

Inferring intentions from natural eye-movements for biomimetic control of prosthetics and wheelchairs

William Welby Abbott

**Brain and Behaviour Lab
Department of Bioengineering
Imperial College London**

This dissertation is submitted for the degree of Doctor of Philosophy

May 2016

In memory of Jean

Declaration of Originality

I hereby declare that unless otherwise acknowledged this thesis and all the work therein is my own original research. Any concepts, ideas, methods or data which I did not develop myself are duly cited and a list references given in the bibliography

Signed: William Welby Abbott

Acknowledgements

Completing this PhD, in combination with the trials and tribulations of life in the last couple of years, has been the most testing thing I have ever done and it is only with the huge support of my family and friends that I made it through. Firstly, my colleagues and now close friends in the lab - Scott, Katya, Ali, Anastasia, Chin-Hsuan, Feryal, Alessandro, Alan, Andreas and Costa. It has been a real pleasure to work with such a brilliant, diverse and supportive group of people. We shared the highs and the lows together during a time I will never forget. I worked particularly closely with Andreas and Costa in our "motion capture huddles". Thanks for all the discussions and brainstorming that always lead to a solution, in the often seemingly impossible challenges of such high-dimensional problematic datasets. Your support has been invaluable, having such close friends to share common frustrations (of which there were many) but also all the laughs and good times that came with it. Thanks of course go to my supervisor Aldo for his energy, enthusiasm and brilliance. It never ceases to amaze me how many plates he keeps spinning at one time. I thank him for the opportunity to work in his lab with the great team he built and all the opportunities to travel and present my work all over the world. My thanks also go to Anil who has been incredibly supportive over the years, with astute advice, many opportunities and the inspiration to teach when assisting on his well honed courses. Also, to Danny and Martin for all the hilarious chats, support and trips together to watch live music and enjoy a beer or two.

I'm also really grateful to my friends outside of the South Kensington bubble who gave me perspective and reminded me about life outside of academia. Special thanks to Aedan for putting up with living with me for most of my PhD, all the eccentricities of our flat and escapades that never ceased to put a smile on my face and take my mind off work. Thanks to Ed and Ani for all the support, amazing summers on the farm and wild weekends when you visited London over the last few years. Thanks to Joe for being a really good friend, particularly over the last year when your support, understanding and insights have been so valued. Finally I have to make special mention of Chezz, Will and Sofia. You have been such good friends to me over the years and I cannot begin to thank you enough for that alone. But in the darkest times over the last year it is you who I knew I could call or go and see no matter what, knowing that you would help me to see things clearer again and put a smile back on my face, I love you guys so much. To Paul for all the escapes to Hulver farm, a sort of "never never land" of fun laughter and happiness. You taught me so much with your ability to fix anything using whatever you had handy - inspiring my frugal approach to engineering. To Pat for all your wisdom, kindness and belief in me.

Lastly and by no means least, my family. I am so lucky to have such a supportive and loving family who I know would do anything for me and are always there to listen. Thanks mum for inspiring me so much with your immense thirst for knowledge, fierce curiosity, immense work ethic, wisdom and deep compassion for those in need. To dad for your calming support, ability to listen and for the balance and perspective on the important things in life. To Tom for your love, support and ability to always keep in touch and think of me. And Rose, you have been my rock. With unending empathy and wise words, always there when I needed you most - I truly would not be where I am today without your love, support and belief in me. I love you all very much.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work

Abstract

The link between perception and action is a fundamental question in neuroscience. To see the world we must look, thus eye movements provide a window into the brain. They reflect our intentions, as the only directly observable behavioural signal that is highly correlated with actions at the task level, and proactive of body movements. Despite this, eye tracking is not widely used as control interface for movement impaired patients due to poor signal interpretation and lack of control flexibility. It is therefore proposed, that tracking the gaze position in 3D rather than conventional 2D provides a considerably richer signal directly relevant for prosthetic control. An ultra-low-cost 3D gaze tracker is developed capable of information transfer rates up to 43 bits/sec, well beyond current invasive and non invasive brain machine interfaces. Further, considering that the eyes make the fastest movements in the body (faster than a heartbeat), reflects the pace of information retrieval required for the brain to coordinate and orchestrate the complex actions of everyday life. Thus eye-movements are naturally disposed to the real-time closed loop control of actuators. This is demonstrated in a large field study involving over 2000 participants playing the arcade game "Pong". While this is a trivial task in comparison to prosthetic control, unlike many brain machine interface applications it requires continuous closed loop control. Pong is proposed as a universally accessible benchmark to compare the performance of interfaces. To achieve this data-logging pong game is developed with a suite of algorithms to analyse the performance and behavioural strategies. This is then used to benchmark the developed gaze input method. In just 30 seconds set up and 5 seconds training time, the majority of subjects were able to successfully play and in some cases win, despite having never using their eyes as a control input. The developed 3D gaze interfaces also allowed this potential to be taken beyond the computer screen to control robotic arms and wheelchairs. Interfaces were designed to minimise the impact on the natural task of the eyes - to sample information from the world. In fact the aim is to base control strategies on these natural eye-movements because they implicitly contain intention information. This has been shown in a large body of literature, however methods are either reductionist lab based studies or require painstaking manual annotation of eye and body movements "in the wild". Thus a new methodology is developed, to record a comprehensive database of human visuomotor behaviour, to inform the development of biomimetic prosthetics. The variance of behaviour is captured, rather than constrained, in natural daily tasks lasting hours rather than minutes. The dataset presented moves towards an embodied approach, building a database of human behaviour more complete, extensive and unconstrained than achieved previously. This leads to new concepts of Embodied Saliency; predicting eye-movements directly from body movements and Embodied Gaze Descriptors; data driven methods for taxonomising the complex interaction between eye and body movement in the wild. Thus by capturing rather than constraining natural behaviour, it is hoped that future biomimetic prosthetics, principled on the full perception action loop, can liberate patients from the constraints of disability.

Contents

1 INTRODUCTION	11
1.1 SUMMARY OF THESIS.....	13
2 EYE-MOVEMENT BEHAVIOUR, TRACKING METHODS AND GAZE BASED INTENTION DECODING.....	16
2.1 THE EYE	16
2.1.1 <i>Anatomy</i>	16
2.1.2 <i>Oculomotor System</i>	17
2.2 EYE MOVEMENT CHARACTERISTICS AND COGNITIVE PROCESSES	19
2.3 SAMPLING THE SCENE: WHAT DRIVES EYE-MOVEMENTS?	21
2.4 KEEP AN EYE ON THE GOAL: OBJECT INTERACTION.....	25
2.5 LOOK WHERE YOU'RE GOING: EYE-MOVEMENTS DURING NAVIGATION	26
2.6 BEYOND TOP-DOWN AND BOTTOM-UP EFFECTS	26
2.7 GAZE CUES FOR INTENTION RECOGNITION AND THE MIDAS TOUCH PROBLEM.....	27
2.8 EYE-TRACKING HISTORY.....	29
2.9 EYE-TRACKING STATE OF THE ART: VIDEO OCULOGRAPHY	30
2.10 GAZE TRACKING: FROM EYE TO GAZE	33
3 GT3D ULTRA LOW-COST 3D GAZE ESTIMATION.....	36
3.1 INTRODUCTION	37
3.2 METHODS.....	38
3.2.1 <i>Ultra-Low Cost Binocular Eye-tracking Hardware</i>	39
3.2.2 <i>Eye-Tracking</i>	40
3.2.3 <i>Calibration for 3D gaze estimation</i>	41
3.2.4 <i>Solution to the Midas touch problem</i>	43
3.3 RESULTS.....	44
3.3.1 <i>Eye-tracking accuracy and precision</i>	44
3.3.2 <i>Human computer interaction</i>	45
3.3.3 <i>Accuracy and precision in 3D tasks</i>	45
3.4 DISCUSSION.....	46
4 LARGE FIELD STUDY OF GAZE BASED CLOSED LOOP CONTROL.....	52
4.1 INTRODUCTION	53
4.2 METHODS.....	55
4.2.1 <i>Public facing BMI arcade machine</i>	55
4.2.2 <i>Subjects</i>	56
4.2.3 <i>"Pong" BMI benchmark</i>	57
4.2.4 <i>Sit, align, scan and play: Set-up, calibrate and control</i>	58
4.2.5 <i>Optimal binocular gaze mapping</i>	58
4.2.6 <i>Data Collection</i>	59
4.2.7 <i>Game Analysis</i>	59
4.2.8 <i>Information Transfer Evaluation</i>	59
4.2.9 <i>Player Strategies feature vector</i>	62
4.2.10 <i>Strategy Clustering</i>	63
4.3 RESULTS.....	65
4.3.1 <i>Public facing auto-experiment booth</i>	65
4.3.2 <i>Calibration Performance</i>	66

4.3.3 Pong game Performance:	67
4.3.4 BMI performance:	68
4.3.5 Gaze control strategies:	71
4.3.6 Modelling Player Strategies	77
4.3.7 Pupil behaviour	81
4.4 DISCUSSION	83
4.4.1 Pong as a BMI benchmark	83
4.4.2 Game Performance	84
4.4.3 Behavioural Analysis	86
4.4.4 Game strategies relation to natural ball catching behaviour	87
5 INTUITIVE GAZE CONTROL: WHEELCHAIRS AND ROBOTIC ARMS	88
5.1 BACKGROUND	89
5.1.1 Eye controlled wheelchairs	90
5.2 METHODS	91
5.2.1 Eye-controlled Arm 1: Novel 3D gaze calibration and control with a robotic arm	91
5.2.2 Eye-controlled Arm 2: Draw by gaze	96
5.2.3 Virtual Wheelchair Platform	99
5.2.4 Real-world gaze controlled wheelchair	103
5.3 OVERVIEW DISCUSSION	105
6 EMBODIED EYE-MOVEMENT BEHAVIOUR IN THE WILD	109
6.1 INTRODUCTION	110
6.1.1 Static fixations and saccades	115
6.1.2 Non-static fixations and saccades	117
6.1.3 Head and trunk tracking "in the wild"	118
6.1.4 Limb Tracking "in the wild"	119
6.1.5 3D gaze behaviour: previous work	120
6.1.6 Specification	121
6.2 METHODS	122
6.2.1 Experimental Protocol	122
6.2.2 Data collection hardware and collection framework	122
6.2.3 Data integration and curation	124
6.2.4 Data Analytics:	127
6.3 RESULTS	128
6.3.1 Data set overview	128
6.3.2 Eye Movement direction and depth Statistics	131
6.3.3 Gaze Event Segmentation	131
6.3.4 Gaze event characteristics: method comparison	134
6.3.5 Gaze event characteristics: setting comparison	135
6.3.6 Blink Characteristics:	136
6.3.7 Gaze event characteristics: across subjects and previous studies	139
6.4 DISCUSSION	141
6.4.1 Cross setting event detection differences	143
6.4.2 Comparing automated coding to hand coding in previous studies	144
6.4.3 Consistency of behaviour across task and setting	145
6.4.4 What we miss in the blink of an eye	146
7 EMBODIED SALIENCY AND GAZE DESCRIPTORS	149
7.1 INTRODUCTION	150

7.2 METHODS.....	150
7.2.1 Linear Regressive Models.....	150
7.2.2 Performance metrics and chance level.....	153
7.2.3 Event triggered average.....	154
7.3 RESULTS.....	154
7.3.1 Gaze-Body interaction characteristics.....	154
7.3.2 Autoregression.....	157
7.3.3 Auto regression with exogenous inputs	158
7.3.4 Exogenous regression.....	158
7.3.5 Regression Performance Comparison.....	161
7.3.6 Gaze Event triggered behaviour.....	164
7.3.7 Saccade triggered grasplets.....	166
7.3.8 Grasplets association with body movements.....	168
7.4 DISCUSSION.....	172
7.4.1 Head-eye coordination: other data driven approach comparison.....	172
7.4.2 Embodied Saliency: the synergies of spatial attention allocation.....	174
7.4.3 Embodied Saliency: other body contributions to spatial attention	175
7.4.4 Embodied Saliency: extending the model.....	176
7.4.5 Embodied Gaze Descriptors: building blocks of visuomotor behaviour.....	177
7.4.6 Grasplets: predictive body movement synergies.....	178
8 CONCLUSION	181
8.1 SUMMARY OF RESULTS.....	182
8.1.1 Gaze estimation beyond the screen.....	183
8.1.2 Natural gaze for instant learning.....	183
8.1.3 Drive and grasp by eye.....	184
8.1.4 Natural behaviour: embodied gaze methodology.....	185
8.1.5 Gaze behaviour in the wild.....	185
8.1.6 Embodied gaze descriptors for prosthetic control.....	186
8.1.7 Embodied Saliency.....	187
8.1.8 Future perspectives.....	187
9 BIBLIOGRAPHY	189
10 APPENDICES	207
10.1 IMAGE PROCESSING IMPROVEMENT AND CALIBRATION AUTOMATION:.....	207
10.2 IMAGE PROCESSING RESULTS.....	217

1 Introduction

The interaction of brain, body and environment to perform the simple tasks of daily life is an immensely complex feat to behold from an engineering perspective. Machine learning and artificial intelligence enables computers to solve extremely complex information search and high-level reasoning problems. However, the ability to physically interact with the world, in ways a two-year-old can, still eludes the capabilities of machines. This is Moravec's paradox, which states *"it is comparatively easy to make computers exhibit adult performance on intelligence tests or playing checkers and difficult or impossible to give them the skills of a one year old when it comes to perception and mobility"*(Moravec, 1988). Though we are closer, this statement still holds almost 30 years on. It is not entirely surprising that great efforts have been made to enable computers to solve the problems we find difficult, rather than the aspects we take for granted. However, when these systems are affected by injury or disease the impact on life is extreme, as is the challenge of restoring them through technical means. This thesis takes a multidisciplinary approach, at the interface between engineering, medicine, and neuroscience to understand our fundamental ability to control movement and thus develop methods to restore this ability in patients with motor disabilities. To do this the intrinsic link between perception and action is targeted to decode movement intentions from eye-movements.

The eyes are the window of the soul. Their almost unnaturally rapid and precise movements have captured our imagination for centuries, with a strong presence in metaphors symbolism and language. They are studied extensively in many fields from neuroscience and psychology, to computer vision and are applied to many applications such as lie detection, advertising, diagnostics and computer control. This thesis focuses on the less utilised application - to control prosthetics and wheelchairs. The approach is not a purely engineering solution, treating the eyes as a tool. Instead technical developments are combined with behavioural experiments towards a new interface that works with the natural function of eye-movements - to guide behaviour. Emphasis is placed on studying natural behaviour in the context of prosthetic use - everyday life. This introductory chapter introduces the three core concepts integral to the unique approach of this endeavour. Firstly, introducing eye-movements as a highly promising, yet underutilised,

approach to prosthetic and wheelchair control. Secondly, the challenges and current approaches to restore movement function to patients who can no longer move due to injury or disease. Thirdly, introducing the importance of studying eye-movements and action "in the wild" outside the confines of the lab.

Controlled movement requires active perception; to see the world we look. Observing the world through discrete, rapid, focussed eye movements (saccades) acting to align the high-resolution central vision area (fovea) of both eyes with an object of interest (fixation point). Visual information is vital to movement planning and control, thus monitoring eye movements gives significant insight into our motor intentions. Humans have evolved highly specialised and efficient mechanisms for targeting information in the environment - precisely where and when it is needed. We look where we are going and make rapid sequences of eye movements to guide complex manipulation of objects (Land and Lee, 1994; Land et al., 1999). Considering the eyes make the fastest movements in the body, faster than a heartbeat, reflects the pace of information retrieval required for the brain to coordinate and orchestrate the complex actions of everyday life. Thus eye-movements represent an important component of the perception action loop, which is the cyclic flow of information between the actor and environment to guide a behavioural sequence towards a goal. In essence the brain can be thought of as a dynamical system mapping sensory inputs to motor actions, thus by understanding the (active) perception-action-loop, eye-movements can be interpreted to provide a high frequency signal directly relevant for neuroprosthetics.

The advancement of neuroprosthetic devices holds the hope to restore vital degrees of independence to patients with neurological and motor disorders, improving their quality of life. While the mechatronics to emulate limb function from a movement or actuation perspective are increasingly sophisticated, patients are unable to utilise this dexterity due to inadequacy of control interfaces. Thus to fully realise the potential of these devices, significant improvements in human machine interface (HMI) technology are required.

An ideal human machine interface requires a signal that can communicate sufficient complexity (information throughput) at a low latency to dynamically guide and respond to environmental changes of everyday life. In addition, for patient acceptance, this interface must be very easy to learn. In modern prosthesis research, the most utilised approach is electromyography (EMG) to measure residual muscle activation, which can be augmented with the promising development of targeted muscle reinnervation (Kuiken TA et al., 2009). For high level spinal injuries where this is not possible, there is huge research into direct brain machine interfaces (BMI) where signals are recorded directly from the motor centres of the brain. Both invasive and non-invasive methods exist, however non-invasive methods such as EEG have insufficient information throughput and too high latencies to be considered realistic solutions at this stage. At the other end of the spectrum, invasive methods involve dangerous surgical implantation of electrodes, expensive pre- and post-operative procedures and suffer from signal degradation with time. These issues are likely to be ameliorated by the steady progress of sensor quality and density, however, this approach is not a practical solution for the near future, particularly for patients where other interfaces are available.

The ocular motor system is innervated from the brainstem and thus eye-movements are retained in spinal injuries and many other afflictions of the peripheral nervous system.

They are the only directly observable behavioural signals that are naturally both highly correlated with actions at the task level, and proactive of our peripheral motor actions (Land et al., 1999). Eye movements offer a highly accurate, low-latency read out. This is because the brain has evolved to minimise the role of noise and delays to meet the tight informational constraints of the perception action loop.

Ideally, a control solution will be multimodal, combining any retained motor function with different interfacing methods to provide the best solution for the individual requirements of the patient. Irrespective of the interfacing method used, to be effective, the prosthesis should mimic natural behaviour as closely as possible to become a seamless part of the perception action loop. This will leverage the natural plasticity of the brain, increase patient uptake and potentially even promote embodiment of the limb. Such biomimetic design requires a deeper understanding of the perception action loop in natural behaviour.

To truly understand, predict and even replicate human behaviour studies must move from the lab and "into the wild". There is a general trend in neuroscience and psychology to ensure the ecological validity of behavioural experiments in the lab (Brabender, 1977). Going beyond this, the idea of cognitive ethology advocates taking behavioural experiments out of the lab entirely (Kingstone et al., 2008). This shift is set against a long tradition of highly reductionist experiments to study sub systems of the brain and behaviour in isolation. Many theories for sensorimotor control are founded on such experiments. However it has been repeatedly observed that when the strict experimental conditions are not met, behaviour becomes unpredictable using these reductionist models (Kingstone et al., 2008; Tatler et al., 2011).

In parallel to this paradigm shift in experimental approach, the idea of embodied cognition has moved to the forefront of thought in computer science and cognitive neuroscience (Gover, 1996). Central to this is the premise, is that the mind is not just a computing component connected to the body that can be studied in isolation, but the body and the environment influence and shape cognition in the brain. Thus as sensor technology becomes mobile, an increased emphasis should be placed on studying behaviour in the wild.

The importance of this paradigm shift has been central in the progression of our understanding of eye-movement behaviour, as will be discussed at length. It is the pioneering studies that monitored eye-movements during everyday activities such as making a cup of tea, driving and playing sport that changed the way we understand eye-movements (Land and Tatler, 2009). Indeed, it is these studies that inspire the approach taken here, because they show that eye-movements are tightly locked, both spatially and temporally, to movement behaviour and thus reflect our intentions.

1.1 Summary of thesis

This thesis attempts to make a holistic contribution to prosthetic and wheelchair control, considering both engineering developments and furthering the understanding of human visuomotor behaviour as equally important components. Working at the interface between engineering and behavioural neuroscience provides complimentary perspectives. Drawing from scientific understanding of behaviour informs the technical development,

while developing data driven technological platforms furthers understanding in science. Figure 1 gives an overview of the main components of the thesis and represents the development cycles at the interface between engineering and neuroscience.

In the initial phase of development, an ultra-low cost binocular eye-tracking system is developed, capable of estimating 3D gaze position (Chapter 2). This approach is motivated by the wealth of literature linking gaze to our actions in everyday life and the physiology of the oculomotor system that enables 3D gaze to be estimated (vergence). The main innovation of this interface is to enable gaze interaction directly in the 3D world, rather than via a computer screen. This method is presented in the context of Brain Machine Interfaces (BMI), using the commonly applied metrics of information throughput and cost to benchmark the system. However much of the BMI benchmarking is applied to discrete control tasks and thus a new continuous control benchmark is proposed in Chapter 3: to play the video game Pong. While this simple computer tennis game seems trivial in comparison to prosthetic control, the fundamental requirement for players to communicate and update movement intention in real time, makes it more relevant than other discrete control benchmarks.

The gaze control interface is benchmarked with Pong in a large field experiment involving over two thousand people. Demonstrating that within 30 seconds, outside of the confines of the lab, people who had never used their eyes as a control input could play the game. The experimental platform developed enabled full post analysis of game behaviour to quantify performance but also to understand the control strategies used. Subjects tended to use similar eye-movement strategies to those reported in literature during ball catching. This simple example highlights the importance of understanding eye-movement strategies in natural behaviour to leverage this inherent behaviour in the control of devices. From this the motivation to control wheelchairs and robotic arms turned into reality. Chapter 4 presents a number of different platforms developed for controlling wheelchairs and robotic arms. While most previous work uses a gaze controlled cursor, via a graphical user interface, the approaches here endeavour to allow direct gaze control in the real world, with minimal impact on natural gaze behaviour. Again these interfaces were developed not only as engineering solutions, but also as experimental platforms, recording behavioural metrics for benchmarking and analysing behaviour for different approaches, including their impact on natural eye-movement behaviour (Chapter 4). In addition, direct interaction with physical devices led to new methods for 3D gaze calibration, leveraging the robotic arm to calibrate the gaze tracker, dramatically improving performance.

With an end-to-end application for gaze controlled robotics and wheelchairs developed the challenge was to further understand natural behaviour in freely behaving healthy subjects to inform the improvement of these interfaces. Towards this, a novel methodology for the embodiment of eye-movement recordings in the wild is presented in Chapter 6. The approach comprehensively captures, rather than constrains, sensory inputs and motor outputs in natural behaviour. To achieve this many technical and algorithmic challenges are met to measure, synchronise and analyse the large volumes of data required to capture the full repertoire of human behaviour. The result is a unique and fascinating database that has applications well beyond the scope of this thesis. New automated analysis methods and observations begin to give quantitative results, building

towards the development of intention decoding and biomimetic prosthetics in Chapter7. In conclusion, Chapter8 summarises this journey, identifying the important findings, the caveats, the questions raised and the future vision for patient centred bioinspired gaze prosthetics.

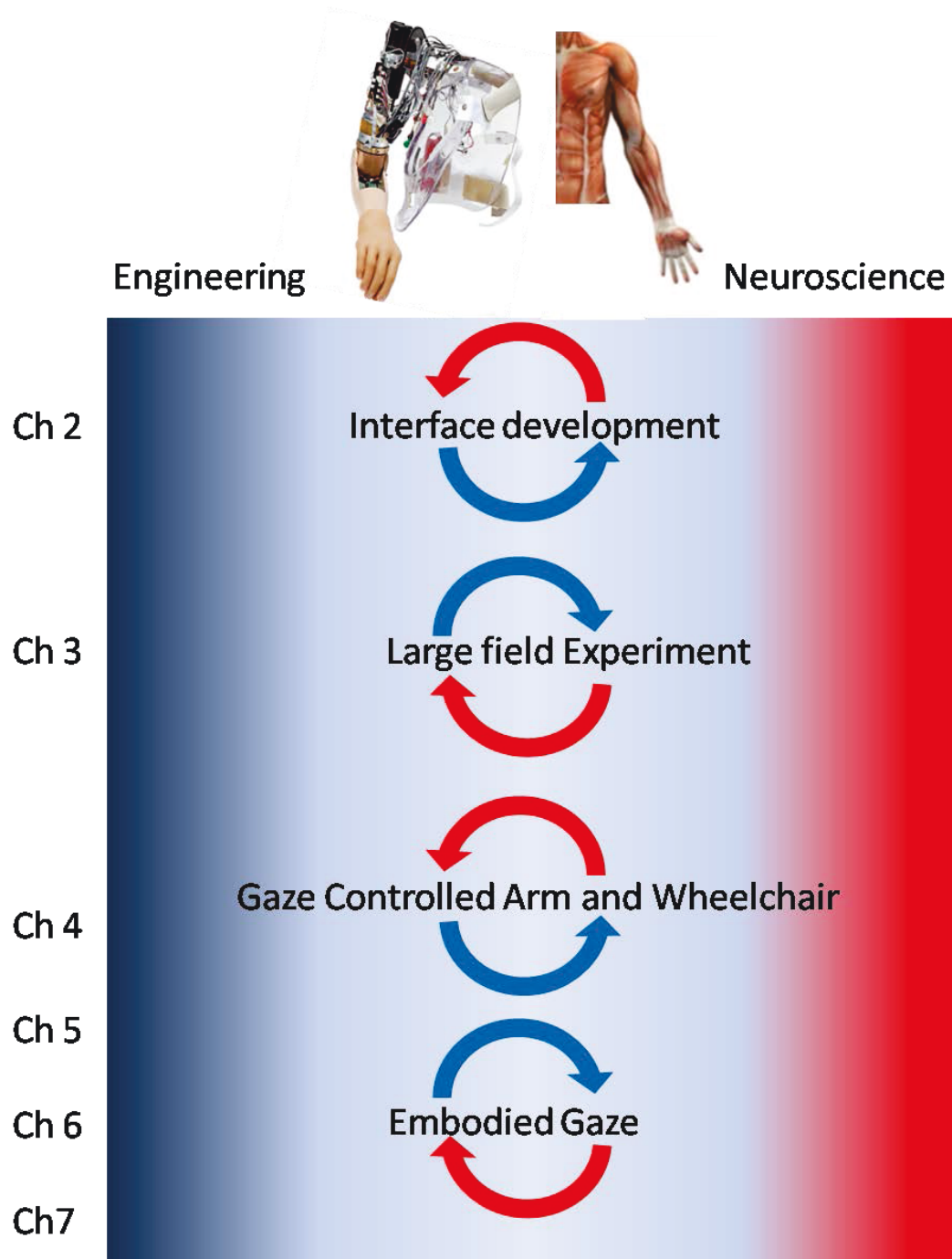


Figure 1: Thesis overview. Representing the four main developments and their relationship at the interface between Engineering and neuroscience.

2 Eye-movement behaviour, tracking methods and gaze based intention decoding

Eye movements are tightly locked to motor behaviour both spatially and temporally. The following chapter presents an overview of eye-movement behaviour and discusses these findings in relation to the neuroscience of perception, action and intention decoding.

2.1 The Eye

2.1.1 Anatomy

The human eye is approximately spherical with an aperture (Pupil), an external fixed lens (Cornea) and an internal lens; the combination of which focus the light onto the photosensitive tissue covering the back hemisphere of the eyeball (Retina). The fovea is a small pit in the retina (~1.5mm diameter, <2 °visual angle) with the highest cone density (peak density of ~200,000 mm⁻²) (Duchowski, 2007). This region corresponds to the central vision, with the “Visual Axis” being defined as the ray connecting the centre of the Fovea with the centre of the visual field. The resolution of the retina drops off rapidly in the first few degrees moving away from the Fovea (Lewis et al., 2003). This can be easily demonstrated by looking at some text and then moving it out of central vision, the text becomes illegible almost immediately. This region of acute vision is extremely small, covering approximately one four-thousandth of the retinal surface (Werner and Chalupa, 2004), and thus the eyes dart around the scene to areas of interest or relevance to our

intended actions. This requirement imposed by the retinal architecture is fundamental to the strong link between eye movements and action.

2.1.2 Oculomotor System

To allow the eyes to move in complex rotational trajectories, facilitating the high-resolution assessment of a visual scene, each eye has 6 muscles to control its movements (Figure 2). Accurate eye movements require very little concentration and in many cases are reflexive. The oculomotor system has a well-defined motor repertoire, which is very similar between individuals and can be categorised into 5 systems that are understood not only in terms of their movements, but the neurophysiological mechanisms that produce them and their perceptual implications.

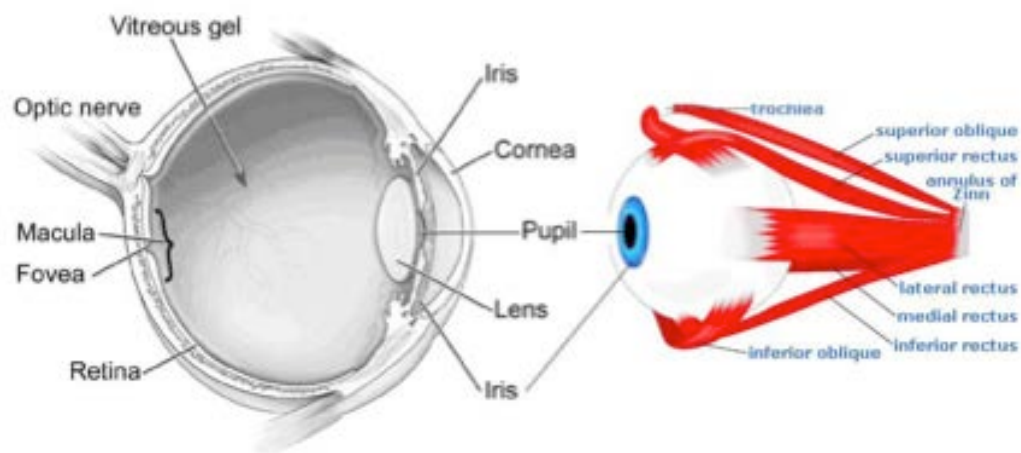


Figure 2. Anatomy of the eye and the extraocular muscles.

1) *Pure Fixation:* To fixate on a region, we must maintain a stable image on the retina to prevent image blur. Just as camera exposure time affects the level of motion blur in an image, the photo receptors in the retina have a certain response time (around 20ms for cones (Friedburg et al., 2004)). This means that for image speeds greater than 2-3 degrees/second we can no longer resolve the higher spatial frequencies (Westheimer and McKee, 1975). Fixations last at least 130ms but vary over a wide distribution, with an average length of around 300ms (Land and Tatler, 2009; Pelz and Canosa, 2001) but can last for several seconds in real-world behaviours (Hayhoe et al., 2003; Land et al., 1999). During fixation there are also three types of involuntary eye movement; Tremor – high frequency (150Hz) low amplitude ($<0.02^\circ$), drift ($<1^\circ$) and microsaccades ($<2^\circ$). Only microsaccades have significant spatial effects on eye-tracking measurements, reaching amplitudes up to 2 degrees (Martinez-Conde et al., 2004).

2) *Image stabilization:* In natural behaviour, humans are highly dynamic, as is their environment and as such during fixations there are two important mechanisms that act to stabilise the retinal image: the vestibular ocular reflex (VOR) and the optokinetic response (OKR). The vestibular system, which detects head movement, is directly coupled to the oculomotor system with a gain ideally close to 1 so that eye movement almost exactly counteracts the head movement. The optokinetic response is a complementary stabilising reflex that compensates global scene velocity detected in the retinal image, rotating the eye to match and thus counteract the retinal image motion. These systems work most

effectively in combination and thus if stimulating the VOR system alone, i.e. in darkness, the gain is usually more like 0.6 (Barr et al., 1976). VOR is feed forward and requires OKR which uses visual feedback to supplement VOR achieving the gains closer to 1. An additional important distinction is that the latency (between head and eye-movement) is an order of magnitude higher for OKR, with 15ms latency in VOR and 150ms for OKR (Land and Tatler, 2009).

3) *Saccadic system*: We must constantly realign our high acuity central vision to analyse our surroundings. If we did this at speeds less than 2-3 degrees per second to prevent any image blur, re-fixation would take tens of seconds, making for dry viewing. Instead we make these transitions very rapidly, reaching angular velocities as high as 900°s^{-1} (Ebenholtz, 2001) p55, but suppress the visual signal during these periods (Bridgeman et al., 1975), a process known as saccadic suppression. Saccades last on average 30ms and tend to be followed by a fixation that lasts at least 130ms but on average 300ms, which means we spend about 1 tenth of our waking day effectively blind, without even considering blink durations (Land and Tatler, 2009) p19.

4) *Smooth Pursuit*: The smooth pursuit system stabilises a small moving stimulus in our field of view by tracking its motion. It is a reflexive action, only occurring with sensory input (can also be auditory or even proprioceptive). It is similar to OKR but follows a small moving target rather than large regions of image motion. Purely smooth pursuit can track targets with angular velocities up to 15°s^{-1} , above which pursuit becomes a combination of smooth movements followed by catch up saccades. When target velocities become greater than 100°s^{-1} purely saccadic eye movements are required to track the stimulus (Land and Tatler, 2009) p23.

5) *Vergence System*: In humans, both eyes are positioned in the front of the head; this means that there is a large overlap in each eye's visual field, giving a very wide binocular field. In this binocular region, the Vergence System aligns the two eyes such that the respective visual axes intersect at the point of interest. Vergence movements change the depth of fixation, driven by disparity between the two eyes and are smooth and relatively slow. They prevent double vision and facilitate depth perception in higher processing centres of the brain. In the visual cortex, the stereo images from the binocular region of the visual field merge to form the image we perceive. For this merged image to make any sense, the visual information reaching each eye must be very similar. A disparity between the images above a certain threshold cannot be resolved and results in double vision. However, because the eyes are a few centimetres apart there will inherently be a small disparity between the retinal image projections. Providing this disparity is below a certain threshold, the images are resolved to a single vision image and the disparity is interpreted to give the perception of depth. The understanding of this mechanism forms the basis for 3D gaze estimation.

These subsystems have been categorised based on reductionist lab paradigms that isolate the different subsystems. Lab based studies tend to fix the head and use simple on-screen stimuli to characterise these subsystems. In natural behaviour however, all these subsystems occur simultaneously, giving a considerably more complex eye-movement repertoire. In addition, in natural behaviour gaze is no longer purely controlled by eye-movements alone but instead it is the coordinated movement of eye, head and body that is

integrated to allocate gaze. These categories however give a foundational understanding of the capabilities and boundary conditions of the eye-movement system, for which the physiology and neurophysiology is well understood. Building on this, huge efforts have been made to understand how the brain allocates overt attention, spatially and temporally, to meet the cognitive demands (for review see (Eileen, 2011)). This line of research was first pioneered by Buswell (1935) and later, more famously, Yarbus (1967) who made the pioneering studies showing that the sequence of eye-movements was different for different cognitive tasks (Buswell, 1935; Yarbus, 1967). The following sections explore findings that relate temporal and spatial statistics of eye-movement behaviour to cognition and action.

