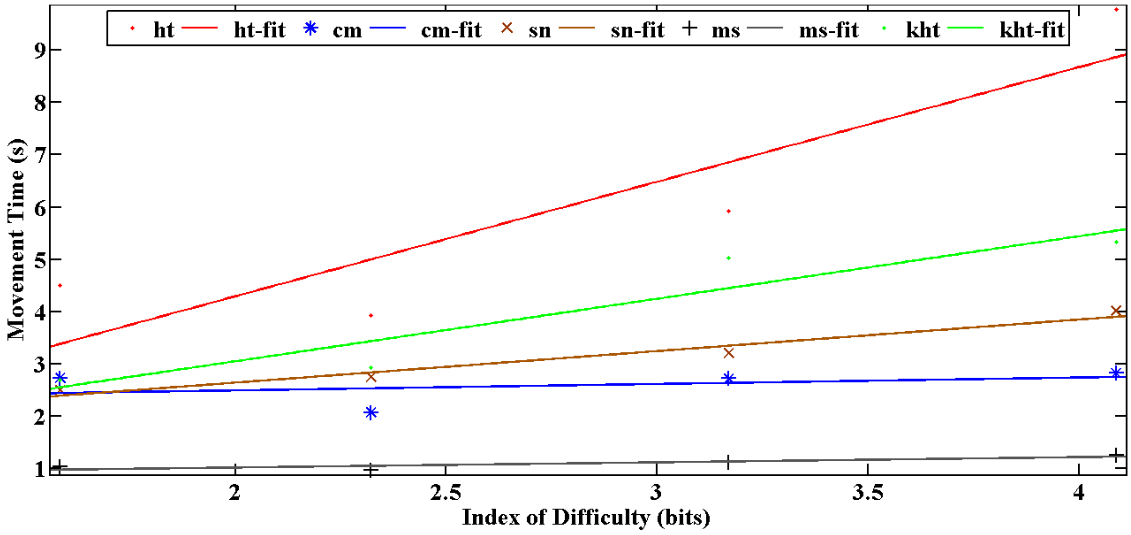


Figure 4.12 Comparison between devices

Figure 4.12, represents the performance of the different assistive device being evaluated based on the usage of two healthy participants. The data from the standard mouse (ms) is also added to view as a baseline. The devices with the lowest MT are the CameraMouse [59] (cm-fit) and SmartNav [8] (sn-fit) as the regression lines obtained for these devices are consistently below the line obtained for the web-camera based head tracker (ht -fit). The performance of the Kinect [14] based head tracker (kht-fit) is better than that of the web-camera based head tracker. In fact the performance of the Kinect based head tracker is closer to that of the SmartNav [8] (sn-fit). This improvement in the performance of the Kinect based head tracker is due to the use of depth data from the sensor.

Table 5.3 shows the performance data i.e. IP and TP_e of the web-camera and Kinect based devices with Dwell, Blink and Eyebrows based facial gesture switches. It also highlights the advantages of using depth data as all the performance measurements of

Table 4.4 Index of Performance (IP) and Effective ThroughPut (TP_e) of tested devices in bits/second

	Dwell		**Blink**		**Eyebrows**	
Device	IP	TP_e	IP	TP_e	IP	TP_e
Webcam	0.48	0.21	0.32	0.15	0.19	0.08
Kinect	**0.68**	**0.30**	**0.61**	**0.28**	**0.20**	**0.09**

the Kinect devices are the highest. The Dwell switch has the best performance, followed by the Blink and Eyebrows movement switch.

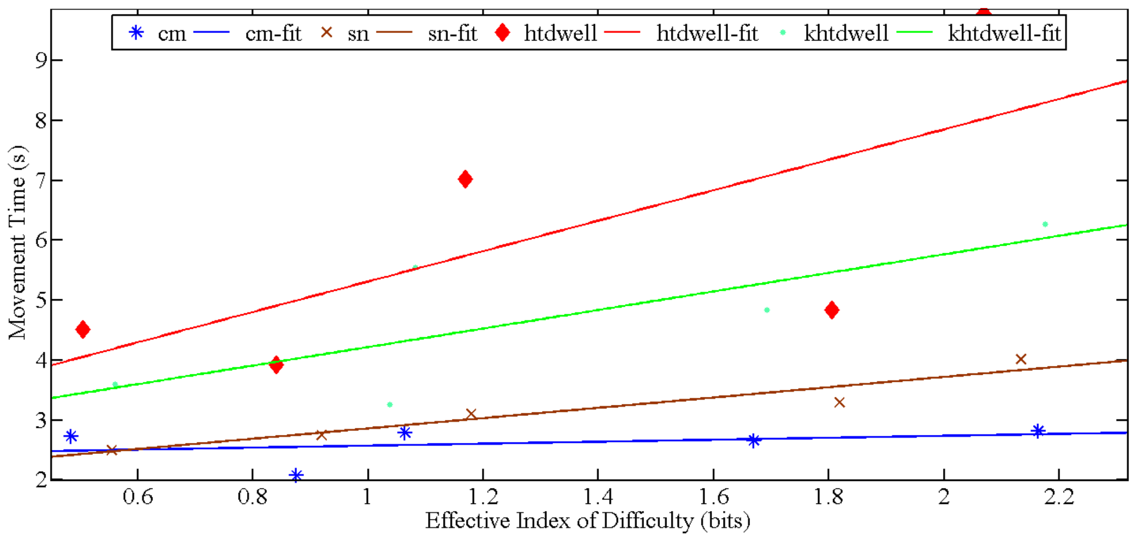


Figure 4.13 Comparison of the commercial devices against the Dwell switch device using the Effective Index of difficulty(ID_e)

Figure 4.13 , the Kinect based head tracker (khtdwell) is better than the 2D head tracker (htdwell) with the Dwell switch. The TP_e of the Kinect head tracker (0.30 bits per second) is nearly a third better than that of the camera-based head tracker (0.21 bits per second) using the Dwell switch as shown in Table 5.3. The performances of the CameraMouse [59] (cm-fit) and SmartNav [8] (sn-fit) are used as a baseline to highlight the progress made during the development of the devices.

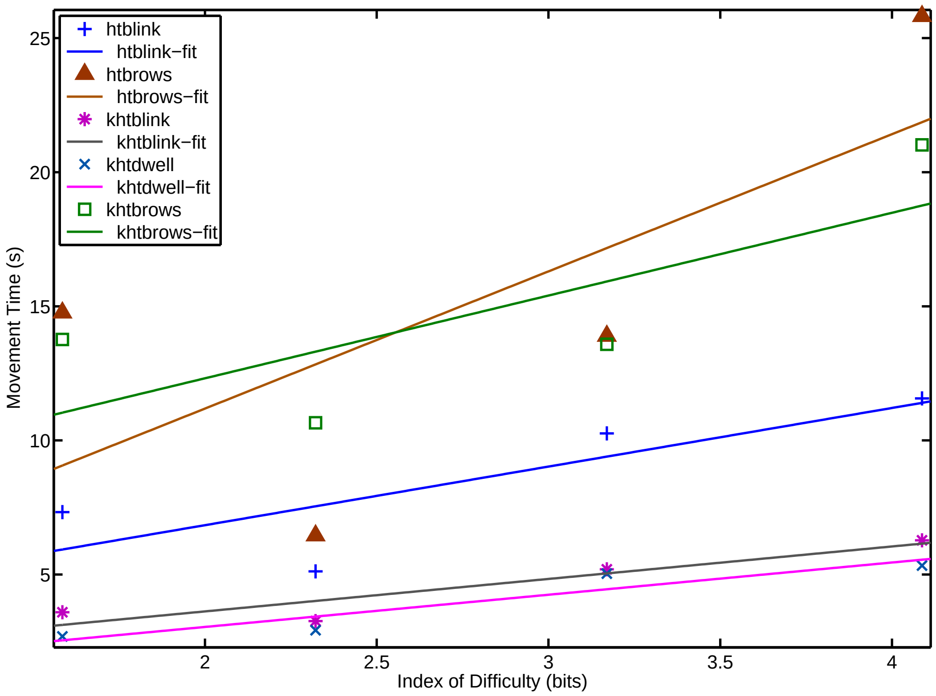


Figure 4.14 Comparison of the 2D and 3D devices using the Index of Difficulty(ID)

In Figure 4.14, the devices with the Eyebrows movement switch have the worst performance. But due to the use of depth data the Kinect based device with the Eyebrows switch is slightly better than the web-camera based one. The Kinect device with the Dwell switch has the best performance, followed by the Kinect device with the Blink switch and the web-camera based device with the Blink switch.

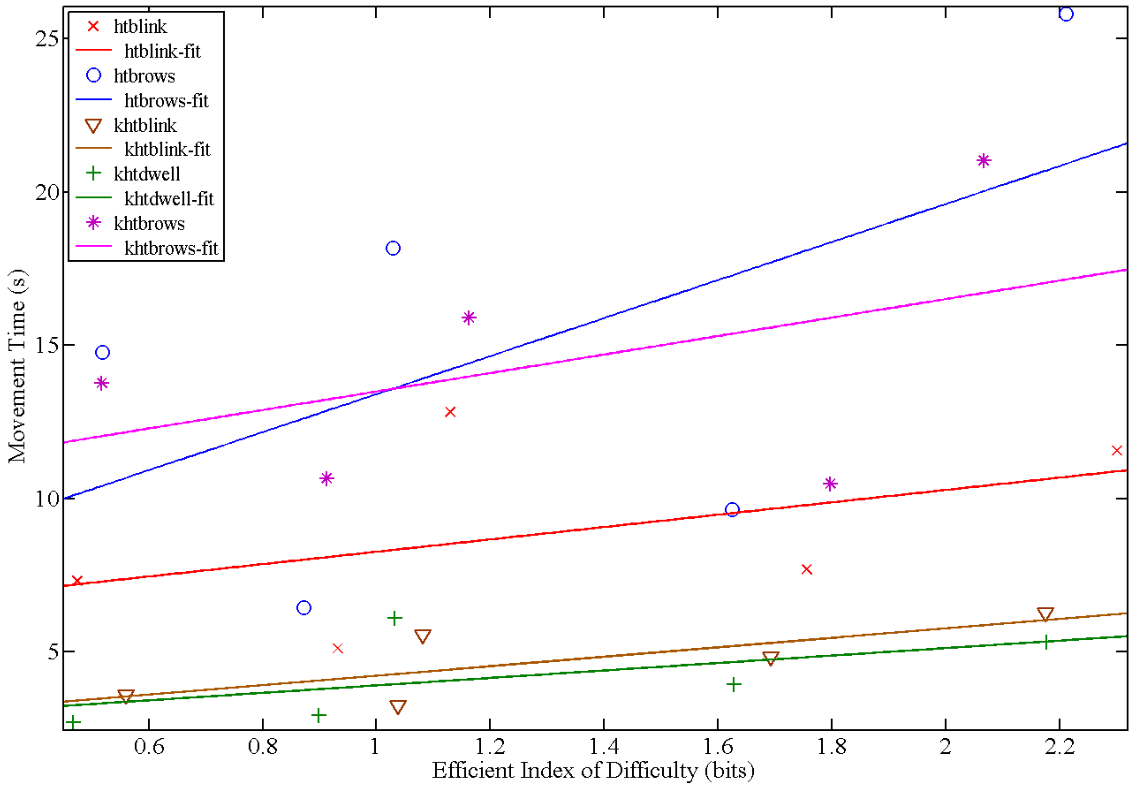


Figure 4.15 Comparison of the 2D and 3D devices using the Effective Index of Difficulty(ID_e)

Figure 4.15 , the Kinect-based dwell switch (khtdwell-fit) and blink (khtblink-fit) is better than the Dwell switch of the vision-based head tracker (ht-dwell-fit). The TP_e of the Kinect [14] head tracker with blink switch (0.28 bits per second) is doubled compared to the effective throughput of the 2D head tracker (0.15 bits per second).

As it can be seen in Figures 4.12, 4.13, 4.14, 4.15, the progress of the devices was followed using the modified Fitts' test. Also, one important point to note is the boost of performance observed in the head tracker with the use of depth data. The Kinect based devices have out performed the web-camera based devices constantly during the investigation. It has clearly been demonstrated that the use of depth data has helped with device performance.

4.2.3 Gesture Algorithm Evaluation

Table 4.5 AUC of off-line Gesture Detection

Algorithm	AUC	TPR	FPR
Blink	**0.684**	0.254	0.106
Eyebrows movement	0.613	**0.425**	**0.037**

From Table 4.5, it can be seen that the Blink detection algorithm has a higher AUC than the Eyebrows movement algorithm but its TPR is lower and FPR is higher. The high AUC of the Blink detection algorithm could be due to its high FPR.

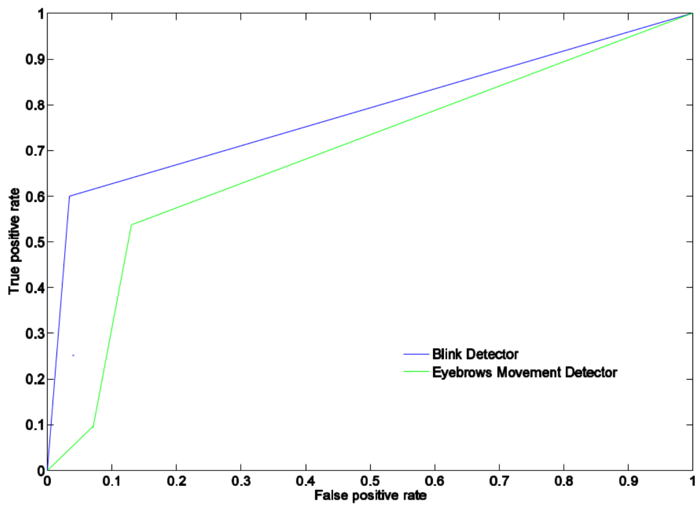


Figure 4.16 ROC curve of gesture algorithm evaluation

In Figure 4.16, it can be seen that the Blink detection performed slightly better than the Eyebrows movement detection algorithm. The Blink detection algorithm was evaluated against the ZJU Blink Database [16]. The ZJU database consists of 80 videos and had to be manually annotated before performing the evaluation. The Eyebrows movement detection algorithm was assessed using the Cohn-Kanade [15] database. The Cohn-Kanade database consists of pictures of people performing various expressions. To be able to test the Eyebrows detection algorithm, the database has to be converted into video file. A total of 458 video files were generated for each folder containing images and because the algorithm requires the gesture to be a neutral face, gesture and then a neutral face again. This was done by adding the first frame of the video as the last frame in the video. For both algorithms, the evaluation was performed 10 times and with a range of threshold from 50% to 95%, and a mean was calculated.

4.2.4 Gesture Error Rate

The experiment was carried out with 22 healthy individuals and one individual with MS. The participants were required to complete the Fitts' test with all 4 devices.

Figure 4.17, is a boxplot of the overall error rate for the different devices. The Kinect-Blink device has the lowest mean error of 0.134 (SD=0.108) and Kinect-EyeBrows has the highest mean error of 0.164 (SD=0.123). The Webcam-Blink and Webcam-EyeBrows have respectively mean error of 0.148 (SD=0.12) and 0.147 (SD=0.123). There is no big difference in the different devices based on the overall error rate, this is to be expected as the tests were performed by the same participants on the different devices. And also that the choice of the device used does not affect the overall error rate.

Figure 4.18 represents the mean error rate at each of the target locations for all the devices. It can be seen that the smallest overall error rate for the devices is the topmost one (4.8%) with ID of 4.088 bits and the highest overall error rate is found near the middle area and is about 25% with an ID of 1.585 bits. The error rates for the outermost locations i.e. with the highest ID (4.088 bits) ranges from 4.8% to 11%. The next inner circle of target locations with ID of 3.170 bits has a range from 10% to 17%. The third inner circle of locations with ID of 2.322 has a range of 15% to 20% and the range for the innermost circle with ID of 1.585 has a range from 16% to 25%. The last remaining location is the home location and it has an error rate of 15%.

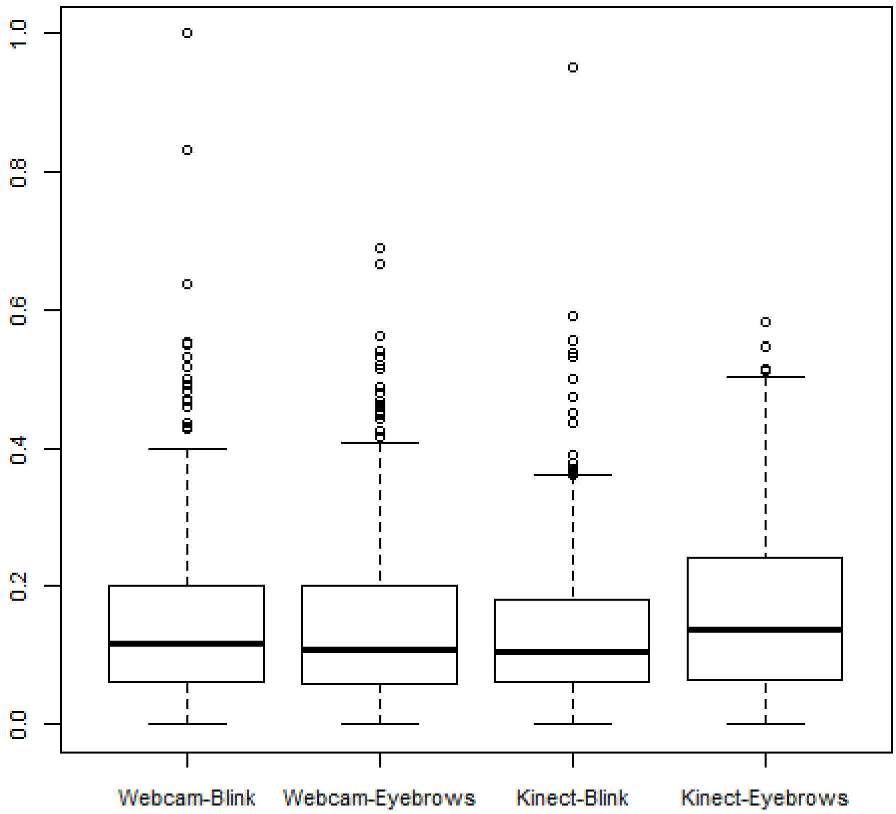


Figure 4.17 Boxplot of Error Rate by Devices

Figure 4.19, represents the error rate of the KentAssist Webcam-Blink device. For this device it can be seen that the smallest error rate is the outermost south-western one (6.0%) and the highest one is found near the middle area and is about 28%. The error rate for the outermost points i.e. with the highest ID seem to have error rate from 6.0% to 14%. The next inner circle of point has a range from 11% to 17%. The third inner circle of points has a range of 13% to 23% and the range for the final inner circle is from 13% to 28%. The innermost point has an error rate of 15%.

Figure 4.20, represents the error rate of the Webcam-EyeBrows device. For this device, it can be seen that the smallest error rate is the outermost south-western one (7.0%) and

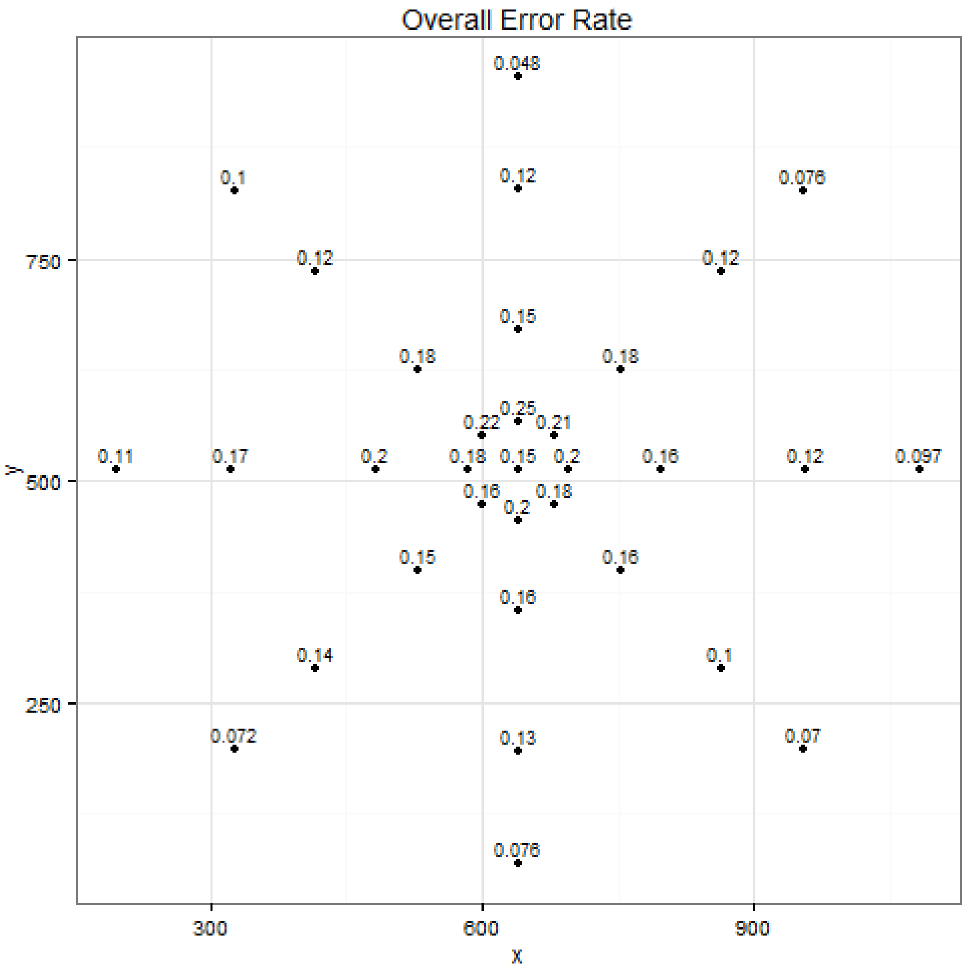


Figure 4.18 Overall Average Error Rate by target location

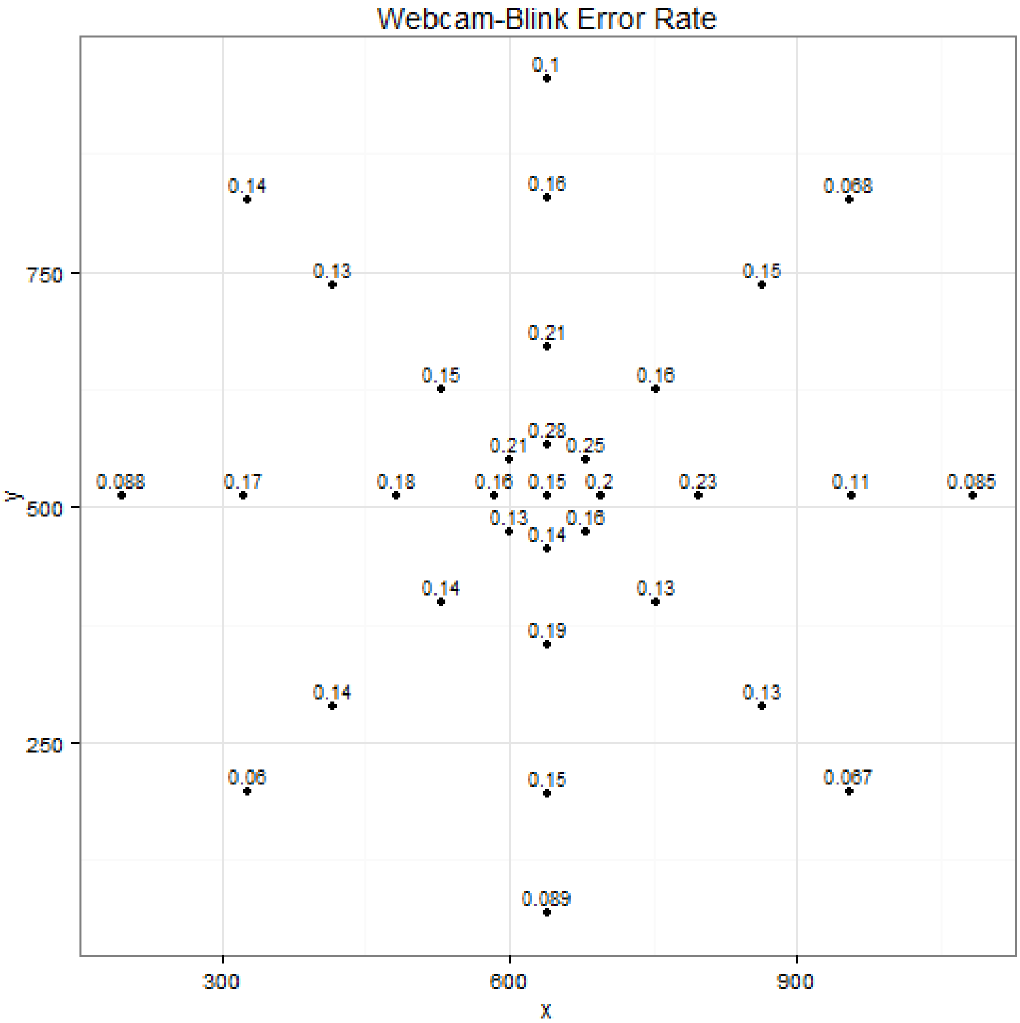


Figure 4.19 Average Error Rate by target location for Webcam-Blink

the highest one is found near the middle area and is about 28%. The error rate for the outermost points i.e. with the highest *ID* seem to have error rate from 7.0% to 16%. The next inner circle of point has a range from 8.4% to 19%. The third inner circle of points has a range of 12% to 22% and the range for the final inner circle is from 14% to 28%. The innermost point has an error rate of 15%.