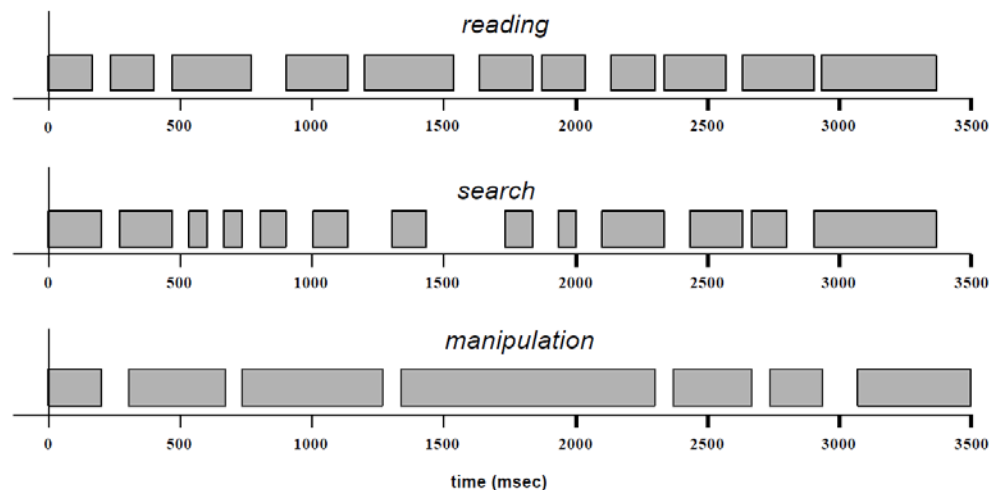


Figure 3. Fixation sequences shown by Pelz et al, in their study of eye-movements during a model building task. The grey bars indicate fixations while spaces between indicate saccades (Pelz et al., 2000).

2.2 Eye movement characteristics and cognitive processes

The general repertoire of eye-movements provides the building blocks of overt attention and informs us where and when information is being acquired (visual intake - fixation and smooth eye-movements) and when it is suppressed during reorientation to new information (saccades). Within these categories there is significant variation in temporal and spatial characteristics. Many studies have shown that fixation length relates directly to the time required to acquire and process the specific information (Droll et al., 2005; Hayhoe et al., 1998, 2003; Land et al., 1999). For example Inhoff and Rayner demonstrated that during reading, fixations are significantly longer on more difficult words (Inhoff and Rayner, 1986). Also, in different tasks such as search, reading and object manipulation, the different cognitive processes involved appear to be reflected in the accompanying fixation duration. Pelz al. showed evidence to support this by tracking eye-movements during a model-building task. Subjects had to read instructions, search for components and manipulate these components to put them together (Pelz et al., 2000). Figure 3 is from their study and shows the temporal fixation characteristics and saccades in the different cases. Search tasks are shown to have the shortest fixations as the eyes scan over the scene rapidly, while object manipulation involves much longer fixations perhaps to extract and process more detailed spatial information per fixation.

Similar differences were also shown by Mills et al. between the free viewing of a scene, and task driven viewing such as memorising or searching a scene (Mills et al., 2011). Search tasks exhibited the shortest fixations compared to memory-based tasks, which may reflect the increased time to extract the information and commit it to memory. While for search behaviour, a rapid sequence of fixations extracts only task relevant information, e.g. the object colour, before quickly moving on. Thus the information processing constraints of the particular task effect the temporal statistics of gaze behaviour. This alone does not disambiguate all activities. For example it is also shown that free viewing of the scene and memory tasks share the characteristic longer fixation durations. To understand why this might be requires consideration of the objective function of the brain in different situations. In the absence of task the inherent minimisation of energy costs would have a similar effect - less unnecessary saccades are made. However search and free-viewing conditions can still be distinguished based on the additional feature of saccade amplitude because it is significantly lower in free-viewing cases.

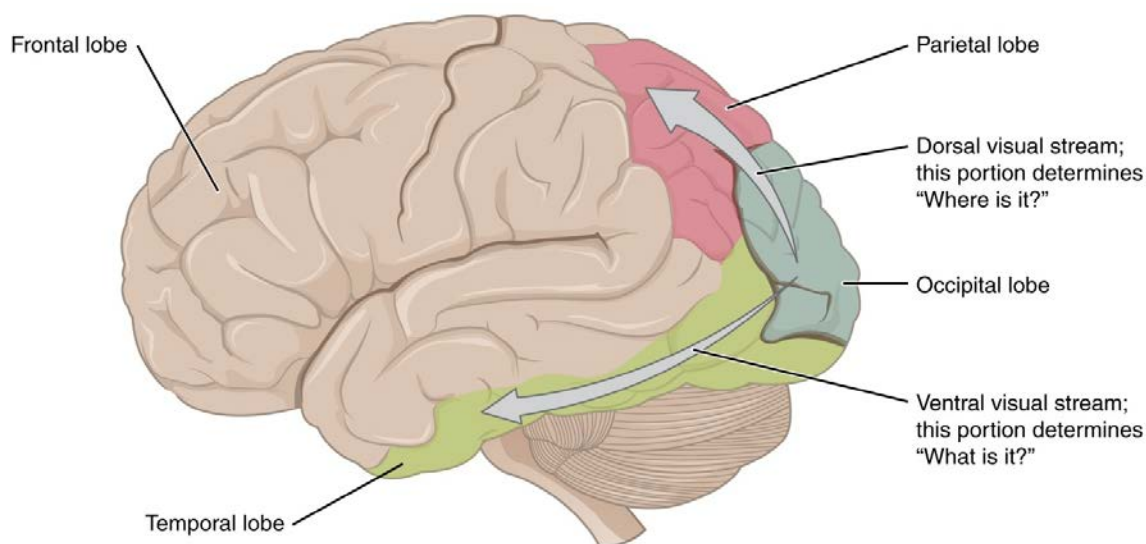


Figure 4: Two stream hypothesis - Ventral (vision for perception) and Dorsal (vision for action) streams. From (OpenStax, 2013).

To relate eye-movement characteristics to visual pathways in the brain Velichkovsky et al (2005) attributed distinct modes of scene viewing to the ventral and dorsal visual streams (see Figure 4). Broadly speaking the dorsal stream is thought to be associated with information about the location and movement of objects (vision for action) while the ventral stream is involved in recognition and discrimination of shapes and objects (vision for perception) (Goodale and Milner, 1992). Velichkovsky et al. extend this distinction to affect the eye-movement characteristics associated with the separate streams (Velichkovsky et al., 2005). Proposing an 'ambient' mode associated with the dorsal stream activity and a 'focal' mode associated with ventral stream activity. It is proposed that ambient processing involves short fixations and large amplitude saccades while focal processing involves longer fixations associated with low amplitude saccades (Unema et al., 2005).

Another key indicator of cognitive processing seems to be pupil dilation (Beatty, 1982; Beatty and Lucero-Wagoner, 2000) and interestingly this response can be proactive to self-initiated actions, by as much as 1.5s (Richer and Beatty, 1985). Care when interpreting

this information must be taken as pupillary reflexes are also caused by fear, pain, anxiety and arousal as well as changes in ambient light levels and accommodation to different depths of focus. Though this complexity would seemingly prevent any inferences being made, perhaps they can be differentiated by their time course and also their correlation with other eye movement characteristics as well as the task context. There are many other studies relating eye movements to cognitive processes – sometimes in surprising situations such as auditory language processing (Allopenna et al., 1998; Eberhard et al., 1995), lying (Dionisio et al., 2001), and even in mental arithmetic which seems to share neural circuitry with areas of the brain involved in calculating eye-movements (Hubbard et al., 2005; Knops et al., 2009; Loetscher et al., 2008). An extensive combination of features used together can be used for interpreting eye movement behaviour. As understanding of eye-movement characteristics becomes clearer, a richer set of features, both spatial and temporal can be extracted to infer cognitive processes and thus contribute to intention decoding from eye-movements.

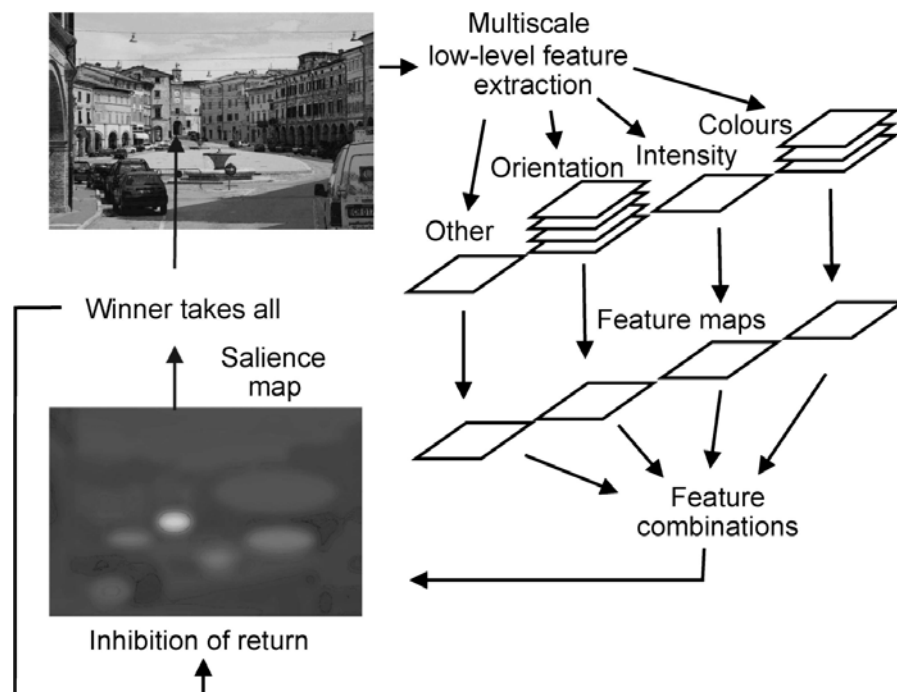


Figure 5. The saliency-driven model for eye movement control. From (Land and Tatler, 2009), based on (Itti and Koch, 2001).

2.3 Sampling the scene: What drives eye-movements?

The studies discussed above show general links between cognitive load and eye-movement features. This relationship provides one aspect of information contained in eye movements for inferring the activity a person is engaged in. This information must also be combined with an understanding of what guides our overt attention in space, i.e. fixation locations, and how this can be related to the task or action we are performing. Key to this issue is what has become known as the saliency argument, which debates how much the control of eye movements is bottom up and based on visual conspicuity or top down and based on cognitive influences and task context – for extensive reviews see (Schütz et al., 2011; Tatler et al., 2011). Visual conspicuity or saliency refers to locations in an image

that “pop-out” from their background due to high contrast in image features such as colour, intensity or orientation. The idea is to apply centre surround contrast for intensity colour and orientation in parallel, at different spatial scales, creating a conspicuity map for each of these features. This closely resembles the parallel processing carried out in the primary visual cortex, reviewed here - (Carandini et al., 2005). These conspicuity maps for each feature are normalised and combined into a saliency map (Itti et al., 1998; Koch and Ullman, 1985). The model then predicts fixations at the most salient feature points in a winner takes all fashion, with a subsequent local inhibition applied to prevent immediate re-fixation. An illustration of the saliency model is shown in Figure 5. This model has received huge attention both because it reflects our neurophysiological understanding of processing pathways and because it is a computational model that can be applied and tested and does, to a certain extent, account for fixations in static scenes (between 0.57-0.68 ROC AUC, where chance is 0.5 see summary table in (Betz et al., 2010)).

Visual saliency clearly holds some predictive power, but there are a number of arguments that further reduce the significance of this already modest predictive power. Firstly other simple known gaze behaviour biases, such as the tendency to fixate at the centre of the screen (central fixation bias) can achieve higher prediction rates than saliency models (Tatler, 2007; Tatler and Vincent, 2009). Secondly, any predictive power can be removed by altering the task (Einhäuser et al., 2008; Foulsham and Underwood, 2008; Henderson et al., 2007). This is not surprising as task is bound to affect where we look, allowing us to obtain the specific information we need. However these feature maps do clearly influence attention, as has been shown by the so called “pop-out” effect that allows us to rapidly detect a target that differs only subtly from distracters in just one feature dimension irrespective of the number of distracters (Treisman and Gelade, 1980). The features used in saliency maps are clearly important, we know that these features are extracted in V1 and the feature conspicuity does represent visually descriptive regions of the image. However as the context and goals become more complex there are other forces at play, acting to guide our attention. Figure 6 is taken from a review by Schütz et al. (2011) and provides a representation of the different control loops that have been shown to contribute to the assignment of our overt attention (Schütz et al., 2011). This includes motor plans, a value function that is tailored to the ongoing task, object recognition and their probable location and finally the low level saliency. Such an integration of top down and bottom up signals has been proposed to occur at higher levels in the visual hierarchy in so called priority maps for which neurophysiological studies are beginning to provide evidence (Schütz et al., 2011). In addition, Follet et al. compared empirical saliency maps of focal and ambient fixations based on their automated classification and found that focal fixations are more associated with visual features than ambient ones, providing further evidence for a distinction between bottom up and top down signals (Follet et al., 2011).

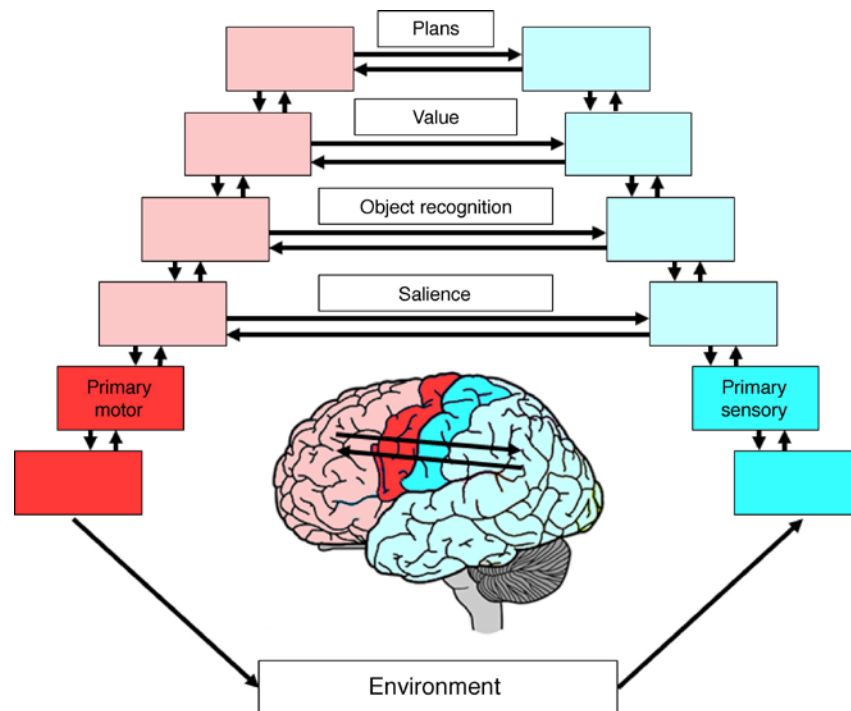


Figure 6.Representation of the contributing control loops proposed to guide saccades. Figure from (Schütz et al., 2011).

An important development in neuroscience and psychology in general is the recent ability to move experiments into ecologically valid environments. Until relatively recently gaze behaviour research was restricted to laboratory settings due to the complexity and non-portability of eye tracking equipment. Ground-breaking work over the last two decades is beginning to shift the focus onto gaze behaviour in natural everyday tasks and contexts such as driving, making a cup of tea (shown in Figure 7), playing cricket or making cucumber sandwiches (or peanut butter and jelly (jam) for the US based studies). Previous work was restricted to static scene viewing, which only represents a very small subset of gaze behaviour. Thus this change in paradigm has brought to light many fascinating insights into gaze behaviour in more relevant contexts, such as when its allocation is based on the coordinated movement of the body and the eyes. It has further demonstrated that the control of where we look is based, overwhelmingly so, on the location of information required by ensuing action sub-goals (Epelboim et al., 1995; Hayhoe et al., 2003; Land and Furneaux, 1997; Land et al., 1999; Patla and Vickers, 1997; Pelz and Canosa, 2001) for extensive reviews see (Schütz et al., 2011; Tatler et al., 2011). In motor actions, we fixate primarily on objects or regions that are directly relevant for task completion – even if this is a location on the featureless (0 saliency) kitchen work top that we plan to place a cup, or the blank wall we expect to observe a squash ball to bounce (Tatler et al., 2011). In addition to this, the temporal arrangements of fixations are directly time locked to the actions they relate to, and thus information is acquired in a “just in time basis” (Ballard et al., 1995). This spatial and temporal relationship between overt attention and action has been shown to be highly consistent across subjects, suggesting a common mechanism. This can be seen in Figure 7, where the location of fixations is represented during the task of filling a kettle for 3 different subjects (from (Land et al., 1999)).

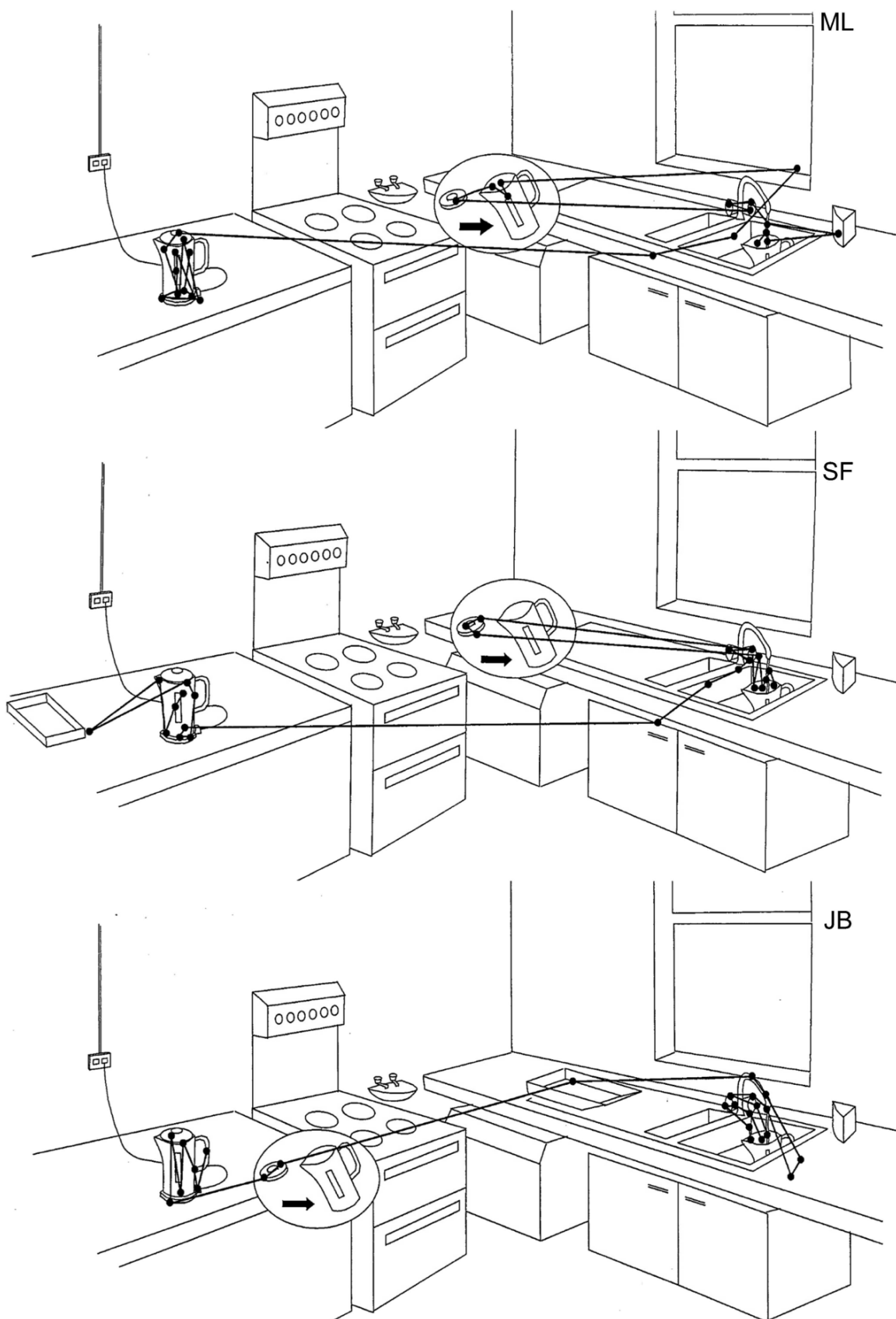


Figure 7.Representation of gaze scan paths for 3 different subjects during a real world kettle-filling task. Shows the consistency of behaviour across subjects. Figure from (Land et al., 1999).

2.4 Keep an eye on the goal: Object interaction

Gaze not only tells us where the object we want to interact with is, but how we look at that object reflects the specific information we require to plan our interaction with the object. The specific information we target about an object can inform the intended action. For example when we look at a bottle of milk on the breakfast table there are key functional areas that give specific information such as the handle orientation to plan a grasp, the level of the milk to plan a pour or a brief glance at the colour of the top to decide between two different types of milk on the table. There are now many studies that illustrate this. Rothkopf et al demonstrated that the action of avoiding an obstacle compared to picking it up resulted in a very different distribution of fixations on the object (Rothkopf et al., 2007). As seen in Figure 8, for the avoid condition, the fixations were clearly distributed on the object edge, presumably to give spatial information about the boundary. While in the “pick up” condition, which required the subject to walk towards the object, the fixations were densely distributed around the object centre. These functional connections between fixation location and action are found repeatedly in the real-world. Here I will briefly discuss some of these observations in the context of object manipulation and navigation.

Gaze distribution on litter in the ‘pickup’ condition Gaze distribution on obstacles in the ‘avoid’ condition

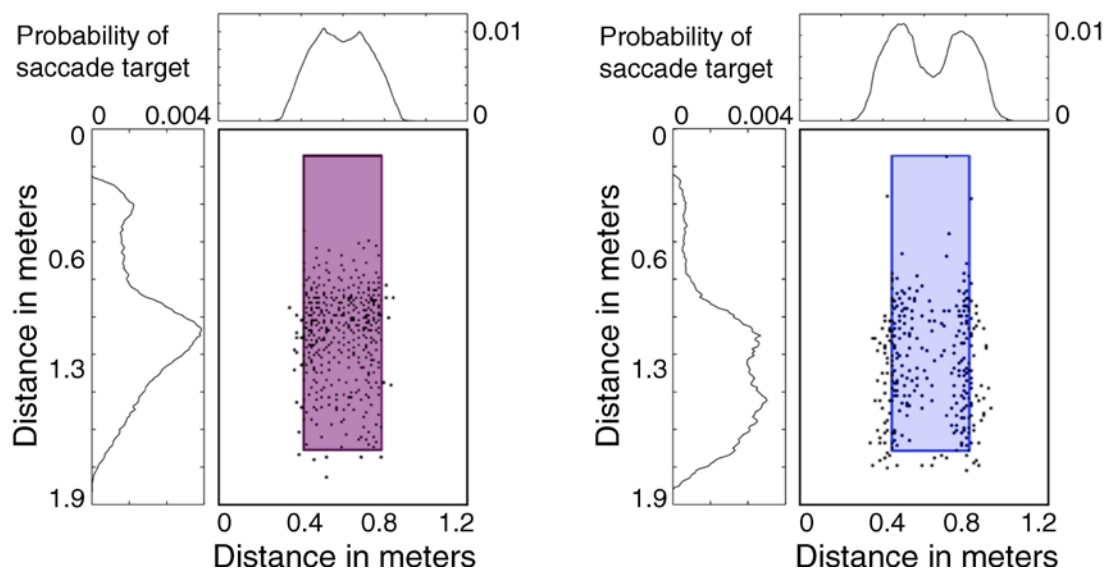


Figure 8. Fixation distribution for the interaction with two different objects in a virtual reality system. The objects are identical in size but the purple object (left) must be “picked up” by walking into it, while the blue object (right) must be avoided as an obstacle. From (Rothkopf et al., 2007)

Well documented fixation patterns that reflect different actions with the same object have also been made for tea making (Land et al., 1999). When the subject puts the kettle on its base the fixations alternate between the base protrusion and the bottom of the kettle, while when pouring from the kettle they fixate on the liquid stream in the receiving vessel. Interestingly, during these various behaviours we never fixate on the hands and once an object is in our hands we don't fixate this either, relying purely on haptic and

proprioceptive information, freeing the gaze to inform subsequent tasks. An illustrative example of this is during eating a piece of food with the hand for which the reach and correct shaping of the hand is visually guided (fixating the piece of food) and then once the food is close enough, the somatosensory system takes over to direct the food into the mouth (Sacrey and Whishaw, 2011). Another interesting aspect of this study is the importance of gaze in pre-shaping the hand for grasping. This has also been shown in a number of other studies (Baldauf and Deubel, 2010; Brouwer et al., 2009a; Johansson et al., 2001). We fixate on specific positions on an object to allow accurate placing of the fingers to grasp the object effectively. This highly specific gaze targeting behaviour, that is proactive of action, gives an idea of the rich information in eye-movements that can be used for inferring intentions in everyday behaviours.

2.5 Look where you're going: Eye-movements during navigation

Vision is also essential for navigation, and gaze precedes us on our chosen path. Land and Lee showed that we reliably look to the lane ahead when travelling straight and to the tangent of the bend when cornering (Land and Lee, 1994). This is consistent between subjects, with three subjects maintaining 50% of fixations within a 3-degree region of each other in a driving experiment. More directly, gaze angle and steering angle were shown to be very similar (Figure 9(a)). This figure shows gaze angle and steering angle during a driving experiment, while Figure 9(b) shows the cross correlation for different time delays, demonstrating that steering follows gaze after around 0.8s (Land and Tatler, 2009).

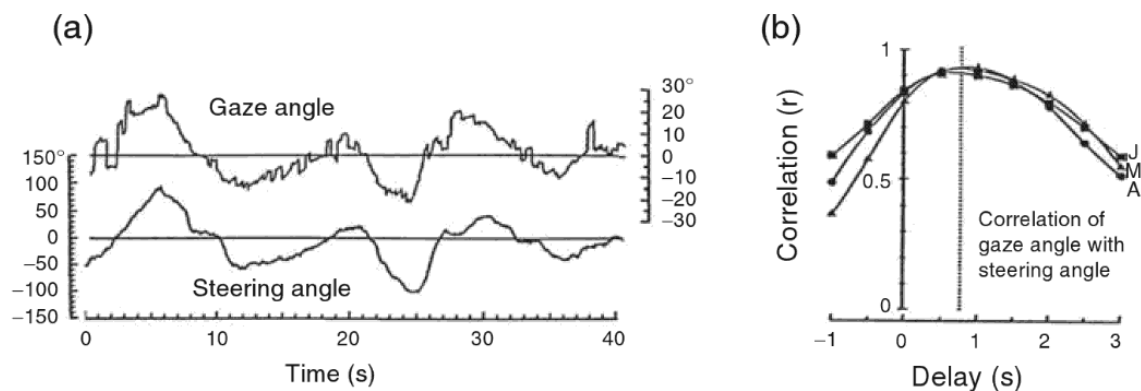


Figure 9. (a) Similarity of driver's gaze angle and steering angle in a driving task. (b) Cross-correlation between gaze angle and steering angle for different delays of steering relative to gaze angle. Each line corresponds to the records of a different subject. From (Land and Tatler, 2009).

2.6 Beyond top-down and bottom-up effects

In addition to this dichotomy between top-down and bottom up saliency, there are other important inputs that have been less well explored. For example, historical gaze has also been shown to bias the allocation of overt attention. It has been claimed that a third historical selection effect must be integrated with top-down and bottom up models (Awh

et al., 2012). This effect is beyond the inhibition of return, acting to reduce the priority of a previously fixated object, but demonstrates that previous biases associated with goals can have a lingering effect on selection bias. Beyond this, to fully capture eye-movements and thus overt attention, there are many other contextual and behavioural considerations. A simple example is the previously mentioned central fixation bias. It is likely that many more biases exist in different behavioural contexts, particularly in free movement behaviour where the state and dynamics of the body must contribute significantly to the eye-movement behaviour as will be explored extensively in Chapter 7. It is often the experimental context and capabilities that shape emerging theories. As eye-movement studies moved out of the lab, the impact of physical task became clear. Now, with the ability to record natural behaviour at an unprecedented resolution, as will be demonstrated, the emergence of all-encompassing data driven models can emerge that work towards capturing the full complexity of behaviour as the integration of these multifaceted and simultaneous drivers of eye-movements and thus attention.

2.7 Gaze cues for intention recognition and the Midas touch problem

Humans effortlessly solve complex problems in social interaction, interpreting and even predicting other people's actions to understand their intentions and state of mind. A simple example is walking along a busy street without bumping into people. We are good at very quickly inferring the behaviour of people walking towards us to avoid them. Eye-movements form a significant component of non-verbal communication for humans, particularly when reading the intentions of others (for overview see (Knapp et al., 2013) p295. This is clear from a very early age in humans, where it has been shown that 12-14 month year old babies were able to predict an adults actions based on gaze direction (Phillips et al., 2002). Access to gaze information has also been shown to significantly improve collaborative tasks in humans (Brennan et al., 2008). It is suggesting that by inferring the intentions and actions, subjects rely heavily on non-verbal gaze communication to coordinate their shared efforts more effectively . In fact in some tasks, the addition of verbal communication actually degrades performance (Brennan et al., 2008). With such a powerful communication mechanism, attempts have been made to facilitate gaze based intention decoding in robots to improve human robot collaboration (for review see (Mavridis, 2015)). For example, in a human pointing task, Renner et al. could predict the target location at the onset of the pointing gesture to within 10cm of the target area in 50% of cases (32x32cm workspace) (Renner et al., 2014). This would enable a robot to react hundreds of milliseconds before the pointing gesture is completed. Though a significant result, it is unclear why the performance was so low given we know it is possible for significantly higher saccadic accuracies to target locations. There are not sufficient details given in the article to ascertain whether this is a problem of their intention decoding method or the lack of accurate gaze information elicited by the subject in the natural pointing task. This is a good example of why it is particularly important to take experiments out of the lab, where very accurate saccadic targeting is observed, and into the wild, to see if the behaviour observed in this context is the same. Only with this understanding can real-world intention decoding algorithms be developed for prosthetic and wheelchair control.

In general, drawing on existing real-world studies while conducting further studies is vital to building effective real-world compatible intention decoding algorithms. As has been discussed, there is a wealth of eye-movement literature that can contribute to this, and the preceding sections give a flavour of the diversity and scope of this research. To relate this to intention decoding in a more structured manner, the descriptive nature of eye-movements can be split into three levels - 1) The aggregate statistics of gaze characteristics that can distinguish between task types at a high level e.g. reading versus searching. 2) The "where" and "when" of interaction, spatial and temporal tagging of the areas or objects we want to interact with, providing a proactive choreography of our motor behaviour. 3) The "what" level of detail, where gaze can inform us the intended action associated with the object of interest. Thus eye-movements clearly hold significant action related information, but to be clear that is just one task of the visual system. We make up to 5 saccades per second, while actions happen with a sub-unit timescale of around 3 seconds. However while it is often reported that only 5-20% of fixations are made to task irrelevant objects in the wild, only about one third of fixations could be clearly associated with the ongoing action (Land et al., 1999). Additionally, we still make saccades even in the absence of action. Thus to refer to gaze interaction literature, the "Midas touch" problem becomes apparent (Jacob, 1990). This is the problem that we make up to 5 eye-movements per second and thus we don't want every eye-movement to elicit an action (an analogy to King Midas for which everything he touched turned to gold). Thus from an intention decoding perspective, how can an algorithm distinguish which fixations contain the specific information we act on and thus make accurate predictions of intent?

As discussed there are differences in eye-movement characteristics in different types of static tasks such as searching versus memorising. These findings have also been reflected in non-static lab experiments, where subtle differences are seen between active and passive tasks associated with objects. Gaze shifts were faster in speed and shorter in duration when observers actively tapped a sequence of 3D targets than when they viewed the sequence passively (Epelboim et al., 1997). Based on the same data Epelboim went on to show the relationship between head and eye-movements changed between the passive and active viewing task, with much more head movements associated with active looking tasks (Epelboim, 1998). Brouwer et al. also showed there were subtle differences between temporal and spatial dynamics of eye-movements between grasping objects and purely viewing them (Brouwer et al., 2009b). The automated distinction between eye-movements associated with vision for action (dorsal stream) versus vision for perception (ventral stream) mentioned previously has also been demonstrated (Follet et al., 2011; Krejtz et al., 2016; Velichkovsky et al., 2014). The study presented by Velichkovsky is particularly relevant as the approach is actually applied to solve the Midas touch problem (Velichkovsky et al., 2014). The question for intention decoding feasibility is how replicable and robust these differences are in the complexity of natural behaviour. To understand this, a new data-intensive approach to record freely behaving subjects in the wild is required. This must fully capture the high dimensional space of human behaviour with sufficient recording durations to further understand the full visuomotor repertoire elicited during daily life tasks. There are many hardware, software and algorithmic challenges to such a feat which are addressed in chapter 6 and 7. While chapters 3-5 attempt to push the boundaries of intention decoding for real-time control, based on the current understanding of visuomotor behaviour.

2.8 Eye-tracking history

The following gives a brief overview of the development of eye-tracking methodologies, for a more detailed review see (Wade and Tatler, 2005; Wade et al., 2003). The rapid eye movements over the scene were described eloquently by Portfield back in 1737 (Porterfield, 1737) as well as the reason for these eye-movements - so that certain parts of an image are "successively received upon the same most sensible part of the retina". The first method that enabled more empirical and accurate descriptions of eye-movements was based on the after image effect towards the end of the 18th century by Wells (Wells, 1792). This effect is based on the residual image perceived after looking at an image for an extended duration. With a new image displayed, this afterimage will then move relative to the new image in concert with eye-movements. This method was the main approach used until late in the 19th century when an additional modality was developed to directly listen to eye-movements. By connecting a rubber tube to the conjunctiva or eye-lid it was possible to directly listen to the eye-movements (Lamare, 1892). The saccade would result in a short clapping sounds attributed to the contractions of the ocular muscles. The first actual recordings of eye-movements were also made at the end of the 19th century by mechanically linking the eye with a kymograph with a plaster cup, lever and bristle such that eye-movements were recorded on the smoked drum. A drawing of such a device is shown in Figure 10, alongside an example of the recording it produced during a reading task (Huey, 1898, 1900). Other alternatives were developed, such as corneal reflection (the Dodge photochronograph (Diefendorf and Dodge, 1908)) or even attaching a mirror to the eye and then using a photographic plate to record the movement of the reflected beam. This method was perfected by Yarbus, with the development of a very light-weight suction cup contact lenses, enabling his pioneering studies already mentioned (Yarbus, 1967).

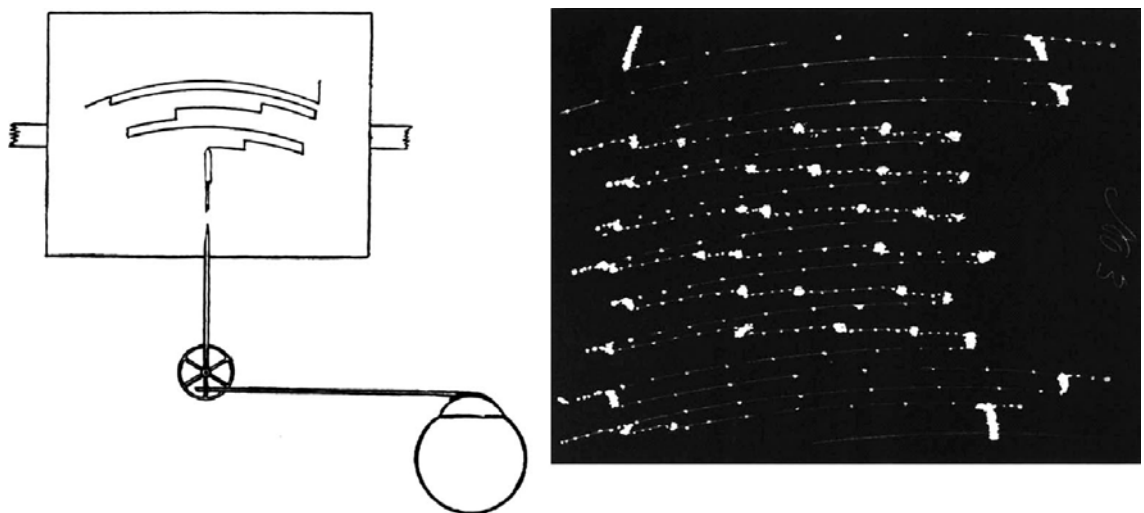


Figure 10. Left, sketch of the late 19th century approach to track eye-movements. The eye is mechanically linked to a kymograph with a plaster cup, lever and bristle such that eye-movements were recorded on the smoked drum. Right, shows an example of then recording made using the apparatus during reading. Figure left from methods paper (Huey, 1898) and right where it is applied to study reading (Huey, 1900).

The alternative method of Electrooculography (EOG) was established by Fenn and Hursh in 1936 (Fenn and Hursh, 1936). This method places electrodes on the skin around the eyes, and the movement of the eye, which is a dipole, induces voltages across these electrodes. This method is less invasive as it requires no contact with the eye and the resulting signal has a direct relationship with eye-movements. However the signal is prone to noise and artefacts from external sources as well as electromyography signals from facial muscles in the area. In addition, the signal is non stationary because the conductance between the skin and electrodes changes with perspiration and temperature and thus regular recalibration is necessary. The method does however have the unique attribute that it can be used to measure eye-movements when the eye-lids are closed and thus is still used widely in sleep studies. However, while a very useful, minimally invasive method, the signal imperfections mean that it is less used for vision based eye-tracking studies these days. Yet another modality emerged in the 60s that embedded a coil in a contact lens within a magnetic field - the scleral search coil (Robinson, 1963). The movement of the eye and thus coil within a magnetic field induces a voltage in the coil and yields very accurate and precise eye-movement measurements. The method has the advantage that ocular torsion can also be measured and this method still remains the gold standard for eye-tracking. However, the most prevalent method of eye-tracking today is video oculography, a method requiring no special contact lens or direct contact with the subject at all.

2.9 Eye-tracking state of the art: Video Oculography

Video-oculography, pioneered by Judd, McAllister & Steel back in 1905, using kinetoscopic photographs (an early form of video) of the eye (Judd et al., 1905). The eye-motion could then be measured manually from frame to frame. This approach was developed considerably over the last century, particularly with the advent of digital video recording and image processing to enable automated online video oculography. This is the most commonly used approach today, but comes in many forms, with a large variability on the hardware configuration and eye-tracking algorithms. The following section expands on this, giving an overview of the different approaches. The eye tracking processes has been split into 7 steps which define the possible hardware and software considerations.

Camera and mounting system: High quality machine vision cameras are used, capable of capturing high frame rates, to allow the rapid saccadic eye movements to be tracked and a high resolution to accurately locate the eye feature are used. The camera can be mounted on a headset or remotely for desk based eye-trackers. Mounting the camera on a headset is slightly intrusive but it allows head-free mobile eye-tracking. While desk based systems tend to fix the head and can achieve considerably higher accuracy and precision. Mounting the camera remotely in this way however means the user's movements are restricted within a certain region.

Eye illumination: The eye must be appropriately illuminated to give clear video imaging and thus efficient image processing. A passive approach relies purely on the ambient light of the user environment. This approach however makes imaging prone to noise from varying light intensities and results in specular eye reflections which make the resolving of eye-features for tracking more problematic. The use of a bulb emitting white light can be used to give better illumination consistency, however this may be distracting to the user.

The more common solution is to use an Infrared (IR) light because the human eye is insensitive to IR light.

Feature definition: To track the eye movement, first a prominent feature must be identified and defined. The key visible features of the eye are the white Sclera, the coloured Iris and the black Pupil. The Sclera strongly reflects both Visual and IR light, while the pupil strongly absorbs both. The Iris however absorbs most of the light in the Visual spectrum (depending on its colour) but absorbs much less in the IR spectrum. Hence under visual light, the most prominent contrast is between the dark Iris and the bright Sclera, while in IR conditions the contrast between the bright Iris and dark Pupil is more pronounced (see Figure 11). The pupil provides the most consistent feature as it is less often obscured by the eyelids (except upon blinking) unlike the iris that tends to be partially obscured to various degrees. An alternative method, known as the Dual-Purkinje-Image is based on tracking the first and fourth Purkinje images. These images are based on reflections from different structures of the eye, namely P1 - the outer surface of the cornea, P2 - the inner surface of the cornea, P3 - the outer surface of the lens and finally P4 - the inner surface of the lens. As the eye-moves the distance between the P1 and P4 images changes and thus measuring this gives a very robust measure of pure eye-rotation that is unaffected by eye-translation. This approach gives the most accurate video-based measurement of eye-movements however it requires a complex optical set-up of lenses and servo controlled mirrors to track these Purkinje images in real-time which is only currently practical for desk based systems.

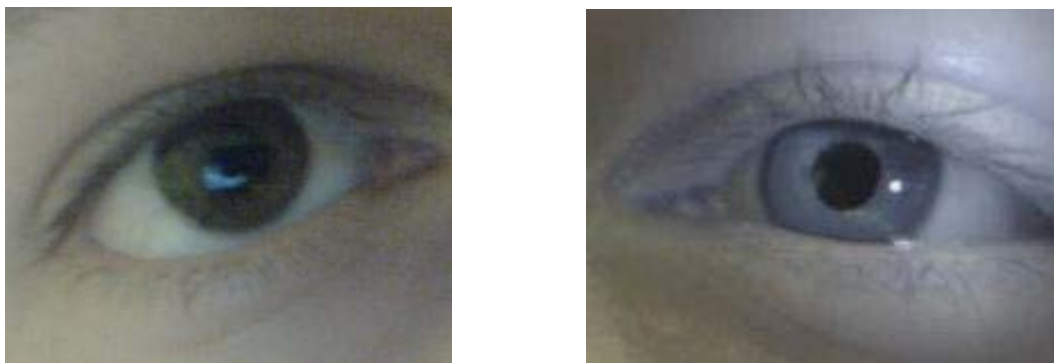


Figure 11. Comparison of eye-imaging under visual light (left) and Infrared light (right).

Image Segmentation: For feature identification, given the strong feature contrast, image processing techniques tend to be based on a greyscale threshold conversion to a binary image. This binary representation, if the correct threshold value is used, effectively extracts the Pupil Iris or Purkinje image, though unwanted features may also be extracted. Edge detection based algorithms have also been applied such as the radial scanning derivative (Li & Parkhurst 2006). This is an interesting technique as it limits the search area by scanning the image radially, starting from the previous frame's pupil centre. By doing this it finds the desired contour efficiently with less susceptibility to noise.

Feature Classification: To ensure the Pupil/Iris feature is correctly chosen, a number of different model based techniques are used. Following threshold based segmentation, a simple method to classify the features, is connected component labelling followed by filtering (Lin et al 2006). This method labels objects in a binary image defined by groups of connected pixels, it then filters unwanted objects based on attributes such as size and axis

ratio (shortest axis divided by longest axis). Another method employed is the elliptical Hough transform (Lee and Park, 2008), which can be used to detect ellipses of certain characteristics. Based on a range of possible pupil ellipse dimensions, the algorithm finds the ellipse of best fit. This method is effective but has a higher computational load because it requires many full image searches as candidate shape models are tested, though it yields the feature centre without further processing.

Feature Centre Extraction: Once the correct feature contour is extracted, its centre coordinates can be found in a number of ways. A basic method just takes the “centre of mass” of the contour (Lin et al 2006). This method is highly error prone, as if any of the feature is obscured (perhaps by Specular reflection) the returned centre will be incorrect. A more robust method fits an ellipse to the contour based on either least squares regression (LSR) or the random sampling consensus (RANSAC) (Li et al., 2005). The RANSAC method is particularly robust to noise as it fits ellipses to many random subsets of the candidate contour points then finds the ellipse of best agreement.

Coordinate reference point: Once an eye-feature coordinate has been extracted, the simplest approach is to then measure its position relative to the eye-image corner. However, to make the measurement more robust to shifts in the camera relative to the eye, it is common take it relative to the eye corner or the IR source’s first Purkinje image (corneal reflection) (see bright spot in IR image - right side Figure 11). With a configuration that mounts the camera and IR LED together, slight relative movement between the head and the camera will be compensated for by a change in the LED Specular reflection and thus the coordinate system origin. This means the system is more robust to slight relative movement between the head and the camera. However using this method doubles the image processing overhead because now the additional LED reflection feature must be found.

Mobile video oculography: A more recent development in eye-tracking technology was to enable tracking eye-movements of freely behaving subjects. A number of steps towards this were made, though the systems were still too bulky to allow behaviour to be considered free. However, by the early 90s, video-recorder technology was compact enough to enable Land to develop a portable eye-tracker recorded the eye image and scene image from the subject’s point of view simultaneously using a split field camera (Land, 1993). While the eye-tracking was still done post hoc frame by frame, the tracking recording was now possible in unconstrained behaviour. This signalled a paradigm shift in our understanding of eye-movement behaviour as has been discussed. Nowadays there are a large number of very compact mobile eye-tracking systems available that automatically track eye-movements and register them to the scene camera view (e.g. SMI Eye-Tracking Glasses II, Tobii Pro Glasses II).

Low-cost video oculography: Eye-tracking technology has advanced considerably over the last few decades with many types of eye-tracker available commercially. However until recently, these eye-trackers were extremely expensive (in excess of £30,000), limiting their application to lab and specialised industrial settings. Low cost eye-tracking systems have been developed using off the shelf web-cams (Li et al., 2006; San Agustin et al., 2009; Schneider et al., 2011). Recently a number of low-cost commercial eye-tracking systems have also been released. The first <£99 eye-tracker was released by Eye Tribe in 2015, quickly followed by the Tobii EyeX. Both of these systems were remote systems in the

form of an eye-tracking bar that was designed to be mounted on the computer monitor. These systems were intended for gaze interaction with the computer screen and thus give a gaze estimation position on the computer screen. The performance of these systems is well below the equivalent research grade remote eye-tracking systems that cost tens of thousands of pounds. With tracking rates around 60Hz rather than 1000Hz and a very noisy gaze signal, these systems cannot be used for eye-tracking research, other than to give an idea of where people are looking. However it marks the beginning of the movement of eye-tracking into consumer market applications such as gaming and human computer interaction. It has opened up the field, bringing eye-tracking to the masses.

2.10 Gaze tracking: from eye to gaze

Once the eye feature coordinate is known (from whichever technique), the gaze position can be estimated in a variety of ways. Most methods focus on 2D gaze estimation, usually on a computer monitor, a head mounted display or a scene view from a head mounted camera. Three dimensional gaze estimation on the other hand combines information from both eyes to obtain additional depth information. This method has receives far less attention in literature but it is a vital development for control of prosthetics in the real world.

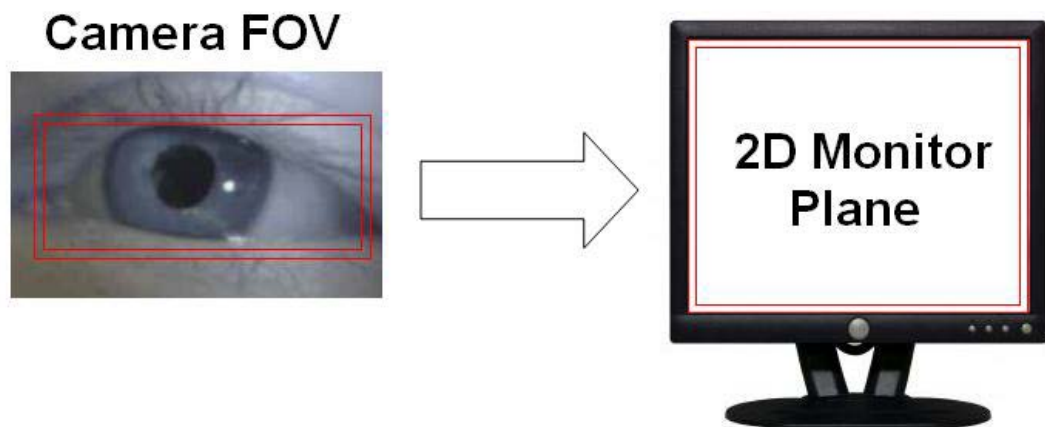


Figure 12. Illustration of the coordinate transform from the eye position in the eye tracking camera field of view (FOV) to the 2D plane of the computer monitor.

2D Gaze Estimation: A coordinate transform is made to relate the pupil position in the Field Of View (FOV) of the eye camera, to the corresponding gaze position in a plane of fixed orientation in space (e.g. computer monitor, see Figure 12). Calibration methods used can be categorised into two major approaches; 3D geometric eye model based or regression based. The geometric model based approach considers the mapping of the 2D pupil position in the camera FOV, to a 3D virtual model of the eye ball (Duchowski, 2007; Hennessey and Lawrence, 2009b; Lee and Park, 2008; Sheng-Wen Shih et al., 2000). From this model, the visual axis can be projected to find its intersection with the coordinate system of the interface (see Figure 13).. The visual axis or gaze vector is the ray that runs between the centre of the fovea in the retina, through the cornea to a gaze fixation point (neglecting the kappa offset between the visual and optical axis). This method is heavily parameterised, requiring information such as the 3D orientations of the camera, monitor

and eyeball to be known as well as estimates of the corneal radius and in some cases even the refractive index of the vitreous fluid (Hennessey & Lawrence 2009b). Additionally the system requires the Purkinje reflection from 3 or 4 LEDs of known position to give a reference pattern in the eye image. It requires a highly controlled experimental set up and is heavily dependent on the pre-calibration measurement accuracy. It has the benefit however that no calibration points are required though performance is significantly improved with a single calibration point. This enables a user specific eye parameter to be estimated (kappa discrepancy angle between the visual and pupillary axis caused by the corneal lens, for additional information see (Hennessey and Lawrence, 2009a).

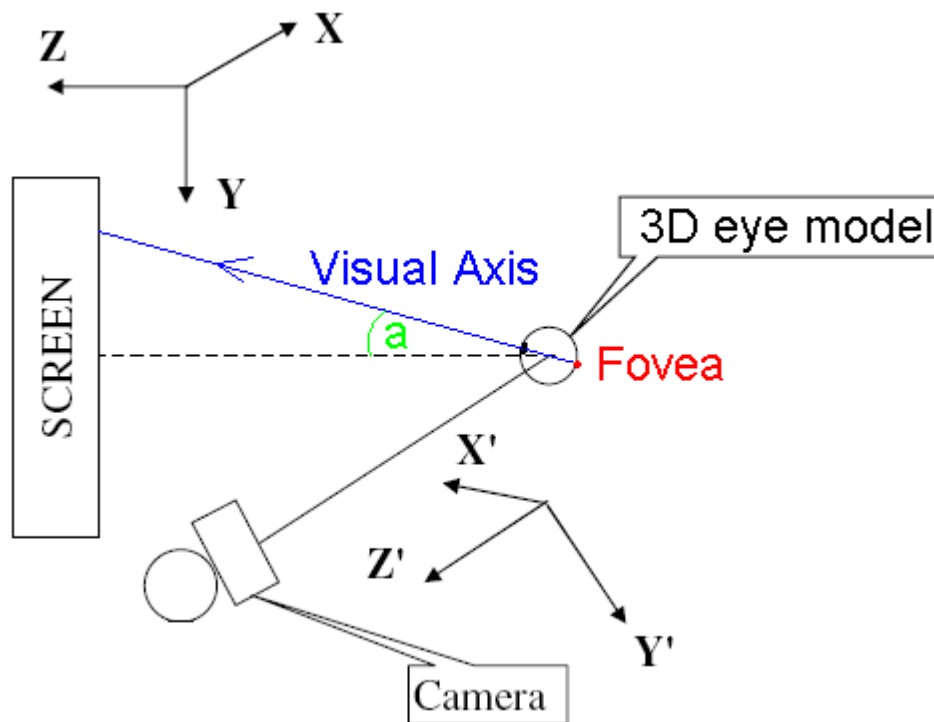


Figure 13. Illustrative example of the geometry based calibration to transform eye-position in the camera FOV to the screen plane. Adapted from (Daunys and Ramanauskas, 2006)

The alternative regression based methods generate a set of points corresponding points between the pupil position in the camera image and the gaze position in the 2D reference frame. These coordinate pairs are obtained by means of a calibration routine. This involves the user to look at each of a set of known points in turn as they are displayed on the screen, recording the pupil coordinates for each. Once calibration points are obtained, providing the points sample the workspace sufficiently, there are many different machine learning approaches that can be applied to successfully infer the mapping (Bishop 2006). Regression based approaches give significant flexibility because no pre-input information about the geometry of the set up is required as this is implicitly inferred from the calibration data. The drawback of this method however is the longer user involved calibration compared to the single calibration point required by geometric model based approaches.

3D Gaze Estimation: Based on knowledge of the Vergence system, if both eyes are monitored, it should also be possible to infer the gaze depth. There are very few papers that perform 3D gaze estimation in the real-world and report the 3D estimation performance. Hennessey and Lawrence claimed to be the “first reported binocular system for estimating the absolute X , Y , Z coordinates of where one is looking in the real 3-D world.” (Hennessey & Lawrence 2009a). Their methodology adopts the geometric model based approach discussed above, obtaining the visual axis of each eye separately. The vergence point can then be inferred by finding the intersection point (or closest point of approach) between the two visual axes. An important difference to note is that this method no longer only requires a single user calibration point as is claimed by its 2D gaze tracking equivalent. It seems that when the geometric model based approach is applied to 3D gaze tracking, a single calibration point is no longer sufficient and the accuracy drops significantly without at least 5 calibration points, and ideally 30 are used. The method therefore requires a means of obtaining 3D calibration points. A Plexiglass screen mounted on aluminium rails is used to provide these 3D calibration points.

In the following chapter a new low-cost binocular head mounted eye-tracking system is presented that outperforms current low-cost systems and has considerably more flexibility and potential for the desired application of prosthetic and wheelchair control. Fundamentally, the system can estimate 3D gaze position in the real-world as will be discussed in detail in the following chapter.

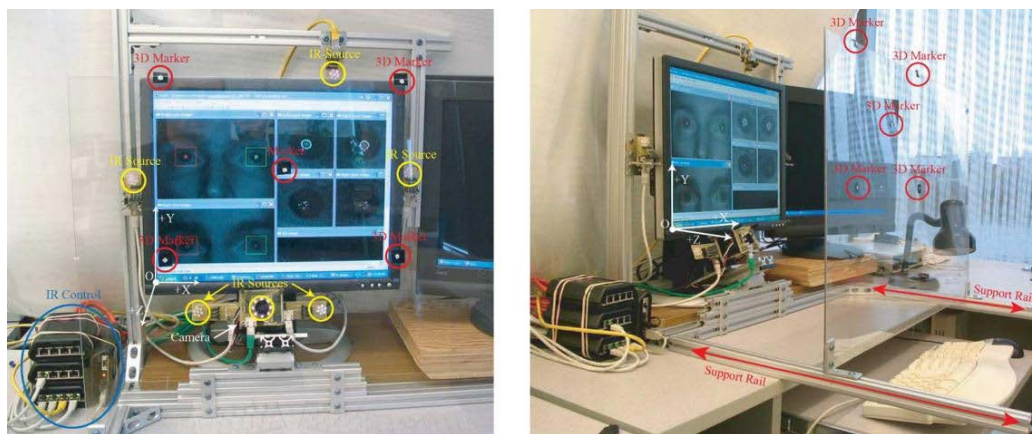


Figure 7: Photographs of the experimental set up used by Hennessey and Lawrence. The Perspex screen mounted on rails allows 3D calibration points to be obtained. Figure adapted from (Hennessey & Lawrence 2009a)

3 GT3D Ultra low-cost 3D gaze estimation

Acknowledgements. This Chapter is a modified version of a paper submitted during the first year of my PhD (Abbott and Faisal, 2012).

Abstract. The ability to move the eyes is often retained by patients with serious motor deficits. Furthermore eye-movements are the only directly observable behavioural signals that are naturally both highly correlated with actions at the task level, and proactive of our peripheral motor actions. Despite this, eye tracking is not widely used as control interface for movement impaired patients due to poor signal interpretation and lack of control flexibility. It is therefore proposed, that tracking the gaze position in 3D rather than 2D provides a considerably richer signal for human machine interfaces by allowing direct interaction with the environment rather than via computer displays. It is demonstrated here that by using mass-produced video-game hardware that an ultra-low cost binocular eye-tracker with comparable performance to commercial systems more than 800 times as expensive is possible. The developed head-mounted system has £30 material costs and operates at over 120 Hz sampling rate with a 0.5-1 degree of visual angle resolution. Both 2D and 3D gaze estimation is performed, controlling a real-time volumetric cursor essential for driving complex user interfaces. This approach yields an information throughput of 43bits/s, more than ten times that of invasive and semi-invasive BMIS that are vastly more costly.

3.1 Introduction

The advancement of Brain Machine Interface (BMI) technology for controlling neuromotor prosthetic devices holds the hope to restore vital degrees of independence to patients with neurological and motor disorders, improving their quality of life. Unfortunately, emerging rehabilitative methods come at considerable clinical and post-clinical operational cost, beyond the means of the majority of patients (Murphy et al., 2016). Here an ultra-low cost alternative is presented: using eye-tracking. Monitoring eye movement provides a feasible alternative to traditional BMIs because the ocular-motor system is effectively spared from degradation in a wide variety of potential users, including those with: 1) Muscular dystrophies and motor neuron disease (Kaminski et al., 1992, 2002), 2) Spinal traumas, because ocular innervation comes from the brain-stem, 3) Paralysis and stroke, when brain lesions occur in areas unrelated to eye movements, 4) Amputees, 5) Multiple Sclerosis and Parkinson's which affect eye movements later than the upper extremities, 6) the rapidly ageing population and associated progressive deterioration of the musculoskeletal system. The ability to control eye-movements can therefore be retained in cases of severe traumas or pathologies in which all other motor functions are lost. Based on disease statistics, within the EU alone, there were over 16 million people in 2005 (3.2% of the population) with disabilities who would benefit from such gaze based communication and control systems (Jordansen et al., 2005).

Humans observe the world through discrete, rapid, focussed eye movements (saccades) acting to align the high resolution central vision area (fovea) of both eyes with an object of interest (fixation point). Visual information is vital to motor planning and thus monitoring eye-movements gives significant insight into our motor intentions, providing a high frequency signal directly relevant for neuroprosthetic control. Eye tracking and gaze-based human-computer-interaction is a long established field, however cost, accuracy and inadequacies of current user interfaces limit the context of use, primarily, to clinical diagnostic and research settings. In addition to cost, even with emerging low cost systems (see Chapter 2) there are remaining issues surrounding the interpretation of eye movement data for effective gaze based interaction systems for everyday patient use. Fundamentally, as mentioned, gaze-based interaction requires more adequate solutions to the Midas touch problem. This is a major issue for existing gaze-based computer interaction, which focus on monocular eye tracking to drive a mouse pointer on the computer screen. The 'select or click' command is usually derived from either blink detection or gaze-dwell time, both of which also occur in natural behaviour and thus require an extended integration time (typically in the order of seconds) to initiate a reliable click.

To address these shortcomings an ultra-low cost binocular eye tracking system is developed here, with similar accuracy to commercial systems and a frame rate of 120Hz. The Midas touch problem is addressed by distinguishing non-behavioural eye winks from behavioural eye blinks, significantly speeding up selection time. This system is subsequently used to control robotic arms and wheelchairs towards restoring independence to severely disabled patients. The major challenge here is to derive a practical control signal from eye movements that meets the interface requirements. This must be achieved without being intrusive to the natural sensory function of the eyes. The aim with the developed system is to allow the user to interact with their surroundings

directly rather than limiting their interactions to via their computer screen. Towards this, the developed "GT3D" (Gaze Track 3D) system derives a signal that provides an information rich signal for inferring user intentions in natural contexts: 3D gaze position.

We interact with a three dimensional world, navigating and manipulating our surroundings. Severe disabilities remove this ability, vital for independence. Gaze based interaction for computer control works towards restoring this by facilitating interaction with the world via a screen based graphical user interface (GUI). Instead here direct 3D gaze interaction is developed towards enabling direct end-point control of motor-prosthetics. With knowledge of both eye positions, gaze-depth information can be obtained because the eye vergence system forces both eyes to fixate on the same object, allowing image fusion and depth perception. The intention-relevant, high-information throughput 3D gaze signal can be applied to tasks such as wheelchair navigation, environmental control, and even the control of a prosthetic arm. Despite the huge potential, 3D gaze estimation has received less attention than the 2D alternative for (mouse) cursor control. A major challenge of gaze estimation, particularly in 3D, is the calibration and adaptation of the estimation system for individual users.

As discussed this was achieved by Hennessey and Lawrence (2009) using a complex lab based system requiring precision measurement, apparatus for generating 3D calibration points (movable Plexiglas screen) and complete stability of the system to function. This precision set-up, required for geometric methods, is not feasible with low cost hardware. Thus an alternative portable, ultra-low cost hardware solution is developed here (GT3D) with a suite of algorithms for the realisation of a system that can estimate the absolute gaze target in X, Y and Z coordinates with an accuracy that rivals present methods. This is achieved without complex configuration routines, the need for 3D display equipment or user-specific details about eye geometry. Gaze interaction directly with the 3D environment has received less attention, though Hennessey and Lawrence (2009) developed the first binocular gaze tracking system for estimating the absolute X, Y, Z coordinates of gaze targets in the real 3-D world (Hennessey and Lawrence, 2009a). To obtain the gaze vectors requires precise positioning of the cameras with full geometric parameterisation of the hardware setup; optical properties and a model of the eye, including the refractive index of the fluid inside the eyeball (vitreous fluid), lens and cornea. Based on the vergence system, a 3D gaze fixation point is then calculated from the gaze vectors' nearest point of approach. This system has only been demonstrated in controlled research environments, possibly because of the strict geometric requirements and detailed modelling of the physical system.

3.2 Methods

The presented system is composed of ultra-low cost imaging hardware and stand-alone software that implements the algorithms developed including methods for 3D and 2D gaze tracking.

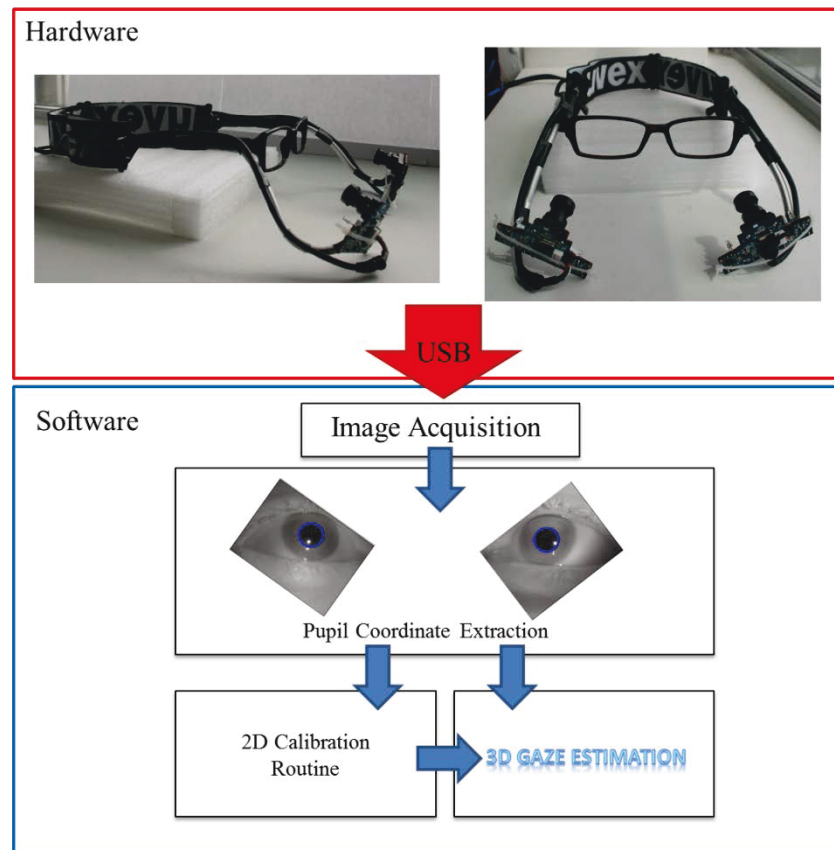


Figure 14 System overview. Hardware: Ultra low-cost head mounted binocular eye tracker built using off the shelf components including two PlayStation 3 Eye cameras (£10 each), two IR LEDs, cheap reading glasses frames and elastic headband support. The cameras are mounted on lightweight aluminium tubing. The hardware total cost is £30. **Software:** The camera frames are streamed at 120Hz via USB to a standard lap-top computer and the pupil positions are extracted using image processing (see Figure 16). A 2D user calibration allows a mapping between pupil and 2D gaze position to be learnt. Using the 2D estimates from both eyes, a 3D gaze estimation can be made by estimating the vergence point.

3.2.1 Ultra-Low Cost Binocular Eye-tracking Hardware

The video based binocular eye tracker shown in Figure 14 uses two ultra-low cost video game console cameras (PlayStation 3 Eye Camera – £10 per unit), capable of 120Hz frame-rate, at a resolution of 320x240 pixels. This is the main cost-reducing step in the system, as typical machine vision cameras operating at this performance are more expensive by two orders of magnitude. To optimise imaging conditions, the camera optics are modified for infrared (IR) imaging at no material cost by removing the IR filter and replacing it with a piece of exposed and developed film negative which acts as a low-cost IR-pass filter. The eyes are illuminated using 2 IR LEDs aligned off axis to the camera, creating a dark pupil effect to enhance the contrast between the pupil and the iris. Chronic IR exposure above a certain threshold leads to retinal damage or the formation of cataracts (Sloney and Freasier, 1973). This threshold has been reported as being between 10 and 20 mW/cm² (Sloney and Freasier, 1973; Sloney et al., 2005). The LEDs used are Optek Gallium arsenide OP165D which produce an irradiance of between 0.28 mWcm⁻² and 1.6 mWcm⁻² (depending on the forward voltage) measured at a distance of 1.5cm. This is well below the safety parameter, especially as it will be mounted at 10cm from the eye. These LEDs

are powered using a USB cable giving a 5 volts supply with up to 500mA of current to be drawn. The driver circuit provides 20mA current to each LED with a forward voltage of 1.6 volts applied. The cameras are head mounted to maximises the eye image resolution and allow unrestrained head movement following calibration. The camera-mounting headset shown in Figure 14 has been designed with off the shelf components costing £10 in total. The system weighs in total 135g and the cameras and their mounting arms exert a moment of approximately 0.1 Nm on the nose. It has been designed to allow four degrees of freedom for adjustment to different users (shown in Figure 15). The images from the cameras are streamed via two USB 2.0 interfaces to a standard lap top computer facilitating an accessible and portable system.

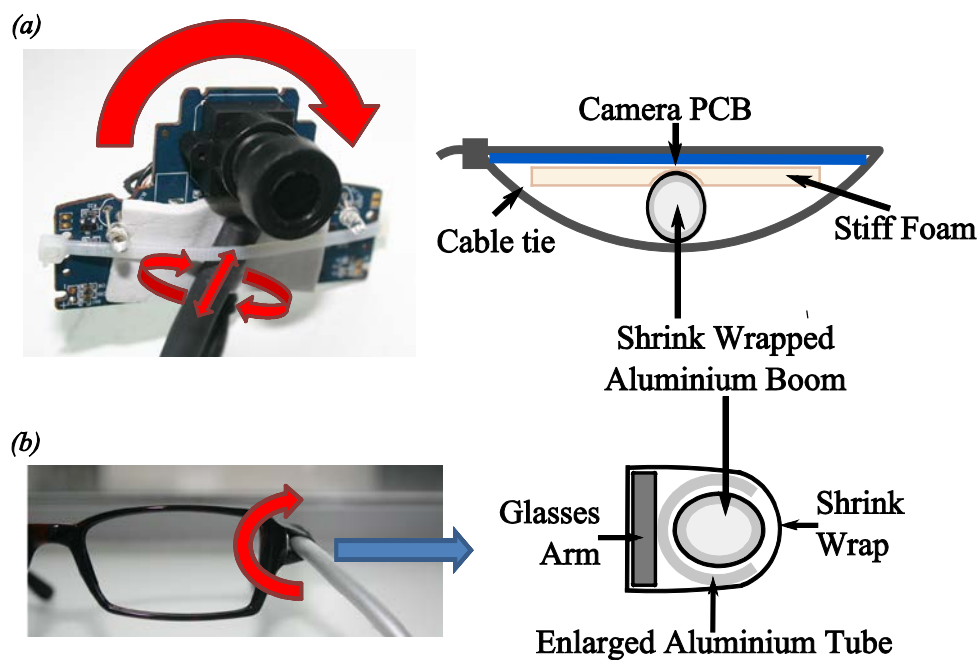


Figure 15. Headset adjustability. Headset design allows the camera position to be adjusted with four degrees of freedom - (a) rotation and translation of the camera on the boom arm and (b) rotation of the boom arm itself. This allows adjustment to customise the system to different users.

3.2.2 Eye-Tracking

The eye-tracking methodology applies standard image-processing methods to locate the pupil centre in each video frame; an overview of this process can be seen in Figure 16. The IR imaging system increases the contrast between the pupil and iris. This allows simple intensity threshold image segmentation, converting the grey-scale image to a binary image (Figure 16(b)). In this single step the data volume per frame is reduced from 230 kilobytes to 9.6 kilobytes; retaining sufficient information to locate the pupil but reducing the subsequent computational load. Due to noise effects and other dark regions in the image, such as shadows and eyelashes, a pupil classification step is made. To reduce the complexity of the classification process, morphological operations of erosion and dilation were applied in a sequence: first “opening” the image, removing the dropout noise; and then “closing” it to fill in any holes in the pupil blob (Figure 16(c)). Connected component labelling is then applied to assign a unique label to the pixels of each candidate pupil region. Subsequently, a shape based filter is applied (Figure 16(d)) to classify the pupil

based on maximum and minimum object size and elongation (axis ratio). The pupil centre is then extracted using least squares regression to fit an ellipse to the classified pupil object contour – this ellipse is shown overlaid on the raw image in Figure 16(e). The x and y coordinate of the ellipse centre (in pixels) are extracted for each eye ellipse as the pupil positions. Further image processing improvements were made to improve robustness to large changes in lighting levels, to avoid the interference of specular reflections and other image artefacts (see appendix 1).