It also shows that when the users move further away from the central location, the error rate is lower. This is because the central location is the ideal location for haar-cascade and the facial features have a high probability of being detected. This in turn

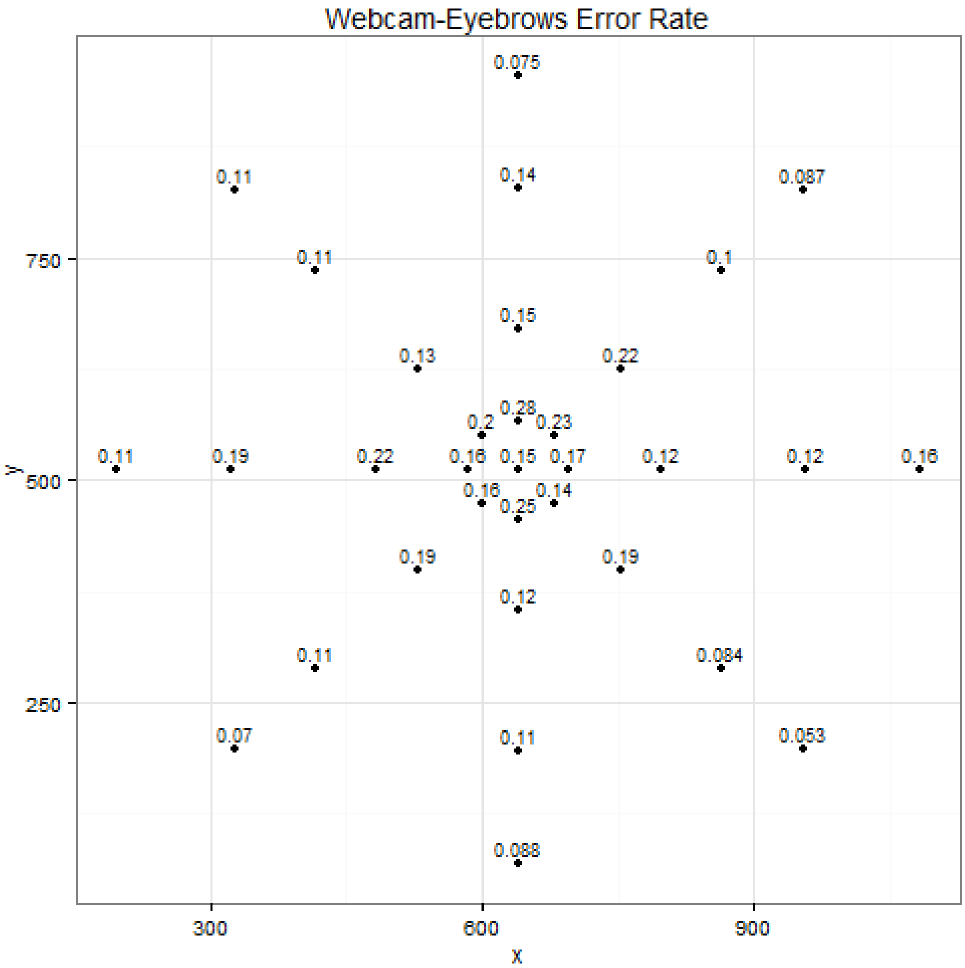


Figure 4.20 Average Error Rate by target location for Webcam-EyeBrows

makes processing time longer as more features are detected. But as the user moves to the outer limit, the pose of the face changes in relation to the sensor and thus for example, the haar-cascade for the head probably fails, and the default setting in this condition is to use the area nearest to the sensor as the region of interest. In this depth defined region of interest if the facial features required for the gesture to be computed is detected then the features are processed and the status of the gesture is updated. In this situation, as only the minimum required features are used, it requires less processing. An example is in the situation whereby only one eye is detected instead of both, the system assumes a blink gesture is detected in only the detected eye, if the state of the eye goes from open to close and to the open state again given that the blink duration threshold is valid.

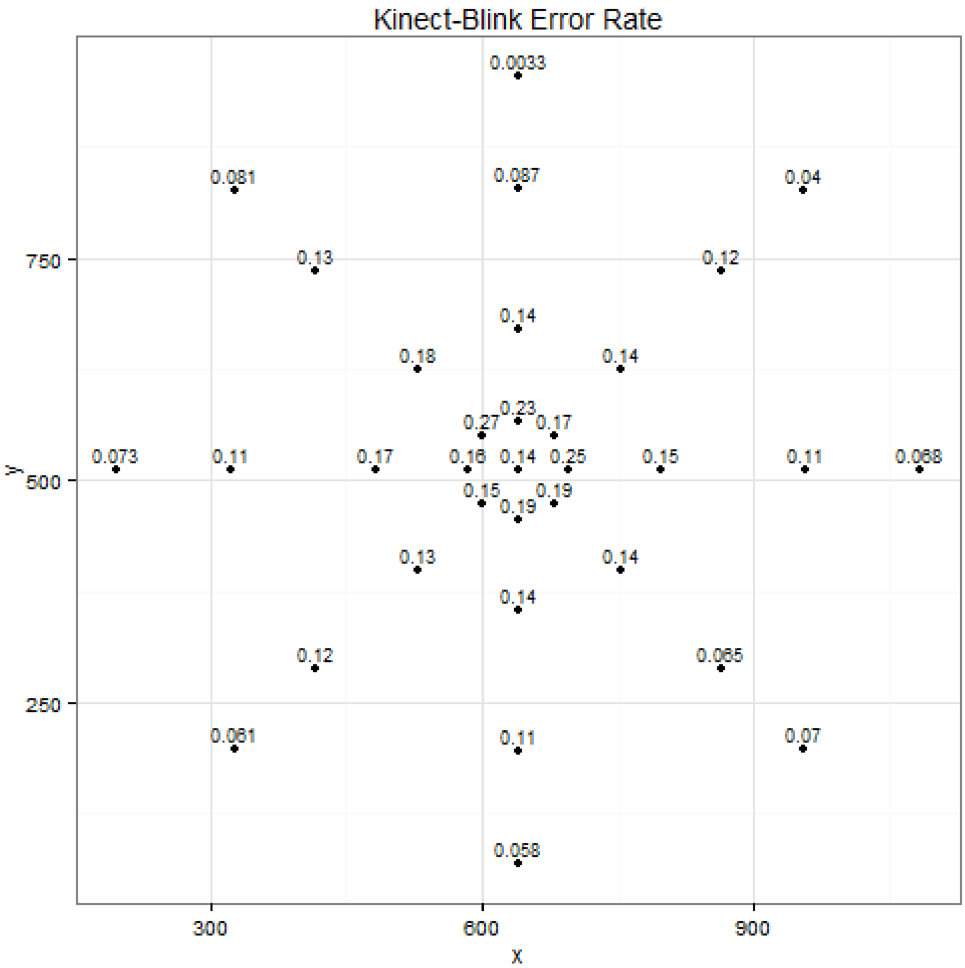


Figure 4.21 Average Error Rate by target location for Kinect-Blink

Figure 4.21, represents the error rate of the Kinect-Blink device. For this device, it can be seen that the smallest error rate is the outermost northern one (0.3%) and the highest one is found in the last innermost circle in the north western direction and is about 27%. For this device too, the further the target locations are from the centre, the smaller the error rate. The error rate for the outermost points i.e. with the highest ID seems to have error rate from 0.33% to 8.1%. The next inner circle of point has a range from 6.5% to 13%. The third inner circle of points has a range of 13% to 18% and the range for the final inner circle is from 15% to 27%. The inner most point has an error rate of 14%.

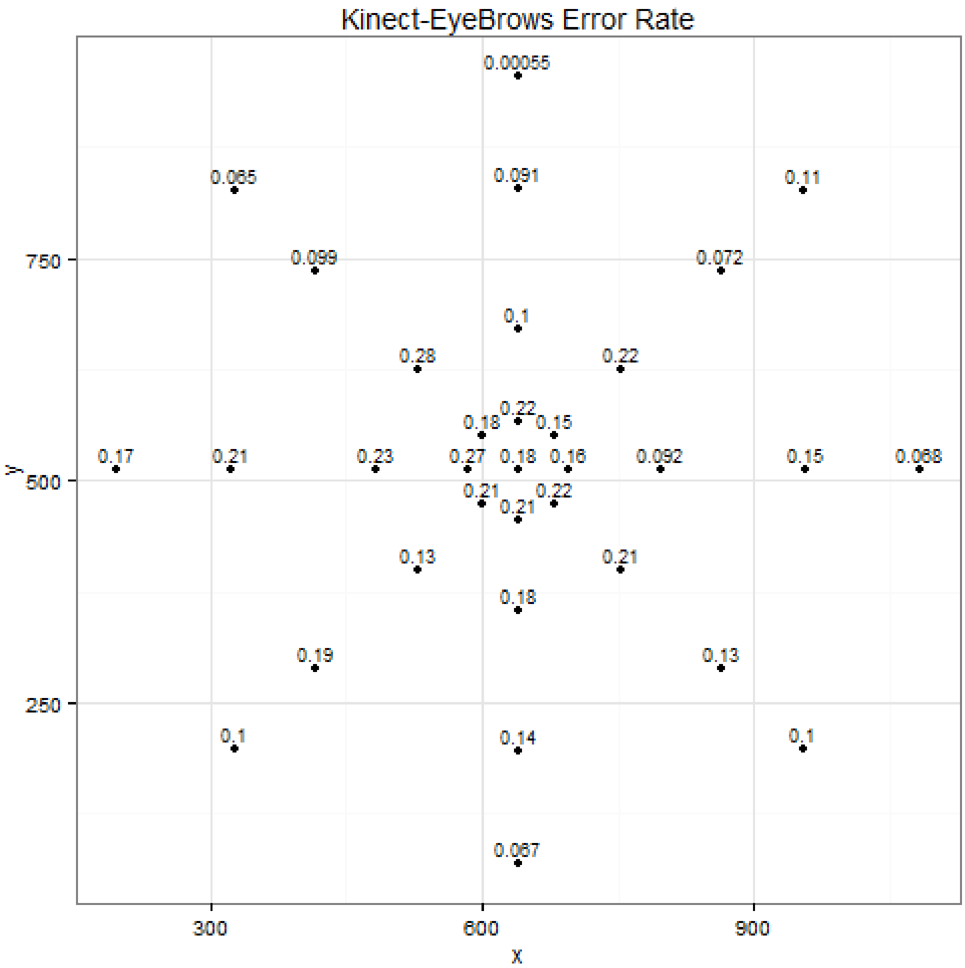


Figure 4.22 Average Error Rate by target location for Kinect-EyeBrows

Figure 4.22, represents the error rate of the Kinect-EyeBrows device. For this device, it can be seen that the smallest error rate is the outermost northern one (0.06%) and the highest one is found in the last innermost circle in the north western direction and is about 28%. The error rate for the outermost points i.e. with the highest *ID* seem to have error rate from 0.06% to 17%. The next inner circle of point has a range from 7.2% to 15%. The third inner circle of points has a range of 10% to 28% and the range for the final inner circle is from 15% to 27%. The innermost point has an error rate of 18%.

4.2.5 Analysis of Variance(ANOVA)

4.2.5.1 Device

Table 4.6 Result of ANOVA of Error rate by devices

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
device	3	0.185	0.062	4.36	0.005	
Residuals	1773	25.028	0.014			
Signif. codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	0.1 ' '	1

As it can be seen in Table 4.6, device has a significant effect on the error rate($F_{3,1773} = 4.36, p$ <0.05). Table 4.7 contains the posthoc [146] analysis and, as it can be seen only the comparison between the Kinect-Blink and the Kinect-EyeBrows device has a significant difference in their means at 95% confidence interval.

Table 4.7 Tukey [146] multiple comparisons of means 95% familywise confidence level of error rate by devices

Pairwise comparison of error rate by devices	diff	lwr	upr	p.adj
Webcam-Eyebrows-Webcam-Blink	-0.001	-0.021	0.019	1.000
Kinect-Blink-Webcam-Blink	-0.014	-0.034	0.006	0.271
Kinect-Eyebrows-Webcam-Blink	0.016	-0.005	0.037	0.188
Kinect-Blink-Webcam-Eyebrows	-0.013	-0.034	0.007	0.346
Kinect-Eyebrows-Webcam-Eyebrows	0.017	-0.004	0.038	0.176
Kinect-Eyebrows-Kinect-Blink	0.030	0.009	0.052	**0.002**

As it can be seen from Table 4.7, only the pair Kinect-Eyebrows and Kinect-Blink has a $p.adj < 0.05$ which indicates that the mean are significantly different. The $diff$ giving the difference in the observed means, lwr represents the lower end point of the interval, upr represents the upper end point and $p - adj$ giving the p-value after adjustment for the multiple comparisons. This is also visible in Figure 4.23 because Kinect-Eyebrows and Kinect-Blink pair do not overlap with the Kinect-Blink - Webcam-Blink and the Kinect-Blink - Webcam-Eyebrows pairs. When, there is no overlapping it means the pairs have a significant difference in their means.

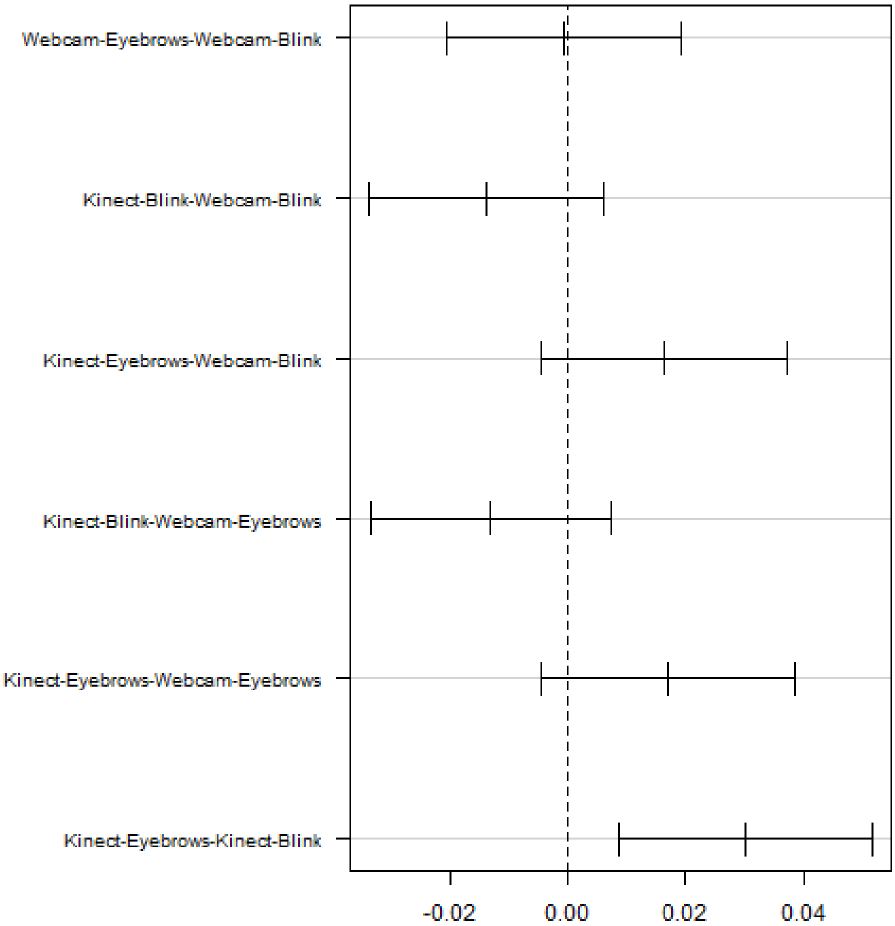


Figure 4.23 Pairwise multiple comparisons of error rate by devices

4.2.5.2 User

Table 4.8 Result of ANOVA for error and user

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
user	7	1.51	0.2154	16.1	<2e-16	***
Residuals	1769	23.70	0.0134			
Signif. codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	0.1 ' '	1

As it can be seen in Table 4.8, device has a significant effect on the error rate($F_{7,1769} = 16.1, p <0$). Table 4.9, contains the posthoc [146] analysis and as it can be seen only the comparison between user 1 and users 3 (3-1), 5(5-1), 6(6-1), 7(7-1), 9(9-1), 10 (10-1) and 11 (11-1), and also user 9 and users 3 (9-1), 5(9-5), 6(9-6), 7(9-7), 10(10-9), 11(11-9); and between users 11 and 10 have a significant difference in their means at 95% confidence interval.

Table 4.9 Tukey [146] multiple comparisons of means 95% familywise confidence level of error rate by users

Pairwise comparison of error rate by users	diff	lwr	upr	p.adj
3-1	-0.057	-0.089	-0.026	**0.000**
5-1	-0.068	-0.099	-0.036	**0.000**
6-1	-0.048	-0.079	-0.016	**0.000**
7-1	-0.065	-0.096	-0.034	**0.000**
9-1	-0.139	-0.188	-0.090	**0.000**
10-1	-0.040	-0.073	-0.008	**0.004**
11-1	-0.083	-0.115	-0.050	**0.000**
5-3	-0.010	-0.041	0.021	0.976
6-3	0.010	-0.021	0.041	0.979
7-3	-0.007	-0.038	0.024	0.997
9-3	-0.081	-0.130	-0.032	**0.000**
10-3	0.017	-0.015	0.049	0.757
11-3	-0.025	-0.058	0.007	0.255
6-5	0.020	-0.011	0.051	0.511
7-5	0.003	-0.028	0.034	1.000
9-5	-0.071	-0.120	-0.022	**0.000**

Table 4.9 Tukey [146] multiple comparisons of means 95% familywise confidence level of error rate by users

Pairwise comparison of error rate by users	diff	lwr	upr	p.adj
10-5	0.027	-0.005	0.059	0.179
11-5	-0.015	-0.048	0.017	0.847
7-6	-0.017	-0.048	0.014	0.701
9-6	-0.091	-0.140	-0.042	**0.000**
10-6	0.007	-0.025	0.039	0.998
11-6	-0.035	-0.068	-0.003	**0.022**
9-7	-0.074	-0.123	-0.025	**0.000**
10-7	0.024	-0.008	0.057	0.309
11-7	-0.018	-0.050	0.014	0.695
10-9	0.098	0.048	0.148	**0.000**
11-9	0.056	0.006	0.106	**0.016**
11-10	-0.042	-0.076	-0.009	**0.004**

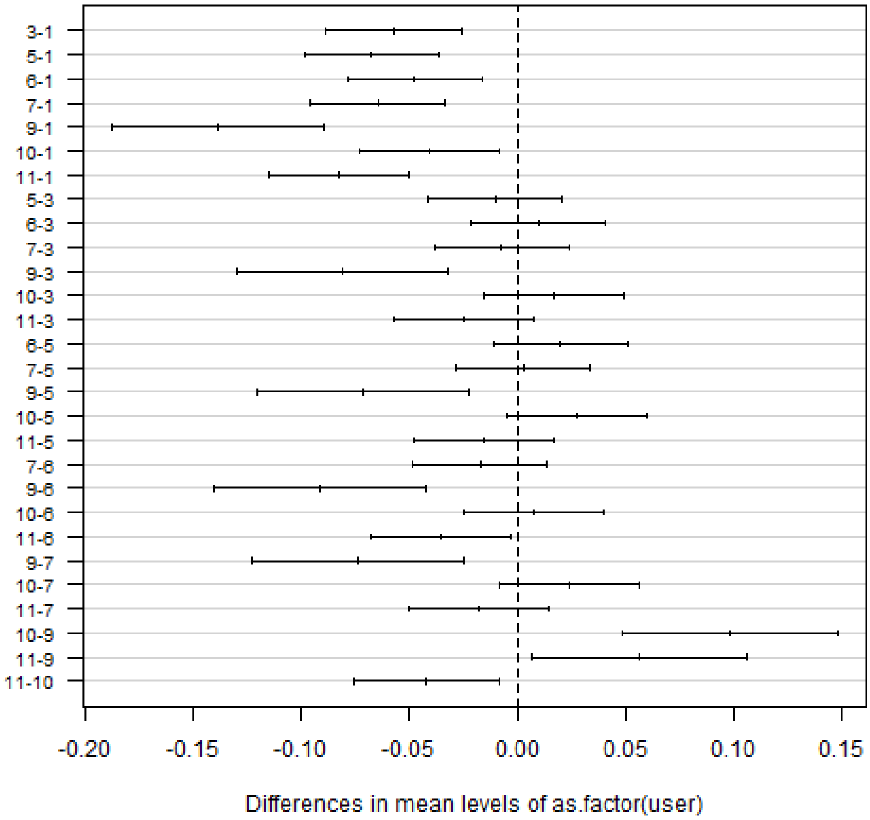


Figure 4.24 Pairwise multiple comparisons of error rate by users

4.2.5.3 Index of Difficulty (ID)

Table 4.10 Result of ANOVA for error and Index of Difficulty

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
id	3	3.29	1.096	88.6	<2e-16	***
Residuals	1773	21.93	0.012			
Signif. codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	0.1 ' '	1

As it can be seen in Table 4.10, ID has a high significant effect on the error rate ($F_{3,1773} = 88.6, p$ <0). Table 4.11, contains the posthoc [146] analysis and as it can be seen all the ID have a significant difference in their means at 95% confidence interval. This is further supported by the fact that in Figure 4.11, each pair has at least one non-overlapping pair such as 2.322-1.585 and 3.17-1.585 pairs, 3.17-1.585 and 4.088-1.585 pairs, 3.17-2.322 and 4.088-1.585 pairs.