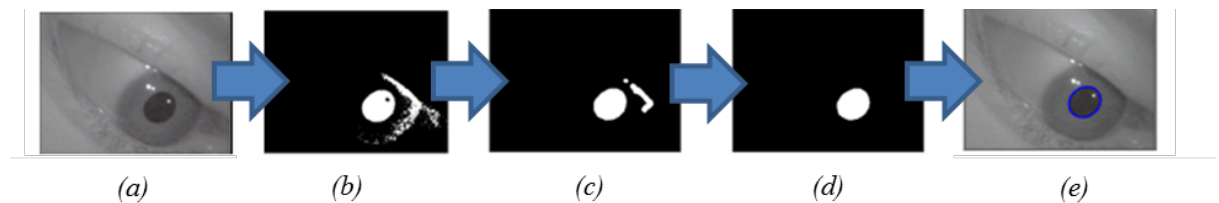


Figure 16. Pupil extraction image processing pipe-line. Images intermediates include: (a) raw greyscale (b) binary (c) noise filtered (d) shape filtered (e) original with extracted ellipse overlaid.

3.2.3 Calibration for 3D gaze estimation.

The pupil positions extracted from the eye images must be related to the gaze position. A purely explicit method requires a rigid system set-up difficult to obtain using low cost hardware, while a purely implicit method requires more involved 3D calibration points. Gaze estimation is achieved in the real 3D environment by combining an implicit step to infer the system parameters with an explicit geometric step to transfer this to a 3D gaze estimate. This involves the calibration of each pupil position to the respective gaze positions on a computer visual display unit (2D calibration). From this the 3D gaze vector of each eye can be found (step 1) from which the 3D fixation point is then calculated (step 2).

Step 1: Calculating 3D gaze vectors using 2D calibration. Calibration to the 2D computer monitor can be made explicitly using the geometry of the system (Hennessey and Lawrence, 2009b; Lee and Park, 2008; Sheng-Wen Shih and Jin Liu, 2004; Sheng-Wen Shih et al., 2000) or implicitly using a calibration routine to infer a mapping between pupil position (in the eye image) and gaze position (on the computer screen) (Brolly and Mulligan, 2004; Morimoto and Mimica, 2005). The implicit mapping provides a more suitable solution because explicit methods require precise geometric knowledge of camera positions, infeasible with the low cost adjustable headset. To learn an implicit mapping, training data is acquired using a calibration routine which displays each point of a 5x5 calibration grid that spans the computer screen. At each calibration point, the pupil location of each eye is extracted from a ten-frame burst and the user's average eye positions are recorded. This reduces noise effects of drift and micro-saccades. The calibration data points to train a linear combination of non-linear basis functions, applying the pseudo inverse (Moore-Penrose pseudo-inverse (Moore, 1920)) to learn mapping weights. Second order polynomial basis functions were found to achieve an optimum trade-off between model complexity and the number of calibration points required to generalise well. Following the 2D calibration routine, the 2nd order polynomial mapping is used to map the position of each eye to the gaze position in the 2D plane of the computer monitor, at a frame rate of 120Hz. When the user fixates in the monitor plane, the gaze

estimates of each eye are approximately superimposed as shown in Figure 17(a). When the user fixates outside of the monitor plane, the 2D gaze estimates diverge as shown in Figure 17(c). This divergence gives depth information as the 2D gaze estimates are effectively the intersection between the gaze vectors and the computer monitor plane (see Figure 17(c)). The gaze vectors are calculated from the 2D gaze estimates x_L, y_L and x_R, y_R (relative to the top left corner of the screen) using equations (2.1) and (2.2). This requires the relative positions of the eyes and monitor to be fixed during calibration and measurements of screen height (S_{height}), eye level (E_{height}), eye to screen ($D_{\text{eye-screen}}$) and inter eye distance (E_{inter}) to be made (see Figure 17(a) and (b)). The eye tracker is head mounted and the system is calibrated with a head-centric coordinate system thus following calibration the user will be free to move their head and the gaze vectors will be relative to the origin which lies between the eyes.

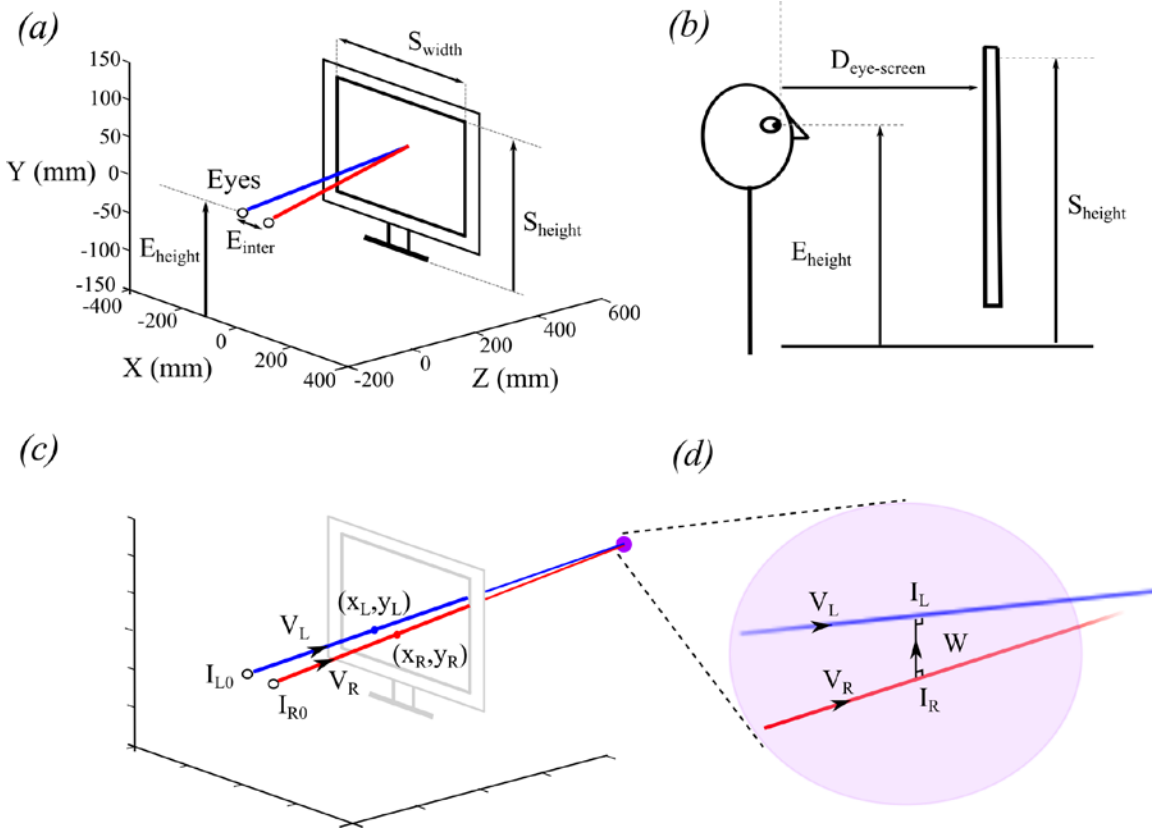


Figure 17. Illustration of the 3D gaze estimation method. (a) 2D Calibration step to relate the pupil positions to their gaze positions in the screen plane. The user is aligned with the horizontal screen centre and must remain stationary during calibration. The measurements shown are required for calibration. (b) Side view of user and computer screen. (c) The left and right gaze estimates on screen are represented by the two dots (x_L, y_L) and (x_R, y_R) and yield the gaze vectors shown (V_L and V_R). (d) The nearest point of approach one each gaze vector is found.

Step 2: Using the 3D gaze vectors to estimate the 3D gaze position. The 3D gaze vectors are used to estimate the 3D gaze position. The 3D gaze position is the vergence point of the 2 gaze vectors. Exact 3D vector intersection is unlikely thus the nearest point of approach on each vector is found. These points are represented by I_L and I_R in Figure 17(d) and are given by the parametric equations (2.3) and (2.4). The positions of the eyes relative to the origin are represented by I_{L0} and I_{R0} as shown in Figure 17(c) while S_L and S_R represent scalars to be found.

$$\vec{V}_L = \begin{bmatrix} x_L - \frac{S_{width}}{2} + \frac{E_{inter}}{2} \\ S_{height} - E_{height} - y_L \\ D_{eye-screen} \end{bmatrix} \quad (2.1)$$

$$\vec{V}_R = \begin{bmatrix} x_R - \frac{S_{width}}{2} - \frac{E_{inter}}{2} \\ S_{height} - E_{height} - y_R \\ D_{eye-screen} \end{bmatrix} \quad (2.2)$$

$$I_L(S_L) = I_{L0} + S_L \vec{V}_L \quad (2.3)$$

$$I_R(S_R) = I_{R0} + S_R \vec{V}_R \quad (2.4)$$

By definition, the nearest points of approach will be connected by a vector that is uniquely perpendicular to both gaze vectors - ($\vec{W}$) shown in Figure 17(d) and equation (2.5). To satisfy this condition the simultaneous equations (2.5), (2.6) and (2.7) must hold. Substituting equations (2.3) and (2.4) into the simultaneous equations and solving for the scalars S_L and S_R , yields the nearest points of approach - I_L and I_R . The 3D gaze estimation (G_{3D}) is then taken as the mid-point of these 2 positions as shown in equation (2.8). This algorithm is performed for each frame in the video streams to obtain a 3D gaze estimate at 120Hz sampling frequency that can be used as a volumetric cursor in the development of advanced user interfaces for neuromotor prosthetics.

$$\vec{W} = I_L - I_R \quad (2.5)$$

$$\vec{W} \cdot \vec{V}_L = 0 \quad (2.6)$$

$$\vec{W} \cdot \vec{V}_R = 0 \quad (2.7)$$

$$G_{3D} = \frac{I_L(S_L) + I_R(S_R)}{2} \quad (2.8)$$

3.2.4 Solution to the Midas touch problem

To address the Midas touch problem the system uses non-behavioural winks to confirm gaze commands. Winks can be distinguished from behavioural blinks by virtue of the binocular eye-tracking feed allowing for much shorter command integration times. In the filtered binary eye image (see Figure 16(d)), when the eye is closed, no pupil object is located by the eye tracker, raising a closed eye flag. Left and right eye winks can be distinguished from simultaneous eye blinks using temporal logic. For example a left wink is defined by the left eye being closed and the right eye being open simultaneously for more than 20 frames. The high frame rate allows for this distinction to be made reliably with low integration times of ~170 milliseconds.

3.3 Results

The three main contributions of this Chapter can be grouped into the following sections: 1. High performance ultra-low-cost binocular eye-tracking system – *Eye-tracking Accuracy and precision*, 2. Solution to the Midas touch problem and continuous control: *Human computer interaction and a BMI benchmark for closed-loop control of devices and* 3. Gaze estimation in the 3D environment: *Accuracy and precision in 3D tasks*.

3.3.1 Eye-tracking accuracy and precision.

To precisely estimate the eye tracking system's accuracy, a subject was calibrated and shown random test points of known 3D locations in three separate trials. Each trial involved a calibration routine that cycled through a 5x5 calibration grid displayed on a computer monitor 50cm from the user's eyes, 15 randomly generated test points then followed this. The results for one trial are shown in Figure 18(a). For each trial a new set of random test points was generated. Each test point appeared in turn and the user looked at the point and hit the space bar, at which point the gaze position was recorded from the real-time data stream. Over all trials a mean Euclidean error of 0.51 ± 0.41 cm (standard deviation) was achieved at a distance of 50 cm which translates into an angular error of 0.58 ± 0.47 degrees.

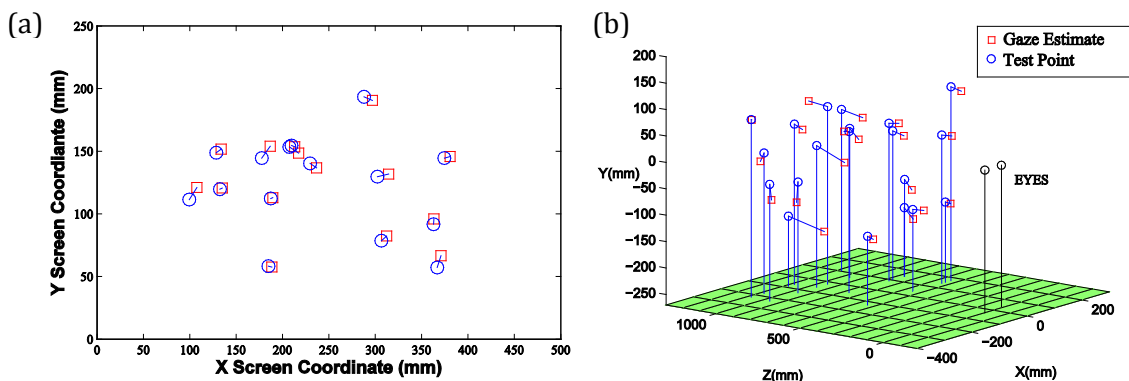


Figure 18 2D and 3D Gaze estimation test points and gaze estimates. (a) 2D gaze estimation: The blue circles represent the 15 randomly generated test positions displayed on the screen and the red squares are the gaze estimate. **(b) Three dimensional plot of gaze estimation results for acalibration test run with test points being displayed at 4 depths – (54cm, 77cm, 96cm and 108cm) following 2D calibration at 54cm.** The blue circles represent the real displayed positions and the red squares represent the gaze estimations.

Table 1 compares the developed system with a commercially available binocular eye tracking system. The developed system's gaze angle accuracy was $0.58 \text{ deg} \pm 0.47$ (mean $\pm$ SD) and is defined as the eye-tracking signal accuracy, namely how precisely the viewing direction of the eye can be determined. This measure is viewing distance and application-independent but has direct implications for both 2D (monocular) and 3D (binocular) gaze target position estimation accuracy. An average of 0.58° is achieved while the reference system's manufacturer (EyeLink II) specifies a typical average accuracy as $< 0.5^\circ$ – but does not provide more specific data or measurement approach. The developed system can perform 2D or 3D gaze estimation, while the EyeLink II, though also a binocular eye-tracker, has software to perform 2D estimation only. The developed system is less than

1/3rd of the mass at 135g compared to the 420g commercial system, and less than 1/800th of the cost with a unit cost of just £30 compared to the £25,000 commercial system cost. Though the tracking range is less for the GT3D system – 6 degree smaller horizontally and 16 degree smaller vertically, this is an image-processing problem that will be solved. The frame rate is also slightly lower at 120Hz compared to the 500Hz of the commercial system, but it is sufficiently high to resolve saccades and gives a frame rate four times that of other low-cost systems.

Table 1 Comparison between the developed system (referred to here as GT3D) and the commercial EyeLink II. Here comparisons are made using the metrics in the EyeLink II technical specifications. More detailed analysis of the 3D performance is shown in table 3.

Metric	GT3D	EyeLink II ^a
Gaze angle accuracy	0.58±0.47°	<0.5 ° ^b
Gaze estimation modes	2D, 3D	2D
Horizontal Range	34 °	40°
Vertical	20°	36°
Headset Mass	135g	420g
Frame Rate	120 Hz	500 Hz
Cost	£30 ^c	£25,000

^aInformation is taken from the SR Research issued Technical specification.

^bFor the EyeLink II The accuracy is expressed as a “Typical Average” in Corneal Reflection mode. No error measurement is given. The mean gaze angle accuracy and standard deviation for the GT3D system averaged over 3 separate trials is reported.

^cPurely material cost. Does not include other overheads of a commercial product

3.3.2 Human computer interaction

With the system in 2D mode, the user can operate a computer, performing such tasks as opening and browsing the web and even playing real-time games. The system does not require a bespoke graphical user interface, operating in a windows environment with an icon size of 3cm². The use of wink commands allows the integration time to be reduced to ~170 milliseconds of wink to make a selection.

3.3.3 Accuracy and precision in 3D tasks

To assess the 3D gaze estimation, the methodology was similar to the 2D experiment but test points were also displayed at different depths. Following the 5x5 2D calibration routine at a depth of 54cm, 5 random test points were generated at 4 depths: 54cm, 77cm, 96cm and 108cm by moving the computer monitor. Over this workspace, the system performed with a mean Euclidian error of 5.8 cm, with a standard deviation of 4.7 cm. The results for this experiment are displayed on a 3D plot in Figure 18(b). The mean absolute error and standard deviation for each dimension is shown in Table 2. The mean depth error (Z in Table 2 and Figure 18(b)) is 5.1cm with a standard deviation of 4.7cm, this

accuracy and precision is 4 times larger than the horizontal and vertical equivalent (X and Y in Table 2 and Figure 18) which explain the considerably higher Euclidean error in 3D compared to 2D gaze estimation. The gaze angle fluctuates around a value of $0.8 \pm 0.2^\circ$ (mean $\pm$ SD) but does not consistently increase with depth. The mean depth error for estimations at each depth increased from 4.6cm at 54cm distance from the face to 6cm at 108cm distance. This is to be expected as a consistent gaze angle error will cause a larger spatial error at deeper depths, particularly in the depth direction.

Table 2 3D gaze estimation performance for the results shown in Figure 18(b) .

	Mean Absolute Error (cm)	Standard Deviation (cm)
x	1.1	0.7
y	1.2	1.1
z	5.1	5.0
Euclidean	5.8	4.7

3.4 Discussion

An ultra-low cost, high-speed binocular eye tracking system capable of 2 and 3D gaze estimation, costing 1/800th of a reference commercial system, yet achieving a comparable eye tracking performance. The description of the system as low cost may seem misleading, as it is purely material cost and does not include all the other overheads of a commercial product (e.g. manufacture, technical support, sales etc). However, this statement is used to emphasise the achievement of the developed algorithms to achieve such a high performance with such low-cost hardware. It is the first binocular eye-tracker built with low-cost off the shelf components able to estimate 3D gaze position in the real world.

The system also drives a mouse-replacement based user-interface for which an improved solution to the “Midas touch problem” is implemented. Tracking both eyes allows for “wink” rather than “blink” detection, decreasing the required selection integration times by a factor of 6. This is because when blink or dwell time is used to make a selection, a blink or dwell must be emphasised with an extended period of time to distinguish commands from normal behavioural blinks and fixations. The system interfaces with the computer operating system via USB, and allows the user to browse the web, type on a visual keyboard and play real-time games, which will be demonstrated in the following Chapter.

In the 3D domain, gaze estimation is directly applicable to motor prosthetics, with the potential to allow patients to interact with their surroundings. The developed system can estimate the absolute real-world 3D gaze position in real time with a performance competitive with research systems. Table 4 shows the performance comparison of the developed system with Hennessey and Lawrence’s system (Hennessey and Lawrence, 2009a). As the table shows, the mean Euclidean error of the developed system is almost 2cm higher. Hennessey and Lawrence calibrate their system using calibration points taken

at both the nearest depth (17.5cm) and the farthest (42.5cm). While for the GT3D system, calibration is made at a single depth of 54 cm and the work space extends out to 108 cm depth from the eyes. The workspace used by Hennessey and Lawrence covers half the depth (25cm compared to 54cm) and though Hennessey and Lawrence do not give the workspace depth from the face explicitly (their coordinate system has its origin at a corner of the computer monitor) it is estimated that the maximum workspace depth is 42.5cm from the eyes (see Table 3 footnote) compared to the 108 cm depth used here. At larger depths the estimation accuracy degrades, as such a more balanced comparison is made by normalising the error and standard deviation by the workspace depth as can be seen in the second column of Table 3. With this metric the GT3D system performs with almost half the normalised Euclidian error of their system. For both methods the error has a large standard deviation, with a magnitude similar to the mean. This variability is partly due to noise in the image sensors and head-set slippage, but may also be due to micro-saccades and drift movements of the eyes. A larger user group study is presented in Chapter 4 with an extension of the GT3D calibration system for robotic arm control.

Table 3 3D gaze estimation performance (referred to here as GT3D) comparison with Hennessey and Lawrence's system(Hennessey and Lawrence, 2009a). Mean Euclidian errors and standard deviation in cm and as a percentage of the workspace depth^a.

	Mean (cm)	Euclidean Error	Mean Error (%) ^a
GT3D	5.8±4.7		5.3±4.4
Hennessey and Lawrence	3.9±2.8		9.3±6.7

^aMeasurements normalised by dividing by the workspace depth over which the methods were tested. GT3D- 108cm from the user. Hennessey and Lawrence - the workspace is described relative to the corner of their computer screen rather than the user's eyes. To allow comparison between the GT3D performance and their data, it is assumed that the subject's head was 60cm away from the screen and given that the depth closest to the screen is 17.5cm away yields a workspace depth of 42.5 cm =60cm-17.5cm.

Though the performance is similar, the system presented here requires only a standard computer monitor as opposed to 3D equipment, no information on eye geometry, optics or precise camera positioning, and following calibration, users have complete freedom to move their heads. The resulting output is a volumetric cursor which can be used for advanced interfaces to allow direct 3D interaction with the world rather than via a computer screen. For use in daily life however there are still a number of issues to be addressed. Firstly, the long term physical stability of a wearable eye-tracker. Over time fine movements of an eye-tracking headset will degrade the calibration accuracy significantly. A potential solution to this is to make regular updates to the calibration without impacting on the user (a novel calibration method proposed in chapter 5 may facilitate this). Secondly, it is important to note that the 3D gaze accuracy, particularly in the depth dimension becomes significantly worse as a function of distance from the user. This means effectively that beyond 1-2 metres, the depth estimation error becomes too high for the measurement to be useful.

It is important to discuss where this limitation comes from and how much it is a result of the measurement system versus the physiology of the eye and oculomotor system. The accuracy achievable by a gaze tracking system will intrinsically be affected by how accurately we fixate on a point of interest. Thus the accuracy of the visual system gives the upper limit for what is achievable. The 3D gaze accuracy will clearly be dependent on the position in space, particularly depth. A model was created to predict this accuracy. The model is based on the assumption that, at best, the fixation point falls somewhere within the fovea (Duchowski, 2007). The fovea makes up approximately 1 degree of viewing angle and thus the fixation accuracy is ± 0.5 degrees (Duchowski 2007). The minimum angle difference required to reliably distinguish between 2 points of fixation is therefore 1 degree. To generate the model, this value was used as the angular increment to make sweeps of the respective visual axes for each eye over the range of the horizontal binocular field of view (average: 114 degrees (Gillam 1996)). The point of intersection for the visual axes was found for each increment and plotted (Figure 19). The separation between these points thus represents the potential error of the vergence system at the particular position. Though a naive model restricted to a single plane, it demonstrates the pattern of accuracy dropping off significantly with distance from the eyes. The depth accuracy will clearly be much poorer than the horizontal and vertical accuracy obtainable. Also important to note is the reduced accuracy between the eyes where the intersection point rapidly tends to infinity (resulting in the gaps in the blue dots directly ahead of the eyes in Figure 19).

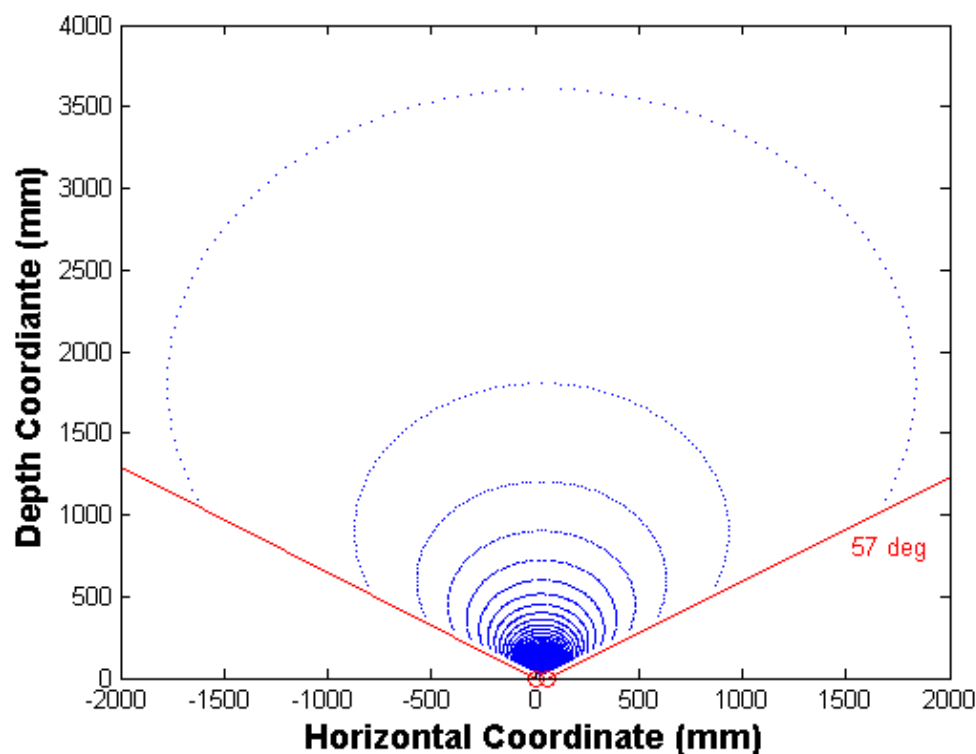


Figure 19. Vergence model to demonstrate upper limit for 3D gaze performance. The red circles represent the eyes, the red lines indicate the binocular range and the blue dots represents the distinguishable fixation positions. The distinguishable fixation positions are based on the intersection points of distinguishable gaze positions based on the assumption that the vergence system, at best, aligns the visual axis within 0.5 degrees of a given target (Duchowski, 2007). Thus for each combination of distinguishable gaze angles, the intersection point is plotted as a blue dot.

Based on this model, depth information is relevant for prosthetic control as the accuracy is very fine within the arm-span. However for wheelchair control, beyond 1 metre the 3D direction vector, rather than absolute depth position should be used. However, the ability to distinguish between when the user is looking into the distance verses nearby can give significant information about the mode of operation i.e. covering long distances verses fine manoeuvres and as the user closes in on a target the 3D accuracy will increase and thus the heading trajectory can be refined online.

One final consideration in the context of prosthetic control is the required system improvement to compensate for head movement. The head-mounted GT3D system calculates the 3D gaze position relative to the head. Thus when controlling a prosthetic, the head direction must be known relative to the torso direction to be able to give a meaningful 3D location to move the robotic arm towards. A solution for this is developed in Chapter 6, using a body sensor network to track the relative positions of the body parts along with the software to integrate this information with gaze information from the eye-tracker. Thus, with the relevant design considerations to the limitations of the current approach, a powerful interface could be developed.

The system is proposed as an alternative and complement to direct brain read out by BMIs. A simple performance metric to compare different BMIs is the information throughput – the rate at which the BMI communication interface can decode information from the brain. The theoretical information throughput achievable with the developed system is calculated and compared to information throughputs presented in an extensive review of BMIs (Tonet et al., 2008). The throughput is calculated as the product of bits communicated per unit command, and the number of unit commands that can be made per second. In the context of the GT3D interface, each fixation can be considered as a unit command. With a sensory estimation error of 1.1cm in width, 1.2 cm in height and 5.1cm in depth (mean absolute error), over a workspace of 47cmx27cmx108cm (width x height x depth), there are 2.04×10^4 distinguishable states giving 14.3 bits of information per fixation. The average human fixation rate is around 3 fixations per second (Land and Tatler, 2009); giving a bit rate of 43 bits/second. This theoretical upper limit is significantly higher than other BMI mechanisms and the signal is obtained non-invasively, for a significantly lower cost. The information throughput reflects the accuracy of the developed gaze-controlled real-time continuous volumetric cursor, which yields a fast control signal with very low latency. Both the speed of information transmission, but also the natural role of gaze in attention and actions the system highly suitable for controlling disability aids such as electric wheelchairs or end-points of prosthetic arms. Figure 20 demonstrates the relationship between estimated treatment cost and bit-rates for different BMI mechanisms, including the developed system, labelled as GT3D in the figure. The treatment costs are estimated based on device cost as well as operational set up and maintenance costs such as surgery and rehabilitation costs (see also Figure 20). The information rates given in (Tonet et al., 2008) may underestimate throughput capacities of the different BMI methods, but at least offer a basic consistent benchmark across different readout technologies. The developed system has an estimated information transfer capacity of 43 bits/second, which is 10 times higher than other invasive BMI approaches (see Figure 20).

BMI information rates from direct recording of neuronal activity are ultimately constrained by noise in the recording systems and the nervous system, itself (Faisal et al., 2008). In particular physical noise sources inside central neurons (Faisal, 2009; Faisal et al., 2002) and peripheral axons (Faisal, 2009; Faisal et al., 2005) will limit decoding performance from limited numbers of independent neuronal sources. Thus, to compensate for noise signal decoders have to observe signals for longer periods of time, thereby increasing response latencies for direct BMIs at the moment. While these issues will be ameliorated by the steady progress of sensor quality and density (Stevenson and Kording, 2011), eye movements already offer a highly accurate, low-latency (and low cost) read out. This is because the brain has already evolved to minimise the role of noise and delays in eye movements, which form an aggregated output of the nervous system. The leap in readout performance (in terms of readout performance and latency) enables closed-loop real-time control of rehabilitative and domotic devices beyond what is achievable by current BMIs: it was estimated that powered wheelchair control requires, on average, 15.3 bits/second and full-finger hand prosthetics require 54.2 bits/second (Tonet et al., 2008). The developed system demonstrated a clear improvement on low-level measures of BMI performance, but such technical measures mask the complexities of learning to use and operating BMIs in the clinic and daily-life. Therefore, a real-world, closed-loop control benchmark is proposed - playing an arcade video game - as a high-level, behaviour-based measure for BMI performance. In the following Chapter this benchmark is applied in a large field study (N=2042) to investigate whether members of the public can calibrate and play the classic video-game "Pong" outside of the stability of lab settings.

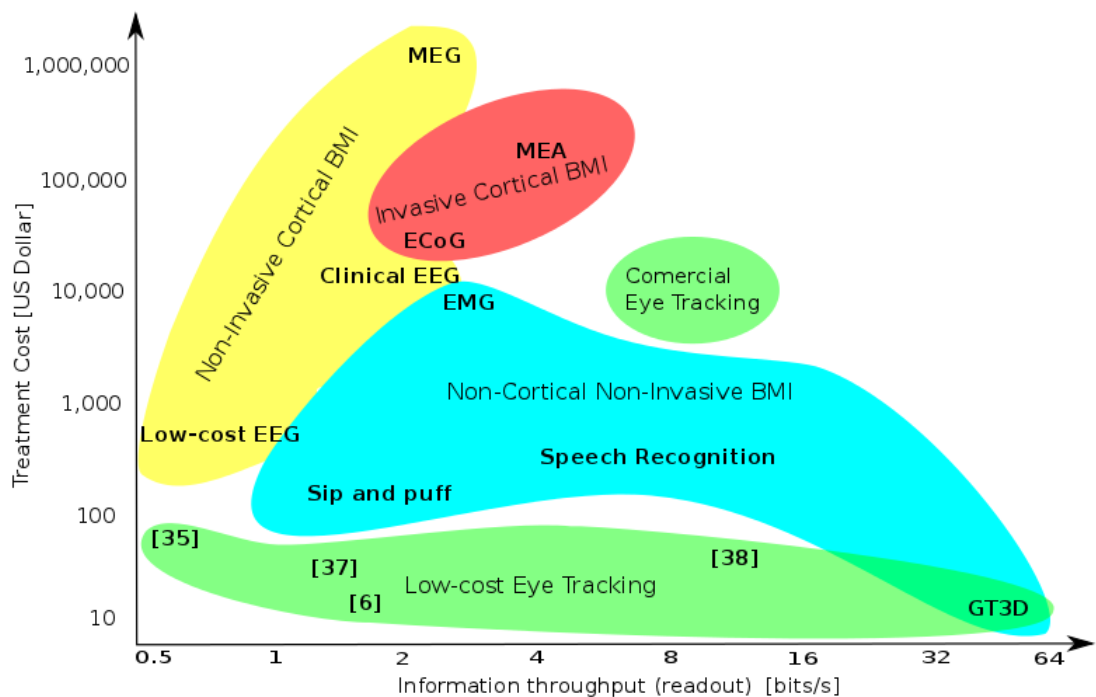


Figure 20 Comparison of different BMI and eye tracking technologies in terms of their treatment and hardware costs (in US Dollars) and readout performance (measured as bits/s). Note, a log10 scale is used for the treatment cost and binary logarithm scale for the bit rate. The bit-rate data invasive and non-invasive BMIs were taken from (Tonet et al., 2008), except where stated otherwise. Treatment costs were taken from published data were available and cited below, or from manufacturer quotes and healthcare providers. GT3D – the developed system (component cost). EMG – Electromyography (cost based on g.tec EMG kit; Guger Technologies, Schiedlberg, Austria). ‘Sip and puff’ – switches actuated by user inhaling or exhaling (system cost from liberator.co.uk). Speech Recognition – Speech actuated commands (cost based on commercial speech recognition system Dragon’s “Naturally Speaking Software”. MEG – Magnetoencephalography ((Ray and Bowyer, 2010)). EEG – Electroencephalography; Clinical EEG (cost based on g.BCI EEG kit, Guger Technologies GmbH, Schiedlberg, Austria), Low-cost EEG (Emotive EEG headset kit, Emotiv, San Francisco, CA), bit rate from (Bobrov et al., 2011). ECoG – Electrocorticography, MEA – Multi-electrode array, cost of clinical research systems is based on Utah electrode-arrays (Blackrock Systems, Salt Lake City, UT) and peripheral equipment plus the preoperative assessment, surgery, postoperative management cost estimated from deep brain stimulation costs (McIntosh et al., 2003). Commercial Eye Tracking costs for 2D gaze tracking (Eyelink II, SR Research, Kanata, Ontario) with bit rate reported in (Tonet et al., 2008). Low-cost Eye Tracking – citations for individual prototype systems and their reported bit rates ((San Agustin et al., 2009); (Lemahieu and Wyns, 2011) bit rate based on 40 characters per second text writing performance times 1 bit entropy per character of English language (Shannon, 1993) yielding 0.67 bits/second; (Rozado et al., 2012a) The system recognizes ten different gaze gestures with an average of 2.5s per gesture, yielding 1.3 bits/second; the system recognises 16 different gaze states at 3 states per second (average number of fixations per second) yielding 12 bits/second.

4 Large field study of gaze based closed loop control

Acknowledgements. Preliminary work associated with this Chapter was published in 2013 (Abbott et al., 2013). Alan Zucconi developed a new version of the classic arcade game Pong (Atari Inc, 1972, USA) which allowed the game states to be logged. The arcade booths were designed and built in collaboration with Haberdashery (creative design company, London, UK) as part of the MRC centenary exhibition "Strictly Science" 2013.

Abstract. Eye movements are currently underutilised in prosthetic, wheelchair and computer control despite their resilience to peripheral motor impairment and the rich information they provide. The brain machine interface community does not widely use this high throughput, non-invasive low cost complement to direct Brain Machine Interfaces. Here, in a large field study (N=2042, 4-87 years, naive to gaze input), 80% of players could calibrate and use the developed gaze based interface to play the real-time video game "Pong" (computer tennis), requiring continuous control and achieving on average bit rates of 300 bits/min. In addition, this performance is achieved with just 30 s set-up (calibration) time and 5s training. This is put in perspective when compared to direct BMIs requiring at least 30 minutes for discrete control and reaching months for continuous control. This game is proposed as a BMI benchmark because, like prosthetics, it requires real-time closed loop control. Additionally it requires no specialist equipment (other than the developed interface method) and thus is accessible to all researchers. The benchmark is presented as an experimental suite, recording comprehensive game data and automated calculation of continuous control information throughput. Post hoc analysis of game data shows subjects use a similar gaze strategy to control the game to those observed in natural eye-movements during ball catching. Thus the interface implicitly leverages pre-learnt behavioural strategies enabling "one-shot" learning.

4.1 Introduction

Significant advances in robotics, demotics and disability aids aim to restore vital levels of independence to patients with motor disabilities, of which there is an ever increasing prevalence in our ageing society. However to fully realise the potential of these devices, significant improvements in human machine interface (HMI) technology are required. An ideal HMI would be based on a non-invasively accessible signal of patient intentions that has a sufficiently high information throughput to control available disability aids (e.g. Wheelchair ~ 15.3 bits/s or hand prosthetic 54.2 bits/s (Tonet et al., 2008)). In addition, for patient acceptance, this interface must be very easy to learn (Murphy et al., 2016; Phillips and Zhao, 1993) and have a minimal latency to enable dynamic control and potentially, embodiment of the device (Antfolk et al., 2013). Towards this, huge resources have been deployed to the development of direct brain machine interfaces (BMI), which have the advantage that they can derive a signal of intent even when the patient has no motor control whatsoever. However this approach comes at considerable clinical and post-clinical cost and is currently unsuitable for use outside of the lab due to either highly invasive set up and training, or very low information transmission bandwidth (Abbott and Faisal, 2012).

This direct BMI avenue of research is vital for furthering neuroscience, but its justification as the only available method for patients with no motor abilities to control computers and disability aids can be misleading. A large proportion of direct BMI approaches actually require the subject to have control of eye-movements to function effectively. For example in commonly used EEG systems based on visually evoked potentials, including P300 event related potentials, the patient must be able to align their gaze with the intended target (Brunner et al., 2010). It is also notable that in recent impressive advances made with intra-cortical neural interfaces (Collinger et al., 2013), the electrodes were implanted next to the frontal eye-fields, an area known to be responsible for voluntary eye movements. In addition, the patient in this study could move their eyes freely and it has been argued that the results achieved may be due to oculo-visual factors, although the evidence is unclear (Tehovnik et al., 2013).

What is clear, from mounting evidence in the BMI field (Abbott and Faisal, 2012; Buckley et al., 2011; Corbett et al., 2012; Onose et al., 2012), as well as the field of eye-movement research (Hayhoe and Ballard, 2005; Land and Tatler, 2009), is that eye-movements provide a vital non-invasive window into the perception-action loop. This huge potential must be utilised more in the BMI field to develop practical and accessible interfaces for patients with severe motor disorders. Eye movements, retained by patients with serious motor deficiencies, paralysis and limb amputation (Kaminski et al., 1992, 2002) are the only directly observable behavioural signals that are naturally both highly correlated with actions at the task level, and proactive of our peripheral motor actions. As a consequence, it provides vital information to decode our intentions in everyday tasks such as object manipulation and driving (Ballard et al., 1992; Land and Lee, 1994). Other behavioural signals such as facial muscles and tongue movement can be used (Mace et al., 2013), however unlike eye-movements, such signals are not naturally indicative of our motor intentions but must be trained and require suppression of otherwise natural movements such as talking and facial gestures. As such a non-invasive and ultra-low cost alternative BMI approach is proposed, the GT3D system, based on intention decoding from 3D gaze

signals. Gaze based approaches offer real-time closed-loop control that outperforms invasive (and non-invasive) BMIs in terms of cost and information read-out data rates (Abbott and Faisal, 2012) and hence has enabled robotic arm control in conjunction with other low data-rate signal sources (EMG(Corbett et al., 2013)EEG (Onose et al., 2012)or tongue-flick-switches(Buckley et al., 2011)). As estimated previously, the GT3D approach could yield bit rates up to 43 bit/s at a system cost of <£30 and this performance is framed within a comparison of BMI approaches in terms of cost and information throughput(Abbott and Faisal, 2012).

This comparison triggers the argument as to whether the use of eye-movements as a control input can be classed as a BMI. To accommodate this yet another acronym has evolved, namely a brain/neural-computer interface (BNCI) that includes other signals not located directly in the brain such as electromyography (EMG), heart rate and eye movements (Brunner et al., 2014; Murphy et al., 2016). While it is argued here that the eyes have a special status as strictly they are an extension of the brain (within the dura mater) and they are the only directly observable behavioural signal that is naturally both highly correlated with actions at the task level, and proactive of our peripheral motor actions. However to avoid getting sidetracked in semantics, the important thing is that the use of eye-movements is already providing input methods, and still has extensive potential to enable patients with severe disabilities to interact with the world. Eye-tracking and even low-cost eye-tracking is nothing new (Li et al., 2006; San Agustin et al., 2009; Schneider et al., 2011) – in fact two commercial companies have recently released low-cost eye trackers costing around £100. The aim here is to promote the application of this powerful interface, together with knowledge from the eye-movement research field, to empower patients with severe movement disabilities to regain control and independence. Towards this it is demonstrated that gaze location, particularly in 3D and at high-data rates (matching those of eye movements), provides a real-time decodable and graded control signal that should be utilised in the BMI or BNMI (brain/neural-machine interface) field.

Having addressed cost, information throughput, latency and minimal invasiveness, an additional barrier for current BMIs is the extensive preparation, set up and patient learning time, which can reach months or even years, posing a major challenge for patient uptake of BMI technology (Murphy et al., 2016). While the best EEG performance to date can achieve control with 3 degrees of freedom, this was only achieved in tightly controlled laboratory conditions and took months of intensive training (McFarland et al., 2010). To provide a practical assessment of emerging BMIs they should be tested outside of the lab. It should include a wide demographic of naïve subjects in the real world, within a limited time-frame. Guger et al. (2003) performed a study using EEG at a public exhibition where after 2 sessions of training (~30 minutes), 93% of subjects (N=99) were able to achieve a classification accuracy above 60% (chance=50%) at an information through put of 1 bit decision in 3.75s – an effective bit rate of 0.46 bits/min(Guger et al., 2003). This required at least 15 minutes for each subject to apply the EEG electrode cap on top of the 30 minutes training, not to mention the removal and the requirement to wash the electrode gel from the hair. More recently, in another study Guger et al. (2009) showed that 72.8% of subjects (N=100) could spell a 5 character word with 100% accuracy at an information transfer rate of 1 character every 28.8 seconds (9.9 bits/min) after 5 minutes of training (Guger et al., 2009). A more recent large study (N=106) study by Allison et al. (2010) used

EEG with steady-state visually evoked potentials (SSVEP) for a spelling task, achieving an average classification accuracy of 96% on a 5.2 bit decision in 21 seconds per decision (13.5 bits/min)(Allison et al., 2010). As mentioned, these approaches require visual focus and the subjects eyes were free to move and thus could be achieved much more efficiently with an eye-tracker.

These 3 BMI studies "in the wild" achieved discrete decisions; left or right in the first and which letter in both the second and third studies. For prosthetic control, in order to give maximal autonomy to the user, continuous real-time control is necessary. The user must have dynamic control to update their control input to react to the dynamic environment. Thus an alternative BMI benchmark task is proposed: the control of the arcade game "Pong", as unlike the above tasks and many other cursor based control tasks used to assess BMI performance in literature, it requires real-time control for successful game play. Commonly, BMIs are assessed for continuous control using one two and even three dimensional cursor tasks in which subjects are presented with a defined target they must move a cursor to in a given time limit (up to 10 seconds) (e.g.(McFarland et al., 2006, 2010)). In the game Pong, the target is the moving ball and thus subjects must be able to dynamically adjust the paddle as their estimate of the ball end point changes. This is much more related to prosthetic control in which the user must be able to adapt to their environment and perturbations in real-time. Expanding on a preliminary small field study(Abbott et al., 2013), a large population of gaze-actuation naïve members of the public (N=2042) attempt to play the BMI benchmark Pong, using eye movements alone. The large subject count is testament to the minimal setup and training our approach requires with just 30 seconds to calibrate and 5 seconds training before the game started.

4.2 Methods

4.2.1 Public facing BMI arcade machine

Gaze controlled arcade booths house the low-cost video-oculography based eye-tracker (as presented in detail in(Abbott and Faisal, 2012)) and the Pong game platform (Figure 21B). The arcade booth frame was custom-designed and built by Haberdashery (London, UK) to create an attractive and engaging interactive exhibit. The booth contained a viewing window to an internal display for the gaze player and an external display for spectators or opponents. The viewing window placed the subjects 60cm from the internal display of resolution 1368x768 pixels and dimensions 41x23cm. The eye-tracker was remotely mounted within the booth, tracking both eyes of the player through the viewing window. These booths were set up at three scientific engagement events and members of the public tried gaze actuation for the first time using our system. An extremely large number of subjects participated (N=2042) outside of the controlled lab situation with no previous experience of using gaze as an input. Details of the gaze tracking system used (<£30) can be found in Chapter 2. The difference here is the eye tracker is remotely mounted inside the arcade booth, rather than on the user's head, and the frame resolution was increased to 640x480 pixels at the cost of reducing the frame rate to 60Hz. This aimed to generalise better between different eye positions and head alignments, thus avoiding per-user camera adjustment.

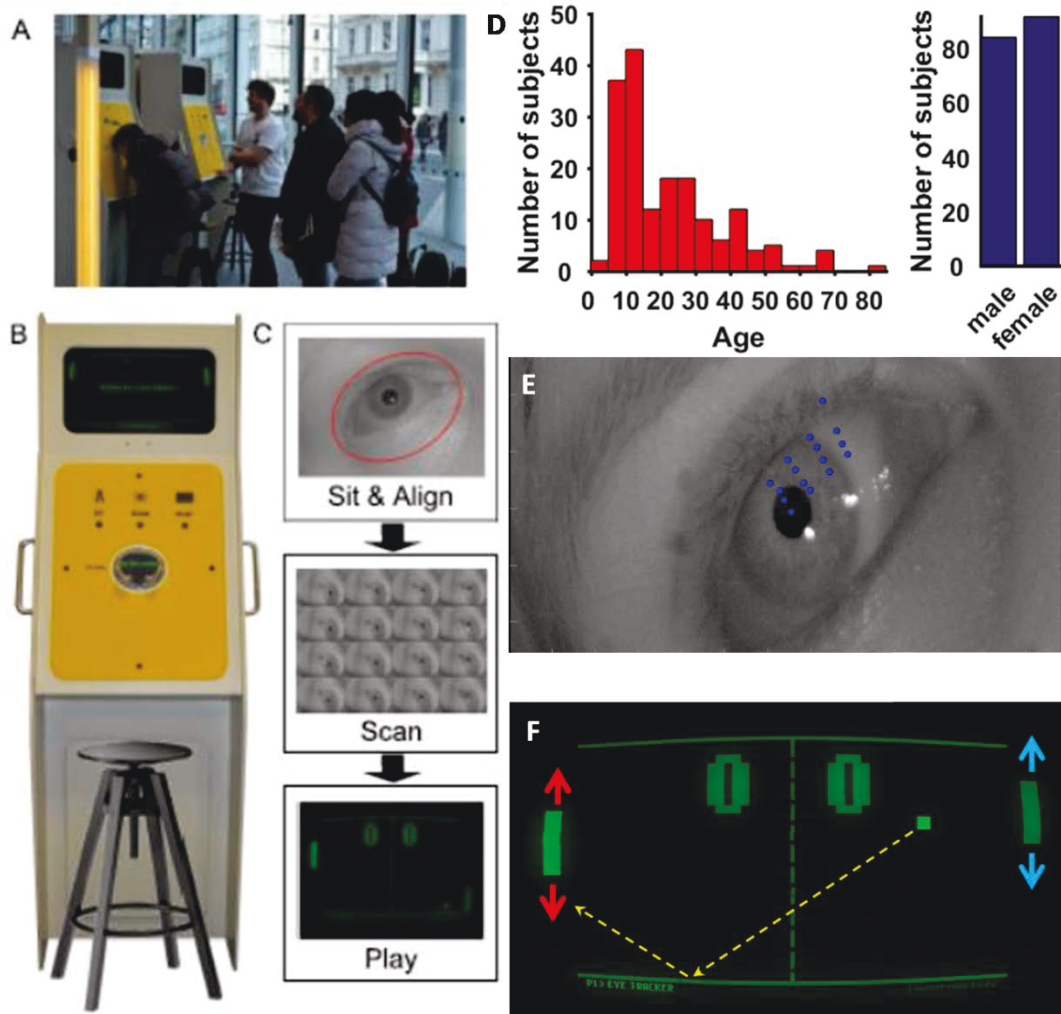


Figure 21: System for field-testing our BMI approach in public engagement of science exhibitions. A) BMI 'arcade booths' at science engagement event, a user leans into the BMI booth to play Pong using their eyes. B) BMI 'arcade booth' with embedded eye-tracking system (inside oval view port). C) BMI calibration process. Sit & Align: User sits down at the booth and onscreen visual feedback instructs to adjust the stool to suit their height. Scan: Automatic gaze calibration (30s duration) starts and eye-images are recorded while user looks at each calibration point. Play: Pong arcade game screen shot with retro-look for user appeal. This screen is visible both through the goggle view in the view port and at the top for spectators. D) Age E) Gender distribution for the subjects for which personal data was recorded and has been entered (N=132). F) Calibration data: example of eye image taken at each calibration point. The blue dots represent the recorded pupil position in pixels for each calibration point. G) Annotated screen shot of game court where the opponent controls the vertical movement of the left hand paddle (green rectangle) up and down as indicated by the red arrows and the opponent controls the right paddle movement indicated by the blue arrows. The green square is the ball and an example trajectory is indicated in yellow.

4.2.2 Subjects

The exhibition was conducted by the medical research council (MRC) who sought the relevant ethical permissions, insurance and public engagement licences. Subjects (N=2042) were self-volunteered from the general population at a science exhibition by simply approaching the booth, queuing and giving their informed consent to storing data collected during their participation. The subject population was highly heterogeneous,

recruited equally from both genders and with ages ranging between 4 and over 80 years of age, with a variety of educational and physical fitness backgrounds. Subject data was stored only when informed consent was gathered. For minors, this consent was sought from their parents. Subjects were previously naïve to using their eye-movements as a control input. Figure 22 gives an overview of the subject data breakdown. As this experiment was outside of lab constraints there was no guarantee subjects would complete the experiment fully, with multiple distracters, misunderstanding of instructions etc. Of the 2042 who approached the booth 1903 completed the 30s calibration routine (for 139 either the calibration was aborted before the end because they left the booth or the system crashed). Of these 1903, 1552 calibrated successfully and for 943 of these the game data was recorded. Of these, 651 subjects played a complete game and had a full data recording.

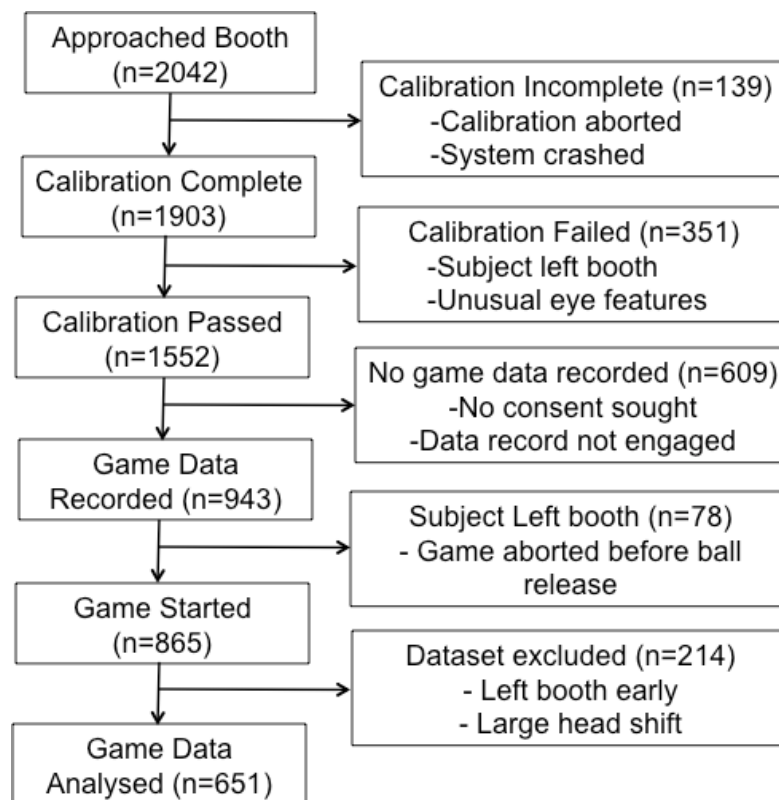


Figure 22: Subject breakdown. Describes the subject data break down from the 2042 subjects that approached the booth.

4.2.3 “Pong” BMI benchmark

The subjects were asked to play the game Pong with their eye movements. This game, a very early computer tennis game, provides a simple and accessible benchmark for closed loop real-time control (Abbott and Faisal, 2012). The subject controls the vertical position of their racket or paddle as it will be referred to here, to meet a moving ball and return it to their opponent. A new version of the game is written to allow full game analytics to be recorded automatically for post-hoc analysis. The latest version of the Pong BMI benchmark software (Figure 21F), presented here, was written by Alan Zucconi in Action

Script 3 for Adobe Flash Player 11 (Adobe Systems Incorporated, San Jose, California) and is rendered at a resolution of 800x600. The game had either a default 1 player mode (eye control against artificial intelligence (AI) computer player) or two-player mode (eye-control against human opponent with up/down game pad). To switch from one player to two-player mode, the human opponent could join at any time during the game, simply by pressing on the external control pad and taking over from the AI. This made the system more engaging for the public with the idea of having the gaze player competing with the conventional input from the hands. The AI controlling the opponent paddle in single player mode has perfect information on the ball position when it is within 200 pixels (10cm) horizontally from the paddle. It can move with a maximum speed of 5 pixels per frame which translates to 12 cms^{-1} . The initial ball speed was also set to 12 cms^{-1} and the starting direction randomly initialised between ± 60 degrees. If moving horizontally this means that the ball takes about 2.5 seconds to travel the game width (30cm). The ball speed increased by 10% with every successful return by either player, until a point is scored at which point the speed is reset back to 12 cms^{-1} for the next point.

4.2.4 Sit, align, scan and play: Set-up, calibrate and control

The system was designed to enable naïve subjects to use the system with little or no assistance. The simple instructions of “sit, align, scan and play” were printed on the front of the booth with pictographic representations. Each subject sat on an adjustable chair, and positioned his or her head in the viewing window. Subjects with corrected vision removed their glasses (but not contact lenses) to position their head comfortably in the viewing window (see Figure 21A,B). When seated comfortably, an automatic calibration routine was initiated with a button press by the subject. This required the subject to look at each calibration point appearing on the screen (4x4 calibration grid) allowing a continuous mapping between the pupil position and the screen gaze position to be learnt by the system (Abbott and Faisal, 2012), see Chapter 3. For this application the vertical component of this gaze signal (60Hz) directly controlled the paddle position during the game. The game launches immediately after calibration with a 5s countdown before the ball is released. In this countdown phase the user has control over the paddle and thus has five seconds to understand this interaction before the ball is released and the game (first to 5 points) starts.

4.2.5 Optimal binocular gaze mapping

The calibration routine provided a correspondence between each pupil position, given by the eye-tracker, to a gaze position on the game screen. This data was used to train a linear combination of nonlinear basis functions (2nd order polynomial, Moore-Penrose pseudo-inverse (Moore, 1920) training) to give a continuous mapping between pupil and gaze (further details found in our previous work (Abbott and Faisal, 2012)). To enhance the system robustness, a binocular eye tracker is used. Although for a 2D gaze estimate a monocular eye-tracker is sufficient, Bayesian sensor fusion is used to combine the binocular gaze feed to improve the gaze estimate. The two gaze signals are automatically combined optimally, based on the reliability of each pupil signal (left and right) during the calibration phase. To obtain this measure of reliability, at each calibration point a burst of ten pupil measurements was taken for each eye and the detection rate was used as a

measure of reliability (ranging from 1 (4x4 grid, 10 samples at each thus, 160/160 pupils detected) to 0 (0/160). This meant that if a pupil wasn't detected consistently the system would automatically rely more on the other eye signal. Calibration is passed if a detection rate greater than 0.8 is achieved in either of the eyes.

4.2.6 Data Collection

From subjects who had given informed consent, age, gender, eye correction, subject calibration data, gaze and game data is recorded where possible. The calibration data includes detected pupil attributes, the eye images taken at each calibration point and if the calibration was successful, the calibration weights mapping pupil to screen gaze for each eye. The game data includes paddle positions, ball position and score, as well as gaze input data. The game data and gaze data were recorded separately at different frame rates. Thus the data was synchronised and the data re-sampled to a union of the timestamps using linear interpolation. This was done so that when analysing game data, timing integrity was preserved which is important in detecting events in the game. Similarly for gaze data, when making time critical analysis on gaze data, it is useful to have the raw gaze data with synchronised game data.

4.2.7 Game Analysis

The data sets collected were analysed to assess calibration quality and control performance. Calibration quality was screened online to reject poor calibrations based on the pupil tracking performance whereby at least one of the eyes had to have a pupil recognition rate of at least 80%. This means at least 128 of the 160 frames taken during calibration must have a pupil detected successfully. To assess control performance post processing extracted the game score, the number of successful ball returns and the accuracy at the point of impact for each subject. To measure the empirical chance level, a further experiment was conducted without player input. The paddle remained static in each trial with three settings; central, top and bottom, with three games each position.

4.2.8 Information Transfer Evaluation

To enable comparison with other BMI studies the information transfer rate (ITR) was calculated using a two alternative definitions. The majority of BMI studies involve discrete target selection while here the task requires continuous control. Thus to compare to the majority of studies, the most widely used method proposed by Wolpaw et al (1998) is employed, as shown in eq. 1 (Wolpaw et al., 1998a). The parameters include N, the number of possible targets, P the probability of correctly selecting desired target and T the time taken to make selection. Though the Pong game requires continuous control, it can be defined as 1D cursor control to hit the target of the ball. The number of possible targets is 437 – the span of the ball position. However to make a successful return, the paddle position must be within half the paddle width. To cover this span the paddle needs to take one of 4 ($472/128=3.7$) positions and thus $N=4$. For this definition to hold, Wolpaw's definition assumes equal target probability - in this case each hypothetical target must have an equal occurrence probability of one quarter ($1/N$). This assumption holds as the ball impact position has an approximately uniform distribution (see Figure 32B).

$$\text{Wolpaw's Information} = \log_2 N + P \log_2 P + (1 - P) \log_2 \left(\frac{1-P}{N-1} \right) \quad (1)$$

$$\text{Fitts's Information} = \log_2 \left(1 + \frac{D}{W} \right) \quad (2)$$

$$\text{ITR} = \frac{\text{Information}}{T} \quad (3)$$

Though the continuous control task can be simplified to a discrete target selection, the time per decision (T), required to give the information transfer rate (eq3), is more difficult to define. In synchronous BMI control T is fixed based on the protocol, while in many asynchronous BMI cursor control tasks the target position is known before the cursor control begins and either the time taken to hit a given target is fixed per trial (McFarland et al., 2006) or is measured (Wolpaw and McFarland, 2004). In the case of Pong, the task is more challenging (and realistic), as the target (ball) is dynamic and players must update their prediction of its end point to make a successful interception. At the most conservative, T could be taken as the time taken for the ball to travel from the opponent's paddle, but this assumes the subject can perfectly predict the end point of the ball's trajectory. More realistically, it is assumed that the subject can perfectly predict the endpoint of a straight trajectory (without rebound) and thus the time since last rebound is taken, whether it is from the opponent paddle or the wall. The game screen and the important measurements are shown in Figure 23. Applying this metric enables comparison with other BMIs, however it does not fully capture the continuous control nature of the task and thus an alternative metric is also applied.

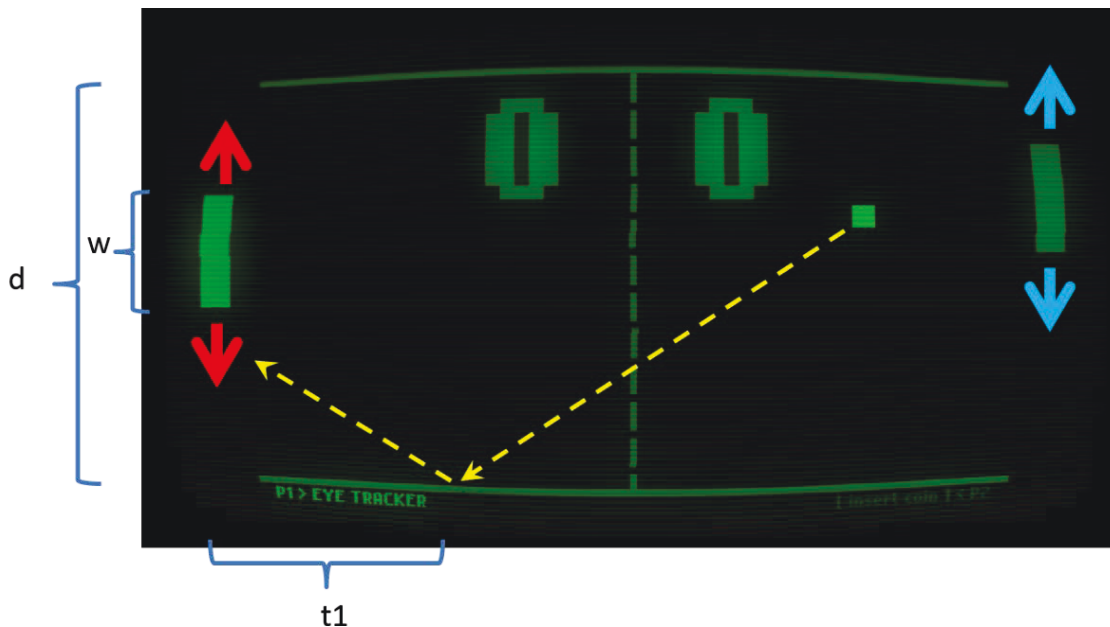


Figure 23: Pong game annotated to demonstrate the variables used for calculating the ITR. (d) is the control width, (w) is the paddle width and t1 is the time since the last bounce (defined T in the information through put calculation in eq.3).

The Fitts's throughput can be used based on the formulation used in (Gilja et al., 2012). According to Fitts's law, human movement can be modelled by analogy to the

transmission of information. It relates human movement to duration and target size. This model has been applied to the assessment of human machine interfaces in the past when developing methods to control a computer (before widespread establishment of the computer mouse) (Card et al., 1978). The Shannon formulation proposed by MacKenzie (1989) is shown in eq. 2 (MacKenzie, 1989). It is based on the distance between current and target position (A) and the size of the target (W) and time taken to select (t). This pointer based task and parameters are shown in Figure 24A. These parameters are formulated to yield an index of difficulty in bits as seen in eq. 2 and the throughput is then found by dividing this by the completion time (eq. 3). For Pong, the ball can be considered as the target and the paddle as the cursor. Though actually, the situation is reversed from Figure 24A i.e. the target (ball) has to end up within the cursor width (paddle), it is equivalent. An additional complication is that the target (ball) is moving and its end point - the actual target for the cursor (paddle), is unknown until the instant of impact. The subject must therefore base their control decisions on a continually updated estimate of the ball end point and thus paddle target. This complication acts to make the calculated throughput a conservative value, underestimating the difficulty of the task.

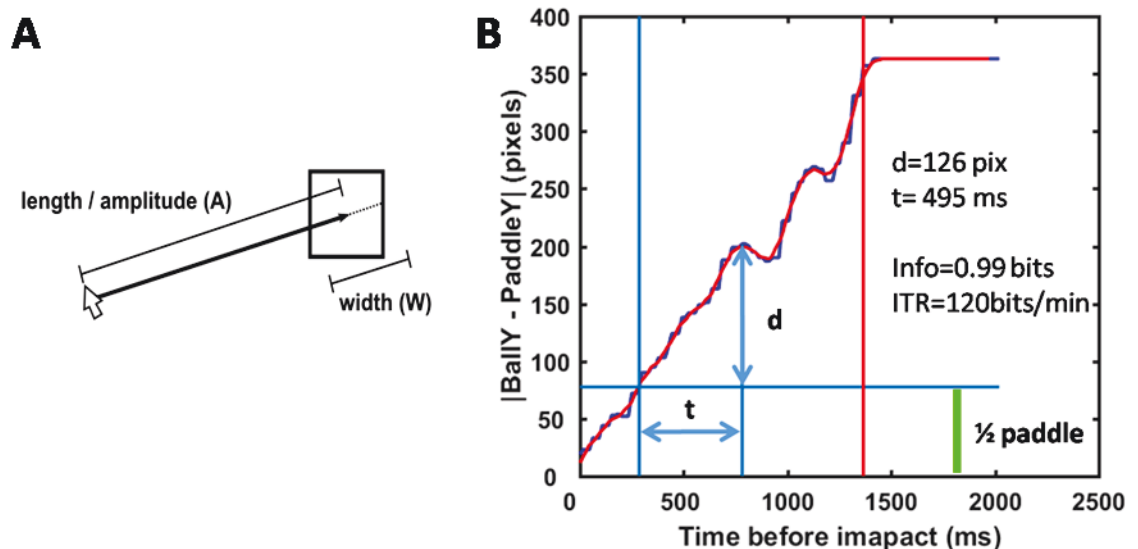


Figure 24: Overview of Fitts's law. A) Diagram showing Fitts's formulation of moving a cursor to a target area. B) Example trace of player returning the ball showing the trajectory of the difference between paddle and balls impact position. The blue trace is raw while the red trace is the low pass-filtered trace to remove high frequency noise. The point at which this trajectory becomes monotonically decreasing is considered the decision point at which the player chooses the target endpoint. From this decision point the target reaching amplitude (A) and the movement time (t) is calculated as shown. The target width is fixed as the paddle width remains constant in the game. To account for the player's performance accuracy, W is set to their mean endpoint error, providing an upper bound of their Fitts's throughput.

This metric can be applied to the game using the real-time data recoded during game play. To calculate Fitts's difficulty, the assumption is made that at a certain point in time, the subject settles on an endpoint target estimate and then moves the paddle towards it. This is identified from the control signal by looking at the paddle end point error trace. An example trace is shown in Figure 24B. Thus the point at which they start to move the paddle towards the target is defined as when the end-point error becomes a monotonically decreasing trajectory towards impact. This is effectively from when the

player only moves their paddle closer to the point of impact. The raw trace (red line) has been low pass filtered to remove high frequency noise that degrades true monotonicity. This approach is a good representation as if the subject begins to move to an incorrect estimate of target endpoint and then corrects this and moves towards the correct target, only the final part of this trajectory will be used to calculate the Fitts's throughput. In addition, the time when the error is within the half-paddle width stops the clock so that if they are highly proactive of the impact and wait with the paddle at the target location, the duration is only taken until they are within the target size (Figure 24). In the above example the distance travelled is 126 pixels (7.25cm) to a target width 128pixels. This yields 1 bit of information in ~0.5s and thus 120 bits/min.

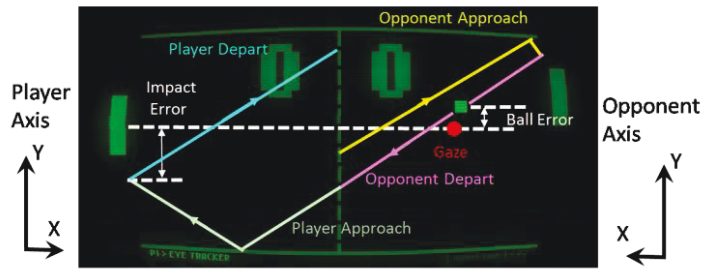
4.2.9 Player Strategies feature vector

Having such a large data set enables data driven approaches to be applied to understand common strategies used by gaze players. To do this involves preparation and pre-processing of the vast quantities of data involved. Figure 25 shows the data analysis pipeline. Firstly, the gaze data is separated into 4 main states, 1) opponent approach, 2) opponent depart, 3) player approach and 4) player depart as shown in Figure 25A. For each of these segments there is a subcategory based on whether a miss or hit has occurred and thus in total 8 categories. For each segment the vertical Y positional data is considered as this is the control dimension for the game. Each ball trajectory may have a different angle and velocity and thus the trajectories have different durations (Figure 25B(i)). To enable comparison, the trajectories are re-sampled (linear interpolation) to a common position stamp rather than time stamp. This effectively time warps the data to a common metric of importance: distance to impact (Figure 25B(ii)). The re-sampled data can then be combined in the same data cube and analysed together (Figure 25B(ii)). Error metrics are extracted from this data cube. The only error metric that actually counts for the performance (miss versus hit) is the error at the instance of impact. However for the analysis of game strategy, the manner in which this goal is reached is of interest. Thus the full segment trajectory, as the ball moves towards impact, is analysed. Two error metrics for the player paddle are defined that can disambiguate proactive and tracking strategies: ball error and impact error. Ball error is the absolute vertical Y error between the paddle and ball. Impact error is the absolute vertical error between the future ball impact position and the current paddle position. Figure 25C demonstrates these error metrics for the examples of pure tracking i.e. the player tracks the ball until impact and purely proactive i.e. the player predicatively puts the paddle at the goal position to make a return, before the ball arrives there. These two error metrics are then concatenated to give an 800 dimensional feature vector that captures player strategy (Figure 25C(iv)). By combining both error metrics disambiguation is possible between situations where the one metric alone would be identical. An example of such a situation is shown in (Figure 25D). Here the two cases are ball moving towards a static paddle, compared to the paddle moving towards the ball (travelling a horizontal trajectory). As can be seen with ball error alone the feature is identical, but combining it with impact error distinguishes the situations.

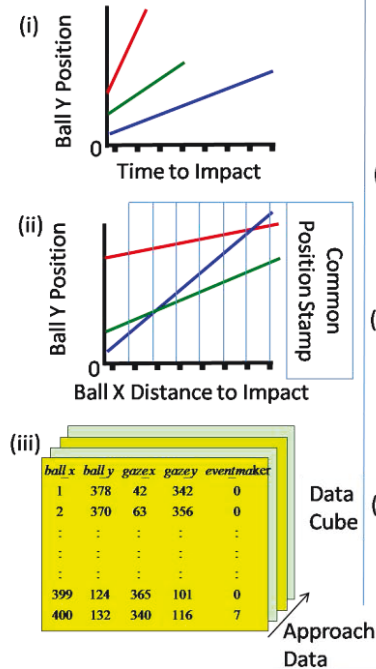
4.2.10 Strategy Clustering

The feature vector can be used to compare behaviour in different segments of play, for example comparing behaviour in the player approach with behaviour in the opponent approach. The global average of strategies across all subjects in these situations gives a very general trend of behaviour, but there is significant variation in behaviour which is hidden in the aggregation. Thus with such large amounts of data the trajectory data is clustered to reveal different strategies. The data is clustered using the unsupervised agglomerative hierarchical approach with Ward's criterion as the objective function. This starts by considering all data points as separate clusters which are sequentially merged. The merging decision is made to minimise the resulting total variance within clusters at each step up the hierarchy. This approach gives an intuitive description of the emergence of separate strategies and families of strategies when it is unknown how many strategies exist. These cluster centre trajectories can be interpreted as stereotyped player strategies and related to different metrics such as whether the strategy resulted in more hits or misses for example.