Table 4.11 Tukey [146] multiple comparisons of means 95% familywise confidence level of error rate by Index of Difficulty

Pairwise comparison of error rate by Index of Difficulty	diff	lwr	upr	p.adj
2.322-1.585	-0.026	-0.045	-0.006	**0.004**
3.17-1.585	-0.070	-0.089	-0.051	**0.000**
4.088-1.585	-0.113	-0.132	-0.093	**0.000**
3.17-2.322	-0.044	-0.064	-0.025	**0.000**
4.088-2.322	-0.087	-0.106	-0.068	**0.000**
4.088-3.17	-0.043	-0.062	-0.023	**0.000**

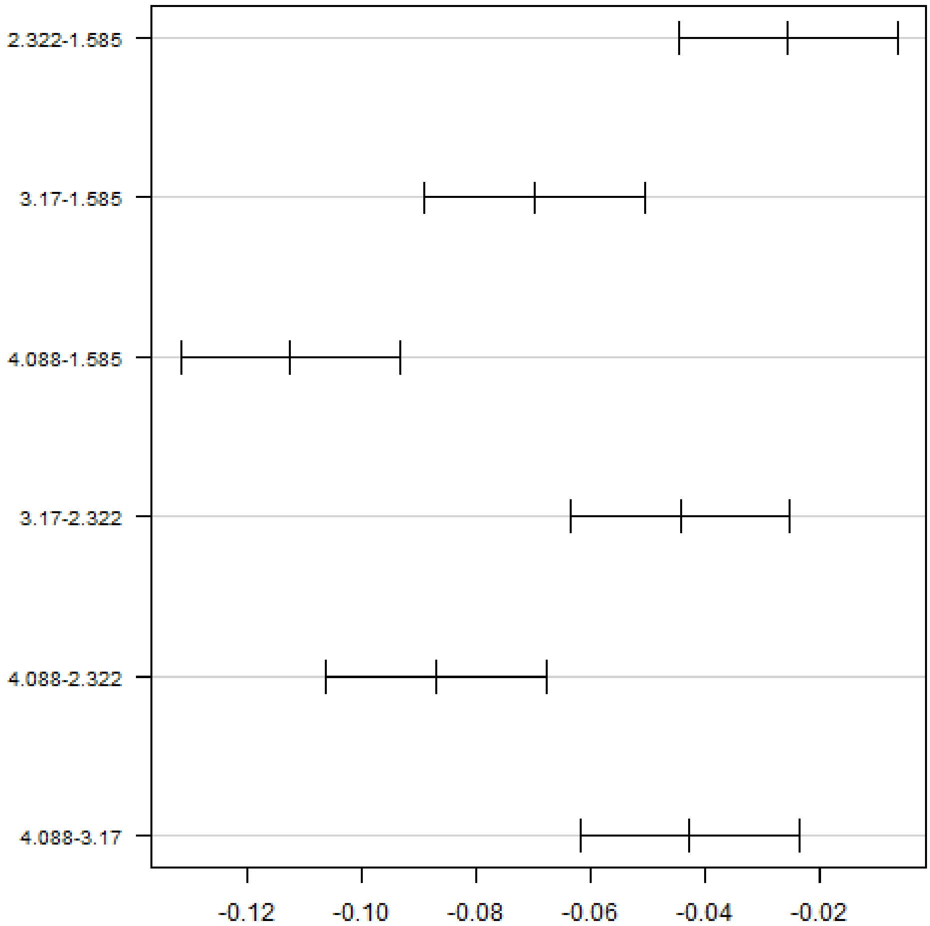


Figure 4.25 Pairwise multiple comparisons of error rate by *ID*

4.2.5.4 Gesture

Table 4.12 Result of ANOVA for error and gesture type

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
gesture	1	0.05	0.0525	3.7	0.055	.
Residuals	1775	25.16	0.0142			
Signif. codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	0.1 ' '	1

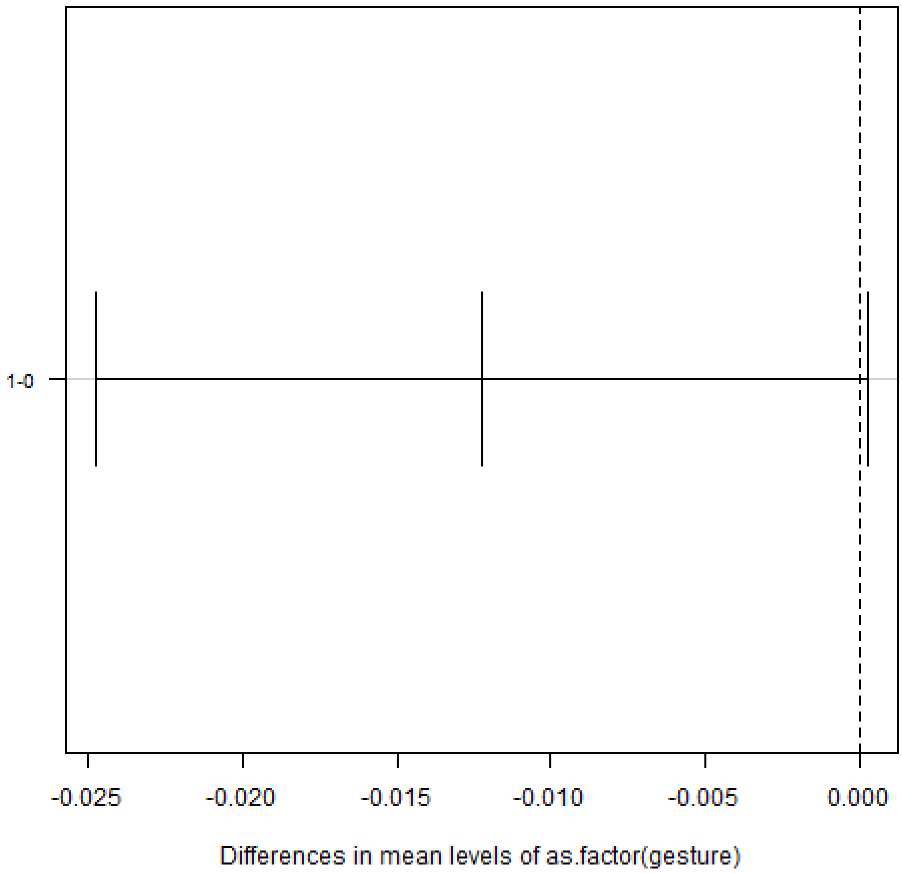


Figure 4.26 Pairwise multiple comparisons of error rate by gestures

Table 4.13 Tukey [146] multiple comparisons of means 95% familywise confidence level of error rate by gesture type

Pairwise comparison of error rate by gesture type	diff	lwr	upr	p.adj
1-0	-0.012	-0.025	0	0.055

As it can be seen in Table 4.12, gesture has a small significant effect on the error rate ($F_{1,1775} = 3.7, p < 0.1$). Table 4.13, contains the posthoc [146] analysis and as it can be seen eye-blink and eyebrows movement has a significant difference in their means.

4.2.5.5 Sensor

Table 4.14 Result of ANOVA for error and Sensor

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sensor	1	0.0	0.00034	0.02	0.88	
Residuals	1775	25.2	0.01420			
Signif. codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	0.1 ' '	1

As it can be seen in Table 4.14, the sensor selected has no significance on the error rate($F_{1,1775}$=0.88,ns). This would indicate that the sensor i.e. the webcamera or the RGB-D sensors, used by the assistive devices had no effect on the error rate. There is not sufficient evidence at a 95% confidence interval to claim that there is a difference in the mean error rate based on the sensor used.

4.2.5.6 Orientation

Table 4.15 shows that orientation of the movement has no significance difference in the mean of the error rate($F_{7,1769}$=0.48,ns) based on Orientation. This would indicates that the direction of the target did not have an effect on the error rate.

Table 4.15 Result of ANOVA for error and Orientation

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
orientation	7	0.05	0.00685	0.48	0.85	
Residuals	1769	25.16	0.01423			
Signif. codes:	0 '***'	0.001'**'	0.01 '*'	0.05 '.'	0.1 ' '	1

4.3 Discussion

In this section, the results obtained from the analysis of the different gesture facial switching mechanisms are discussed.

4.3.1 Gesture Error Rate

The mean error rate for each of the devices tested, shows that the target location associated with the ID of 4.088 bits has the lowest error rate. The lowest error rates for the Web-Blink, Web-Eyebrows, Kinect-Blink and Kinect-Eyebrows are respectively 6%, 7%,4% and 0.05%, and occur at orientations 135^o,135^o,45^o and 0^o. The highest error rates for most of the devices occur at locations associated with ID of 1.585 bits, except for the Kinect-Eyebrows device which occurs at a location associated with an ID of 2.322 bits. The highest error rates for all the Web-Blink, Web-Eyebrows, Kinect-Blink and Kinect-Eyebrows are respectively 28%, 28%,27% and 28%, and occur at orientations 0^o,0^o,227^o and 227^o. For the home location for the different devices, the error rate is between 14% and 18%. It can also be seen that as the ID increases, the error rates seem to be decreasing.

4.3.2 Analysis of Variance(ANOVA)

From the analysis of the variance of the different parameters it can be seen that Orientation and Sensor type are not significant at a 95% confidence interval. This indicates that there is not enough evidence to determine the effect of Orientation and Sensor type on the error rate. The fact that the Orientation has no effect on the error rate, could also point to the fact that the facial gesture algorithm is not affected by the pose of the face while performing the gesture. However, the device - a device is a combination of a sensor and a switching mechanism - used to perform the gesture is significant i.e. there is a difference in the means of the error rates at a 95% confidence interval. The error rate obtained depends on the device used to perform the switching task. The user who performed the gesture is also highly significant at a 95% confidence interval. This could point to the fact that although all the users were healthy, there was still a significant difference in the error rates of the different users. As seen in section 4.3.1, at a 95% confidence interval, there is sufficient evidence to indicate a relationship between the error rates and the ID.

4.4 Conclusion

Assistive devices need to be fine-tuned to the residual abilities of disabled individuals. The Fitts' Test may be used to determine the optimum device for an individual and the best settings for the screen of the individual. This would reduce the time to find a well suited device for an individual, increase the usability of the individual and increase adherence by patients/carers. The test could also be used to monitor the progress of the user. The modified Fitts' Test used in this has been validated by the fact that both commercial devices(CameraMouse [59] and SmartNav [8]) evaluated have the best performance.

It can be seen that the addition of the depth data to the RGB image enables the removal of the background distractions (based on the distance to the sensor) and thus reduces the noise in the data. The inclusion of a sensor capable of measuring depth information from structured illumination (such as Kinect for Windows [14]) is shown to have the potential in this case to enhance assistive device performance. From our RGB-D system, the TP_e of dwell clicking increased by a third (from 0.21 to 0.30 bits per second) and that of blink clicking doubled (from 0.15 to 0.28 bits per second) compared to our 2D system. The eyebrows detection algorithm has the worst performance as it can be seen from the increased MT. The eye blink detection algorithm performance is quite near to the performance of the Dwell click switch using depth.

In the experiments carried out in this chapter, the Kinect based devices have had a better performance than the web camera based system. The Kinect has been used in a number of systems [147–153] since its release. In Saini et al. [147] a low cost game based stroke rehabilitation framework was proposed to support hand and leg rehabilitation. The framework consisted of a game design with a standard angle based representation of body motion during motion so as to improve the accuracy of the exercises presented to the patients. The Kinect was used as a balance virtual rehabilitation system by stroke patients suffering from hemiparesis in Llorens et al. [148]. In Roy et al. [149], a system called Kinect-o-Therapy used the Kinect to implement an exercise based rehabilitation system. The system tracks the exercise of the patients and provides them with feedback. A system called REMOVIEM using the Kinect to provide motor rehabilitation exercises to persons suffering from MS was proposed by Lozano-Quilis et al [150]. In Chang et al. [151], the Kinect is used in a task prompting system called the Kinempt. The system was used to enable four participants with cognitive impairments to train for a job. The Kinect was also used for sign language recognition [152, 153]. In Halim and Abbas [153], the Kinect was used to detect and recognise sign language using image processing and pattern, and translate it into speech. The system was reported to have an accuracy of 91% for sign language detection. In Oszust and Wysocki [152], a Polish sign language word recognition system was proposed. The nearest neighbour classifier with Dynamic Time Wrapping (DWP) was used with features extracted from the skeletal data from the Kinect depth data and skin colour of the hand region. The system had an accuracy rate of 89.33%. Most of the Kinect based systems have relied most on the skeletal data provided by the device to recognise gestures performed by the users.

Based on the investigation into the gesture detection performed while evaluating the KentAssist devices, it was found that there might be strong significant evidence of a relation between the error rate and the ID of the task being performed. It was also found that there is no sufficient evidence at a 95% confidence interval to point to a relationship between the orientation in which the gesture was being performed and thus, this might indicate that the gesture detection algorithm is robust to facial changes due to orientation. It was also found that the gesture being performed has a moderate effect on the error rates but not the sensor used to capture the gesture. In the next chapter, the preliminary experiment comparing commercially available devices to KentAssist is performed using Fitts' test.

The contribution of this chapter is on the off-line evaluation of the proposed eye blink and eyebrows detection algorithm on two facial gesture databases namely the Cohn-Kanade [15] database and the ZJU Blink Database [16]. It also introduces the GER as a performance measurement of the gesture recognition system. The GER is obtained by the analysis of the data captured during the data capture sessions to obtain the error rate for all the target locations. The performance of the three switching mechanisms - Dwell click, eye blink clicking and eyebrows movement clicking - using a web camera and the Kinect is evaluated and analysed. This chapter also provides the data obtained from the KentAssist headtracker system. It is evaluated and compared with existing systems such as CameraMouse [59] and SmartNav [8]. The proposed system should also be tested in varying illumination condition to further test the robustness of the system. Further investigation into the effectiveness of the facial gestures is carried out in the next chapter.

Chapter 5

Facial Gesture Switch

In this chapter, the evaluations of the 2D and 3D head tracking system with facial gesture switches are discussed. The contributions discussed in this chapter are the evaluation of an eye blink and an eyebrow movement switch using either a web camera or an RGB-D sensor, and the creating of a dataset for facial gestures and head pointing with both colour and depth images. The devices are evaluated based on the experiments discussed in Chapter 3. The aim of the experiment carried out in this chapter is to investigate the effect of the Eye Blink and Eyebrows switch on the performance of devices. It was observed that both Kinect [14] devices have lower MT and higher IP and TP_e than the web camera based devices. The usability assessment survey, suggests that there is a significant difference in eye fatigue experienced by the participants; blink gesture was less tiring to the eye than eyebrows movement gesture. This demonstrates that both the selected sensor and facial gesture have an impact on the performance of the assistive devices.

In 5.1 contains information on head tracking and different technologies used to track the head, Fitts' Law, the evaluation methods and our approach using both depth and vision images are presented. In 3.6, the procedures and the data collected are discussed. In 5.2, the results obtained are presented. Section 5.3 contains the analysis and discussion of the result obtained. Finally, section 5.4 contains the conclusions of the investigation.

5.1 Design

In this section we are looking at the design of the head tracker. The parameters of the Fitts' test, the stimuli displayed to the participants and the sensors used during the data capture session are presented. The data obtained from the Kinect [14] sensor and the depth-based segmentation carried out are also discussed.

5.1.1 Overview

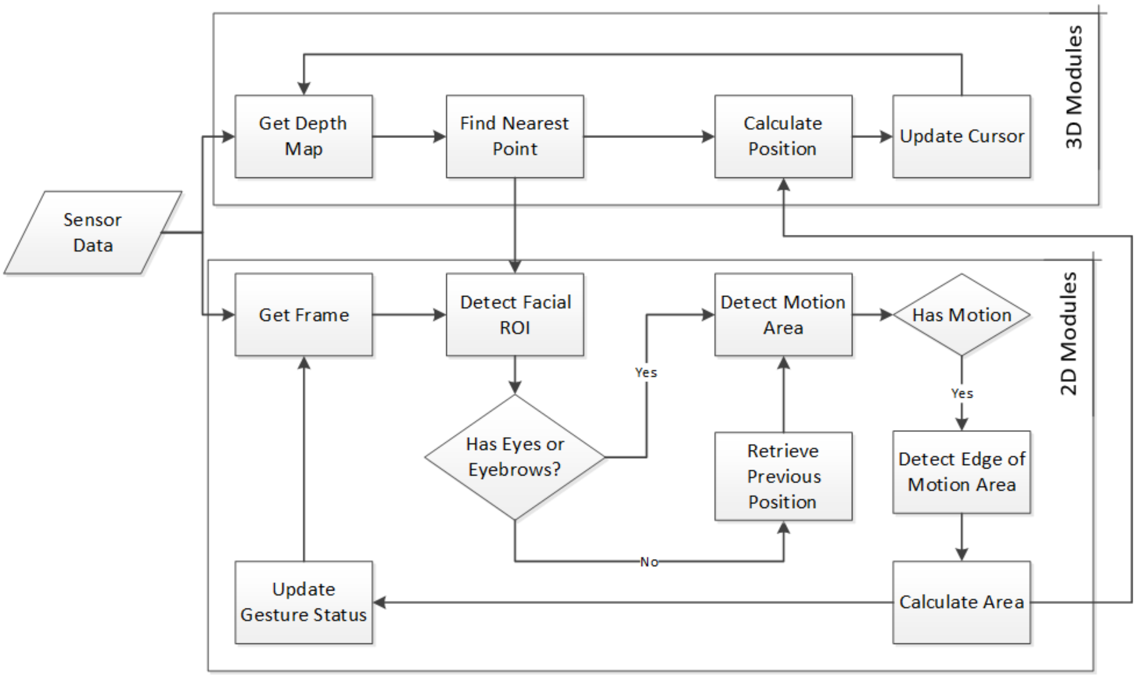


Figure 5.1 Algorithm to detect Blink and Eyebrows Movements.

Figure 5.1 shows an overview of the 3D head tracking and facial gesture recognition system. The facial gesture recognition system is similar for both the 2D vision system and the 3D Kinect system. Depth data is used only to filter the region of interest when processing the facial image - only objects within a meter of the 3D sensor were included in the region of interest and all other background was removed before further processing. The facial gesture recognition system used the RGB data from the sensors. Facial areas of interest such as the head, eyes region, left eye and right eye are detected using a Haar-Cascade [24].

5.1.2 Fitts' Test

Table 5.1 Parameters for Fitts' Test

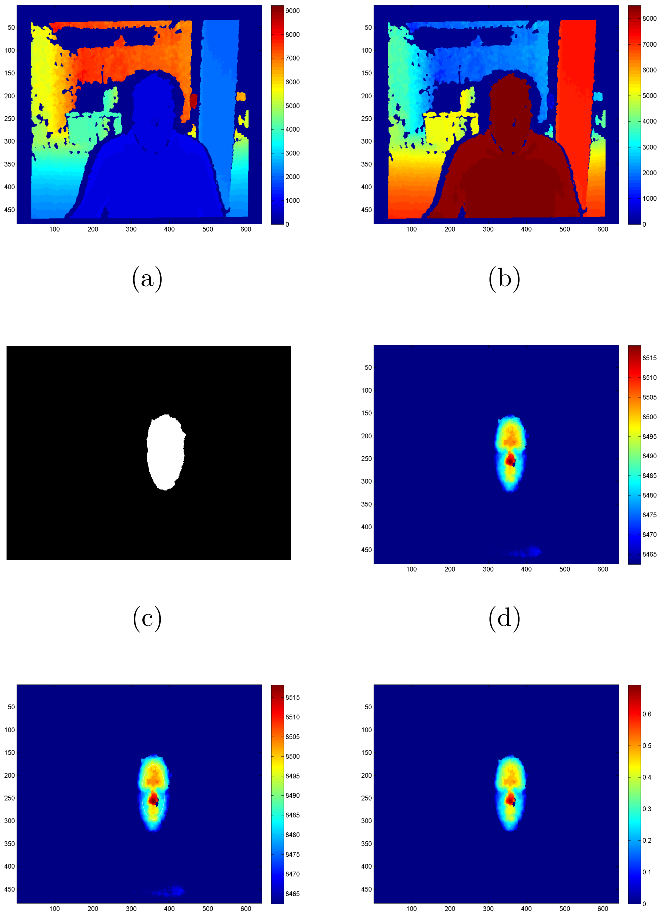
Width (W)	Distance (D)	Index of Difficulty(ID)
25	50	1.585
50	200	2.322
25	400	4.088
50	400	3.170

Fitts' Test is conducted in blocks of 4 sequences with 8 different orientations for each device as shown in Figure 3.4. Both the orientation and difficulty index of the target within each block are randomly selected. The approach is used to prevent the users from anticipating the location of the next target and thus ensure the independence between subsequent target selections.

There is a rest period of 3 seconds (3000 milliseconds) between each block of test. The rest period is to reduce the fatigue a user might experience during a test and to prevent the current test from influencing the following one.

5.1.3 Depth Data

The depth data obtained from the Kinect [14] sensor is used to reduce the search area for the different Haar-Cascade features. This reduces the computational load and avoids background distractions, such as people, movements, and changes in lighting and hence increase the performance. A mask for the object closest to the sensor is created from the depth data. Unlike the method presented in Section 4.1.6.2 this method does not rely on the object - in our case the head of the user - being within 1000mm of the sensor. An example is shown in Figure 5.2.



(a) (b)

(c) (d)

(e) (f)

Figure 5.2 Adaptive Segmentation (a) Original Depth data (b) Inverted Depth data (c) Mask of the nearest object to sensor (d) Region of interest (e) Ellipse denoting face area within the ROI (f) Normalised ROI.

5.2 Result

5.2.1 Analysis of Gesture

The blink and eyebrows movement gesture switches were evaluated using a Receiver Operating Characteristic (ROC) curve. The data used for the evaluation was obtained during the data capture. The threshold used in this setting is respectively 90% and 95%

for the blink and eyebrows movement algorithms. The ROC curve plots the True Positive Rate(TPR) against the False Positive Rate(FPR) of the algorithm.

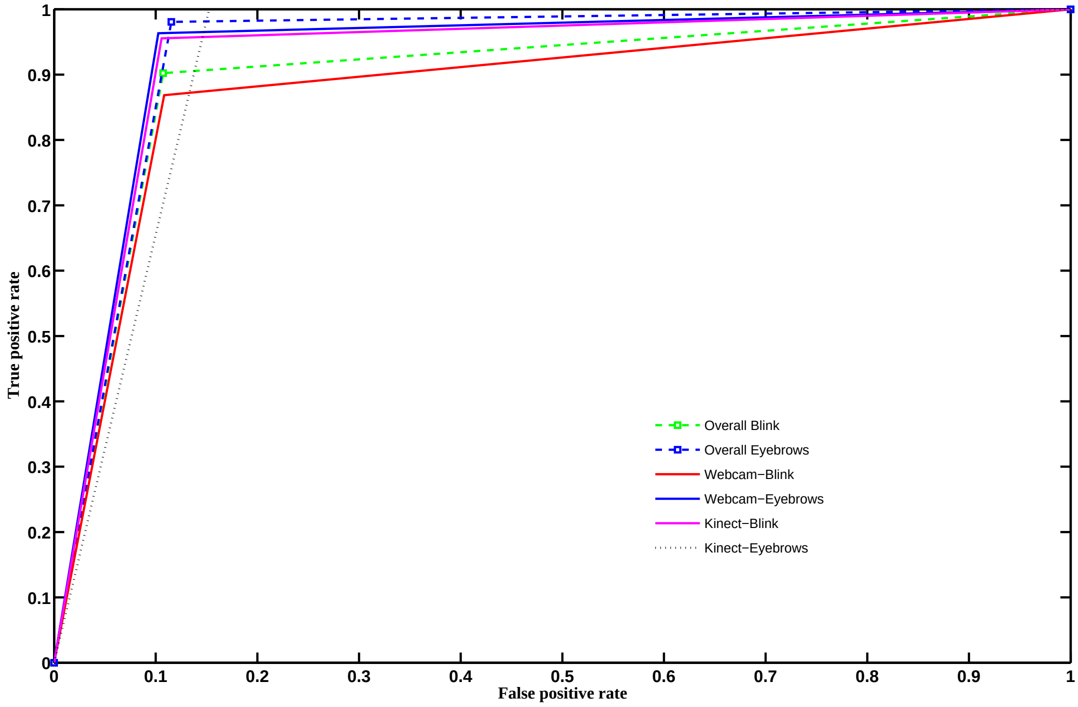


Figure 5.3 ROC for online gesture detection

Table 5.2 AUC of on-line Gesture Detection

Device	AUC
Overall Blink	0.897
Overall Eyebrows	0.933
Webcam-Blink	0.880
Webcam-Eyebrows	0.930
Kinect-Blink	0.925
Kinect-Eyebrows	0.924

From Table 5.2, the Webcam-Eyebrows has the highest AUC, followed by the Kinect-Eyebrows, the Kinect-Blink and finally the Webcam-Blink device. This data was obtained during our data capture session, and consists of data from 22 healthy individuals and 1 person with MS. Using the data captured and the *perfcurve* function in Matlab [140].