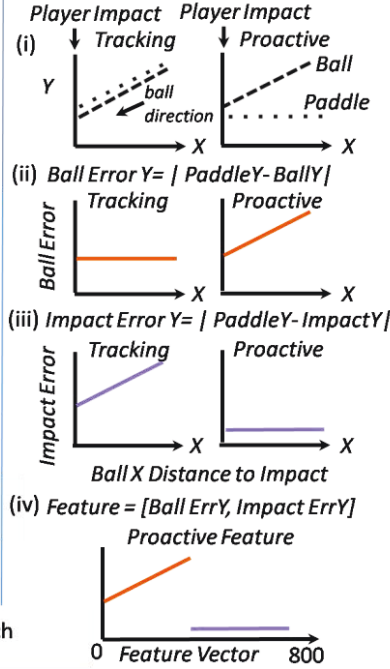
A. Data Segmentation:



B. Data Interpolation:



C. Error Metrics:



D. Two Error Metrics Enable Disambiguation:

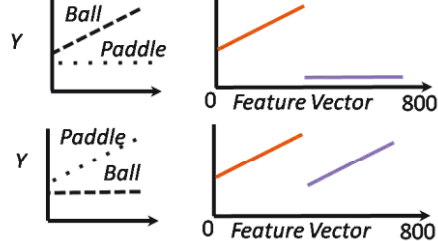


Figure 25: Data analysis Pipeline. A. Data Segmentation: Game data is separated into 4 main states, 1) opponent approach, 2) opponent depart, 3) player approach and 4) player depart. For each of these there is a subset of miss or hit. B. Data interpolation: The paddle control dimension is analysed (vertical - Y). (i) Ball Y position against time for 3 trajectories with different ball speeds and directions - thus different durations. B(ii) Trajectories re-sampled (linear interpolation) to a common position stamp. B(iii) Re-sampled data combined into a data cube for analysis. C. Error Metrics: C(i) Ball position (dotted) and Paddle position (dashed), vertical Y against horizontal X for the player tracking the ball with the paddle until impact (Tracking) and where the player predicatively puts the paddle at the impact position (Proactive). C(ii) Ball error is the absolute vertical Y error between the paddle and ball, this is shown for the tracking and proactive traces. C(iii) Impact error is the absolute vertical error between the ball impact position and the paddle position as the ball approaches. C(iv) Error metrics concatenated to create a feature vector that describes the player strategy. D. Combined error metrics disambiguates situations that one metric alone could not. The example here is when the ball moves towards the paddle versus the paddle moving towards the ball which travels in a horizontal trajectory.

4.3 Results

4.3.1 Public facing auto-experiment booth

A public facing experimental platform has been developed that automates data collection via an arcade booth, allowing members of the public to walk up and within 30s experience eye control for the first time. Figure 21A shows the booth in action and Figure 21B and C shows the booth and the simple “sit”, “scan” and “play” interface. The platform provides vital public engagement and assessment of the approach in an unconstrained environment and subject demographic.

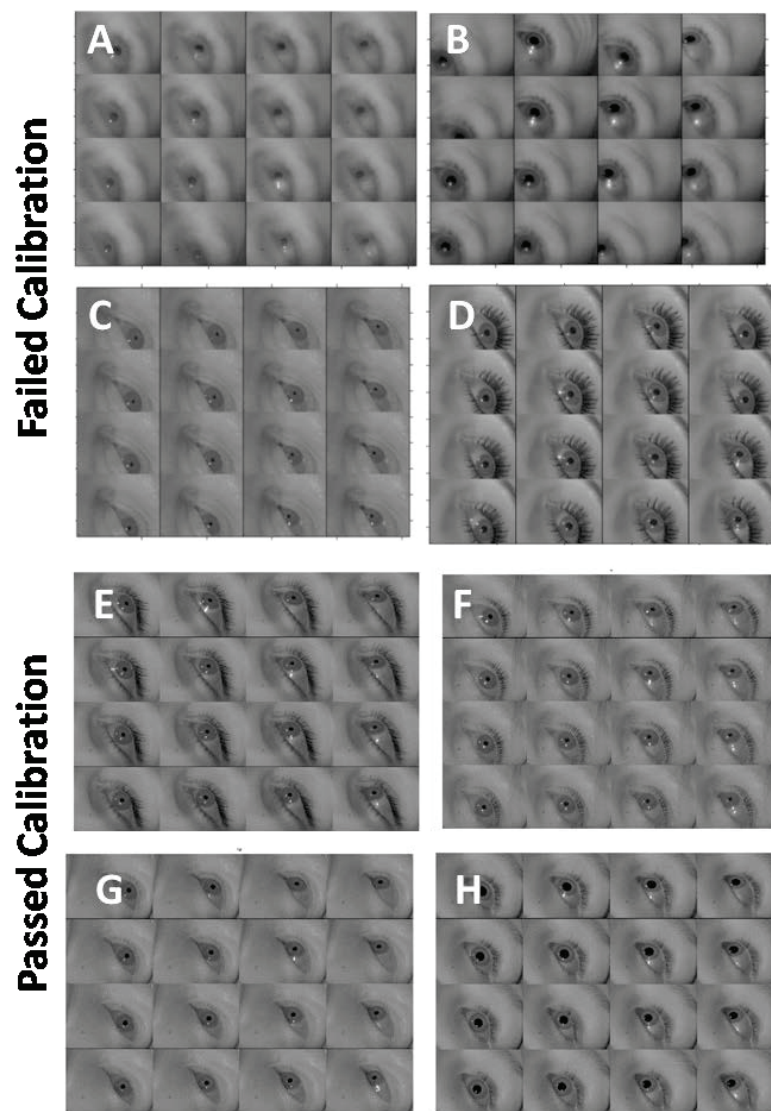


Figure 26 Examples of failed calibration (A-D) and passed calibration (E-H): A-D Failed due to A) Steaming up the viewing visor, B) excessive head-movements, C) Very small pupils, D) Strong eye-make up. E-H Passed despite E) Less extreme eye-make-up and F-H) Varying pupil size, eye positioning, eyebrow/eyelash thickness.

4.3.2 Calibration Performance

This study provided a large field verification “in-the-wild” (outside of controlled lab conditions), of the developed eye-tracking image processing algorithm and the task-level approach to BMI control using gaze intention decoding. The age and gender of sampled subjects is presented in Figure 21D. With this large demographic, out of all subjects to approach the booth (N=2042), 1903 completed the calibration phase, of which 82% calibrated successfully on a single calibration attempt. Figure 21E represents an example of the calibration data stored for each of these subjects which includes calibration images from both eyes and pupil location for each point in a 4x4 calibration grid. For those who failed to calibrate, the stored images can be used to identify generalisation issues across a very large range of eyes. The reasons for calibration failure include both behavioural and detection problems. Behavioural examples include steaming up the viewing visor as shown in Figure 26A, excessive head-movements (Figure 26B) or leaving the booth before calibration was finished. Detection failure also occurred when the subject wore strong eye-makeup (Figure 26D) or had particularly extreme eye features such as a very small pupil as shown in Figure 26C. However as shown in Figure 26E-F, a wide range of eye-types and slightly less strong eye-makeup could be accommodated automatically by the system. For those that failed, the large database of images can be used to improve the system and image processing algorithms to account for all eye types and orientations in the substantial sample of over 400,000 eye images.

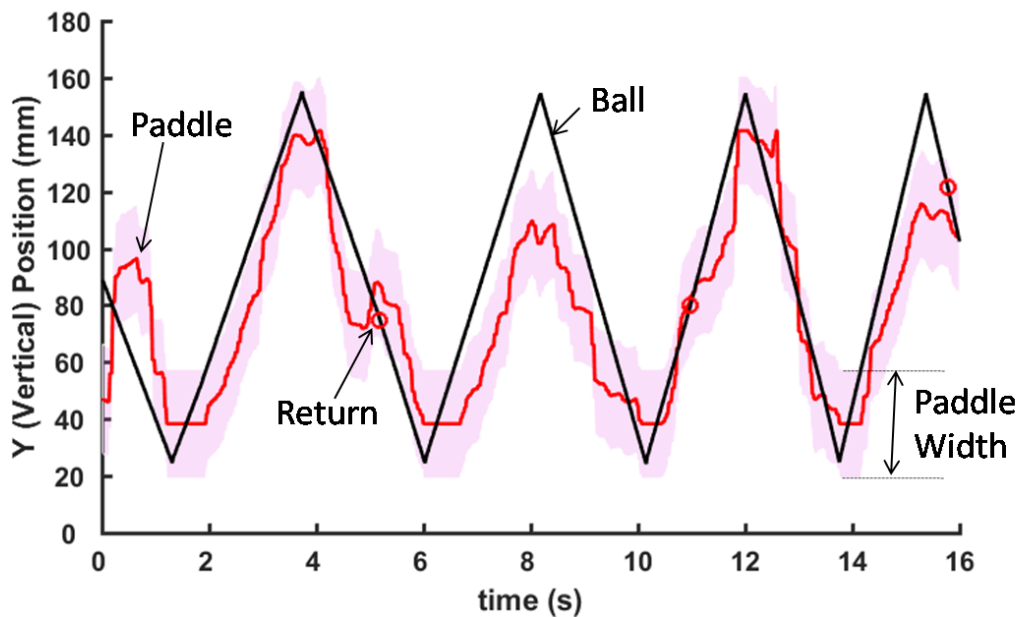


Figure 27: Task-level BMI control data recorded during the arcade game Pong from the first instant of the user's first interaction. Vertical position of the ball (black line), continuously closed-loop controlled position of BMI controlled paddle (red line). The shaded area represents the span of Pong paddle reach. Red dots mark contact of the Pong ball with the paddle. Note, time axis indicates the actual seconds since the first use of the system for the subject, preceded by a calibration stage (30s) where subjects received no feedback on eye position and a 5 s countdown before ball release, when subjects receive paddle control feedback.

4.3.3 Pong game Performance:

Of the 1552 subjects that passed the calibration phase, game and gaze data for 651 subjects' data is analysed. This includes the full state representation of the game (25Hz) and gaze location (60Hz). Figure 27 shows an example data trace of the vertical ball position (black line) and the player's paddle position (red line) recorded during the subject trial. The figure represents the immediate intention control achieved from the instant the ball is released. Prior to this, the subject has had just 5 seconds to understand the control input they are making. The subject effectively controls the paddle (red line) with their vertical gaze position and intercepts the ball from the start. Each ball interception, or *return* as it will be referred to hence forth, is marked by a red circle in Figure 27. These returns are counted for each subject to provide a performance metric. For the studied subject group (N=651 see Figure 22 in methods for full breakdown), games lasted on average 73 ± 33 s (mean $\pm$ SD) and players returned 6.54 ± 6.1 shots successfully, compared to an average 10.9 returns made by the opponent. The empirical chance level is estimated to be 2.6 ± 2.5 (see Figure 28A,B, empirical chance explained in Methods p59). In addition 17% actually managed to beat their opponent, returning on average 10.4 shots for these games. These high level metrics are across a large demographic (age 4–84) thus Figure 28B shows the performance break down in terms of successful returns for 3 age categories: children (0-10 N=24), adolescents and adults (11-60 N=105) and seniors (>60, N=3). All categories performed better than chance on average, with the adult and adolescents category performing particularly well. Overall the player accuracy (ball paddle error on miss/return impact) was on average 76.0 ± 70.3 pixels, where the required accuracy to make a return was actually 76.5 pixels. The absolute error distribution for all players is shown in Figure 28C, showing the returns in green and the misses in black. The long tailed distribution clearly shows the player's performance is not due to chance. The distribution of subject mean error over the game is shown in Figure 28D.

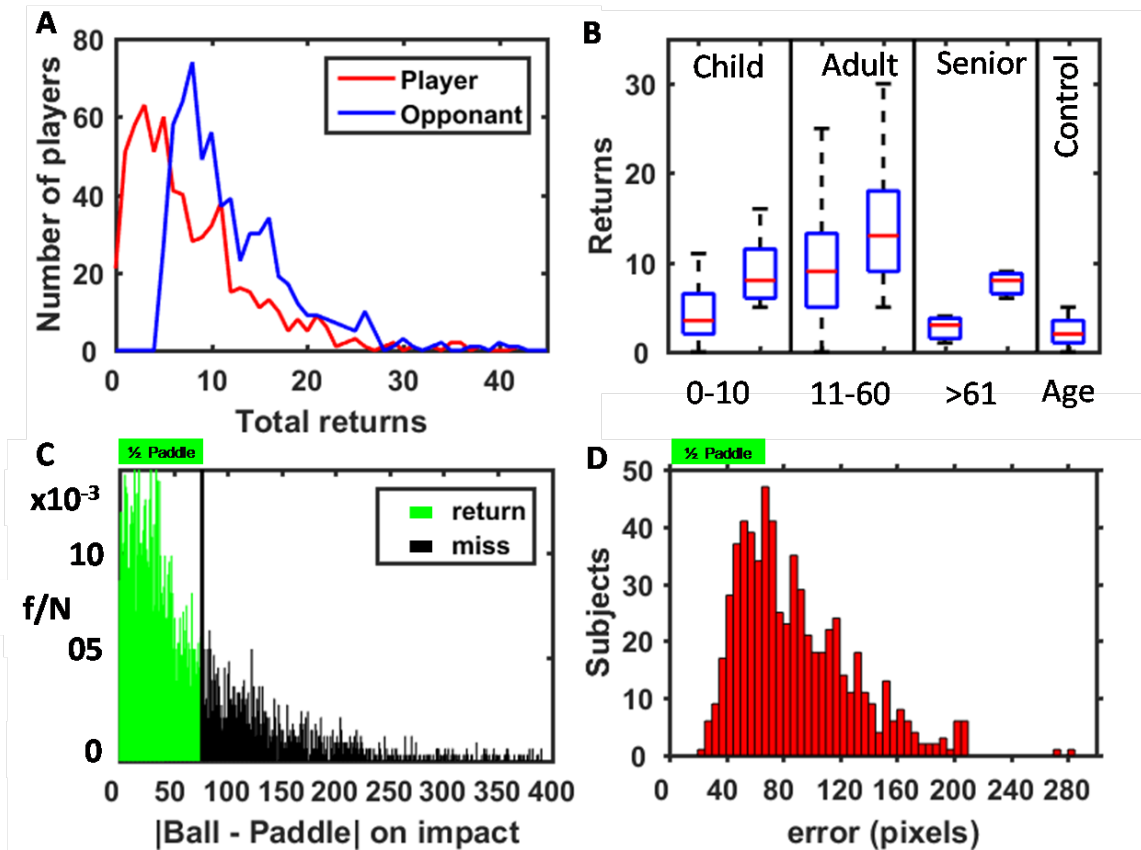


Figure 28: A) Distribution of balls returned per game for each user(red) and opponent (blue). B) Total returns for each subject categorised by age (data available for 103 subjects) with the corresponding opponent scores as well as the returns achieved without input, representing chance level. C) Absolute Ball paddle error with trajectories separated into misses (black) and returns (green). D) Distribution of subject mean error for full game (game height is 600pixels).

4.3.4 BMI performance

To enable comparison with existing large field studies the results for the analysed 651 subjects are converted into information transfer rate (ITR). As discussed in the Methods, information transfer rate has a number of forms and definitions and thus the raw performance data is presented (to enable alternative interpretations) and then two alternative ITR metrics from literature are applied. The first definition, proposed by Wolpaw et al. (1998) for discrete control tasks, requires 1) number of possible states, 2) the error rate and 3) the time required to communicate a decision (see Methods, (Wolpaw et al., 1998a)). To apply this to the Pong, a continuous control task, the control is simplified with the assumption that N discrete paddle positions are required to return all ball positions and thus, the game height divided by the paddle span giving $N=4$ ($472/128=3.7$). The distribution of subject return probability based on the number of balls returned over the total chances is shown in Figure 29A. For the time required to communicate the decision, the time since the last bounce is measured, assuming the subjects are able to perfectly predict the ball impact from this point (see t_1 in methods Figure 23). The distribution over this communication time is shown in Figure 29B. Combining this yields the ITR distribution of subjects performance in bits per minute as shown in Figure 29C. There is a large range of performance levels, as would be expected of

such a large and unconstrained field study. Only 20 (3%) subjects playing couldn't communicate any commands, while on average subjects achieved 37 bits/min, with some achieving over 200 bits/min. This performance can be put into context by comparing it to the 3 large field studies mentioned previously. Figure 31A shows both the peak and average performance for these large field studies (Allison et al., 2010; Guger et al., 2003, 2009), where the method developed here is referred to as GT3D. For both the average and peak performance, the GT3D system achieves considerably higher average performance of 37 bits/min, compared to 13.5 bits/min by the closest competitor (Allison et al., 2010). With the peak performance of 281 bits/min, compared to a peak performance of 15bits/min for (Allison et al., 2010). The gaze GT3D performance is achieved in a set up and training time 60 times smaller than the fastest method compared (Figure 31A).

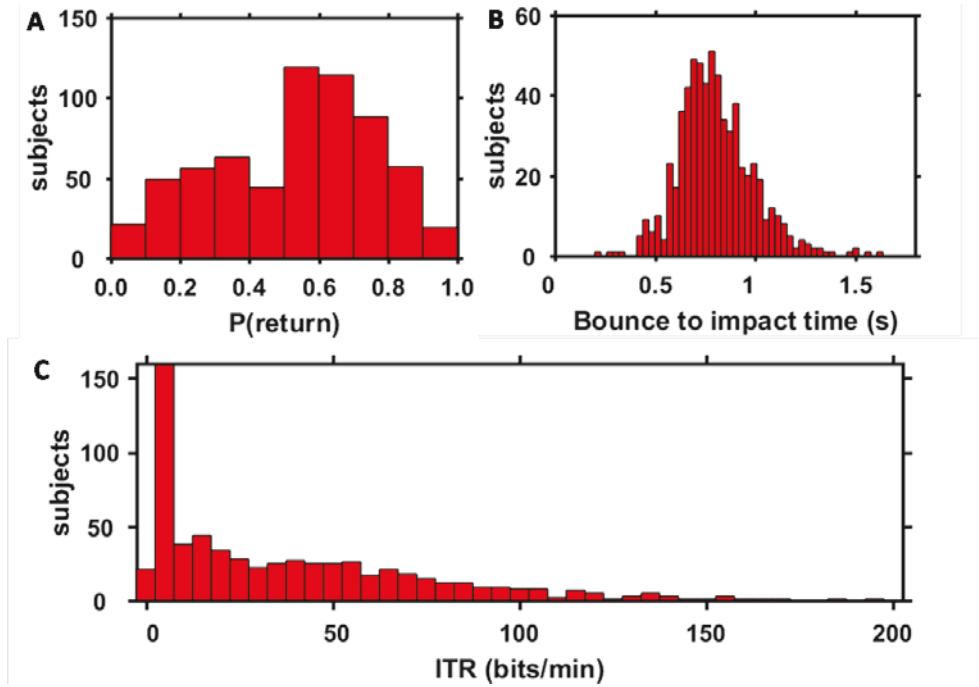


Figure 29: Wolpaw's Information throughput component variable distributions across subjects. A. Probability of making a return, based on number of returns divided by the number of chances. B. Time per decision is taken as the duration from last bounce. C. Information throughput distribution based on Wolpaw's definition, where the, the number of discrete states $N=4$. Note the first bin purely 0 ITR. Mean ITR 37 bits/min.

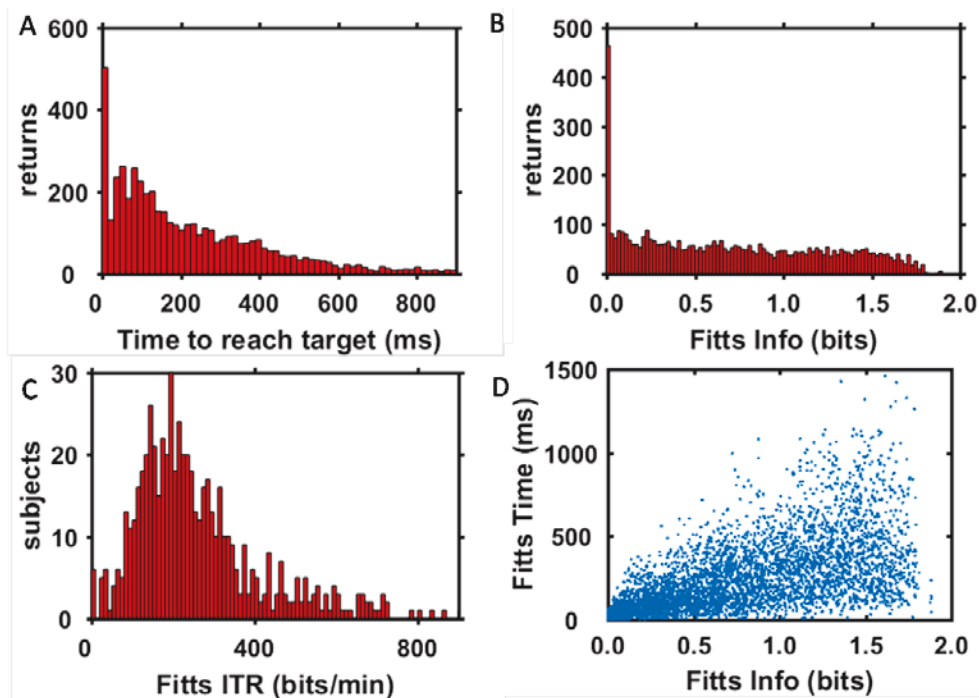


Figure 30. Fitts's Information Throughput (ITR) component variable distributions. A. Time to reach target. Estimated from when the subject starts moving towards the ball impact position (see Methods section Figure 24B). B. Fitts's information based on the distance between the ball and paddle when the player starts moving the paddle towards the target impact point (again see Methods section Figure 24B). C. Distribution of average Information Throughput Rate (ITR) across subjects based on Fitts's Law definition (Fitts's Info (B) divided by the time to reach the respective target (A)). Mean ITR 300 bits/min. D. Fitts's information against Fitts's time (time to reach target shown in A) The correlation coefficient is 0.83, suggesting the task obeys Fitts's law.

The second method to calculate information throughput, more suited to a continuous control task, is based on Fitts's law for assessing the difficulty of a pointing task. It is based on 1) the distance between current and target position, 2) the size of the target (paddle width) and 3) the time taken to select the object. For a deeper discussion on how these parameters are extracted from the game data see the methods section and Figure 24. The distribution of Fitts's information or difficulty (as defined in Figure 24 and eq. 2) per return is shown in Figure 30B with the distribution of time taken to reach the target also shown (Figure 30A). Combining this information gives the Fitts's information throughput distribution shown Figure 30C. The task obeys Fitts's law reasonably well, with the Fitts's information and the time to reach target (Fitts's time) having a strong correlation (0.83, as shown in Figure 30D). The Fitts's ITR distribution across subjects (Figure 30C), centres on a mean information throughput of 300 bits/min, reaching transfer rates over 1000 bits/min. This performance is compared to three other BMI cursor control studies (Ganguly and Carmena, 2009; Gilja et al., 2012; Kim et al., 2008; Taylor et al., 2002) that have been quantified using Fitts's information throughput by Gilja et al (2012) (Gilja et al., 2012). Figure 31B shows the performance comparison of the mean Fitts's information throughput for the 5 studies, showing again that the study presented here achieves significantly higher performance. The set up and training duration cannot be shown in Figure 31B because it is 6 orders of magnitude higher for the invasive direct brain machine interfaces, involving months between electrode implantation and the study even commencing.

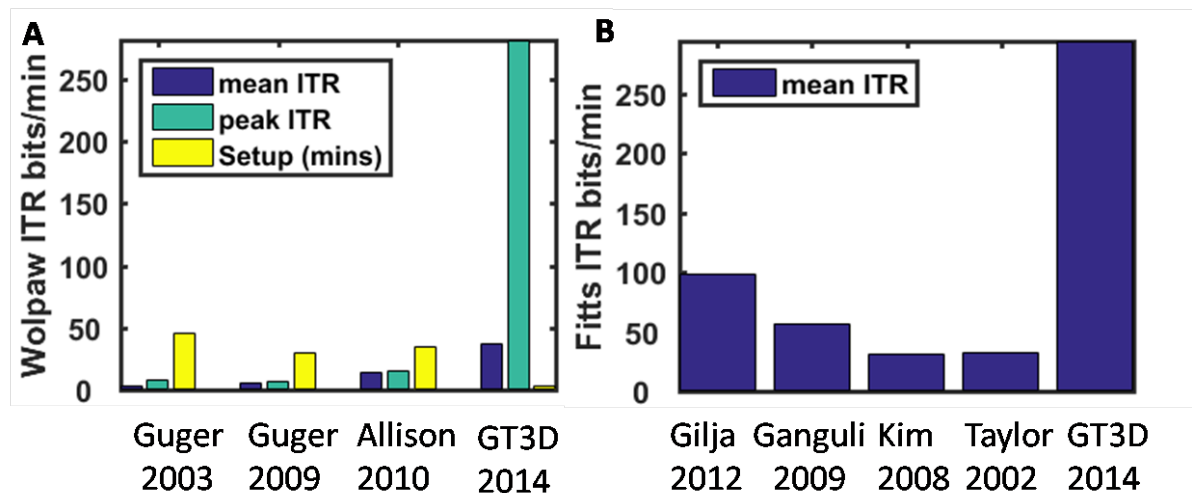


Figure 31: Performance comparison using two definitions for information throughput with other studies. A) Comparison with large field studies of BMI approaches using conventional Wolpaw definition for ITR. Slow SCP N=99 (Guger et al., 2003), P300, N=100 (Guger et al., 2009), SSVEP, N=106 (Allison et al., 2010). B) Comparison with invasive approaches achieving a significantly higher ITR and based on an alternative ITR interpretation based on Fitts's information.

4.3.5 Gaze control strategies:

The rich data set, with such a large sample size (N=651), allows subtle effects in gaze behaviour conditioned on the game states to be observed. Figure 32A shows the probability distribution of gaze allocation across all subjects while Figure 32B shows the ball distribution across all games. The subjects spend the majority of the game looking at the left hand side of the screen (their control zone), suggesting they do not simply track

the ball position, which is much more uniformly distributed across the screen. It is interesting to understand how the gaze behaviour is different when the ball is moving towards the player versus towards their opponent. To understand the distinction further and to understand different control strategies employed, the gaze input (vertical gaze position) is analysed with respect to the ball position. The data preparation is explained in detail in the Methods section and Figure 25.

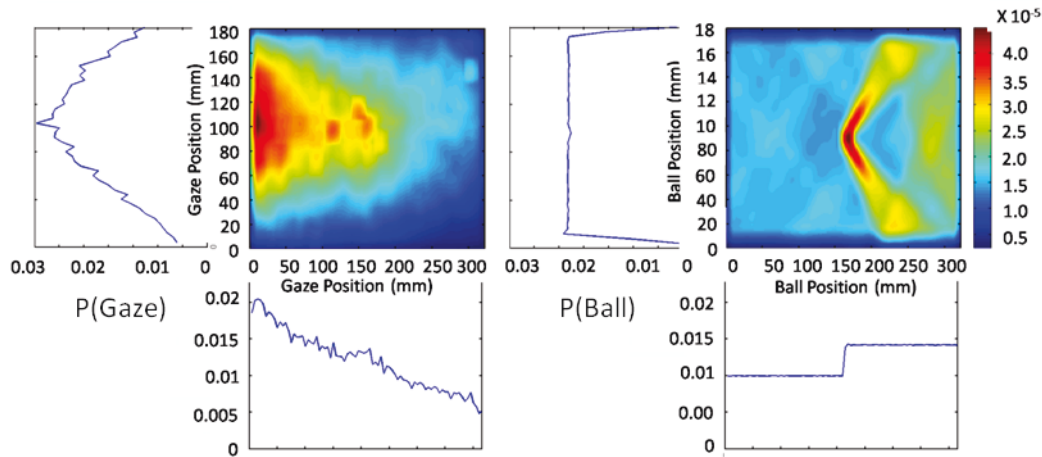


Figure 32: Probability distribution over the game court for A) Gaze position and B) Ball position. Calculated based on all subjects with game data (N=651) during game play. 2D histogram smoothed with a 5x5 Gaussian, sigma=2. The one dimensional marginal histograms are shown adjacent to their corresponding dimension.

An important question is whether subjects just track the ball or are they actively trying to hit the ball back, i.e. is the behaviour different when the ball is moving towards the player paddle compared to when it is moving towards the opponent's? If this is the case it would suggest they learn to or can implicitly control the task at hand and thus demonstrates the intuitive nature of gaze control. The data is separated into player and opponent half in Figure 33. The ball-paddle characteristics (ball paddle error (Figure 33A) and ball impact error (Figure 33B)) for these two groups are shown. Note for both the situation of the ball approaching player (red) and ball approaching opponent (magenta), the relationship between player paddle (eye controlled) and ball is analysed. The difference between the two sides is very clear. For the player side (red line) the ball-paddle error rapidly decreases towards impact (Figure 33A), while on the opponent side, little trend is observed. For the endpoint error (Figure 33B), on the player side it decreases rapidly as the ball moves towards impact, suggesting tracking rather than proactive behaviour. On the opponent side there is also some degree of paddle movement towards the endpoint but it is much more localised to the point of impact. Combining this relationship with the lack of tracking shown in Figure 33A it suggests there is some movement towards the impact point, but not in concert with the ball. Thus overall, significantly more ball tracking behaviour is observed on the player side than on the opponent side. However this is looking at global averages and thus may hide different component strategies.

The global average of strategies across all subjects gives a very general trend of behaviour and clearly distinguishes the player's gaze behaviour when receiving versus observing the ball impact. However, there is significant variation in behaviour (~ 100 pixels SD) which is hidden by the aggregation. Thus this variation is explored by clustering the trajectory data.

To do this each trajectory becomes one data point in an 800 dimensional space, where 800 is the length of the concatenated ball paddle error and impact paddle error (see Methods, Figure 25). All approach trajectories are pooled (i.e. both gaze behaviour during the ball approaching player trajectory and when the ball approaches the opponent) and thus clustering is completely unsupervised. The Dendrogram for the hierarchical clustering is shown in Figure 34. To give a reference, at the top of the figure two extremes strategies are shown: 1) Purely proactive, i.e. the player estimates the impact point of the ball and moves the paddle there proactively of the ball arriving. This gives a decreasing paddle-ball error as the ball moves in towards the proactively placed paddle, while the impact error is low (assuming a good endpoint estimate) and constant. 2) Purely tracking, i.e. the player looks at the ball and tracks its movements and consequently the paddle tracks the ball. This results in a low and constant paddle-ball error with a decreasing ball-impact error which decreases as the ball moves towards the impact point.

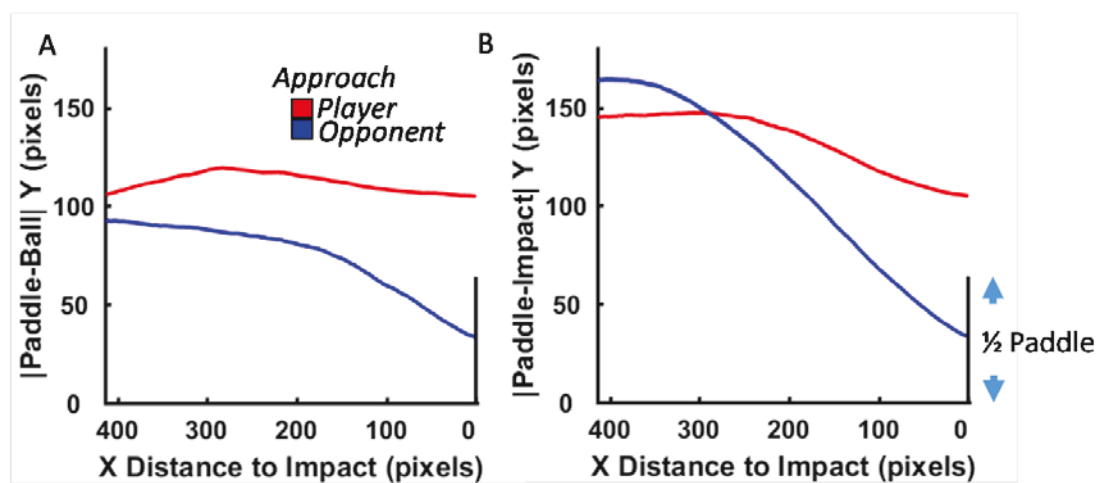


Figure 33: Global Average Strategy for making a return. A. Average absolute player paddle ball error for returns on the player side and opponent side. B. Average absolute player paddle impact position error for returns on the player side and opponent side. Note standard error is shaded but very small over such large numbers of trajectories.

Referring again to Figure 34, at each node in the hierarchy the cluster centre is plotted. In addition the proportion of player trajectories and opponent trajectories within the class is shown. Of the player trajectories, the proportion of misses versus returns is also shown for each cluster. For each leaf node (at the bottom of the hierarchy), these metrics are plotted as well as the overall cluster membership frequency (bar chart). In general, purely proactive strategies emerge, which are more likely on the opponent side, purely tracking strategies, which are more prevalent on the player side, a combination of the two (potentially) and finally what appears to be preparation strategies which are most prevalent on the opponent side. The following describes a walk down the hierarchy to describe the emergence of this behavioural taxonomy.

The global strategy, seen at the top of the dendrogram in Figure 34(1) resembles tracking, with a flat ball-paddle error and decreasing endpoint error. Overall there are slightly more hits than misses and slightly more trajectories approaching the opponent. Moving down the dendrogram, this cluster is split into a predominantly proactive (2) strategy in the left branch and a predominantly tracking (3) strategy on the right branch. Following down the right hand branch first, as it is closest in Euclidean pixels to the global strategy, the

tracking like strategy (3) is observed. The tracking error is on average higher than a purely tracking approach and it also has a slight increase towards impact. The proportion of misses and hits remains similar with a slight increase in the proportion of misses. Continuing down the right side of the dendrogram, a strikingly clear tracking strategy (4) and what appears to be a strategy more proactive in nature (5) emerge. The tracking strategy is more likely on the player side than the opponent's and results in more hits than misses. This strategy, when followed to the leaf nodes (6,7,8), contains two clear tracking strategies (6,8) which differ only in the initial amplitude of the impact error. This suggests the starting position of the ball in relation to the paddle is larger. In addition the tracking error increases slightly towards impact, similarly to the great grandparent node (3). This may be explained by a bounce occurring close to impact position making tracking more difficult. Again with all members of this tracking strategy family (6,8), there are more hits than misses, with this behaviour occurring predominantly on the player side. The other strategy in this leaf node triplet (7) is neither tracking or proactive, tending to occur predominantly on the opponent side, though when it occurs on the player side it is much more likely to result in a miss. The shape of this cluster seems to suggest an accurate sampling of gaze position half way along the trajectory (perhaps at a bounce point) but low accuracy at impact point. This may suggest preparatory sampling of ball position in preparation to predict the return bounce from the opponent.

Retracing back to the third level and instead dropping down the right hand branch to (5) the cluster centre looks proactive with a constant ball-impact error, but rather than a decreasing ball-paddle error this actually increases towards impact. This suggests the paddle has moved to the opposite location to the impact position. Interestingly, this strategy is much more predominant on the opponent side which may be an effect of the player moving their paddle to receive the ball currently being returned by the opponent. If this strategy is used on the player side, inevitably a miss is considerably more likely. Following this branch to the leaf nodes, three predominantly opponent side strategies appear (10,11,12), that if used on the player side result in a miss, but on the opponent side may indicate preparation to receive the ball.

Finally, returning to the root of the dendrogram and then tracing down the left hand side to (2), the very clearly proactive strategies emerge (14,15) with the tracking error decreasing to impact, while the impact error remains relatively constant at a low level of error. Of these two strategies the first (on the left hand side) occurs with higher probability on the opponent side and when occurring on the player side has a high probability of return. The other proactive strategy occurs even more predominantly on the opponent side and the impact error begins to increase towards impact. Given the majority of these trajectories arise on the opponent side, the error may begin to increase as they begin to prepare to receive the opponents return. Finally in addition to these two proactive strategies, a strategy which resembles neither purely tracking nor proactive behaviour emerges, with broadly flat trajectories in both the ball-paddle and ball impact error. This may indicate a combination of the two strategies, for example the maintaining the paddle position half way between predicted impact and current ball position would create such a shape. In addition, the shape of this trajectory also suggests the situation where the ball moves on a more horizontal trajectory and thus the change in vertical position is less extreme.

To summarise, 4 main families of strategies emerge, the purely proactive (14,15), the purely tracking (6,8), a combination of the two (16) and finally the preparation strategy (7,10,11,12) (or miss strategy when applied on the player side). The purely proactive strategies occur predominantly when the ball is moving towards the opponent while the tracking strategies occur more predominantly when the ball moves towards the player. This approach is supervised and puts no assumptions on the data allowing these data driven strategies and associations to emerge.

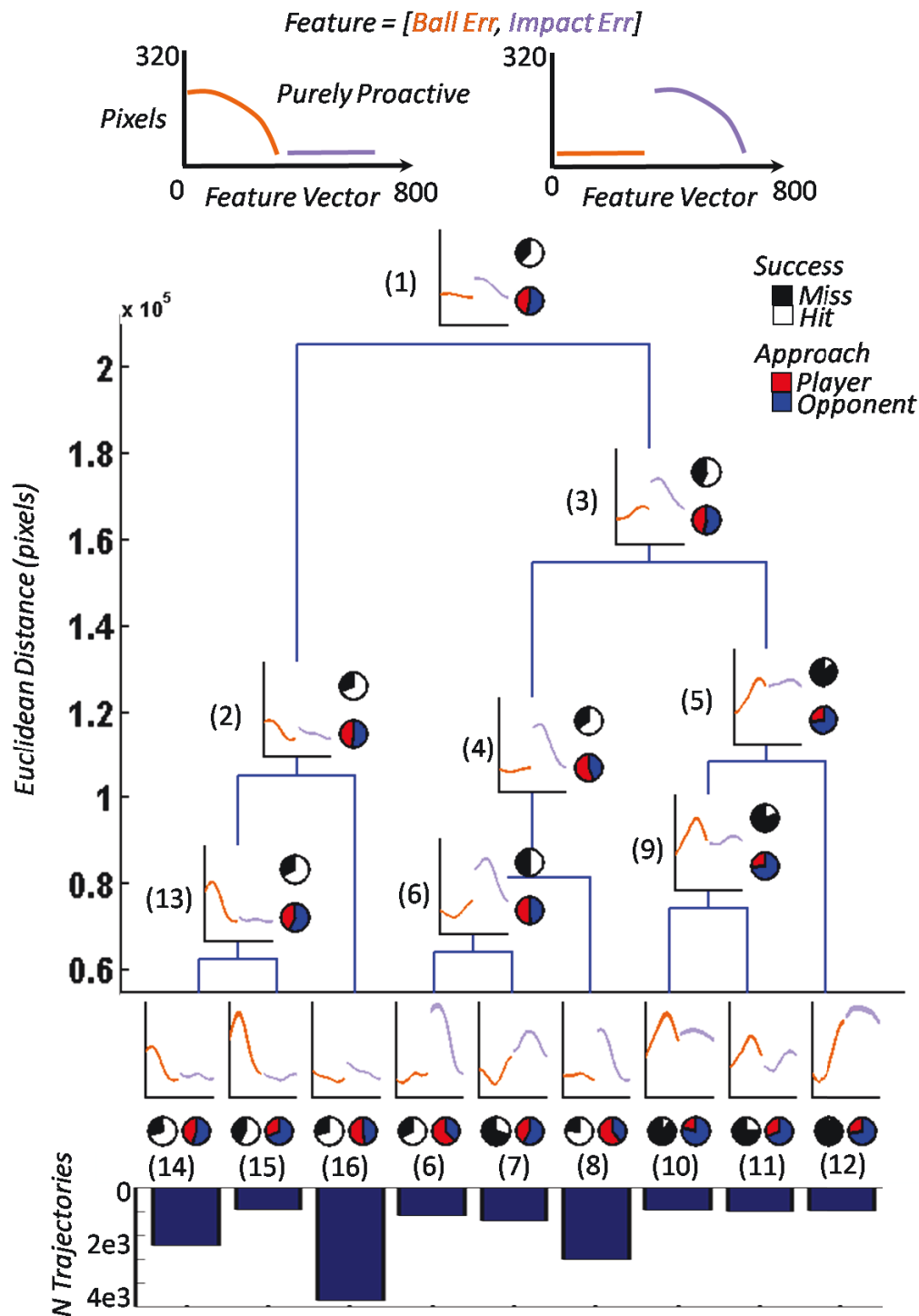


Figure 34: Player Strategy Dendrogram. A. Idealised strategy cartoon showing the purely proactive and purely tracking response in terms of the feature vector. (see the Methods and Figure 25). B. Dendrogram with cluster centres, pie charts showing proportion of players vs opponent side trajectories and the proportion of player side misses and returns. Note: the feature vector is always based on player paddle behaviour, no analysis of the opponent strategy is being made here. The numbering is used within the text to explain the taxonomy of strategies. The bar chart represents the frequency of each leaf node class membership. To summarise, 4 main families of strategies emerge, the purely proactive (14,15), the purely tracking (6,8), a combination of the two (16) and finally the preparation strategy (7,10,11,12) (or miss strategy when applied on the player side).

4.3.6 Modelling Player Strategies

Having identified these main strategies in an unsupervised manner, these findings allow a modelling approach to fit the observed behaviour with a linear combination of purely proactive and tracking strategies. The outcome gives a finer measure of behaviour than a cluster assignment. The modelling process is demonstrated in Figure 35A. For each ball trajectory, the ideal feature vector ($|ball\ error|$, $|impact\ error|$) for purely tracking (X_1) and purely proactive (X_2) strategies is calculated as shown in Figure 35A (refer to methods Figure 25 for feature vector explanation). The actual subject trajectory feature vector is then fitted to be a proportional combination of the proactive and tracking feature vectors, this linear model is expressed in eq. 4.

$$\begin{aligned} \alpha X_1 + (1 - \alpha) X_2 + c + \varepsilon &= \hat{X} \\ \varepsilon &\in \mathcal{N}(\mu, \sigma) \\ c &\in \mathcal{R} \end{aligned} \tag{4}$$

For each trajectory the scalar parameters α and c are fitted while the residual ε is assumed to be normally distributed. The parameters are fitted by minimising the absolute error between the model and the actual feature vector. This was chosen rather than least squares to give a more robust fit. Figure 35B shows an example ball trajectory (black line) and paddle movement (blue line). For this trajectory the feature vector is shown (blue line) in the second column with the ideal proactive feature vector (green line) and the ideal tracking feature vector (red line) and the model fit (magenta). The mixture coefficient (α) and the quality of the fit (r^2) are also shown. This example (Figure 35B) is for a tracking strategy and α is 0.96, thus indicating a predominantly tracking trajectory. Further examples include (Figure 35C), an example of a highly proactive strategy with α fitted as 0.25. and (Figure 35D), an example of a strategy that has both tracking components ($\alpha=0.4$). The relative mixture coefficient (α) of tracking versus proactive can then be compared between the opponent side and the player. Figure 36A shows the relative frequency distribution of R^2 values for the player (gaze control - red) and the opponent (computer or joypad - blue). In general the distribution is similar with more R^2 greater than 0.5 for both player and opponent, though the modelling fit is slightly better for the player side. The alpha distributions are compared for the player versus the opponent, where only trajectories with a model fit R^2 greater than 0.5 are considered. This is shown in Figure 36B and shows more higher values of alpha (suggesting tracking) for the player side than the opponent side (suggesting proactive) and thus confirms what was observed in the unsupervised clustering results.

Overall there is a tendency towards tracking to hit the ball and proactivity when observing the ball being hit. However, there are also learning effects observed as the α value decreases over trials. This can be seen in Figure 36C where the average α z-score is plotted against the approach number, irrespective of whether a miss or a hit occurs. The average z-score starts higher on average and then reduces with experience. This is observed in both player and opponent trajectories. This suggests initially subjects track more but then become increasingly proactive with experience. This does not negate the result that the player side is predominantly tracking and the opponent side is predominantly proactive as the changes are subtle, though the effect is clear. This suggests very fast learning, after just 3 or 4 trials the strategy levels off.

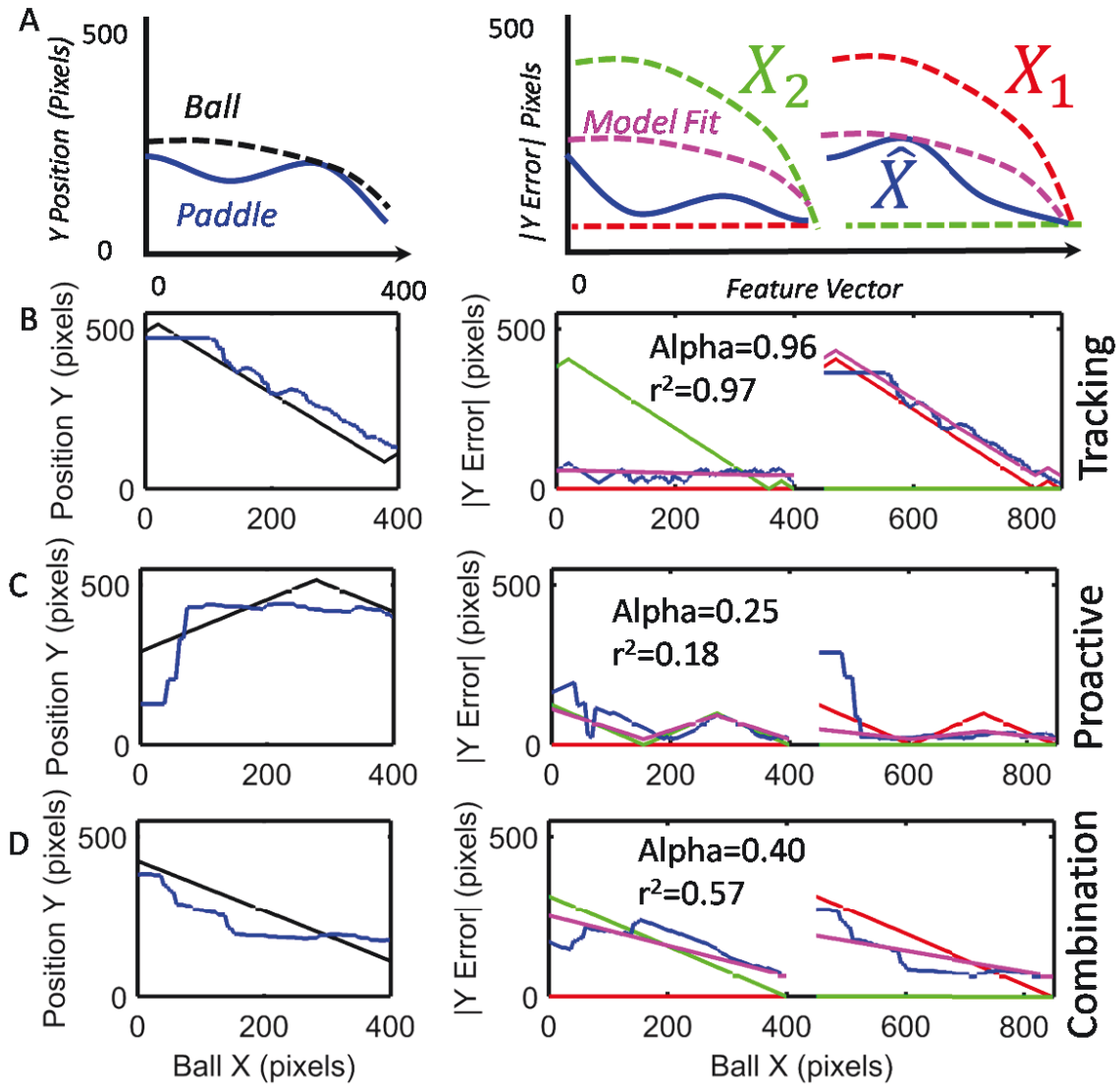


Figure 35. Modelling gaze strategy. A. Cartoon demonstrating the fitting of the linear model. The raw paddle (blue) and ball trace (black) is shown in the left hand plot, this is then transformed into the feature vector (blue) in the right hand figure where the purely tracking (X₁-red) and purely proactive (X₂-green) are calculated per trajectory and the linear model (magenta). Examples of the fit observed for different traces are shown: B. Tracking, C Proactive and D a combination.

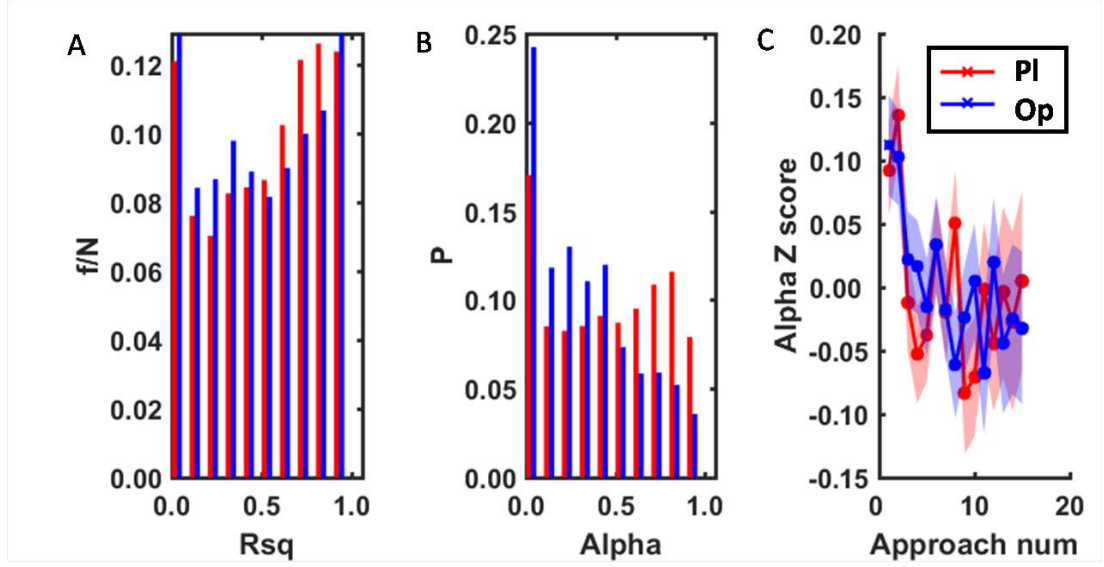


Figure 36: Modelling results. A. Distribution of R^2 for the player side (red) and opponent side (blue) trajectories. B. Alphas distribution of alpha fitted for all trajectories with $R^2 > 0.5$. Alpha is the tracking coefficient (refer to equation 4) and thus high alpha suggests a predominantly tracking strategy. C. Learning curve showing the change in alpha z-score for each approach number, averaged across all players.

Additionally, based on the model fit the performance of the player can be predicted. Figure 37 shows the steps to achieve this. For each subject the residuals between each trajectory and the fitted model is combined to give a distribution over residuals as shown in Figure 37A. A normal distribution is fitted to this distribution as shown in Figure 37B and eq. 5. Using this residual the probability of return is estimated as the integral of this distribution over the paddle width (eq. 6). This is based on the assumption that the residual represents the subject's control error. Based on this return probability, given the number of trials per subject, the expected number of returns can be estimated based on the maximum likelihood of the binomial distribution for these parameters (eq. 7). Figure 37C shows the prediction performance, with the actual returns for each subject plotted against the predicted returns. The prediction is highly correlated with the actual return number, with $R^2 = 0.94$. However, it is important to compare this to chance because the number of returns is proportional to the number of trials (a parameter of the binomial distribution). Thus to control for this, two methods were used to assign probabilities randomly to each subject (instead of using the residual). The first method samples probabilities from a uniform distribution for each subject and calculates the maximum likelihood of the binomial distribution for their respective number of returns. This results in the prediction performance of 0.72 and is shown in Figure 37D. The second method randomly samples from the subjects' estimated return probabilities and thus the same probability distribution is sampled from but the association with the subject is lost. Again the similar R^2 of 0.72 is achieved.

$$\text{Residual} = \mathcal{N}(\mu, \sigma) \quad (5)$$

$$P(\text{Return}) = \int_{-\frac{1}{2}pw}^{\frac{1}{2}pw} \mathcal{N}(\mu, \sigma) \quad (6)$$

$$\text{Performance} = \max_k B(a; n, p) \quad (7)$$

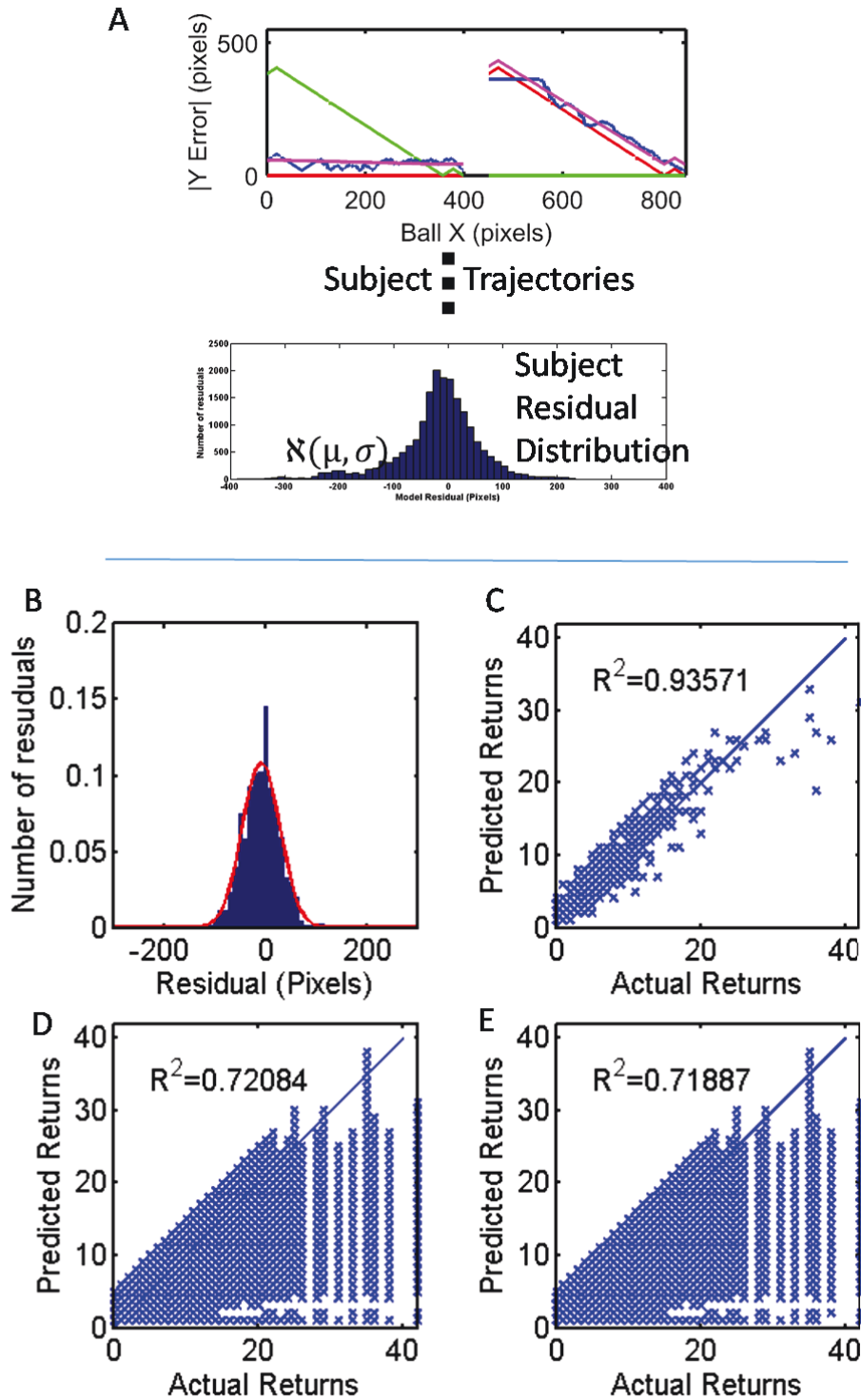


Figure 37: Model Performance Prediction. A. Residual distribution. For each subject the residuals between each trajectory and the fitted model is combined to give a distribution over residuals as shown. Using this residual the probability of return is estimated as the integral of this distribution over the paddle width. B Example of subject residual and fitted normal distribution (red line). C. Actual returns for each subject plotted against the predicted returns. The prediction is highly correlated with the actual return number, with $R^2=0.94$. D. Control 1: Return probability sampled from a uniform distribution. E. Control 2: Randomly samples from the return probabilities calculated and thus the same probability distribution is sampled from but the association with the subject is lost.

4.3.7 Pupil behaviour

The pupil behaviour, an indicator of cognitive load and arousal (Beatty, 1982; Beatty and Lucero-Wagoner, 2000), also shows a clear difference during player versus opponent returns. The z-scored pupil diameter (standard error shaded) is shown for the different segments of the game in Figure 38. The pupil behaviour as a return is made is shown in Figure 38A (left eye) and B (right eye), with the magenta trajectory representing opponent return and the red trajectory during player return. Again a clear difference between pupil behaviour in the active control versus passive observation can be observed. The paired t-test significance of this difference is indicated by a * at each position along the trajectory. For the player return, the pupil dilates increasingly towards impact and for a period afterwards before dropping off again. This is in line with the expected dilation of the pupil with increasing cognitive load and arousal and the slight delay in this response causing the continued increase for a short period after impact. The opposite relationship is observed during opponent returns where the minimum for pupil dilation is slightly after the point of impact, suggesting least arousal and cognitive load. Surprisingly, the pupil behaviour even shows a significant difference during active control for trials when the participants missed compared to where a successful return is made as shown in Figure 38C (left eye) and D (right eye). This may indicate the level of engagement in the task being lower for the missed balls compared to the returned.

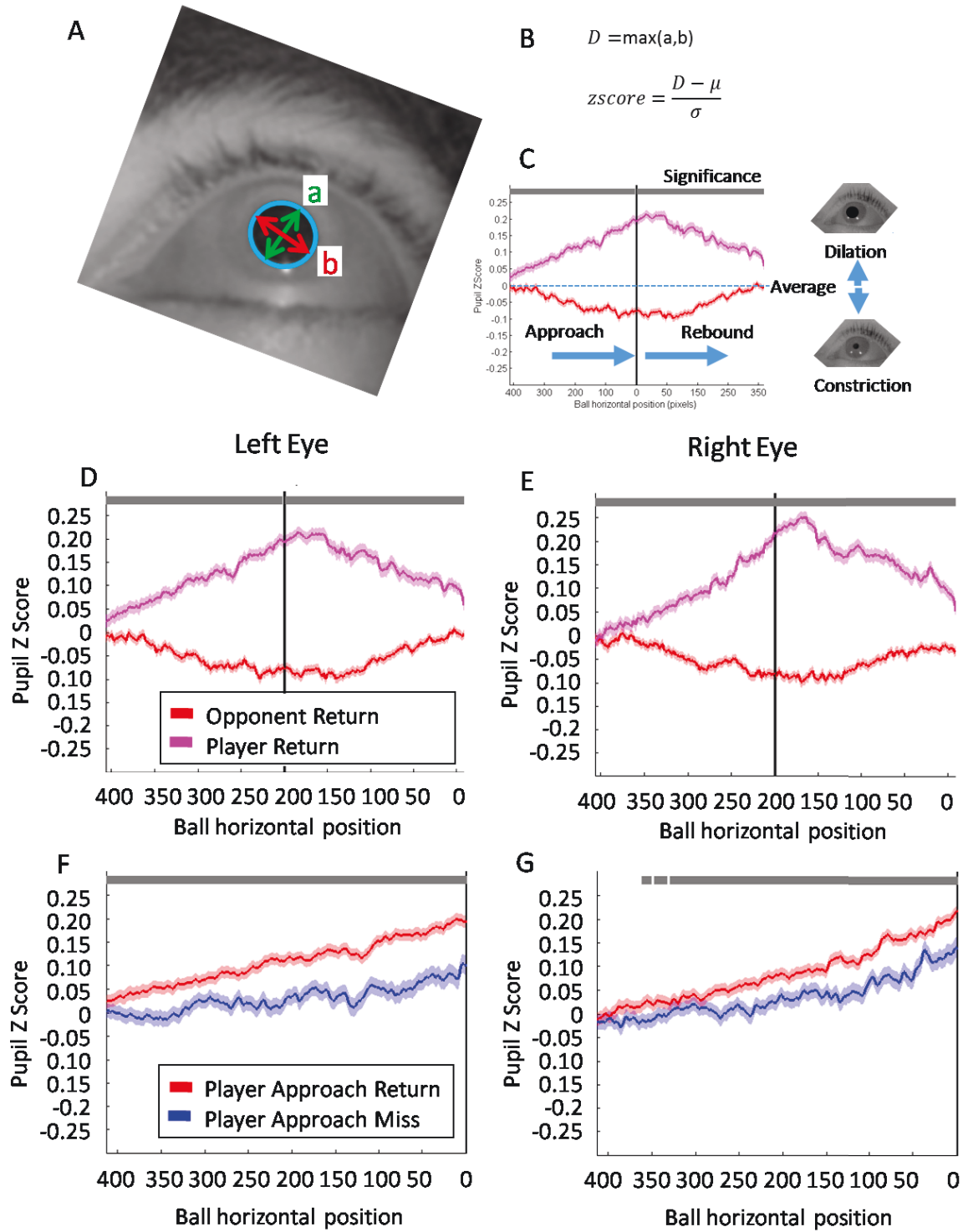


Figure 38: Pupil dilation relative z score for the different segments of the game. A. Pupil image and the extracted ellipse characteristics a and b. B. The maximum ellipse axis is always used to minimise perspective effects and this is z scored across each trajectory. C. Cartoon showing the interpretation of the following plots. D. Left eye z score averaged during all opponent returns and all player returns. E. Similar to D for right eye. F. Left eye z score averaged during all player returns and all player misses. G. Similar to F for right eye.

4.4 Discussion

The direct applicability of the gaze-based alternative to direct brain machine interfaces is demonstrated with the public facing experimental suite allowing over 2000 people to use the system outside the confines of the lab – the largest scale study of this nature. The participants were from a very broad demographic with participants aged from 4 to 80 years of age. Unlike previous real-world studies, the participant's task involved real-time control of a one-dimensional cursor to play the game Pong. Playing Pong requires closed loop control to move a paddle to hit a moving ball of uncertain target end-point. This makes the task much more relevant for prosthetic and wheelchair control than discrete tasks often used to test BMIs. The public facing "citizen science" platform is fully automated with no requirement of an experimenter to apply specialist equipment and the system automatically recorded the game state at 25Hz and the eye-movements simultaneously at 60Hz, providing rich information for post-hoc analysis.

The developed low-cost eye-tracker (GT3D) automatically calibrated over 80% of subjects successfully, with a novel method that optimally combined information from the two eyes to estimate on screen gaze position. For the 20% that failed calibration, reviewing the eye-images showed that this was largely due to subjects moving, the interface visor steaming up or in some cases due to strong eye-makeup or unusual eye-characteristics such as particularly small pupils. While the performance is already promising given the real-world application outside of the lab, the study also recorded eye images to allow future improvements to the eye-tracking generalisation. The database collected contains over 400,000 eye-images recorded for 1089 of the subjects for which 132 have age, gender and presence of contact lens information. In addition to improving the image processing and eye-tracking performance, this data set also has known gaze locations on screen and thus can be used to improve the mapping between pupil and gaze position on screen. The unnatural requirement to keep the head very still during play certainly affected the performance and thus it is important incorporate head compensation in future work.

4.4.1 Pong as a BMI benchmark

There have been a number of examples of pong being used to demonstrate BCI control. For example, Goebel et al. used functional magnetic resonance imaging (fMRI) to enable two subjects to play pong against each other (Goebel et al., 2004). It is difficult to make comparisons to this study because, although the stunt was famous and often cited, there was no paper published - just a conference abstract. Other examples of brain controlled pong exist, however in the majority of approaches, the BCI didn't directly control the paddle position. Instead, control was shared with either an additional player or an AI player that directed the paddle position, while the BCI player's concentration level modulated the paddle size and speed. Thus if their concentration lapsed, the paddle would become small and slow, thus making it very difficult to hit the approaching ball. This approach has been show-cased at a number of scientific engagement events, one of which I witnessed first-hand at the British Neuroscience (BNA) festival in 2013. The application of BCIs to control pong has also been mentioned in a number of publications e.g. (Blankertz et al., 2006, 2006; Krepki, 2004), however very little details were given as to how this was achieved or what the performance was. The only known work to give specific details of BCI control with pong was by Babiloni et al. (Babiloni et al., 2007). In this study, involving

a longitudinal training period spread over a month, the best 2 of 20 subjects went on to play pong with a performance of 96%. The training involved 5-10 sessions, lasting less than 2 hours each (factoring in time for electrode application etc). In this implementation the BCI input did directly control the paddle position. The parameters of the game are not reported (such as ball speed, game dimensions and paddle width) and no in depth analysis is made. Thus although the game has been used to demonstrate BCI performance on a number of occasions, it has never been used to give a clear and transparent benchmark for closed loop real-time control. Thus, here we propose using the game as a BMI benchmark, with a highly data-driven approach to evaluate, quantify and compare performance with a universally accessible benchmark. Other researchers can easily test the BCI they have developed with the data-logging version of pong developed here. This gives a benchmark that is directly relevant to prosthetic control with a format that enables clear comparisons between alternative approaches in a relevant real-time control task.

4.4.2 Game Performance

The proposed Pong benchmark developed provides 4 metrics to compare performance. These include the game parameters of successful returns and the game score as well as two alternative data-driven information throughput measures. This enables the performance to be compared to other BMI studies. Players were able to return on average 6.54 ± 6.1 shots successfully, compared to the empirical chance level of 2.6 ± 2.5 (see Figure 28A,B). With the highest performing subjects reaching return counts above 40 despite having never controlled an actuator before using their eyes. In addition 17% actually managed to beat their opponent, returning on average 10.4 shots for these games. To put the game performance into the context of other BMI studies, this translates to an average information throughput of 37bits/min based on Wolpaw's definition (Wolpaw et al., 1998b), almost three times larger than the highest performing large field study of EEG in literature (Allison et al., 2010). The system took less than a minute between the subject sitting down and being in control of the game, some 60 times faster than the EEG approaches. In addition, the tasks presented in the previous large field studies were discrete selection tasks and thus a further comparison is made to BMI studies with continuous cursor control tasks. An alternative ITR metric for continuous control is applied based on Fitts's law as applied by (Gilja et al., 2012). This metric goes much further towards capturing the difficulty of the Pong continuous control task and uses the raw control data to calculate it. This represents a novel method to capture an unconstrained control task with a data-driven information throughput mechanism. Using this metric the performance was on average 300 bits/min, again almost three times larger than the highest performing study to also apply this metric (Gilja et al., 2012). Gilja et al (2012) involved an invasive primate BMI, with a set up and training time 6 orders of magnitude higher. In addition, the task in this study was constrained with 6 fixed deterministic targets in the classical centre out reaching task and thus there was no target uncertainty.

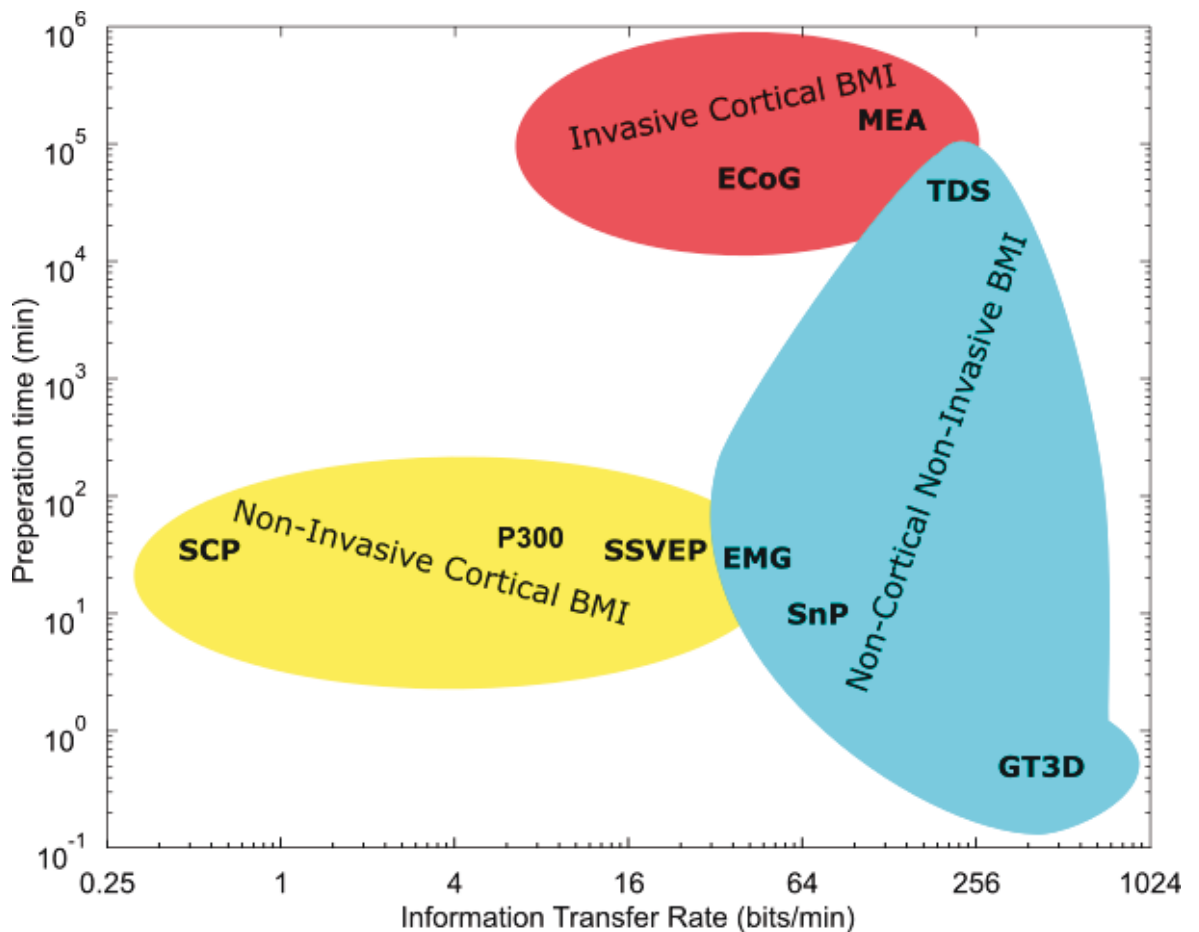


Figure 39: Comparing information throughput with training time. Data based on the following publications: Slow SCP, N=99 (Guger et al., 2003), P300, N=100 (Guger et al., 2009), SSVEP, N=106 (Allison et al., 2010), EMG (Kamavuako et al., 2014), SnP and TDS (Kim et al., 2013), MEA (Gilja et al., 2012) (all with $N < 10$) and finally the approach presented in this chapter ($N=651$).