5.2.2 Fitts' Test

The experiment was carried out with 22 healthy individuals and one person with disabilities, who completed the tests with at least two devices. The MT in Fitts' Test is the time taken to move to the target location from the starting point and performing the task. To be able to compare the devices and the effect of the facial gesture, the task has been broken in two - Task 1 and Task 2. Task 1 involves moving the cursor to the target location using the movement of the head. Task 2 encapsulates Task 1 and also involves selecting the target by using one of the facial gestures as a switching mechanism. In Figure 5.4, the Kinect-eyebrows device has a lower MT than the Kinect-blink device for an ID greater than 1.9 bits. Overall for Task 1, it can be seen that the Kinect-Eyebrows device has the lowest MT, followed by the Kinect-Blink, the Webcam-Blink and finally the Webcam-EyeBrows, which took the maximax time to complete.

Table 5.3 Overall Index of Performance (IP) and Effective ThroughPut (TP_e) of tested devices

	Task 1		**Task 2**	
Device	IP	TP_e	IP	TP_e
Webcam-Blink	0.36(R^2=0.98)	0.74	0.32(R^2=0.78)	0.41
Webcam-Eyebrows	0.39(R^2=0.90)	0.67	0.55(R^2=0.21)	0.37
Kinect-Blink	0.5(R^2=0.89)	0.89	0.45(R^2=0.78)	0.6
Kinect-Eyebrows	0.68(R^2=0.81)	0.95	0.67(R^2=0.48)	0.64

From Table 5.3, it can also be seen that both IP and TP_e for moving the cursor to the designated target (Task 1) were better than that of combination of moving and the click action using the different facial gestures for all devices. This is to be expected as the clicking/selection method has an effect on the performance and efficiency of the system used. Also, the IP and TP_e of the 3D Kinect system were better than that of the 2D

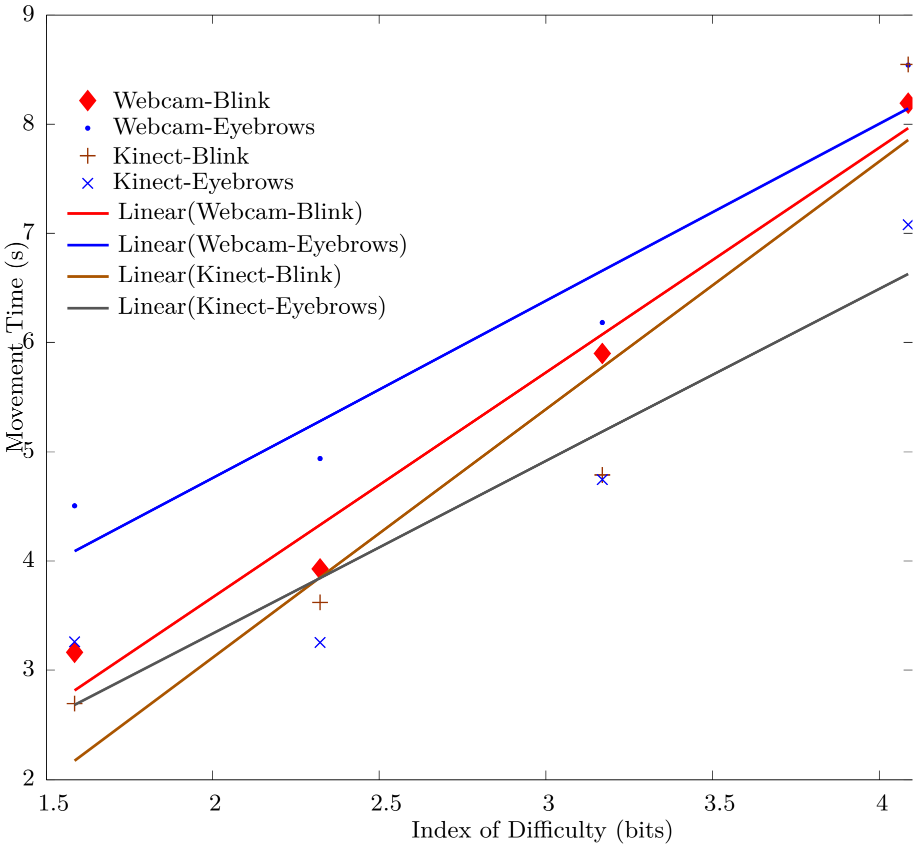


Figure 5.4 Fitts' Test result for Task 1 (Movement to target).

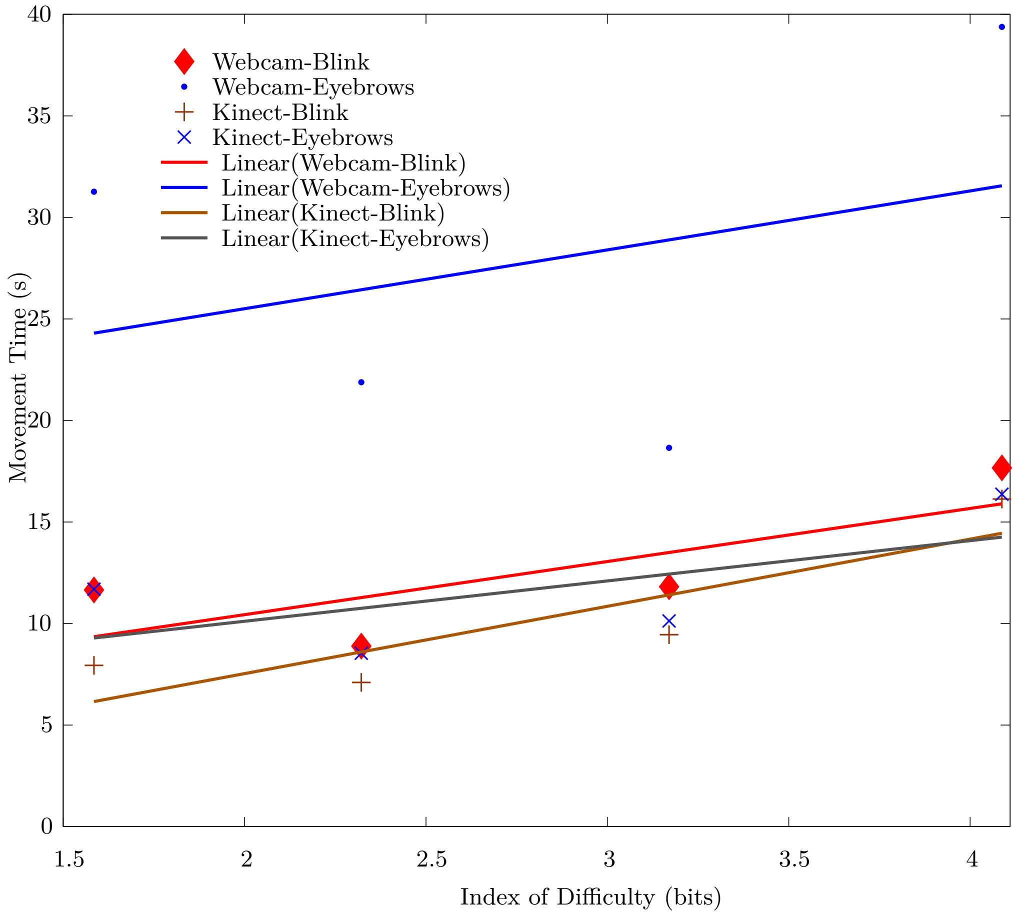


Figure 5.5 Fitts' Test result for Task 2 (Performing the Facial Gesture).

Vision system. R^2 is the coefficient of determination and measured as a percentage of how well the data fits the linear model [154]. In the following table (Table 5.4) the IP for different devices are presented when performing Task 1 and a combination of Task 1 followed by Task 2 for different target orientations.

Table 5.4 IP of Task 1(bits/second)

Device/Orientation	0	45	90	135	180	225	270	315
Webcam-Blink	0.51	0.56	0.42	0.52	0.48	0.4	0.49	0.56
Webcam-Eyebrows	0.94	0.29	1.43	0.67	0.84	1.57	0.83	0.57
Kinect-Blink	0.27	0.57	0.49	0.5	0.3	0.35	0.68	0.77
Kinect-Eyebrows	0.56	0.51	1.26	0.35	0.53	1.23	1.94	0.65

Table 5.5 IP of Task 2(bits/second)

Device/Orientation	0	45	90	135	180	225	270	315
Webcam-Blink	0.63	0.91	0.44	0.12	0.28	0.38	0.7	0.47
Webcam-Eyebrows	0.34	-0.35	-0.73	3.74	0.1	0.13	0.22	1.07
Kinect-Blink	0.29	0.88	0.21	0.37	0.12	0.3	0.37	0.52
Kinect-Eyebrows	0.31	1.54	0.43	0.22	0.21	-5.41	2.82	1.1

Table 5.6 TP_e of Task 1(bits/second)

Device/Orientation	0	45	90	135	180	225	270	315
Webcam-Blink	0.67	0.88	0.67	0.8	0.81	0.59	0.88	0.69
Webcam-Eyebrows	0.96	0.62	0.61	0.71	0.62	0.64	0.62	0.62
Kinect-Blink	0.82	0.8	0.91	0.88	0.73	0.83	1.04	0.98
Kinect-Eyebrows	0.85	0.99	1.12	0.78	0.79	0.9	1.39	0.92

Table 5.7 TP_e of Task 2(bits/second)

Device/Orientation	0	45	90	135	180	225	270	315
Webcam-Blink	0.41	0.45	0.39	0.34	0.47	0.33	0.44	0.41
Webcam-Eyebrows	0.68	0.34	0.32	0.35	0.35	0.3	0.35	0.36
Kinect-Blink	0.45	0.54	0.63	0.64	0.51	0.62	0.71	0.64
Kinect-Eyebrows	0.54	0.55	0.68	0.44	0.54	0.63	1.06	0.67

A one-way ANOVA test was performed on the TP_e for the different orientations and gestures of both Task 1 and Task 2. For the comparison by orientations, p<0.01 (p=0.001 and p=0.007) for Task 1 and Task 2, it can be said that there is significant difference between the mean of the different orientations. For the comparison by gesture, only Task 2 had p<0.01 (p=0.0093). This indicates that there is a significant difference between the mean of the TP_e based on the gesture being performed. The mean of Task 2 is greater due to the increased challenge of both moving and selecting/clicking.

5.2.3 Usability Assessment Survey

Table 5.8 shows the result of the most significant questions from the usability survey. The questionnaire had a 7 Likert scale [124] but it was reduced to only 3 levels due to the number of participants in the survey. The importance of the questions was determined by performing a Kruskal-Wallis test [155, 156]. Kruskal-Wallis test is similar to the one way ANOVA test but instead of using mean, it uses the ranks of the data.

Table 5.8 Main items of user perception of device (%)

Questions	Webcam-blink	Webcam-eyebrows	Kinect-blink	Kinect-eyebrows
4. Eye Fatigue				
Low	37.5	25	75	25
Neither low nor high	25	0	12.5	25
High	37.5	75	12.5	50
8. Physical Effort				
Low	75	25	50	37.5
Neither low nor high	12.5	12.5	25	12.5
High	12.5	62.5	25	50
9. Mental Effort				
Low	37.5	12.5	50	25
Neither low nor high	25	12.5	25	37.5
High	37.5	75	25	37.5
10. Smoothness				
Rough	25	62.5	12.5	12.5
neither rough nor smooth	50	25	37.5	50
Smooth	25	12.5	50	37.5

5.3 Discussion

5.3.1 Fitts' Test

Using facial gestures as a switch is possible in real time but the use of such gestures may cause a drop in the overall IP of the devices. IP and TP_e values in Table 5.3 using the four different devices were obtained with participants successfully reaching and selecting all targets. As it can be seen in results for the overall IP (Table 5.3), the R^2 value which represents the goodness of fit of the fitted line for the Kinect 3D system is greater than 0.7 i.e. the line accounts for more than 70% of the variance. In contrast, the Webcam-EyeBrows device R^2 is 0.21, and thus accounts for only 21% of the variance. This could also indicate that the presence of outliers has a large influence on the fitted line and thus the gradient. As the IP calculation from Equation (3.6) is based on the inverse of the slope, it is also being influenced by outliers at very low and very high indices of difficulty. It should be borne in mind that each of the points in Figures 5.4 and 5.5 are obtained from the mean of data obtained from 22 users and 8 directions giving 64 data points. In the presence of such outliers relying on TP_e as a measure of performance might be better.

It can be seen that there is a decrease in the TP_e of all the four different devices after the switching action is performed. The decrease in the TP_e of the 2D Vision system is 45% and 44% for the blink and eyebrows devices respectively. Similarly, the decrease in the TP_e of the 3D Kinect system is 32% and 35% for the blink and eyebrows devices respectively. The higher total TP_e value indicates that the Kinect system, utilising 3D information, has resulted in better performance when the two tasks of moving and selecting are combined and thus improved the ease of use of the system as a whole. It has also been shown that the TP_e for Task 2 based on gesture are from different populations - with eyebrows having a higher mean TP_e. There is no evidence to support a difference in performance based on sensor or device. This also supports the impact of the improvement in performance of the gesture detection algorithm.

In addition, the facial gesture detection rate affected the MT for the different devices. In this implementation of the Fitts' test, the tasks were considered completed only when the switch was activated and click action was performed. The participants were allowed to repeat the gestures until the click action was detected and thus this caused the movement time to increase.

5.3.2 Usability Assessment Survey

From Table 5.8, it can be seen that the response to Question 4 of the Device Assessment Survey, which was the feedback on the state of eye fatigue, is the most significant with $p<0.05$ ($p=0.0185$) based on gestures and significant with $p<0.1$($p=0.0770$) based on device used, but not on sensor used. This indicates a difference in the median of the feedback answers given by the participant based on the gesture they were performing. Therefore, it confirms the survey data response for Question 4, Table 5.8, where low and neither low nor high eye fatigue is around 75% for blink devices and 37.5% for eyebrow devices. Thus, it can be seen that the users considered that the blink gesture was significantly less tiring than the eyebrows gesture.

For both Question 8 and 9 with $p<0.1$ ($p=0.0691$ and $p=0.0794$) based on gestures, it can be seen that there is a significant difference in the median in the answers given by the participants between blink and eyebrows movement. From Table 5.8 for Question 8, where low and neither low nor high physical fatigue is around 81.3% for blink devices and 43.8% for eyebrow devices. Thus, it can be seen that the users experienced the blink gesture as being significantly less physically fatiguing than the eyebrows gestures. From Table 5.8 for Question 9, low and neither low nor high physical fatigue is around 68.8% for blink devices and 43.8% for eyebrow devices. Thus, it can be seen that the users experienced the blink gesture as being significantly less mentally fatiguing than the eyebrows gestures. This may be explained by the blink gesture being more natural and instinctive than eyebrows movement.

Regarding smoothness of cursor movements (Question 10 Table 5.8 and with $p<0.1$ ($p=0.0519$), based on the different sensor groupings), smooth and neither rough nor smooth is reported to be around 56.3% for web camera-based devices and 87.5% for Kinect-based devices, thus it can be seen that the users experienced significantly smoother movement for the cursor with the Kinect-based devices than with the web camera based devices. This could be due to the use of 3D information to reduce the computational load of subsequent stages in the processing leading to a smoother overall experience with a reduced cursor delay.

5.4 Conclusion

Both Kinect devices have lower MT and higher IP and TP_e than the webcam based devices thus showing that the introduction of the depth data had a positive impact on the head tracking algorithm. This could be explained by the ability to throw away unnecessary data at an early stage in processing using depth information and thereby speeding up subsequent stages to create a smoother experience for the users. The usability assessment survey, suggests that there is a significant difference in eye fatigue experienced by the participants; blink gesture was less tiring to the eye than eyebrows movement gesture. In this work we have looked at only blink and eyebrows movement gestures. Further work will have to be carried out on additional gestures such as mouth opening/closing and tongue movement. We now intend to conduct translational research with neurological patients.

Chapter 6

Conclusions and Future Work

In this chapter, the outcome of the research carried out and potential recommendations for future work that can be investigated are discussed.

6.1 Conclusions

From the preliminary results obtained, assistive devices need to be fine tuned to the residual abilities of disabled individuals. The Fitts' Test may be used to help determine the optimum device for an individual and the best settings such as screen resolution, cursor speed etc. This would reduce the time to find the best well-suited device for an individual, increasing both the usability of the device and adherence by patients/carers. The test could also be used to monitor the progress of the user, but tests with individuals with impairments over a period would be required to be able to validate this process.

Additionally, it has been shown that the use of the depth data improved the use of the device. The use of the depth data with the RGB image enables the removal of the background distractions (based on the distance to the sensor) and thus reduces the noise in the data. The inclusion of a sensor capable of measuring depth information from structured illumination (such as Kinect for Windows [14]) is shown to have the potential to enhance the performance of vision based assistive device. From our RGB-D system, the TP_e of dwell clicking increased by a third (from 0.21 to 0.30 bits per second) and that of blink clicking doubled (from 0.15 to 0.28 bits per second) compared to our 2D system. The eyebrow detection algorithm has the worst performance as it can be seen from the increased MT. During the experiments, it was found that the eye blink detection algorithm performance was quite near to the performance of the dwell click switch using depth. The proposed system should also be tested in varying illumination conditions to further test the robustness of the system.

Both Kinect devices have lower MT and higher IP and TP_e than the web-camera based devices, thus showing that the introduction of the depth data had a positive impact on the head tracking algorithm. This could be explained by the ability to throw away unnecessary data at an early stage in processing using depth information and thereby speeding up subsequent stages to create a more smooth experience for the users. The usability assessment survey, suggest that there is a significant difference in eye fatigue experienced by the participants; blink gesture was less tiring to the eye than eyebrows movement gesture. In this work, we have looked at only blink and eyebrows movement gestures, further work will have to be carried out on additional gestures such as mouth opening/closing and tongue movement. We now intend to conduct translational research with neurological patients to varying degrees in order to validate the system, its efficacy and utility.

The Kinect-Blink device was also recently used to control a motorised wheelchair. The motor-powered wheelchair was the outcome of the team at the University of Kent participation in the COALAS - Cognitive Assisted Living Ambient System - project. A demo of the head controlled wheelchair was presented at a meeting for COALAS project.

6.2 Future Work

This section provides recommendations for future work that can be carried out to further the work started in this thesis. One of the main areas of focus should be to perform the experiments with individuals with impairment. One of the limitations of this work is, only one individual with impairments was recruited to participate in the experiments.

More individuals with impairments should be recruited to trial the different devices and also to test our facial gesture algorithms in real world situations.

A larger database of depth facial gesture data should be created. The database created contains the data of only 22 individuals performing at least three gestures - blink, eyebrows movement and mouth movement (opening and closing). More types of gestures such as tongue movement could be captured to improve the database. In our current investigation only a small set of users was used to obtain the data used in our experimentation. A larger set of individuals should be tested to get a better idea of the performance of the facial gesture recognition algorithm.

Adaptivity of the devices should be investigated. Using calibration, the participants were able to use their range of neck movement to control the cursor over the entire screen. The adaptivity of the devices should be improved to enable the system to select the best gesture for the user from a range of available gestures. For example this could enable the system to select eye blink over eyebrows movements, if the eyebrows were to be covered by hair over the forehead.

The effect of hair i.e. both facial hair and hair from the top of the head on the Kinect depth data should be further investigated.

Improvements to the accuracy and the performance of the facial gesture switching algorithms should be investigated. Improving the accuracy of the blink detection algorithm could greatly improve the accuracy of the eye blink switching mechanism and also the performance of the head tracking system. And similarly improving the accuracy of the eyebrows gesture detection algorithm could greatly improve the accuracy of the eyebrows switching mechanism, as well as the performance of the head tracking system. Potential venues of investigation could be the location of the sensors i.e. placing the sensor below the monitor instead of on top, or using multiple sensors at the same time.

Further investigation of the proposed modified Fitts' Test as a monitor of the progression of the state of the impairment should be carried out. This is the use of the modified Fitts' Test in a longitudinal study where individuals with impairments are tested regularly over a period of time. The tests should be conducted along with a QUEST 2.0, PIADS or ICF questionnaire to investigate the impact on any change in the progression of the impairment and whether it can be detected in both the questionnaires and in the Fitts' Test.

The performance of Mouth Opening and Closing switching algorithms should be investigated and compared to the performance in terms of both device performance and user accessibility with the eye blink and eyebrows movement algorithms. Both of the current gestures investigated in the this work are upper facial gestures, the comparison with a lower facial gesture could offer some additional insights in the usability of facial gestures as switching mechanisms for assistive devices.

6.3 Concluding Remarks

The aim of this rsearch was to develop a low cost Head Tracking system. The KentAssist devices are low cost devices which can be used with either a web camera or the Kinect [14]. The performance of the devices was evaluated using a modified Fitts' Test. It was shown that the performances of the devices using the Kinect were better than the web camera based devices. The enhanced performance of the Kinect based devices indicated that the addition of depth data could have contributed to the improved performance. Both an objective and subjective evaluation of the assistive devices were performed with the help of both healthy and disabled participants. Both an eye blink and eyebrows movement switching mechanisms were developed and evaluated. Also, a dataset for facial gestures with both colour facial and depth images was created. The participants were requested to point at different locations on the screen and perform timed gestures consisting of blinking, eyebrows movement and mouth closure gestures.

References

[1] World Health Organization, "World report on disability," WHO, Geneva, Tech. Rep., 2011. [Online]. Available: http://whqlibdoc.who.int/publications/2011/9789240685215_eng.pdf

[2] Royal College of Physicians, *Medical rehabilitation in 2011 and beyond.- Report of a working party*. London: RCP: Royal college of Physicians & British Society of Rehabilitation Medicine, 2010, no. November 2010.

[3] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni, B. Gannon, D. H. Jones, P. Jennum, A. Jordanova, L. Jönsson, K. Karampampa, M. Knapp, G. Kobelt, T. Kurth, R. Lieb, M. Linde, C. Ljungcrantz, A. Maercker, B. Melin, M. Moscarelli, A. Musayev, F. Norwood, M. Preisig, M. Pugliatti, J. Rehm, L. Salvador-Carulla, B. Schlehofer, R. Simon, H.-C. Steinhausen, L. J. Stovner, J.-M. Vallat, P. Van den Bergh, P. V. den Bergh, J. van Os, P. Vos, W. Xu, H.-U. Wittchen, B. Jönsson, and J. Olesen, "Cost of disorders of the brain in Europe 2010." *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*, vol. 21, no. 10, pp. 718–79, Oct. 2011. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21924589

[4] Human Rights Council, “Report of the Special Rapporteur on the rights of persons with disabilities, Catalina Devandas-Aguilar,” Office of the High Commissioner on Human Rights, United Nations, Tech. Rep. February, 2015.

[5] United Nations and Union, Inter-parliamentary, *From Exclusion to Equality: Realizing the Rights of Persons with Disabilities: Handbook for Parliamentarians on the Convention of the Rights of Persons with Disabilities and Its Optional Protocol.* UN, 2007.

[6] United Nations, “Convention on the Rights of Persons with Disabilities Committee on the Rights of Persons with Disabilities Implementation of the Convention on the Rights of Persons with Disabilities Initial reports submitted by States parties under article 35 of the Convention,” no. October, pp. 1–59, 2013.

[7] UNCRPD, “Convention on the Rights of Persons with Disabilities,” p. Article 24.1, 2006. [Online]. Available: http://www.ohchr.org/EN/HRBodies/CRPD/Pages/ConventionRightsPersonsWithDisabilities.aspx#26 [Last Accessed: 2015-08-26]

[8] SmartNav. [Online]. Available: http://www.naturalpoint.com/smartnav/ [Last Accessed: 2012-03-02]

[9] Tobii Technology. [Online]. Available: http://www.tobii.com/ [Last Accessed: 2014-03-06]

[10] F. Loewenich and F. Maire, “Hands-free mouse-pointer manipulation using motion-tracking and speech recognition,” in *Proceedings of the 2007 conference of the computer-human interaction special interest group (CHISIG) of Australia*

on Computer-human interaction: design: activities, artifacts and environments - OZCHI '07. Association for Computing Machinery (ACM), 2007. [Online]. Available: http://dx.doi.org/10.1145/1324892.1324955

[11] R. Bates and H. Istance, "Why are eye mice unpopular? a detailed comparison of head and eye controlled assistive technology pointing devices," *Universal Access in the Information Society*, vol. 2, no. 3, pp. 280–290, 2003. [Online]. Available: http://dx.doi.org/10.1007/s10209-003-0053-y

[12] N. Ragot, G. Caron, M. Sakel, and K. Sirlantzis, "Coalas: A eu multidisciplinary research project for assistive robotics neuro-rehabilitation," in *IEEE/RSJ IROS Workshop on Rehabilitation and Assistive Robotics: Bridging the Gap Between Clinicians and Roboticists, Chicago, USA*, 2014. [Online]. Available: http://users.eecs.northwestern.edu/~argall/14rar/submissions/ragot.pdf

[13] E. Catley, S. P. Guness, K. Sirlantzis, G. Howells, and S. Kelly, "A novel depth-based head tracking system for proportional wheelchair control," in *TechAAL 2015*, 5 Nov. 2015.

[14] "Kinect for Windows." [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/ [Last Accessed: 2013-02-19]

[15] T. Kanade, J. Cohn, and Y. T. Y. Tian, "Comprehensive database for facial expression analysis," *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=840611

[16] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *2007 IEEE 11th International Conference on Computer Vision*. Institute of Electrical & Electronics Engineers (IEEE), 2007. [Online]. Available: http://dx.doi.org/10.1109/iccv.2007.4409068

[17] K. Francis, J. R. Bach, and J. A. DeLisa, "Evaluation and rehabilitation of patients with adult motor neuron disease." *Archives of physical medicine and rehabilitation*, vol. 80, no. 8, pp. 951–963, Aug. 1999. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10453774

[18] L. Holper, M. Coenen, A. Weise, G. Stucki, A. Cieza, and J. Kesselring, "Characterization of functioning in multiple sclerosis using the ICF," *Journal of Neurology*, vol. 257, no. 1, pp. 103–113, Jan. 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19756827

[19] World Health Organization, "Atlas: Country Resources for Neurological Disorders," 2004. [Online]. Available: http://www.who.int/mental_health/neurology/neurogy_atlas_lr.pdf?ua=1

[20] P. Rosenbaum, N. Paneth, A. Leviton, M. Goldstein, M. Bax, D. Damiano, B. Dan, and B. Jacobsson, "A report: The definition and classification of cerebral palsy April 2006," *Developmental Medicine and Child Neurology*, vol. 49, no. April, pp. 8–14, 2007.

[21] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2001.

[22] R. Lienhart and J. Maydt, "An extended set of Haar-like features for rapid object detection," *Proceedings. International Conference on Image Processing*, vol. 1, pp. 900–903, 2002. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1038171

[23] F. C. Crow, "Summed-area tables for texture mapping," *ACM SIGGRAPH Computer Graphics*, vol. 18, no. 3, pp. 207–212, 1984.

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[25] L. Sigal, S. Sclaroff, and V. Athitsos, "Skin color-based video segmentation under time-varying illumination," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 7, pp. 862–877, Jul. 2004. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2004.35

[26] A. R. Smith, "Color gamut transform pairs," *ACM SIGGRAPH Computer Graphics*, vol. 12, no. 3, pp. 12–19, aug 1978. [Online]. Available: http://dx.doi.org/10.1145/965139.807361

[27] M. Jones and J. Rehg, "Statistical color models with application to skin detection," in *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*. Institute of Electrical & Electronics Engineers (IEEE), 1999. [Online]. Available: http://dx.doi.org/10.1109/cvpr.1999.786951

[28] F. Dadgostar and A. Sarrafzadeh, "An adaptive real-time skin detector based on Hue thresholding: A comparison on two motion tracking methods," *Pattern Recognition Letters*, vol. 27, no. 12, pp. 1342–1352, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865506000134

[29] Y.-T. Pai, S.-J. Ruan, M.-C. Shie, and Y.-C. Liu, "A simple and accurate color face detection algorithm in complex background," in *2006 IEEE International Conference on Multimedia and Expo*. IEEE, Jul. 2006, pp. 1545–1548. [Online]. Available: http://dx.doi.org/10.1109/ICME.2006.262838

[30] B. Li, X. Xue, and J. Fan, "A robust incremental learning framework for accurate skin region segmentation in color images," *Pattern Recognition*, vol. 40, no. 12, pp. 3621–3632, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320307002105

[31] A. Soetedjo and K. Yamada, "Skin color segmentation using coarse-to-fine region on normalized RGB chromaticity diagram for face detection," *IEICE Transactions on Information and Systems*, vol. E91-D, pp. 2493–2502, 2008. [Online]. Available: http://kjs.nagaokaut.ac.jp/yamada/papers/e91-d_10_2493.pdf

[32] H. M. Sun, "Skin detection for single images using dynamic skin color modeling," *Pattern Recognition*, vol. 43, no. 4, pp. 1413–1420, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320309003665

[33] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, Jan. 2002. [Online]. Available: http://www.springerlink.com/content/r0833464v0675146/

[34] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," *UMass Amherst Technical Report*, no. January, 2010. [Online]. Available: http://works.bepress.com/erik_learned_miller/55/

[35] Y. Ban, S.-K. Kim, S. Kim, K.-A. Toh, and S. Lee, "Face detection based on skin color likelihood," *Pattern Recognition*, vol. 47, no. 4, pp. 1573–1585, Apr. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S003132031300455X

[36] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition*. IEEE Comput. Soc. Press, 1994. [Online]. Available: http://dx.doi.org/10.1109/icpr.1994.576366

[37] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, jan 1996. [Online]. Available: http://dx.doi.org/10.1016/0031-3203(95)00067-4

[38] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures*, vol. 6. venus.inrialpes.fr, 2004. [Online]. Available: http://venus.inrialpes.fr/jlc/papers/Pointing04-Gourier.pdf

[39] M. Stegmann, B. Ersboll, and R. Larsen, "Fame-a flexible appearance modeling environment," *IEEE Transactions on Medical Imaging*, vol. 22, no. 10, pp. 1319–1331, oct 2003. [Online]. Available: http://dx.doi.org/10.1109/tmi.2003.817780

[40] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) database," in *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002*, no. 1. Ieee, 2002, pp. 53–58. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1004130

[41] M. Weber, "Frontal face dataset," *California Inst. of Technology, http://www. vision. caltech. edu/html-files/archive. html*, 1999.

[42] R. Frischholz, "Bao face database at the face detection homepage," 2012.

[43] D. R. Iskander, S. Mioschek, M. Trunk, and W. Werth, "Detecting eyes in digital images," in *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings.*, vol. 2, Jul. 2003, pp. 21–24 vol.2. [Online]. Available: http://dx.doi.org/10.1109/ISSPA.2003.1224805

[44] K. Peng, L. Chen, S. Ruan, and G. Kukharev, "A robust agorithm for eye detection on gray intensity face without spectacles," *Journal of computer science and technology*, vol. 5, 2005. [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/9584

[45] A. De Santis and D. Iacoviello, "Robust real time eye tracking for computer interface for disabled people." *Computer methods and programs in biomedicine*, vol. 96, no. 1, pp. 1–11, Oct. 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19406500http://www.scopus.com/inward/record.url?eid=2-s2.0-67650688592&partnerID=40&md5=bf2a30f28f750459e25657c9a216a338

[46] T. Rajpathak, R. Kumar, and E. Schwartz, "Eye Detection Using Morphological and Color Image Processing," *Proceeding of Florida Conference on Recent Advances in Robotics*, pp. 1–6, 2009.

[47] S. Han, S. Yang, J. Kim, and M. Gerla, "EyeGuardian," in *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications - HotMobile '12*, 2012, p. 1. [Online]. Available: http://www.hotmobile.org/2012/papers/HotMobile12-final72.pdfhttp://dl.acm.org/citation.cfm?id=2162090http://dl.acm.org/citation.cfm?doid=2162081.2162090

[48] AT&T and Cambridge University Engineering Department, "The Database of Faces or formerly known as The ORL Database of Faces." [Online]. Available: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html [Last Accessed: 2015-05-04]

[49] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," *Proceedings of 1994 IEEE Workshop on Applications of Computer Vision*, pp. 138–142, 1994.

[50] N. T. S. C. (1951-1953), *[Report and Reports of Panel No. 11, 11-A, 12-19, with Some supplementary references cited in the Reports, and the Petition for adoption of transmission standards for color television before the Federal Communications Commission*, ser. Report, National Television System Committee (1951-1953), Ed., 1953, no. pt. 11. [Online]. Available: https://books.google.mu/books?id=pAw2AQAAIAAJ

[51] National Television System Committee (1951-1953) and D. G. Fink, Eds., *Color television standards: selected papers and records*, ser. McGraw-Hill television series. New York: McGraw-Hill, 1955.

[52] K. Grauman, M. Betke, J. Lombardi, J. Gips, and G. Bradski, "Communication via eye blinks and eyebrow raises: video-based human-computer interfaces," *Universal Access in the Information Society*, vol. 2, no. 4, pp. 359–373, nov 2003. [Online]. Available: http://dx.doi.org/10.1007/s10209-003-0062-x

[53] D. Anson, G. Lawler, A. Kissinger, M. Timko, J. Tuminski, and B. Drew, "Efficacy of three head-pointing devices for a mouse emulation task," *Assistive Technology*, vol. 14, no. 2, pp. 140–150, Dec 2002. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10400435.2002.10132063

[54] R. Cloud, M. Betke, and J. Gips, "Experiments with a camera-based human-computer interface system," *Proceedings of the 7th ERCIM Workshop "User Interfaces for All," UI4ALL 2002*, pp. 103–110, 2002. [Online]. Available: http://cstest.bc.edu/~gips/UI4ALL-2002.pdfhttp://www.cs.bu.edu/faculty/betke/papers/Cloud-Betke-Gips-UI4ALL-2002.pdf

[55] M. Ashdown, K. Oka, and Y. Sato, "Combining head tracking and mouse input for a GUI on multiple monitors," in *CHI '05 extended abstracts on Human factors in computing systems - CHI '05*. Association for Computing Machinery (ACM), 2005. [Online]. Available: http://dx.doi.org/10.1145/1056808.1056873

[56] W. Akram, L. Tiberii, and M. Betke, "A Customizable Camera-Based Human Computer Interaction System Allowing People with Disabilities

Autonomous Hands-Free Navigation of Multiple Computing Tasks," in *Proceedings of the 9th conference on User interfaces for all*, ser. ERCIM'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 28–42. [Online]. Available: http://portal.acm.org/citation.cfm?id=1783789.1783793

[57] C. A. M. Pereira, R. Bolliger Neto, A. C. Reynaldo, M. C. D. M. Luzo, R. P. Oliveira, R. B. Neto, and M. C. de Miranda Luzo, "Development and evaluation of a head-controlled human-computer interface with mouse-like functions for physically disabled users," *Clinics*, vol. 64, no. 10, pp. 975–981, Jan. 2009. [Online]. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-59322009001000007&lng=en&nrm=iso&tlng=enhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2763072&tool=pmcentrez&rendertype=abstract

[58] M. Betke, *Intelligent Interfaces to Empower People with Disabilities.* Boston, MA: Springer US, 2010, pp. 409–432. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-93808-0_15

[59] M. Betke, J. Gips, and P. Fleming, "The camera mouse: visual tracking of body features to provide computer access for people with severe disabilities," *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 10, no. 1, pp. 1–10, Mar. 2002. [Online]. Available: http://dx.doi.org/10.1109/TNSRE.2002.1021581

[60] D. O. Gorodnichy and G. Roth, "Nouse 'use your nose as a mouse' perceptual vision technology for hands-free games and interfaces," *Proceedings from the*

15th International Conference on Vision Interface, vol. 22, no. 12, pp. 931–942, Jan. 2004. [Online]. Available: http://www.sciencedirect.com/science/article/B6V09-4CGMFJJ-4/2/d0bb649628b49f47edc22aeb90eb407b

[61] J.-S. Jang and T. Kanade, "Robust 3D head tracking by online feature registration," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2008, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4813307

[62] D. G. Lowe, "Distinctive image features from Scale-Invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[63] R. Kjeldsen, "The IBM HeadTracking Pointer: improvements in vision-based pointer control." *Disability and rehabilitation. Assistive technology*, vol. 3, no. 4, pp. 208–220, 2008.

[64] P. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement." *Journal of experimental psychology*, vol. 47, no. 6, 1954. [Online]. Available: http://psycnet.apa.org/journals/xge/47/6/381/

[65] ISO TC 159/SC 4, *ISO 9241-9:2000, Ergonomic requirements for office work with visual display terminals (VDTs) - Part 9: Requirements for non-keyboard input devices*. International Organization for Standardization, 2002. [Online]. Available: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=53372&published=on

[66] F. Loewenich and F. Maire, "A Head-Tracker based on the Lucas-Kanade optical flow algorithm," in *Proceedings of the 2006 conference on Advances in Intelligent IT: Active Media Technology 2006*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 19 May 2006, pp. 25–30. [Online]. Available: http://portal.acm.org/citation.cfm?id=1566561.1566569

[67] B. D. Lucas, T. Kanade, and Others, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, vol. 81, 1981, pp. 674–679. [Online]. Available: http://www.ri.cmu.edu/pub_files/pub3/lucas_bruce_d_1981_1/lucas_bruce_d_1981_1.ps.gz

[68] S. Epstein, E. Missimer, and M. Betke, "Using kernels for a video-based mouse-replacement interface," in *Personal and Ubiquitous Computing*, vol. 18, 2014, pp. 47–60.

[69] K. R. Park, M. C. Whang, and J. S. Lim, "A study on non-intrusive facial and eye gaze detection," in *Advanced Concepts for Intelligent Vision Systems*, ser. Lecture Notes in Computer Science, J. Blanc-Talon, W. Philips, D. Popescu, and P. Scheunders, Eds. Springer Berlin Heidelberg, 20 Sep. 2005, pp. 52–59. [Online]. Available: http://link.springer.com/chapter/10.1007/11558484_7

[70] R. Adolphs, "A landmark study finds that when we look at sad faces, the size of the pupil we look at influences the size of our own pupil," *Social cognitive and affective neuroscience*, vol. 1, no. 1, pp. 3–4, 3 Jun. 2006. [Online]. Available: https://academic.oup.com/scan/article-lookup/doi/10.1093/scan/nsl011

[71] N. A. Harrison, T. Singer, P. Rotshtein, R. J. Dolan, and H. D. Critchley, "Pupillary contagion: central mechanisms engaged in sadness processing," *Social cognitive and affective neuroscience*, vol. 1, no. 1, pp. 5–17, Jun. 2006. [Online]. Available: http://dx.doi.org/10.1093/scan/nsl006

[72] J. T.-Y. Wang, M. Spezio, and C. F. Camerer, "Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in Sender-Receiver games," *The American economic review*, vol. 100, no. 3, pp. 984–1007, 1 Jun. 2010. [Online]. Available: http://www.ingentaconnect.com/content/aea/aer/2010/00000100/00000003/art00012

[73] C. M. Privitera, L. W. Renninger, T. Carney, S. Klein, and M. Aguilar, "Pupil dilation during visual target detection." *Journal of vision*, vol. 10, no. 10, p. 3, 2010. [Online]. Available: http://journalofvision.orgwww.journalofvision.org/content/10/10/3.short

[74] M. Castelhano, M. Wieth, and J. Henderson, "I See What You See: Eye Movements in Real-World Scenes Are Affected by Perceived Direction of Gaze," *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, pp. 251–262, 2008. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-77343-6_16

[75] M. S. Castelhano, M. L. Mack, and J. M. Henderson, "Viewing task influences eye movement control during active scene perception," *Journal of vision*, vol. 9, no. 3, pp. 6.1–15, 13 Mar. 2009. [Online]. Available: http://dx.doi.org/10.1167/9.3.6

[76] J. San Agustin, H. Skovsgaard, J. P. Hansen, and D. W. Hansen, "Low-cost gaze interaction," in *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09*, vol. Boston, MA, Apr. 2009, p. 4453. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-70349173044&partnerID=40&md5=072fc5d6e24d13181309472d769ded36http://portal.acm.org/citation.cfm?doid=1520340.1520682

[77] C. N. Chan, S. Oe, and C. S. Lin, "Active eye-tracking system by using quad PTZ cameras," in *IECON 2007 - 33rd Annual Conference of the IEEE Industrial Electronics Society*. ieeexplore.ieee.org, Nov. 2007, pp. 2389–2394. [Online]. Available: http://dx.doi.org/10.1109/IECON.2007.4460366

[78] S. Asteriadis, K. Karpouzis, and S. Kollias, "Feature extraction and selection for inferring user engagement in an HCI environment," in *Human-Computer Interaction. New Trends*. Springer Science + Business Media, 2009, pp. 22–29. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-02574-7_3

[79] J.-Y. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm," *Intel Corporation*, vol. 5, pp. 1–10, 2001. [Online]. Available: http://robots.stanford.edu/cs223b04/algo_affine_tracking.pdfhttp://trac.assembla.com/dilz_mgr/export/272/doc/ktl-tracking/algo_tracking.pdf

[80] Y.-P. Lin, Y.-P. Chao, C.-C. Lin, and J.-H. Chen, "Webcam mouse using face and eye tracking in various illumination environments," *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and*

Biology Society. Conference, vol. 4, pp. 3738–3741, 2005. [Online]. Available: http://dx.doi.org/10.1109/IEMBS.2005.1617296

[81] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 696–706, May 2002. [Online]. Available: http://dx.doi.org/10.1109/34.1000242

[82] J. Magee, M. Scott, B. Waber, and M. Betke, "EyeKeys: A Real-Time Vision Interface Based on Gaze Detection from a Low-Grade Video Camera," *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004.

[83] S. Bernet, D. Lefloch, C. Cudel, and M. Basset, "Rapid calibration for eye tracking," in *21st International Conference of the Association Francophone dInteraction HommeMachine IHM 09*, vol. Grenoble, Oct. 2009, pp. 173–176. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-73249120804&partnerID=40&md5=cc1fa5333b90269caf54540d0cda1713

[84] D. Hansen, J. Agustin, and A. Villanueva, "Homography normalization for robust gaze estimation in uncalibrated setups," *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*, vol. 1, no. 212, p. 13, 2010. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1743666.1743670$\delimiter"026E30F$nhttp://dl.acm.org/citation.cfm?id=1743670

[85] T. Nagamatsu, J. Kamahara, and N. Tanaka, "Calibration-free gaze tracking using a binocular 3D eye model," in *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '09, vol. Boston,

MA. New York, NY, USA: ACM, 2009, pp. 3613–3618. [Online]. Available: http://doi.acm.org/10.1145/1520340.1520543

[86] J. Y. Kaminski, D. Knaan, and A. Shavit, "Single image face orientation and gaze detection," *Machine Vision and Applications*, vol. 21, no. 1, pp. 85–98, 2009. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-70350046976&partnerID=40&md5=3ee59666875102361a14dbfc1ed5ed83

[87] C. N. Karson, "Spontaneous eye-blink rates and dopaminergic systems." *Brain : a journal of neurology*, vol. 106 (Pt 3), pp. 643–53, Sep. 1983. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/6640274

[88] G. F. Wilson, "An analysis of mental workload in pilots during flight using multiple psychophysiological measures," *The International Journal of Aviation Psychology*, vol. 12, no. 1, pp. 3–18, Jan 2002. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/S15327108IJAP1201_2

[89] K. L. Schmidt, S. Bhattacharya, and R. Denlinger, "Comparison of Deliberate and Spontaneous Facial Movement in Smiles and Eyebrow Raises." *Journal of nonverbal behavior*, vol. 33, no. 1, pp. 35–45, Mar. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2843933&tool=pmcentrez&rendertype=abstract

[90] K. L. Schmidt, Z. Ambadar, J. F. Cohn, and L. I. Reed, "MOVEMENT DIFFERENCES BETWEEN DELIBERATE AND SPONTANEOUS FACIAL EXPRESSIONS: ZYGOMATICUS MAJOR ACTION IN SMILING," *Journal*

of nonverbal behavior, vol. 30, no. 1, pp. 37–52, 2006. [Online]. Available: http://dx.doi.org/10.1007/s10919-005-0003-x

[91] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 2. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2005, pp. 568–573 vol. 2. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.297

[92] A. Królak and P. Strumiłło, "Eye-blink detection system for human–computer interaction," *Universal Access in the Information Society*, vol. 11, no. 4, pp. 409–419, 2 Oct. 2011. [Online]. Available: http://link.springer.com/article/10.1007/s10209-011-0256-6

[93] J. O. Wobbrock, J. Rubinstein, M. W. Sawyer, and A. T. Duchowski, "Longitudinal evaluation of discrete consecutive gaze gestures for text entry," in *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, ser. ETRA '08. New York, NY, USA: ACM, 2008, pp. 11–18. [Online]. Available: http://doi.acm.org/10.1145/1344471.1344475

[94] M. Chau and M. Betke, "Real Time Eye Tracking and Blink Detection with USB Cameras," *Boston University Computer Science*, vol. 2215, no. 2005-12, pp. 1–10, 2005. [Online]. Available: http://dcommon.bu.edu/xmlui/handle/2144/1839http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.3862&rep=rep1&type=pdf

[95] L. Yunqi, Y. Meiling, S. Xiaobing, L. Xiuxia, and O. Jiangfan, "Recognition of eye states in real time video," in *2009 International Conference on Computer Engineering and Technology*, vol. 1. Ieee, Jan. 2009, pp. 554–559. [Online]. Available: http://dx.doi.org/10.1109/ICCET.2009.105

[96] E. Missimer and M. Betke, "Blink and wink detection for mouse pointer control," in *Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '10. New York, NY, USA: ACM, 2010, pp. 23:1–23:8. [Online]. Available: http://doi.acm.org/10.1145/1839294.1839322

[97] L. Del Val, M. I. Jiménez, A. Alonso, R. De La Rosa, A. Izquierdo, and A. Carrera, "Assistance system for disabled people: A robot controlled by blinking and wireless link," in *Communications in Computer and Information Science*, vol. 111 CCIS, 2010, pp. 383–388. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-642-16318-0_45

[98] A. Krolak and P. Strumillo, "Vision-based eye blink monitoring system for human-computer interfacing," in *2008 Conference on Human System Interactions*. Institute of Electrical & Electronics Engineers (IEEE), may 2008. [Online]. Available: http://dx.doi.org/10.1109/hsi.2008.4581580

[99] S. Deb and S. Deb, "Designing an intelligent blink analyzer tool for effective human computer interaction through eye," *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*, pp. 1–5, Dec. 2012. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6481806

[100] M. Su, C. Yeh, and S. Lin, "An implementation of an eye-blink-based communication aid for people with severe disabilities," *Audio, Language and ...*, pp. 351–356, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4590090

[101] I. S. MacKenzie and A. Behrooz, "Blinkwrite: efficient text entry using eye blinks," *Universal Access in the Information Society*, vol. 10, no. 1, pp. 69–80, Mar 2011. [Online]. Available: http://link.springer.com/10.1007/s10209-010-0188-6

[102] D. J. Higginbotham and G. W. Lesher, "The ambiguous keyboard," *SIG 12 Perspectives on Augmentative and Alternative Communication*, vol. 13, no. 1, pp. 12–16, 2004. [Online]. Available: http://dx.doi.org/10.1044/aac13.1.12

[103] I. S. Mackenzie and T. Felzer, "SAK: Scanning ambiguous keyboard for efficient one-key text entry," *ACM Trans. Comput. -Hum. Interact.*, vol. 17, no. 3, pp. 11:1–11:39, Jul. 2010. [Online]. Available: http://doi.acm.org/10.1145/1806923.1806925

[104] T. Felzer, I. S. MacKenzie, P. Beckerle, and S. Rinderknecht, "Qanti: A Software Tool for Quick Ambiguous Non-standard Text Input," in *Computers Helping People with Special Needs*. Springer Science + Business Media, 2010, pp. 128–135. [Online]. Available: http://link.springer.com/10.1007/978-3-642-14100-3_20

[105] W. O. Lee, E. C. Lee, and K. R. Park, "Blink detection robust to various facial poses," *Journal of Neuroscience Methods*, vol. 193, no. 2, pp. 356–372, Nov 2010. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165027010004942

[106] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, Jun 1994, pp. 593–600.

[107] G. P. Nam, B. J. Kang, and K. R. Park, "Robustness of face recognition to variations of illumination on mobile devices based on SVM," *KSII Transactions on Internet and Information Systems*, vol. 4, no. 1, pp. 25–44, 2010. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-77649174653&partnerID=tZOtx3y1

[108] O. Tuisku, V. Surakka, T. Vanhala, V. Rantanen, and J. Lekkala, "Wireless face interface: Using voluntary gaze direction and facial muscle activations for human-computer interaction," *Interacting with Computers*, vol. 24, no. 1, pp. 1–9, Jan 2012. [Online]. Available: http://iwc.oxfordjournals.org/cgi/doi/10.1016/j.intcom.2011.10.002

[109] J. Lombardi and M. Betke, "A camera-based eyebrow tracker for hands-free computer control via a binary switch," in *Proceedings of the 7th ERCIM Workshop, User Interfaces For All (UI4ALL 2002)*, 2002, pp. 199–200. [Online]. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:A+Camera-based+Eyebrow+Tracker+for+Computer+Control+via+a+Binary+Switch#0

[110] G. E. Lancioni, D. Bellini, D. Oliva, N. N. Singh, M. F. O'Reilly, and J. Sigafoos, "Camera-based microswitch technology for eyelid and mouth responses of persons with profound multiple disabilities: two case studies." *Research in*

developmental disabilities, vol. 31, no. 6, pp. 1509–1514, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S089142221000140X

[111] G. E. Lancioni, D. Bellini, D. Oliva, N. N. Singh, M. F. O'Reilly, R. Lang, and R. Didden, "Camera-based microswitch technology to monitor mouth, eyebrow, and eyelid responses of children with profound multiple disabilities," *Journal of Behavioral Education*, vol. 20, no. 1, pp. 4–14, Mar 2011. [Online]. Available: http://link.springer.com/10.1007/s10864-010-9117-2

[112] D. J. Ward, A. F. Blackwell, and D. J. C. MacKay, "Dasher—a data entry interface using continuous gestures and language models," in *Proceedings of the 13th annual ACM symposium on User interface software and technology*. New York, New York, USA: ACM, 1 Nov. 2000, pp. 129–137. [Online]. Available: http://portal.acm.org/citation.cfm?doid=354401.354427

[113] J. Abascal and C. Nicolle, "Moving towards inclusive design guidelines for socially and ethically aware hci," *Interacting with Computers*, vol. 17, no. 5, pp. 484–505, Sep 2005. [Online]. Available: http://iwc.oxfordjournals.org/cgi/doi/10.1016/j.intcom.2005.03.002

[114] I. S. MacKenzie, "Fitts' law as a research and design tool in human-computer interaction," *Human-Computer Interaction*, vol. 7, no. 1, pp. 91–139, Mar 1992. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327051hci0701_3

[115] J. Accot and S. Zhai, "Beyond Fitts' law: models for trajectory-based HCI tasks," *Proceedings of the SIGCHI conference on Human*, pp. 295–301, 1997. [Online]. Available: http://dl.acm.org/citation.cfm?id=258760

[116] S. A. Douglas, A. E. Kirkpatrick, and I. S. MacKenzie, "Testing pointing device performance and user assessment with the iso 9241, part 9 standard," in *Proceedings of the SIGCHI conference on Human factors in computing systems the CHI is the limit - CHI '99*. ACM Press, 1999, pp. 215–222. [Online]. Available: http://portal.acm.org/citation.cfm?doid=302979.303042

[117] F. Hwang, S. Keates, P. Langdon, and P. Clarkson, "A submovement analysis of cursor trajectories," *Behaviour & Information Technology*, vol. 24, no. 3, pp. 205–217, 2005. [Online]. Available: http://dx.doi.org/10.1080/01449290412331327474

[118] R. Bates and H. O. Istance, "Why are eye mice unpopular? a detailed comparison of head and eye controlled assistive technology pointing devices," *Universal Access in the Information Society*, vol. 2, no. 3, pp. 280–290, 6 Aug. 2003. [Online]. Available: http://link.springer.com/article/10.1007/s10209-003-0053-y

[119] D. W. K. Man and M. S. L. Wong, "Evaluation of computer-access solutions for students with quadriplegic athetoid cerebral palsy," *American Journal of Occupational Therapy*, vol. 61, no. 3, pp. 355–364, may 2007. [Online]. Available: http://dx.doi.org/10.5014/ajot.61.3.355

[120] X. Zhang and I. S. MacKenzie, "Evaluating Eye Tracking with ISO 9241 - Part 9," in *International Conference on Human-Computer Interaction (HCI)*, vol. 4552 / 200. Springer-Verlag, 2007, pp. 779–788.

[121] R. Javanovic and I. S. MacKenzie, "MarkerMouse: Mouse cursor control using a head-mounted marker," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6180 LNCS, no. PART 2, 2010, pp. 49–56. [Online]. Available: http://www.springerlink.com/index/A0452P042145V152.pdf

[122] H. Day and J. Jutai, "Measuring the psychosocial impact of assistive devices: The PIADS." *Canadian Journal of Rehabilitation*, vol. 9, no. 2, pp. 159–168, 1996.

[123] C. Gummesson, I. Atroshi, and C. Ekdahl, "The disabilities of the arm, shoulder and hand (DASH) outcome questionnaire: longitudinal construct validity and measuring self-rated health change after surgery," *BMC musculoskeletal disorders*, vol. 4, no. 1, p. 11, 16 Jun. 2003. [Online]. Available: http://dx.doi.org/10.1186/1471-2474-4-11

[124] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932. [Online]. Available: http://psycnet.apa.org/psycinfo/1933-01885-001

[125] L. Demers, R. Weiss-Lambrou, and B. Ska, "Development of the quebec user evaluation of satisfaction with assistive technology (quest)," *Assistive Technology*, vol. 8, no. 1, pp. 3–13, Jun 1996. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10400435.1996.10132268

[126] ——, "The quebec user evaluation of satisfaction with assistive technology (QUEST 2.0): an overview and recent progress," *Technology and disability*, vol. 14, no. 3, pp. 101–105, 2002. [Online]. Available: http://content.iospress.com/articles/technology-and-disability/tad00095

[127] G. Salivia and J. P. Hourcade, “Identification of pointing difficulties of two individuals with Parkinson’s disease via a sub-movement analysis,” in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA ’11*. ACM, 2011, p. 137. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1979742.1979636

[128] J. P. Hourcade, K. Perry, and A. Sharma, “PointAssist: helping four year olds point with ease,” in *Proceedings of the 7th international conference on Interaction design and children - IDC ’08*. ACM, 2008, pp. 202–209. [Online]. Available: http://dl.acm.org/citation.cfm?id=1463757

[129] G. Salivia, W. Hall, and J. P. Hourcade, “PointAssist : Assisting Individuals with Motor Impairments,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 1213–1222. [Online]. Available: http://dl.acm.org/citation.cfm?id=2466157

[130] M. R. Williams and R. F. Kirsch, “Evaluation of head orientation and neck muscle EMG signals as command inputs to a human-computer interface for individuals with high tetraplegia,” *IEEE transactions on neural systems and rehabilitation engineering: a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 16, no. 5, pp. 485–496, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1109/TNSRE.2008.2006216

[131] J. P. Hourcade, C. M. Nguyen, K. B. Perry, and N. L. Denburg, “Pointassist for older adults,” in *Proceedings of the 28th international conference on Human factors in computing systems - CHI ’10*, 2010,

p. 1115. [Online]. Available: http://dx.doi.org/10.1145/1753326.1753494http://dl.acm.org/citation.cfm?id=1753326.1753494

[132] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. ieeexplore.ieee.org, Jun. 2010, pp. 94–101. [Online]. Available: http://dx.doi.org/10.1109/CVPRW.2010.5543262

[133] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, Jul. 2005, pp. 5 pp.–. [Online]. Available: http://dx.doi.org/10.1109/ICME.2005.1521424

[134] P. M. Fitts and B. K. Radford, "Information capacity of discrete motor responses under different cognitive sets," *Journal of experimental psychology*, vol. 71, no. 4, pp. 475–482, Apr. 1966. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/5909070

[135] "Logitech Pro 9000." [Online]. Available: http://support.logitech.com/product/quickcam-pro-9000 [Last Accessed: 2015-04-15]

[136] Microsoft Corporation, *Visual Studio 2010*. One Microsoft Way,Redmond: Microsoft Corporation, 2009.

[137] "Emgu CV: OpenCV in .NET (C#, VB, C++ and more)." [Online]. Available: http://www.emgu.com/wiki/index.php/Main_Page [Last Accessed: 2015-04-13]

[138] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[139] G. Bradski, "OpenCV Library," 2014. [Online]. Available: http://opencv.org/

[140] MATLAB, *version 7.12.0.635 (R2011a)*. Natick, Massachusetts: The MathWorks Inc., 2011.

[141] C. E. Shannon, "The mathematical theory of communication. 1963." *MD computing computers in medical practice*, vol. 14, no. 4, pp. 306–17, 1948. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9230594

[142] I. S. MacKenzie and W. Buxton, "A tool for the rapid evaluation of input devices using fitts' law models," *ACM SIGCHI Bulletin*, vol. 25, no. 3, pp. 58–63, Jul 1993. [Online]. Available: http://portal.acm.org/citation.cfm?doid=155786.155801

[143] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun 2006. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S016786550500303X

[144] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press (CUP), 2004. [Online]. Available: http://dx.doi.org/10.1017/cbo9780511811685

[145] E. Malis and M. Vargas, "Deeper Understanding of the MJO," *Science*, vol. 318, no. 5857, pp. 1691b–1691b, dec 2007. [Online]. Avail-

able: http://hal.archives-ouvertes.fr/inria-00174036/,http://www.sciencemag.org/cgi/doi/10.1126/science.318.5857.1691b

[146] J. W. Tukey, “Comparing individual means in the analysis of variance,” *Biometrics*, vol. 5, no. 2, pp. 99–114, 1949. [Online]. Available: http://www.jstor.org/stable/3001913

[147] S. Saini, D. R. A. Rambli, S. Sulaiman, M. N. Zakaria, and S. R. Mohd Shukri, “A low-cost game framework for a home-based stroke rehabilitation system,” in *2012 International Conference on Computer and Information Science, ICCIS 2012 - A Conference of World Engineering, Science and Technology Congress, ESTCON 2012 - Conference Proceedings*, vol. 1, 2012, pp. 55–60. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6297212

[148] R. Lloréns, M. Alcañiz, C. Colomer, and M. D. Navarro, “Balance recovery through virtual stepping exercises using kinect skeleton tracking: a follow-up study with chronic stroke patients,” *Studies in health technology and informatics*, vol. 181, pp. 108–112, 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22954838

[149] A. K. Roy, Y. Soni, and S. Dubey, “Enhancing effectiveness of motor rehabilitation using kinect motion sensing technology,” in *2013 IEEE Global Humanitarian Technology Conference: South Asia Satellite (GHTC-SAS)*, Aug. 2013, pp. 298–304. [Online]. Available: http://dx.doi.org/10.1109/GHTC-SAS.2013.6629934

[150] J. Lozano-Quilis, H. Gil-Gomez, J. Gil-Gomez, S. Albiol-Perez, G. Palacios, H. Fardoum, and A. Mashat, “Virtual reality system for multiple sclerosis

rehabilitation using KINECT," in *International Conference on Pervasive Computing Technologies for Healthcare and Workshops*, 2013, pp. 366–369. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=6563968

[151] Y.-J. Chang, L.-D. Chou, F. T.-Y. Wang, and S.-F. Chen, "A kinect-based vocational task prompting system for individuals with cognitive impairments," *Personal and Ubiquitous Computing*, vol. 17, no. 2, pp. 351–358, 17 Feb. 2013. [Online]. Available: http://link.springer.com/10.1007/s00779-011-0498-6

[152] M. Oszust and M. Wysocki, "Polish sign language words recognition with Kinect," in *2013 6th International Conference on Human System Interactions, HSI 2013*. Ieee, jun 2013, pp. 219–226. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6577826

[153] Z. Halim and G. Abbas, "A Kinect-Based Sign Language Hand Gesture Recognition System for Hearing- and Speech-Impaired: A Pilot Study of Pakistani Sign Language," *Assistive Technology: The Official Journal Of RESNA*, vol. 27, no. 1, pp. 34–43, 2015. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10400435.2014.952845

[154] D. R. Cox and E. J. Snell, *Analysis of binary data*. CRC Press, 1989, vol. 32.

[155] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, p. 583, dec 1952. [Online]. Available: http://dx.doi.org/10.2307/2280779

[156] W. J. Conover, *Practical nonparametric statistics*, 3rd ed., ser. Wiley series in probability and statistics. New York, NY [u.a.]: Wiley, 1999. [Online]. Available: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+24551600X&sourceid=fbw_bibsonomy

[157] World Health Organization, "Towards a common language for functioning , disability and health ICF," *International Classification*, vol. 1149, pp. 1–22, 2002. [Online]. Available: http://www.who.int/classifications/icf/training/icfbeginnersguide.pdf

[158] L. Ng and F. Khan, "Use of the international classification of functioning, disability and health to describe patient-reported disability: A comparison of motor neurone disease, Guillain-Barré syndrome and multiple sclerosis in an Australian cohort," *Disability and Rehabilitation*, vol. 34, no. 4, pp. 295–303, Feb. 2012. [Online]. Available: http://informahealthcare.com/doi/abs/10.3109/09638288.2011.606345

Glossary

interocular	the distance between the two pupils of a person
musculoskeletal	relating to or denoting the musculature and skeleton together
neurodegenerative	progressive loss of structure or function of neurons, including death of neurons. Many neurodegenerative diseases including ALS, Parkinson's, Alzheimer's, and Huntington's occur as a result of neurodegenerative processes.

Appendix A

Data Capture Forms

This section contains:

- the consent form, to declare willingness to participate in the experiment,
- the Participant Information Sheet with all the information about the experiment being conducted, and
- the Device Assessment Questionnaire.

These forms have been approved by the University of Kent Ethics committee.

A.1 Consent Form

CONSENT FORM

Title of project: Can Automatic Facial Gesture Recognition Techniques be developed as an input device to Electronic Assistive Technologies for Severely Disabled Patients?

Name of investigator: Shivanand P. Guness

Participant Identification Number for this project:

Please initial box

1	**I confirm I have read and understood the information sheet dated 31st January 2013 (version 1.0) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.**	☐
2	**I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason. *(Email:spg23@kent.ac.uk).***	☐
3	**I understand that my responses will be anonymised before analysis. I give permission for members of the research team to have access to my anonymised responses.**	☐
4	**I agree to take part in the above research project.**	☐
5	**I agree to video and sensor data to be captured during the data capture session to be used in publications, presentations, posters, publicity and web presentations.(Note that the University cannot accept any liability for the acquisition or use of this data from the Web by third parties)**	☐
6	**I would like to receive information on the publication of this data.**	☐

________________ ________________ ________________

Name of participant   Date   Signature

________________ ________________ ________________

Name of person taking consent   Date   Signature
(if different from lead researcher)
To be signed and dated in presence of the participant

________________ ________________ ________________

Lead researcher   Date   Signature

Copies:

When completed: 1 for participant; 1 for researcher site file; 1 (original) to be kept in main file

A.2 Participant Information Sheet

Aim

The aim of the experiment is to capture facial gestures. The facial gestures to be captured can be divided into 2 categories upper facial gestures such as eye blink, eye winks and eye brows movement and lower facial gestures such as the opening-closing of the mouth, tongue protrusion and tongue movement.

Objectives

The objective of this experiment is to collect videos of healthy individuals performing a number of gestures. The gestures to be collected are:

1. Blink.
2. Left wink.
3. Right wink.
4. Left eyebrow movement.
5. Right eyebrow movement.
6. Both eyebrows movement.
7. Mouth open/mouth closed.
8. Tongue protrusion.

For participant wearing glasses the gestures will be captured both with and without glasses.

Set-up

As it can be seen in **Fig. 1**, the Kinect sensor will be placed on top of the monitor in a central location. The participant will be asked to sit in front of the monitor before starting the capture procedure. The experiment will be carried out in an unrestrained but fixed posture and pose in the lab with normal/natural lighting. The participants will be asked not to move during the gesture sessions. The participants will also be placed at a distance of 30-60cm from the sensor. The participant is expected to follow an on-screen stimulus during the data capture session. The session is likely to last no more than 1 hour.

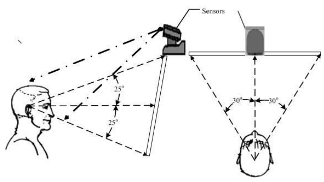


Fig. 1. : Top and side views of the experimental setup

Target population

The targeted population is adults from 18 to 60 years old. The experiment is going to be carried out on a population size of approximately 10 participants. Participation to this experiment is voluntary

Data

For each participant, each gestures are going to be collected five times at 3 different time intervals (500ms, 750ms, 1000ms) with ambient lighting or variable lighting. The gestures are to be collected for the frontal face where the stimulus of the gestures will appear in the centre of the screen. The stimulus will be moved in different orientation around the central location and the gesture will be captured at each location. For facial gestures such as eye blink and eye winks an audio cue in the form of a computer beep will be issued to the user to inform them to open their eyes. The audio cue will be used also in cases where a participant is performing the gesture with glasses on.

For each gesture captured session for a participant an RGB video, a depth map video, and a point cloud video will be captured. A data file in the format of a csv file will also be stored for each gesture captured. The file will contain the id of the user, the gesture being captured, the sequence of the gesture, the time interval, the coordinates of the mid-point area being sampled, the maximum depth, the minimum depth, the average depth of the sampling line, the number of pixels being sampled, the depth of each pixel being sampled both from the depth map and the point cloud. The reaction time of the user i.e. the time taken from the time the stimuli are presented to the user and the user started to perform the gesture will be recorded.

A.3 Questionnaire

Trial Number:

Device:

Time and date:

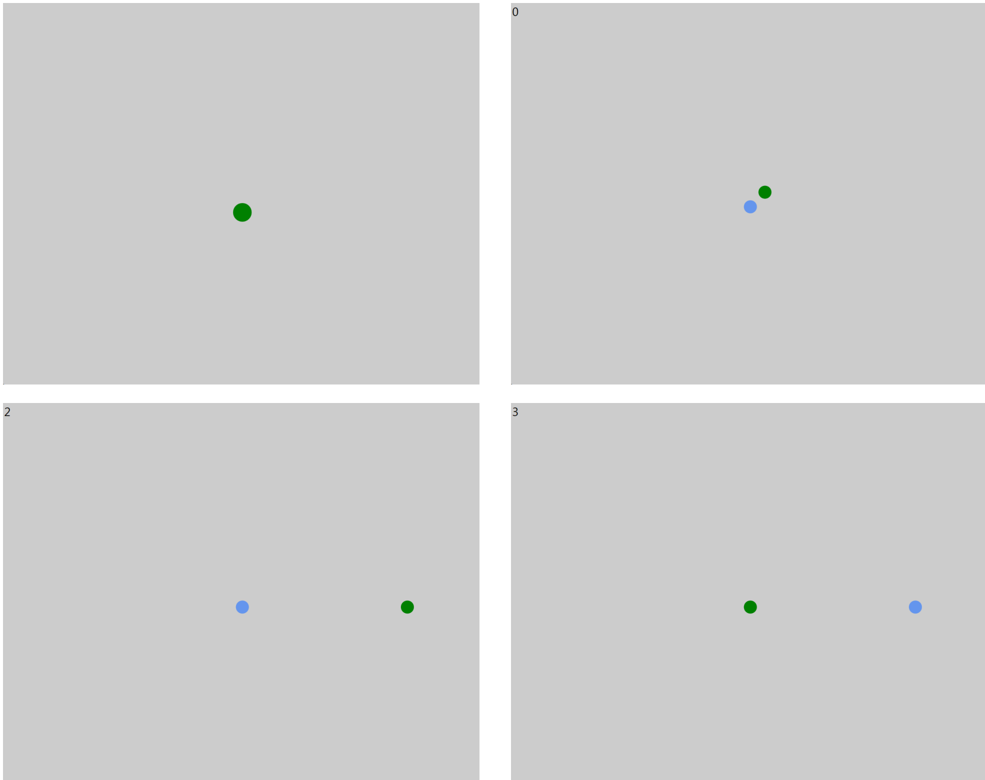
Q1. Overall (compare to the mouse) *(1 = worst and 7= better)*						
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Q2. Shoulder Fatigue *(1 = low and 7= high)*						
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Q3. Neck Fatigue *(1 = low and 7= high)*						
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Q4. Eye Fatigue *(1 = low and 7= high)*						
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Q5. Operation Speed *(1 = slow and 7= fast)*						
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Q6. Target Selection *(1 = easy and 7= difficult)*						
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7
Q7. Accurate Pointing *(1 = easy and 7= difficult)*						
○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7

Q8. Physical Effort *(1 = low and 7= high)* 1 2 3 4 5 6 7
Q9. Mental Effort *(1 = low and 7= high)* 1 2 3 4 5 6 7
Q10. Smoothness *(1 = rough and 7= very smooth)* 1 2 3 4 5 6 7
Q11. Was the target size (1= difficult to track and 7=very easy to track) 1 2 3 4 5 6 7
Q12. Overall the input device was *(1 = very difficult to use and 7= very easy to use)* 1 2 3 4 5 6 7

Appendix B

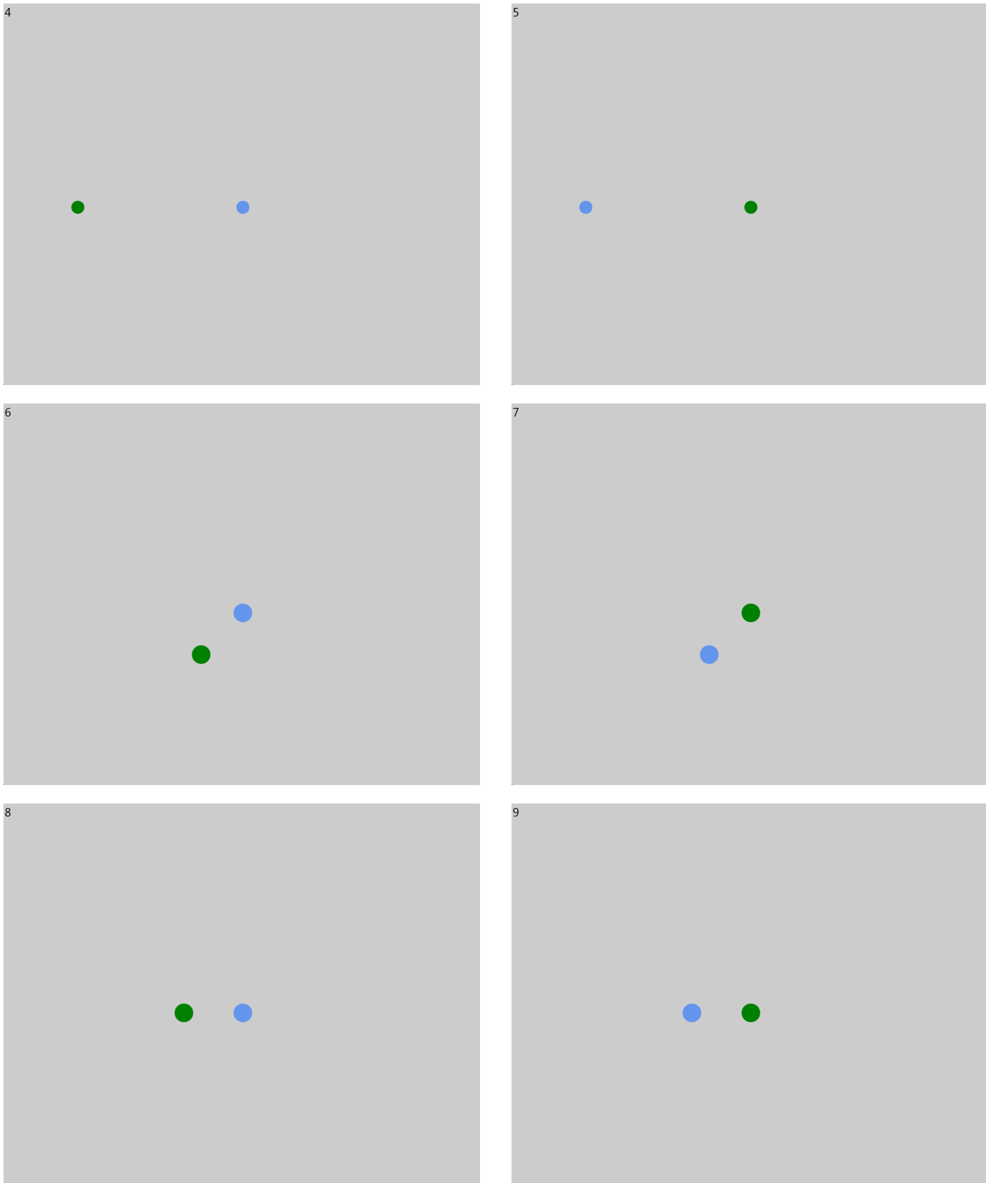
Stimuli used in Fitts' Test

Table B.1 Example of the location of the Stimuli



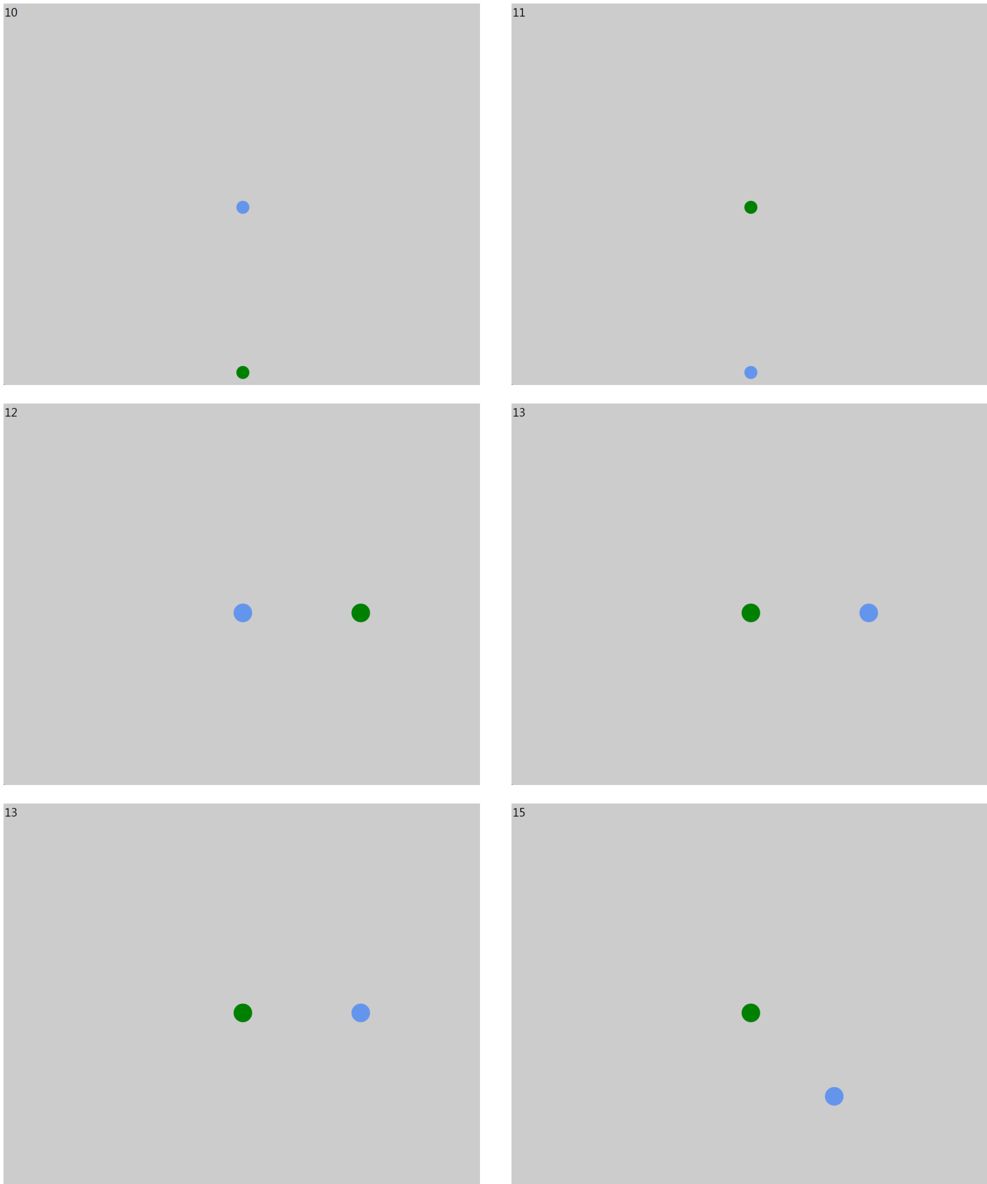
Continued on next page

Example of the location of the Stimuli



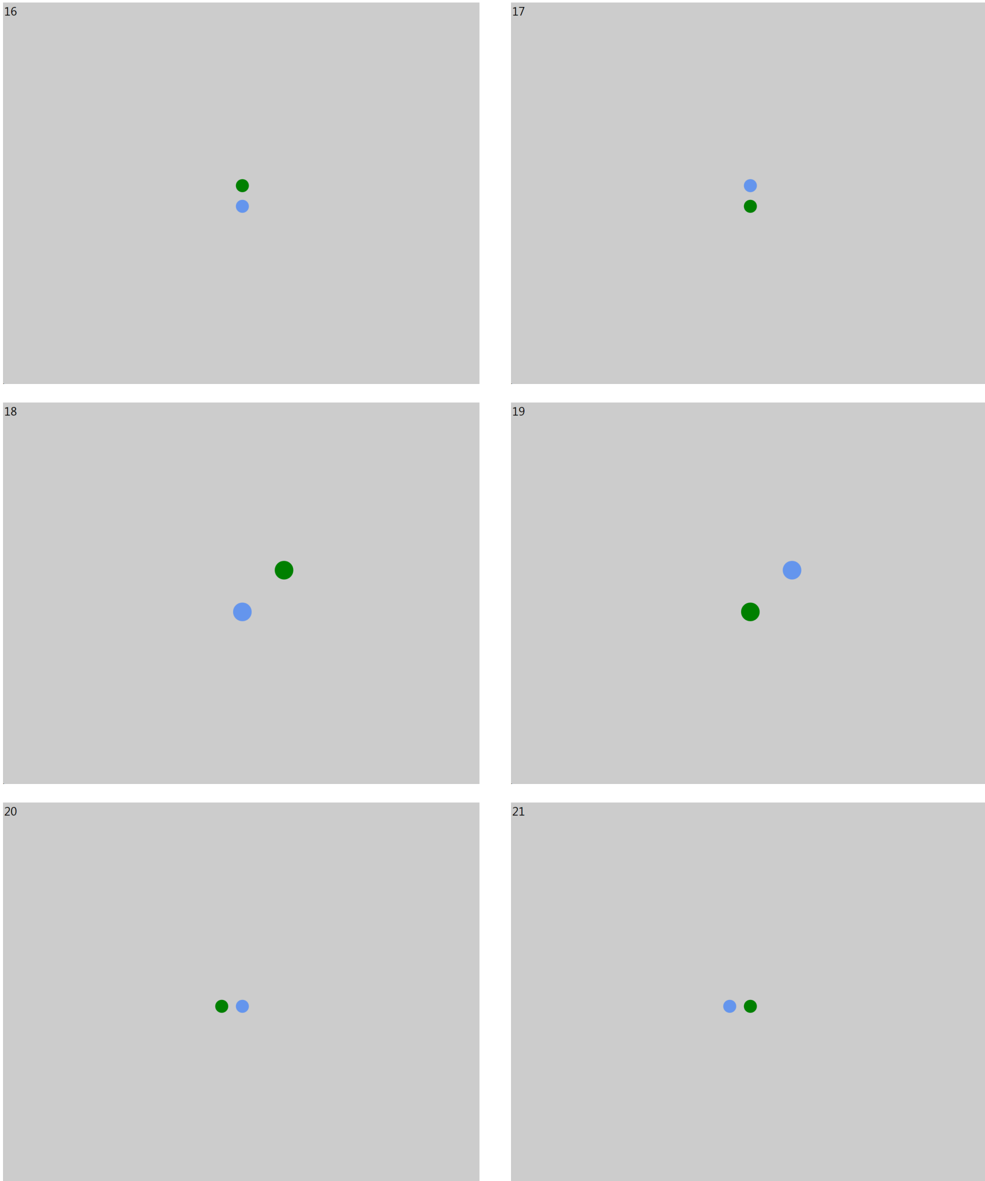
Continued on next page

Example of the location of the Stimuli



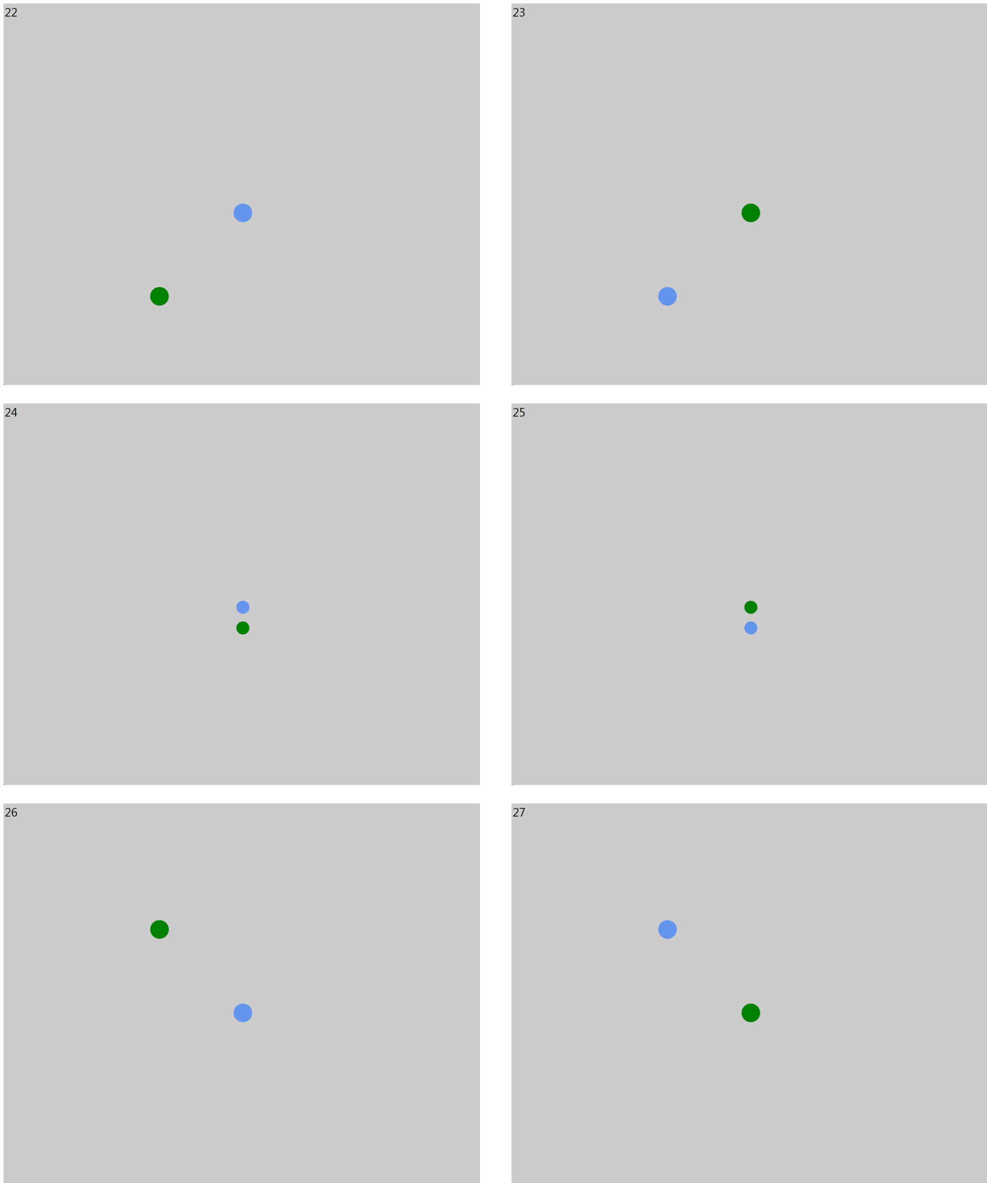
Continued on next page

Example of the location of the Stimuli



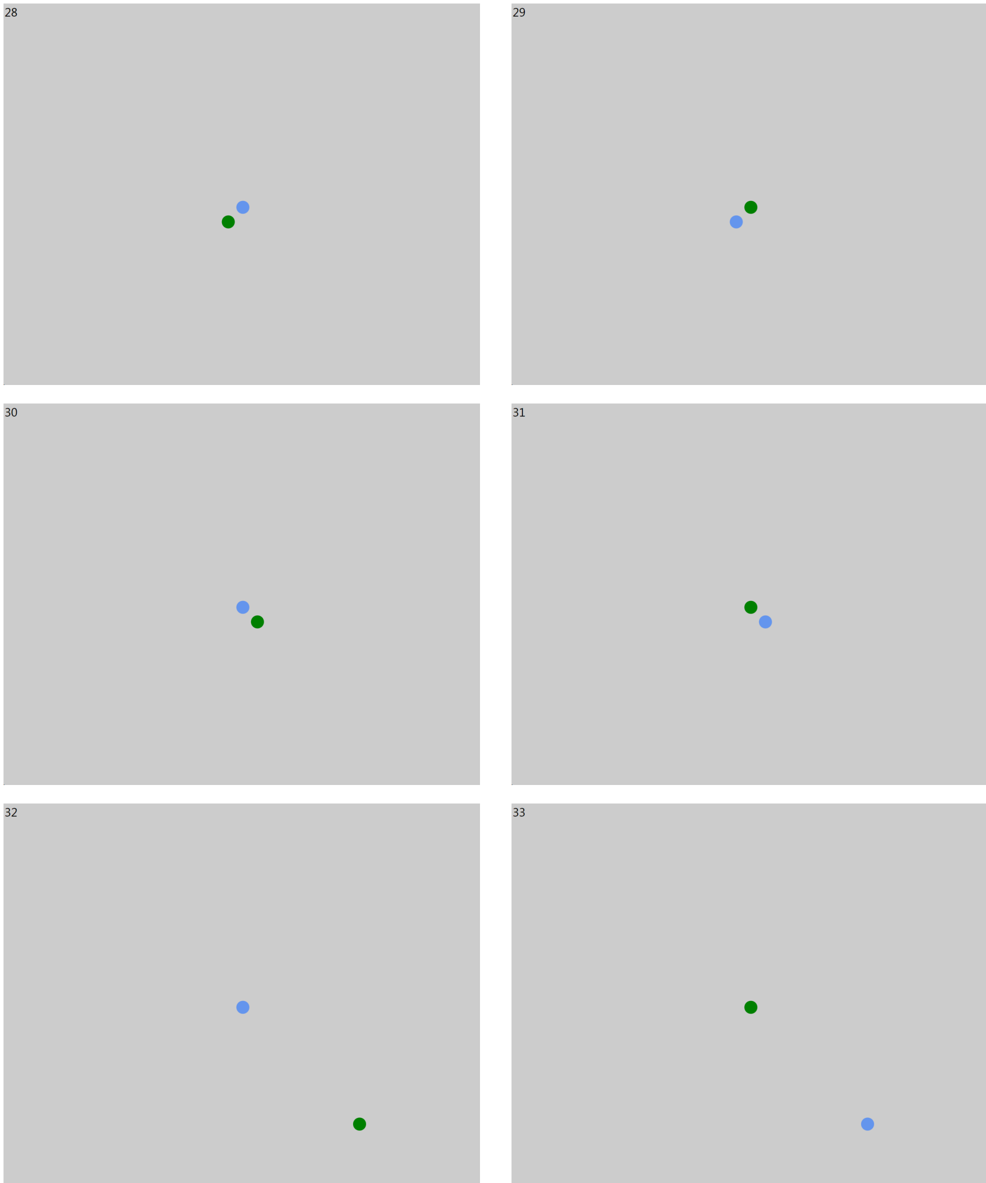
Continued on next page

Example of the location of the Stimuli



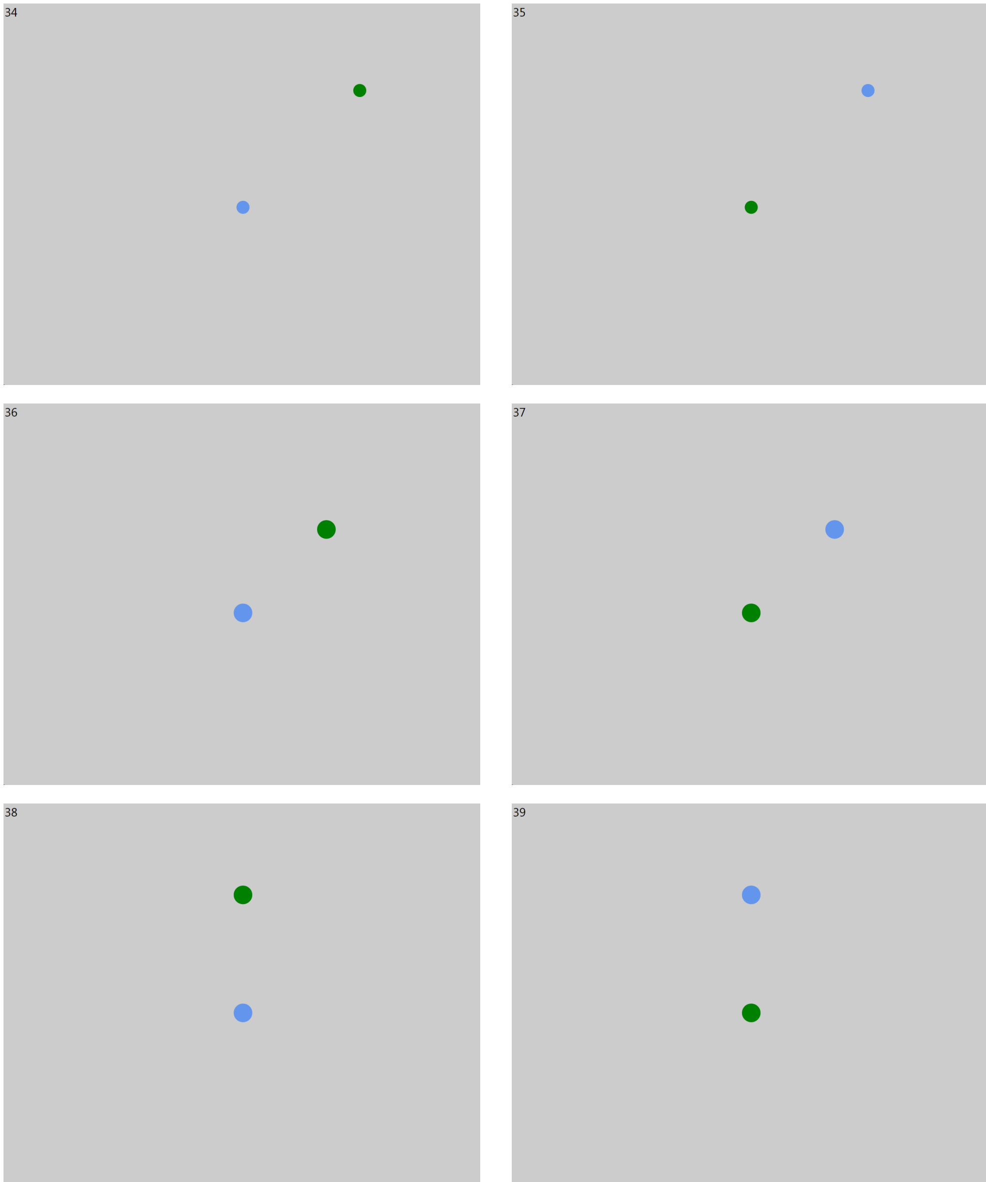
Continued on next page

Example of the location of the Stimuli



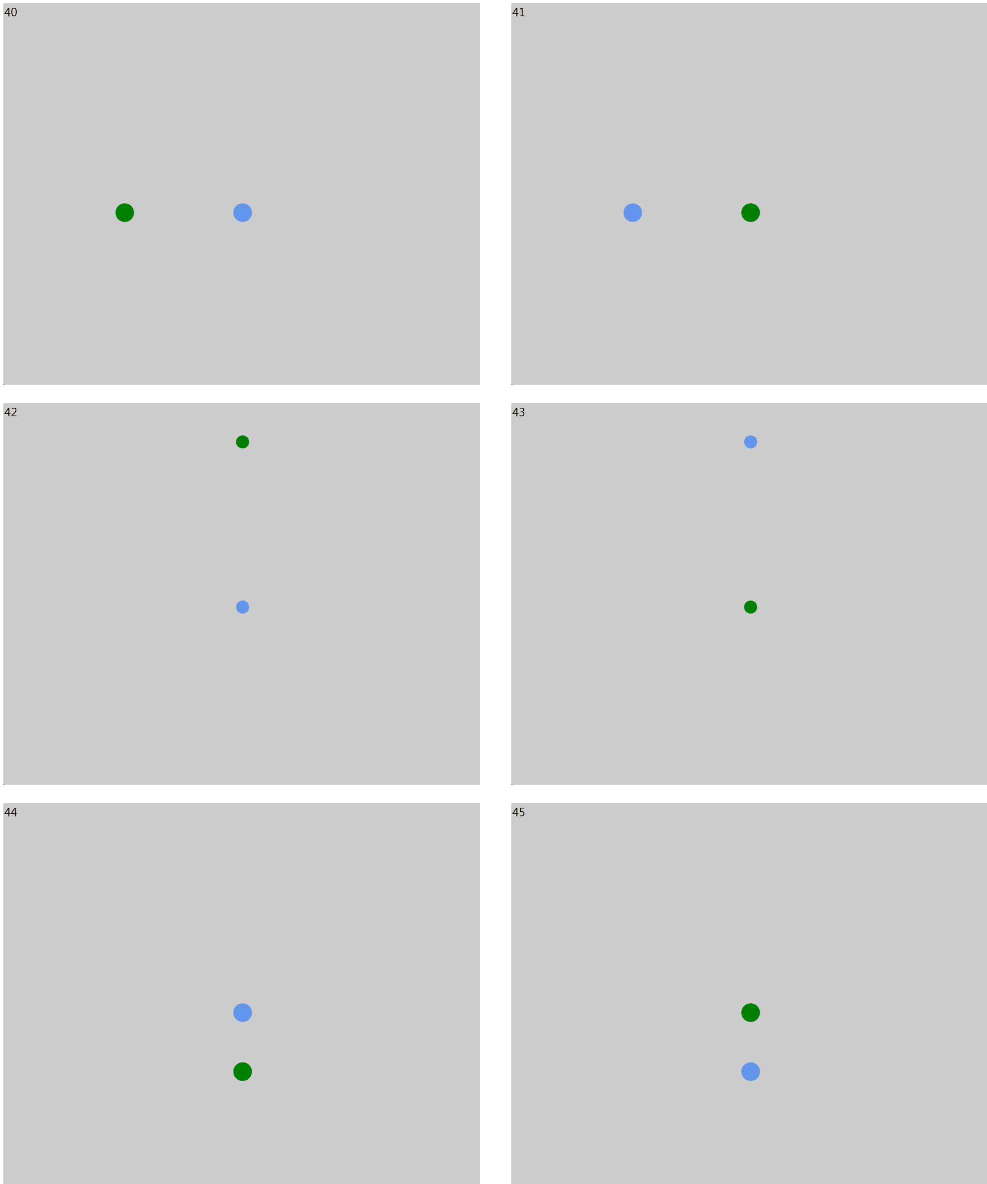
Continued on next page

Example of the location of the Stimuli



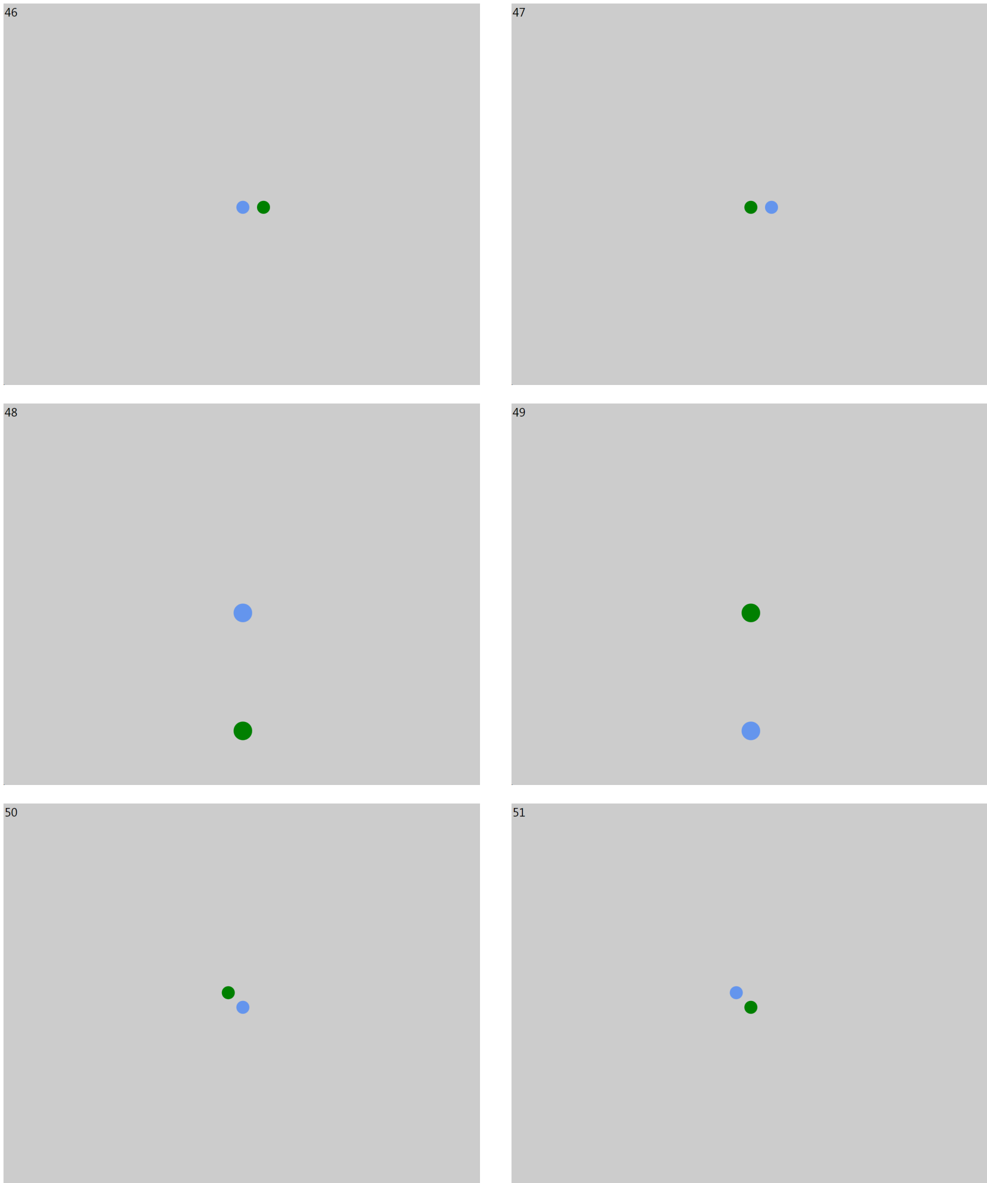
Continued on next page

Example of the location of the Stimuli



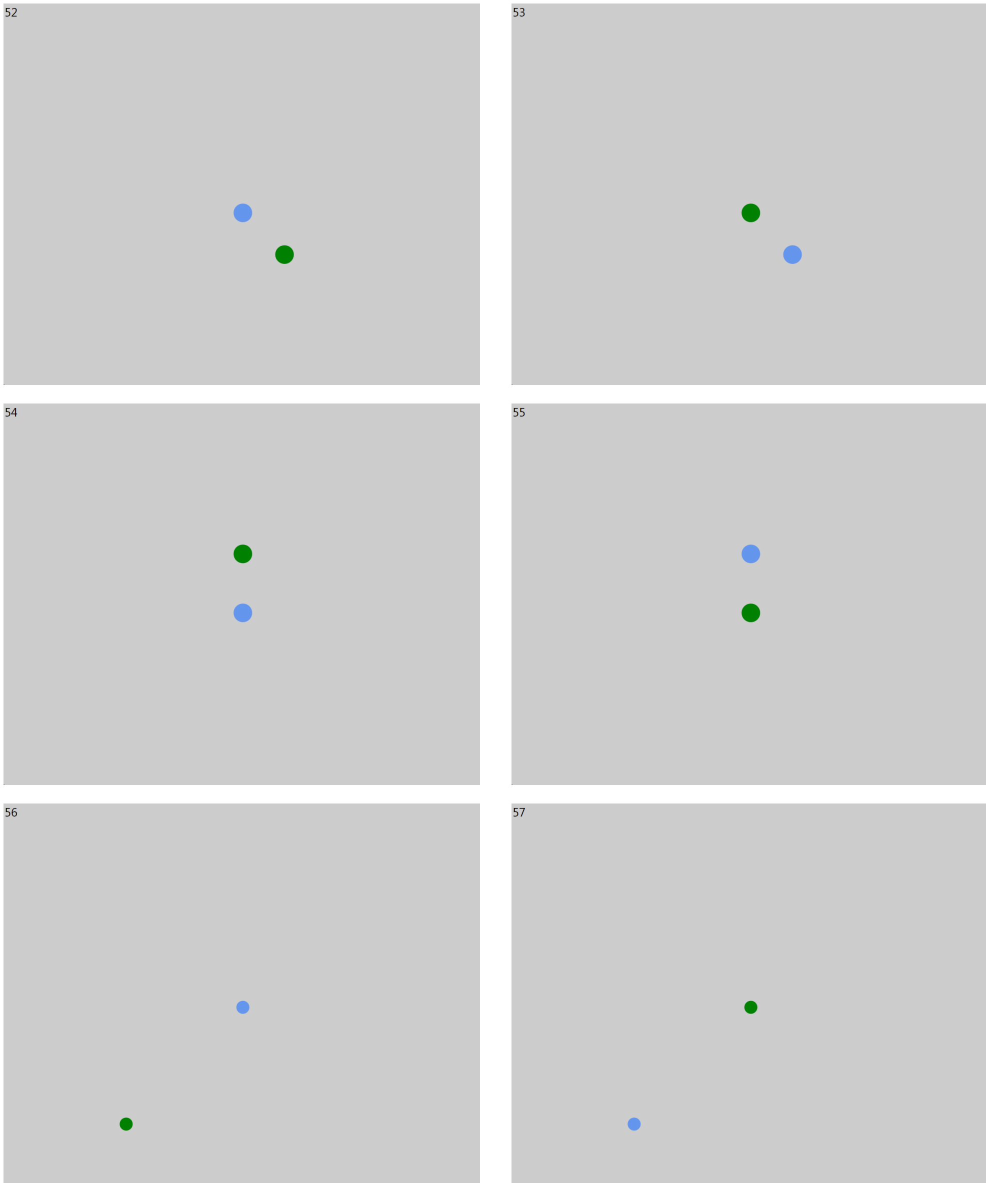
Continued on next page

Example of the location of the Stimuli



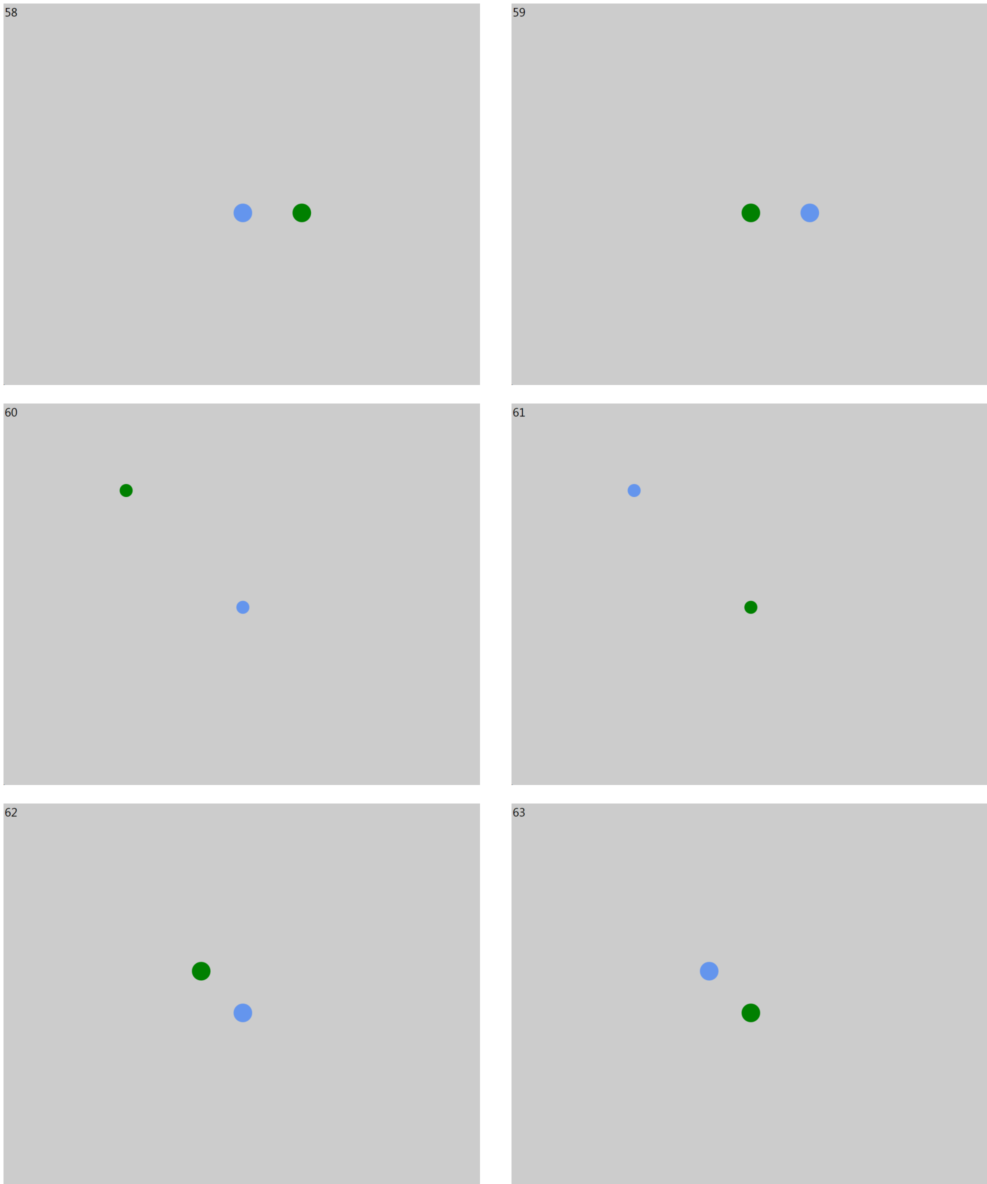
Continued on next page

Example of the location of the Stimuli



Continued on next page

Example of the location of the Stimuli



Appendix C

Disabilities and Impairments

The table below provides us with an overview of the impairments associated with MS and MND based on the ICF. The ICF published by the WHO provides a standard language and framework for the description of health and health-related states [157]. It provides a classification of health and health-related domains. The domains help to describe changes in body function and structure, the level of capacity of a person with a health condition in a standard environment and the level of performance of the individual in their environment.

The different domains of the ICF are the body function, the body structure, activity, participation and environmental factors. The body function domain relates to the physiological and psychological function of the body. The body structure domain refers to the different parts of the body such as the limbs and organs i.e.the anatomical parts of the body. The WHO has defined activity as "The execution of a task or action by an individual", thus the activity domain refers to the ability of the individual to execute certain essential tasks such as reading, writing etc. Participation is defined by the WHO as "Involvement in a life situation". The participation domain refers to the ability of the individual to handle life situation such as to have interpersonal interactions and relationships [157].

In [18, 158], surveys were conducted with questionnaires to find the difficulties or limitations encountered by persons with MS and MND based the ICF framework. Using the ICF framework enables the comparison of the issues encountered by people with the same disabilities or different disabilities. In Table C.1, a subset of the body functions from the ICF framework which was found to affect the people in the surveys [18, 158]. The tables denote both the number of person who experienced difficulties of the function but also the percentage of the population in the survey is denoted in brackets.

Table C.1 ICF: Body Function

Body Function		MS[18]	MS[158]	MND[158]
b126	Temperament and personality functions	98 (47%)		
b130	Energy and drive functions	146 (71%)	98 (97.0%)	38 (86.4%)
b134	Sleep functions	140 (68%)	84(83.1%)	22(50.0%)
b140	Attention functions	131 (64%)	66 (65.5%)	
b144	Memory functions	51 (24%)	62 (61.4%)	
b152	Emotional functions	68 (33%)	97(96%)	25(56.8%)
b210	Seeing functions	154 (72%)	47 (47.5%)	
b235	Vestibular functions	78 (38%)	71(70.3%)	
b280	Sensation of pain	113(49%)	76 (75.2%)	24(54.5%)
b310	Voice			19(43.1%)
b440	Respiration			20 (45.5%)
b455	Exercise and tolerance function		97 (96.0%)	34 (77.3%)
b510	Ingestion			32 (72.7%)
b515	Digestive functions	56 (24%)		
b525	Defecation			15(34.1%)
b620	Urination functions	167 (81%)	94(93.1%)	15(34.1%)
b640	Sexual functions	83 (40%)	57 (56.4%)	15(34.1%)

b710	Mobility of joint functions	170 (82%)		
b730	Muscle power functions	194 (95%)	96 (95.1%)	42 (95.5%)
b735	Muscle tone functions	130 (63%)	94 (93.1%)	40 (90.1%)
b740	Muscle endurance function		93 (92.1%)	
b760	Control of voluntary movement function		66 (65.3%)	
b765	Involuntary movement functions	88 (43%)		36(81.2%)
b770	Gait pattern functions	184 (90%)	99 (98.0%)	18 (40.9%)
b780	Sensations related to muscles and movement functions	167 (80%)		

Table C.2 ICF: Body Structure

Body Structure		MS[18]	MS[158]	MND[158]
s110	Structure of brain	204 (100%)	100 (99.0%)	43 (97.7%)
s120	Spinal cord and related system	191 (93%)		43 (97.7%)
s2	Structures of eye, ear and related structure	150 (69%)		
s610	Structure of urinary system	164 (80%)	93 (92.1%)	
s630	Structure of reproductive system	54 (26%)		
s720	Structure of shoulder region	55 (25%)		
s730	Structure of upper extremity	121 (56%)	44 (43.6%)	41 (93.1%)
s740	Structure of pelvic region	69 (33%)		
s750	Structure of lower extremity	154 (73%)	97 (96.0%)	37 (84.1%)
s760	Structure of trunk	97 (42%)	85 (84.2%)	18 (40.9%)

Table C.3 ICF: Activity and participation

Activity and participation		MS[18]	MS[158]	MND[158]
d155	Acquiring skills	69 (34%)		
d160	Focusing attention		70 (69.3%)	
d166	Reading	143 (69%)		
d170	Writing	115 (56%)		34 (77.3%)
d175	Solving problems	82 (40%)	34 (33.6%)	
d177	Making decisions	104 (51%)	59 (58.4%)	
d210	Undertaking a single task			25 (56.8%)
d220	Undertaking multiple tasks	84 (41%)	88 (87.1%)	40 (90.1%)
d230	Carrying out daily routine	170 (83%)	80 (79.2%)	35 (79.5%)
d240	Handling stress and other psychological demands		101 (100.0%)	31 (70.5%)
d330	Speaking			36 (59.1%)
d350	Conversation			18 (40.1%)
d410	Changing basic body position			30 (68.2%)
d430	Lifting and carrying objects	121 (59%)	53 (52.5%)	37 (84.1%)
d440	Fine hand use	138 (67%)	51(50.5%)	39 (88.6%)
d445	Hand and arm use	145 (71%)	37(36.6%)	28 (63.6%)
d450	Walking	186 (91%)	101(100.0%)	40 (90.1%)
d455	Moving around	181 (88%)	99 (98.0%)	41 (93.2%)
d460	Moving around in different locations	164 (80%)		
d465	Moving around using equipment	143 (70%)	98 (97.0%)	38 (86.4%)

d470	Using transportation	142 (69%)	100(99.0%)	40 (90.1%)
d475	Driving	138 (65%)	99(98.0%)	39 (88.6%)
d510	Washing oneself	74 (36%)	41 (40.6%)	39 (88.6%)
d520	Caring for body parts	70 (34%)	40 (39.6%)	37 (84.1%)
d530	Toileting	69 (34%)		39 (88.6%)
d540	Dressing	57 (28%)		40 (90.1%)
d570	Looking after ones health	42 (20%)	88 (87.1%)	34 (77.3%)
d620	Acquisition of goods and services	181 (88%)	92 (91.1%)	31 (40.3%)
d630	Preparing meals	113 (55%)	89 (88.1%)	39 (88.6%)
d640	Doing housework	186 (91%)	94 (93.1%)	39 (88.6%)
d650	Caring for household objects	143 (70%)	84 (83.2%)	38 (86.4%)
d660	Assisting others	130 (63%)	87 (86.1%)	34 (77.3%)
d720	Complex interpersonal interactions			22 (50.0%)
d730	Relating with strangers			29 (65.9%)
d740	Formal Relationships			20 (45.5%)
d750	Informal social relationships			27 (61.4%)
d760	Family relationships		73 (72.2%)	37 (84.1%)
d770	Intimate relationships	112 (54%)	61(60.4%)	33 (75.0%)
d825	Vocational training	57 (28%)		
d830	Higher education	86 (42%)		
d845	Acquiring, keeping and retaining a job		73 (72.2%)	15 (34.1%)
d850	Remunerative employment	153 (75%)	90 (89.1%)	25 (56.8%)

d865	Complex economic transactions	69 (34%)		
d870	Economic self-sufficiency	118 (58%)	84 (83.2%)	22 (50.0%)
d910	Community life	187 (87%)	79 (78.2%)	33 (75.0%)
d920	Recreation and leisure	185 (90%)	97 (96.0%)	37 (84.1%)
d930	Religion and spirituality	40 (20%)		21 (47.7%)
d940	Human rights	62 (30%)		

Table C.4 ICF: Environmental

Environmental		MS[18]	MS[158]	MND[158]
e110	Products or substances for personal consumption	126 (61%)	101 (100%)	43 (97.7%)
e115	For personal use in daily living			31 (70.5%)
e120	Products and technology for personal indoor and outdoor mobility and transportation	125 (61%)	91 (90.1%)	39 (88.6%)
e150	Design, construction and building products and technology of buildings for public use	78 (38%)		
e155	Design, construction and building products and technology of buildings for private use	86 (42%)		
e210	Physical geography		39 (38.6%)	14 (36.4%)
e225	Climate	142 (69%)	99 (98.0%)	
e240	Light	122 (60%)		
e250	Sound	72 (35%)		

e310	Immediate family	205 (100%)	45 (44.5%)	43 (97.7%)
e315	Extended family	199 (97%)	42 (41.6%)	
e320	Friends	198 (97%)		
e325	Acquaintances, peers, colleagues, neighbours and community members	168 (82%)		
e330	People in positions of authority	77 (38%)		
e340	Personal care providers and personal assistants	48 (23%)		16 (36.4%)
e355	Health professionals	143 (70%)		
e360	Other professionals	41 (20%)		
e410	Individual attitudes of immediate family members of extended family members	205 (100%)		
e420	Individual attitudes of friends	199 (97%)		
e440	Individual attitudes of personal care providers and personal assistants	53 (26%)		
e450	Individual attitudes of health professionals	151 (74%)		
e455	Individual attitudes of health-related professional	40 (20%)		
e460	Societal attitudes		31 (30.7%)	
e465	Social norms, practices and policies	95 (46%)		

e525	Housing services, systems and ideologies	76 (37%)		
e540	Transportation services, systems and policies	62 (30%)	68 (67.3%)	34 (77.3%)
e570	Social security services, systems and policies	70 (34%)		
e575	General social support services, systems and policies	76 (37%)		
e580	Health services, systems and policies	133 (65%)	79 (78.2%)	40 (90.9%)
e585	Education and training services, systems and policies	51 (25%)		
e590	Labour and employment services, systems and policies	107 (52%)		