Other important considerations for the practicalities of clinical BMI application, as well as information throughput, include the invasiveness to the patient and the time and effort required to set up and learn the interface. Previously a comparison of different BMI approaches has been presented based on cost versus information throughput (Abbott and Faisal, 2012). This comparison demonstrated that the GT3D approach had a very low cost per bit of information throughput ($\$0.71 \text{ bit}^{-1}$) compared to EEG ($\5000 bit^{-1}) and intracortical electrode implants ($\$25,000 \text{ bit}^{-1}$). Here a comparison of the same technologies is presented, but this time the relationship between set-up and learning time against information throughput is shown (Figure 39). This gives the eye-controlled approach a training time per bit of information throughput as $3.4 \times 10^{-3} \text{ minutes bit}^{-1}$, 2 orders of magnitude less than EEG ($6.7 \times 10^{-1} \text{ minutes bit}^{-1}$) and 6 orders of magnitude less than intracortical electrode implants ($1.6 \times 10^3 \text{ minutes bit}^{-1}$). Although this metric is an oversimplification, it illustrates the important balance that must be struck for real patient acceptance. The amount of negative impact due to invasive surgeries, recovery time, aesthetic impacts and of course training time must be far outweighed by the functional restoration the interface provides.

One additional real-world consideration that is particularly important with gaze control is the Midas touch problem. This study involves a control task unaffected by the so called

“Midas touch problem” which describes the problem of distinguishing between eye-movements intended as active control commands versus those of passive observation. In the game Pong, the paddle moved in concert with the subject’s gaze throughout the game, which functions perfectly well in this context but would be less desirable when say controlling a wheelchair. Here the gaze is tasked with many other purely observational activities such as visual search and the attention may be drawn to external distracters unrelated to the active control of the wheelchair. As such in this situation, the Midas touch problem must be addressed. To do this eye-movement behaviour must be automatically recognised as either active control or passive observation to switch gaze control on and off. As discussed, the standard solutions include gaze dwell–(Jacob, 1990), which is still being optimised (Penkar et al., 2012); gaze blink to click (Majaranta and Rähä, 2002) and the wink variant developed here, presented in the previous Chapter and here (Abbott and Faisal, 2012). In addition many alternatives have been suggested such as face gestures such as smiling (Rantanen et al., 2012), teeth clicking (Zhao et al., 2012) or tongue movements (Buckley et al., 2011). Alternatively, learnt gaze gestures are used, such as looking at all 4 screen corners (Drewes and Schmidt, 2007), tracing out patterns (Rozado et al., 2012b) or looking off screen (Istance et al., 2008) to convey commands (Hyrskykari et al., 2012). However these gaze gestures can be difficult to learn and significantly increase the cognitive load on the user. With the eye-movement data and game states, the behaviour during gaze control can be analysed towards understanding characteristics of behaviour that distinguish observing the ball to gather information and that of actively trying to move the paddle to a desired location.

4.4.3 Behavioural Analysis

Post-hoc analysis of the eye-movement behaviour during game play shows that there is a significant difference in eye-movement behavioural strategies and pupillary responses between active control to hit a ball, versus passive observation of the opponent hitting the ball. This powerful result could provide a contributing signal to distinguish active control from passive observation and thus towards an implicit solution to the Midas touch problem. To achieve a reliable intention-decoding engine that can automatically detect the state of the subject, the features observed here would need to be combined with additional features of gaze characteristics. Though a simple game, there are many gaze behaviours involved including relocating gaze to the ball (based on parafoveal information or predictive information), extracting event timing and position information from key points such as bounces and during trajectories to enable end point prediction. The dynamics of the eye-movements can be broken into saccades, fixations and smooth pursuits in this task and the relative proportions of these behaviours, as well as their characteristics may indicate the which of these processes they are involved in. Combining these features with an understanding of the task and the behavioural components (semantics) discussed as well as their probable sequence (grammar), an algorithm could be developed that sampled the gaze information only at the time instance most likely to convey positional intention. To provide more insights towards this the following relates the behavioural strategy findings here with real-world eye-movement behaviour. One important caveat to consider here is that subjects fixed their head still during the experiment and thus the eye-movement repertoire elicited bears a pale resemblance to the complexities of real-world

head free behaviour. This behaviour must be understood to take such an approach into real-world applications, which is the subject of Chapters 7 and 8.

4.4.4 Game strategies relation to natural ball catching behaviour

The behaviour is analysed to distinguish gaze strategies when the ball is moving towards the player's paddle versus the opponent's. Through unsupervised methods different strategies emerge. In general when receiving the ball the player tracks the ball towards impact, while when the opponent receives the ball the player tends to move the gaze proactively of the ball. This conclusion has been drawn based both on unsupervised clustering of behavioural strategies as well as modelling based on these emerging strategies. These findings can be related to eye-movement studies involving natural ball catching (with the hands). Hayhoe et al. (2005) studied the task of catching a tennis ball that bounced once before the catch was made. They found that the subjects fixated the throwers hand, then just above future bounce point (average 53ms before bounce event) followed by smoothly pursuing the ball into the hands (Hayhoe et al., 2005). This reflects what is observed: proactive fixations to optimally gather information at the right time to then ensure the gaze is in the optimal place to initiate smooth pursuit of the ball to guide catching. Further it has also been shown that prediction is improved with this pursuit activity (Spering et al., 2011), suggesting this is an important component of the catching process.

A key limitation of such large field studies is the subject interaction is very short and thus there is not time to run additional comparisons and controls. Having observed the similarity between eye-based control strategies and natural eye-movements during catching, future studies should probe this comparison more directly with controlled lab experiments. It would be interesting to directly compare the eye-movements during manual game-play (hand-control input), eye-input and pure observation of game play. This would help to dissociate the eye-movement behaviour that is purely driven by visual stimulus from that associated with visuomotor control. However, given the behavioural strategies observed in this study, combined with the apparent one shot learning, the evidence suggests that the control mechanism applied here is natural and intuitive because it leverages the already learnt eye-movements from real life. This may make the task seem trivial - subjects do what they do naturally with little to learn, but it is exactly this approach that makes the interface so intuitive and effective. The interface facilitates natural gaze strategies to control the paddle. This captures and informs the essence of the proposed approach for gaze interaction for prosthetics and wheelchair control. The following Chapter takes gaze interaction away from the computer screen and into the real world for the control of wheelchairs and robotic arms using natural eye-movement strategies.

5 Intuitive gaze control: Wheelchairs and Robotic Arms

Acknowledgments. This work has been the collaboration over a number of Masters and Bachelors group projects I supervised. Various aspects of the wheelchair have been worked on by Sofia-Ira Ktena, Martin Seitz, Cecile Abdo, Kirubin Pillay, Giuseppe Pietro Gava, Samuel Oxley, Philipp Brunnbauer, Cyril Deroy and Maria Freitas da Silva. With work published on the virtual wheelchair: (Ktena et al., 2015). The Robotic arm projects have involved Marcel Admiraal, Shashin Lad, Pablo Tostado and Sabine Dziemian. With work published on both arm platforms (Tostado et al., 2016) and (Dziemian et al., 2016). This is not "first author" work however significant time was spent supervising these students and the work contributes important proofs of concept to the thesis. For brevity the important outcomes are summarised in a single Chapter.

Abstract. Previously gaze interaction has been predominantly screen based. While there are applications to real world tasks, these are often either gaze interaction via a graphical user interface (GUI) or involve unnatural gaze gestures that detract from natural gaze behaviour. Here the developed eye-tracking system (GT3D) is applied to controlling robotic arms, a virtual wheelchair and a real wheelchair. In addition, two recently released commercial low cost eye-trackers (£100) are tested. These interfaces are developed to utilise gaze information in the environmental context rather than via a computer screen GUI, with efforts to leverage, rather than impact on, the natural information gathering task of gaze allocation. Controlling a robotic arm enables a novel 3D calibration method to be developed, using the arm to calibrate the tracker - improving 3D gaze estimation by >60%. The virtual wheelchair demonstrates the instant ability to control a wheelchair with gaze and demonstrates the benefits of working with natural gaze behaviour, rather than

learning new gaze gestures. Each system is developed to test different gaze control strategies, with care taken to enable accurate behavioural, contextual, and performance data to be collected. Future application of these experimental suites will gather more extensive gaze behaviour during their operation with able-bodied inputs (e.g. joystick). This will allow new interfaces to be developed, conditioned on the natural gaze behaviour in the contexts of operation. The following summarises the current progress, with each experimental suite as a separate section that briefly summarises methods, results and conclusions.

5.1 Background

The GT3D system has been presented at length; here it is applied to the control of robotic arms and wheelchairs. Other research groups are working in this area of development but a key differentiation is that here; gaze interaction is in the real world, rather than via a computer GUI. The interface methods developed here aim to utilise natural eye movements rather than requiring new gaze gestures to be learnt. We naturally look where we are going and at objects we wish to grasp, providing highly informative information that can be extracted for controlling prosthetics and wheelchairs.

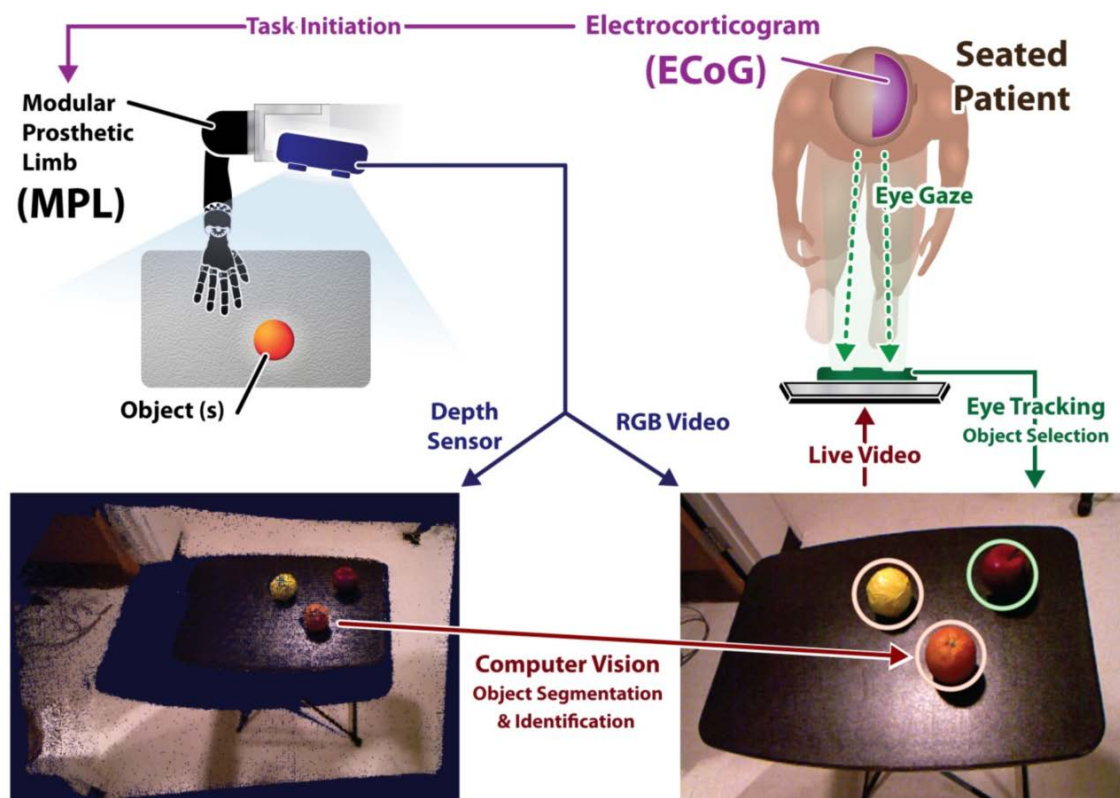


Figure 40. Example of hybrid ECoG and Eye controlled prosthesis from (McMullen et al., 2014).

Buckley and colleagues used gaze position to control a virtual arm prosthesis (Buckley et al., 2011) as a proof of concept. They demonstrated the system using a simple block sorting game. A number of virtual objects of different colours and shapes are presented which must be grasped with the correct grip type (out of 3) and then moved into the correct

colour coded position. The object selection is made using gaze position on the visual display (essentially a discrete decision between 3 options), and a tongue control input to choose the grasp type and initiate the grasp onset. A similar approach was used by Loconsole et al. to control a neuro-rehabilitation robotic arm guide (Loconsole et al., 2011). The target positions are chosen based on gaze proximity to an object in the scene camera and with an object selected, the environment camera is used to locate the given object in 3D to provide a target location to the robotic arm control system. Other promising approaches combine 2D gaze with either electromyography (EMG) (Corbett et al., 2012, 2013) or non-invasive electroencephalography (EEG) (Frisoli et al., 2012) and even electrocorticography (ECoG) (McMullen et al., 2014). The latter of these approaches is illustrated in Figure 40. Here the user interacts via live feed of the scene they wish to interact with displayed on a computer screen. Their on-screen gaze position is used to select the object they wish to grasp while the ECoG input is used to solve the Midas touch problem - initiating the grasp action. These systems provide promising solutions, however they all use 2D gaze estimation. They either use the gaze input via an onscreen graphical user interface or the 2D gaze input is with respect to a planar workspace. Thus in the work that follows, a major innovation is the application of 3D gaze for 3D endpoint control. In addition, a novel method of using the robotic arm to calibrate the GT3D system is developed which significantly speeds up calibration while dramatically improving performance.

5.1.1 Eye controlled wheelchairs

Gaze control of wheelchairs has been attempted with a number of different strategies. The majority of methods apply an indirect control mechanism, which again requires conscious eye commands to select movement directions displayed on a screen based graphical user interface (GUI) as shown in Figure 41A (from (Wästlund et al., 2010)). Variations on this theme can be found here (Barea et al., 2003; Gips, 1997; Lin et al., 2006). An example of the GUI used by Gips is shown in Figure 41B.

The major issue with this type of system is the conflict it creates between the eyes' natural sensory function and the control input. Attention must constantly switch between the GUI and the environment. As Lin et al. reports, this leads to significant fatigue (Lin et al., 2006). It is highly impractical as the user's attention is taken away from the environment they wish to navigate in order to make discrete screen based commands, before looking back to the scene to gain feedback. An improvement to this overlaid the control GUI on a video feed of the scene ahead (Wästlund et al., 2010) – this is shown in Figure 41C. Other systems remove the need for a GUI, using learnt command regions in the users field of view (Arai and Mardiyanto, 2011; Lin et al., 2006) or simply converting horizontal gaze angle deviation to control the direction of wheelchair motion (Barea et al., 2003). However these systems do not consider their impact on natural gaze behaviour during navigation and have not been benchmarked to assess their performance in comparison to other methods. Thus the aim is to develop a benchmark system and to compare existing approaches with the more natural approach devised here.

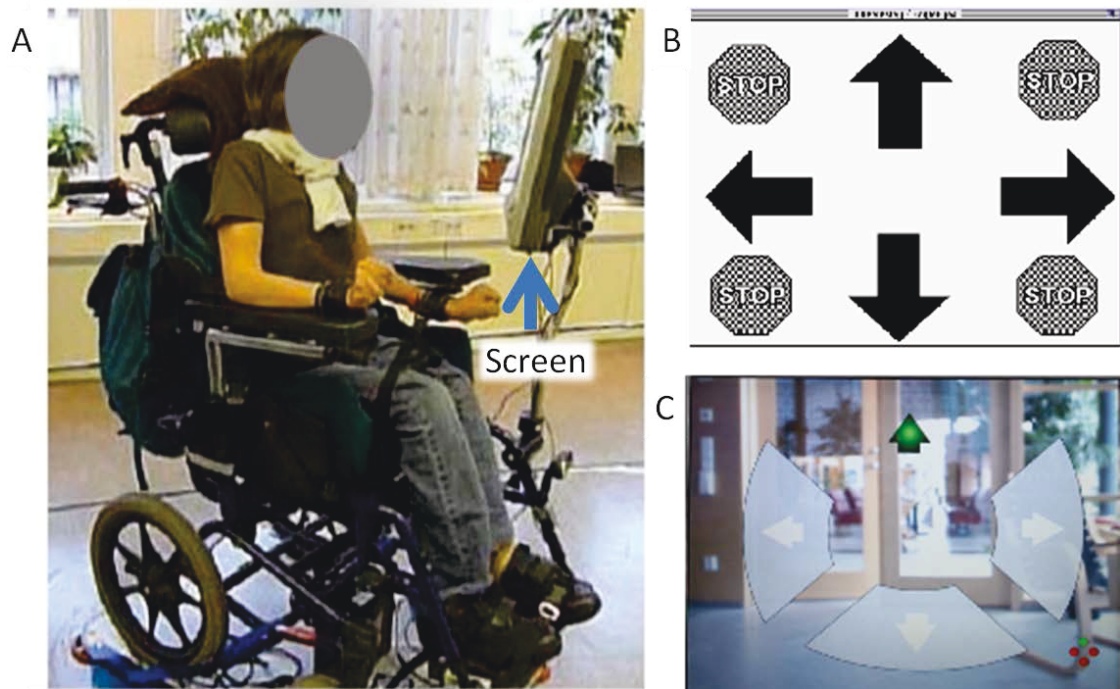


Figure 41. Gaze controlled wheelchair examples. A. Photo from (Wästlund et al., 2010) showing a user (anonymised) controlling the wheelchair via the computer screen (mounted ahead of user). B. Control GUI where gazing at the direction arrows moves the wheelchair accordingly (Gips, 1997). C. Eyes up GUI control interface which overlays the direction commands on a video feed of the world. This allows the user to monitor the scene and make control commands simultaneously. From (Wästlund et al., 2010).

5.2 Methods

The methods for four platforms are described: 1) 3D gaze calibration and control with a robotic arm, head free gaze control of a robotic arm, control of a wheelchair in virtual reality training environment and control of a wheelchair in the real world. The methods and results are summarised together to avoid confusion between the different platforms.

5.2.1 Eye-controlled Arm 1: Novel 3D gaze calibration and control with a robotic arm

The potential for 3D gaze interaction has been discussed at length with the presentation of the GT3D system in Chapter 2. Thus to make this reality, initial efforts took a low-cost robot system (Widow X) and integrated the GT3D to enable robotic end-point control in 3D. Figure 42 gives an overview of the hardware set-up. The integrated gaze controlled arm system consists of the remotely mounted GT3D cameras, three commercial OptiTrack cameras for motion capture and a WidowX robot arm (5 degrees of freedom (DOF), 0.4 m operation radius, 0.5Kg payload, measured endpoint error <2cm). The interface between the separate hardware and software systems is written in C++. In Figure 42A and B the front and side view of Widow X robotic arm, remotely mounted GT3D system and chin rest are shown. Figure 42C shows the subject in situ. At this stage a head fixed remote set-up was used to first characterise what is feasible without the added complication of head movement compensation, but in future iterations the optical tracking system can be extended to track head movements. Currently the robot workspace is monitored by 3 OptiTrack cameras as shown in Figure 42D. With the low cost robot used in this study, the

timing information and inverse kinematics of the robotic arm could not be relied on to accurately estimate endpoint (mean Euclidean error $\sim 2\text{cm}$) in a time synchronised manner and thus an optical tracking system is integrated to give accurate (1mm resolution) arm end-point information.

5.2.1.1 Novel high throughput 3D Calibration process

Calibration is normally achieved through a grid of calibration points to cover the workspace. To achieve this in a 3D workspace becomes impractical - with the standard 16 point grid becoming 64 points. This would be a very long and frustrating calibration process for a user, taking at least 2 minutes. To alleviate this issue, smooth pursuit eye-movements are used to rapidly cover the whole space in a fraction of the time, while collecting hundreds of calibration points. The robotic arm enables this new calibration approach, and thus a novel method to gather calibration data using smooth arm movements is used to gather approximately 600 calibration points per minute (compared to 16 in 30s with conventional screen calibration). Thus to collect comparable data-points with the conventional method would take almost 20 minutes, a completely impractical duration for the subject. In this pre-patient trial, $N=7$ able bodied subjects were recruited to take part in the study. Each subject was seated and placed their head on the chin-rest. Following fine adjustment of the eye-tracker the calibration routine was initiated. This involved the robot arm tracing out a predefined trajectory to cover the workspace. The subject then tracked an LED on the arm endpoint with their eyes. The pupil positions for each eye are recorded by the GT3D system and the OptiTrack records the robot arm end point. Figure 42C shows an example trajectory followed by the robot during calibration. The number of acquired points usually ranges between 500 and 1500 points, depending on the trajectory selected.

5.2.1.2 Gaussian Processes Regression

With these volumes of data more complexity can be captured in the model used to map between pupil and 3D gaze location without the risk of over fitting. Thus Gaussian Processes (GP) regression was used as it enables a data driven approach with little assumptions and constraints on the mapping. In addition it is probabilistic and thus it gives a confidence interval, which is very useful to filter anomalous results from the control signal. The calibration data is divided into training (70%) and test data (30%). Training is performed using the Matlab Gaussian Processes Regression for Machine Learning (GPML) toolbox (Rasmussen and Williams, 2006). The GP model selected for the system is comprised by (1) Linear Mean, (2) Matern covariance, (3) Gaussian Likelihood and (4) Exact Inference method. Once the hyper parameters are learnt, simple regression may be applied to infer the 3D gaze estimate from new eye-data. Further details and mathematical definitions can be found (Rasmussen and Williams, 2006; Tostado et al., 2016).

5.2.1.3 3D gaze interaction

Following calibration, the 3D gaze is estimated in real-time, providing a volumetric cursor for the control of the robotic arm. To solve the Midas touch problem (not all eye-movements should initiate an actuation command), dwell time is used to command the arm to move to a fixated location and a wink command opens or closes the gripper

depending on the current state. The dwell command is based on fixations that are detected when the pupil centre dispersion is less than 15 pixels for at least 500ms. Importantly the pupil deviation, rather than gaze deviation, is used as this gives a more reliable signal independent of the gaze depth. The dwell threshold can be adjusted on a per user basis and while this served this proof of principle, a more sophisticated solution is required in the future for long term operation.

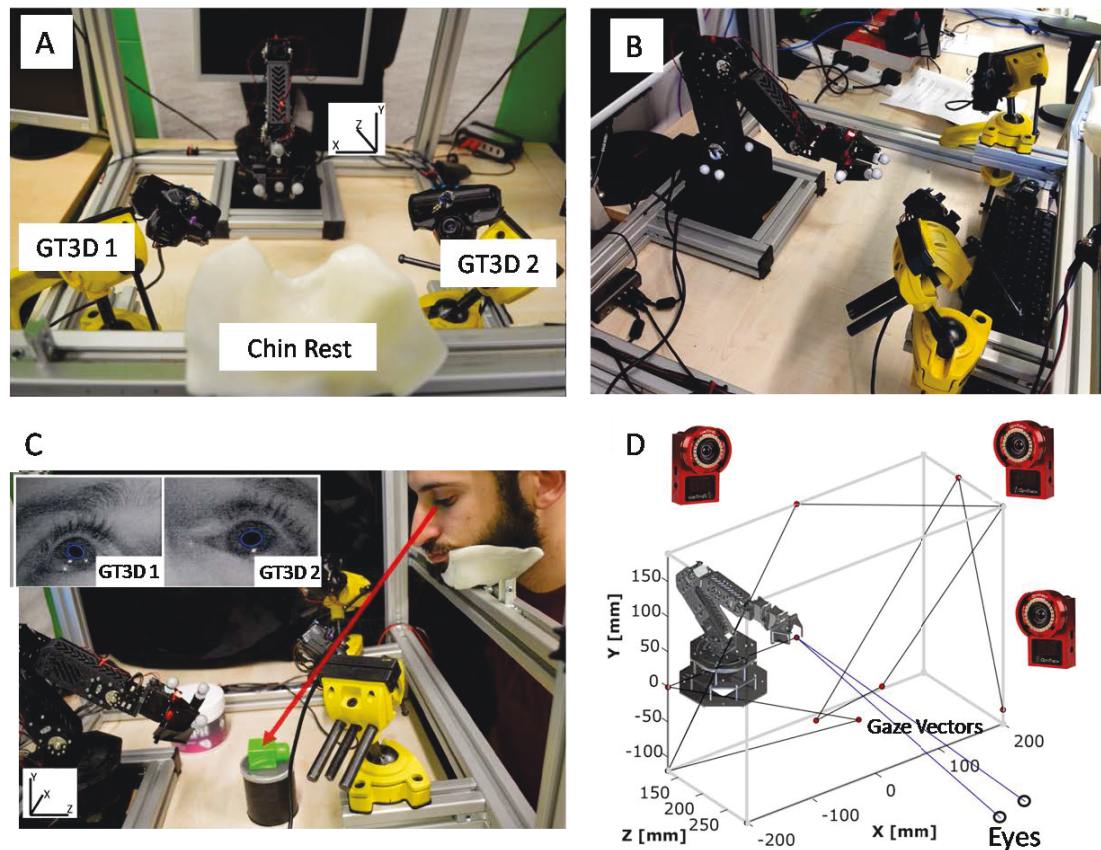


Figure 42: System Overview: A and B. Front and side view of widow ex robotic arm, remotely mounted GT3D system and chin rest. C. Side view with subject in situ during object interaction and GT3D feeds inlaid. D. Diagram from similar perspective to photograph in B. Diagram of robotic arm workspace from similar perspective to photograph B. Shows Optitrack camera system, an example calibration trajectory and the extracted gaze vectors.

5.2.1.4 Results

The 3D gaze estimation achieves an average Euclidean error $<2\text{cm}$ across 7 subjects. Figure 43A shows a 3D representation of the performance, showing the ground truth target positions (blue) and the 3D gaze estimate for each (red). The estimates are clearly very good in most cases, with a few situations where the estimates errors are large. Looking at the dimensions separately (Figure 43B,C,D - X,Y,Z respectively) it can be seen more clearly that the estimates are better in the X (horizontal) and Y (vertical) dimensions. While in the depth direction errors are larger, with larger GP confidence intervals. This is also reflected in the average performance across all subjects, where the depth error is larger and has a larger variability. The more extreme outliers tend to have a larger GP confidence interval and thus these can be filtered from the control signal.

5.2.1.5 Discussion

This system is the first published system to control a robotic arm end-point in 3D. The experimental suite developed enables a novel 3D calibration routine that uses the robotic arm itself to calibrate the 3D gaze with a rapid movement protocol. This overcomes the issue that with a pure 3D regression between eye and gaze, the number of calibration points becomes cubic to cover the additional depth dimension. In the only previous study that estimates (and reports) the 3D accuracy of gaze in the real-world, calibration involves a complex lab set-up with 2D calibration grids that can be moved in depth (Hennessey and Lawrence, 2009a). This hugely increases the calibration overhead on the user, making this approach cumbersome and unlikely to be acceptable to the user outside of the lab. Thus to overcome this, the novel approach developed uses a smooth arm trajectory to cover the workspace while simultaneously gathering pupil positions as the subject tracks the arm trajectory. This enables hundreds of calibration points to be collected in the same time it would take to calibrate a 2D system. This approach not only solves the calibration problem, but goes beyond it to collect much larger quantities of training data than conventional approaches, these new quantities are sufficient to train "big data" machine learning algorithms.

Based on this new approach the 3D gaze accuracy and precision is improved by approximately 70% and 60% respectively compared to the original GT3D algorithm ($1.6 \pm 1.7\text{cm}$ compared to $5.1 \pm 5\text{cm}$). In addition, these results make considerable improvement on Hennessey and Laurence's study (2009), where they achieved $3.93 \pm 2.83\text{cm}$ Euclidean error, with a considerably more expensive eye-tracking system and cumbersome calibration set-up. This powerful approach marks a big step to making gaze controlled prosthetics realisable. Firstly and fundamentally, it enables interaction with the 3D world as opposed to the confines of the computer screen. Secondly, the novel method for rapid gathering of calibration points enables a more data driven Bayesian approach, without a time consuming calibration. In the time taken to gather the traditional point by point 4x4 calibration grid ($\sim 30\text{s}$), this system gathers 300 data-points.

One major limitation of this approach however is that the subject remains static, with their head fixed. Thus while providing a proof of concept, it is important to understand the difficulties of translating the approach into the real-world where eye-movements, in combination with head movements are considerably more complex. Thus there remains significant challenges, both technical and scientific to take the promising approach into the real-world. Efforts to overcome this are made in real-world eye-tracking studies in Chapters 7 and 8. In addition, a head free approach is explored in the following section.

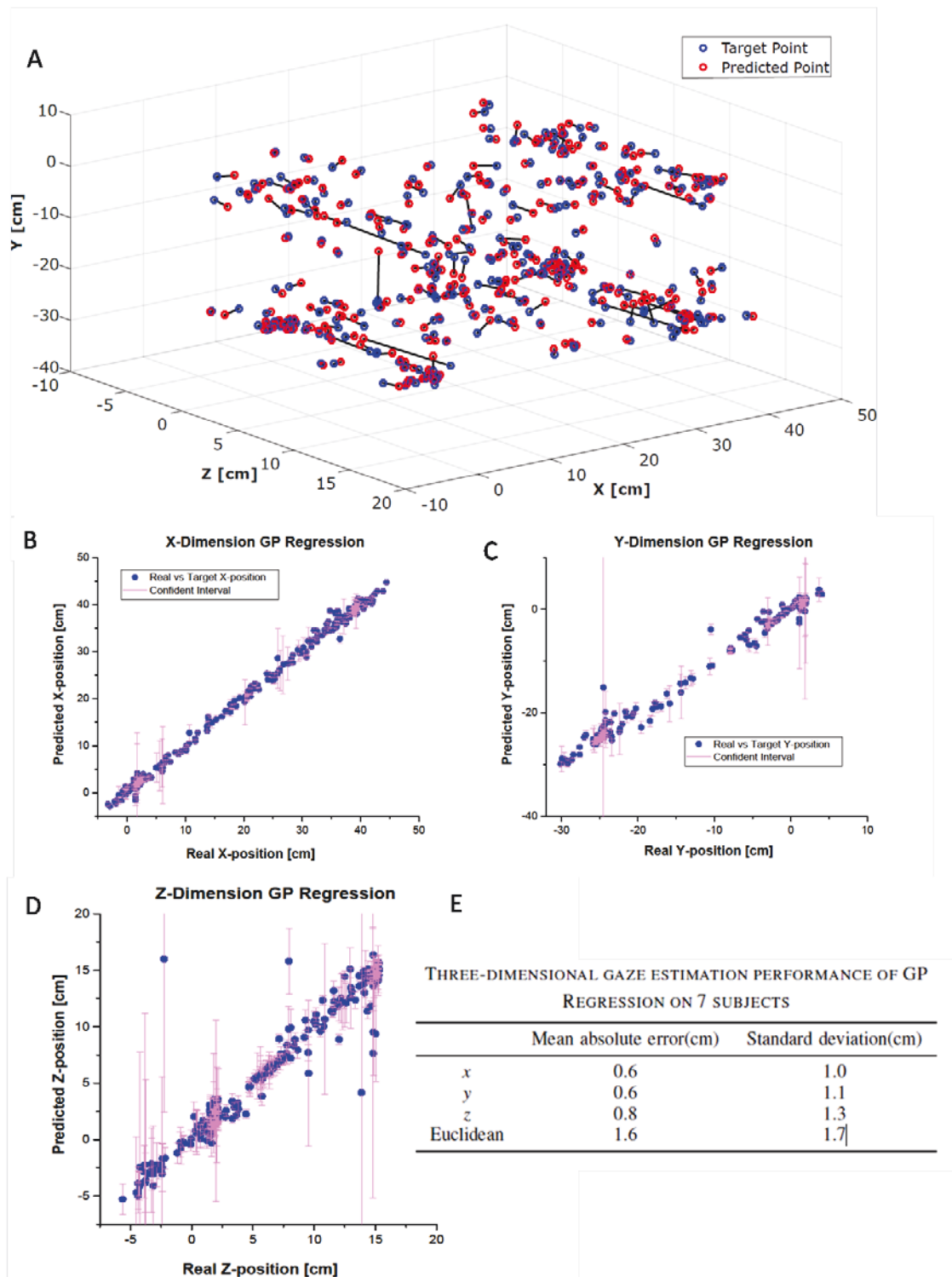


Figure 43: 3D gaze estimation performance: A. 3D gaze estimation results for 1 subject's test data (70% test and 30% training). Blue points represents the ground truth position and the red points represent the estimated positions. The black lines connect the associated ground truth and estimated points. B. X dimension ground truth versus estimated gaze position. The GP confidence interval is also represented with the magenta error bars. C and D: The same graph for the y and z dimensions, where x is horizontal, y is vertical and z is depth (refer to Figure 42). E. Summary of 3D estimation performance averaged across all subjects. The mean absolute error and the standard deviation of this error is shown for each dimension individually as well as the Euclidian error.

5.2.2 Eye-controlled Arm 2: Draw by gaze

The second robotic arm controlled system uses a considerably more sophisticated robotic arm (Universal robotics UR10, 6 DOF, 1.3m operation radius, 0.1mm position repeatability, 10Kg payload) and the head free remote eye-tracking system (Tobii Eye X, remote, 55Hz). Importantly, though an industrial robot, the UR10 is designed for collaborative operation with humans. Figure 44 shows an overview of the hardware set-up and system architecture. The subject sits at the robot workspace as shown in Figure 44A without any head worn apparatus and with their head free to move. An overview of the system architecture is shown in Figure 44C-F. The eye-tracker (Figure 44C) has an API (application programming interface) from which data is streamed to the main developer computer with a Linux operating system. The gaze data is processed (Figure 44E) and when a high level movement command (new intention position) is issued it is passed to ROS (Robotic Operating System) which plans and executes the appropriate robot movement (Figure 44F). The proof of concept task for the 3D robot control is to write letters in pen, requiring 2D information to make the character shapes, and a third input to move the pen to and from the paper to separate characters.

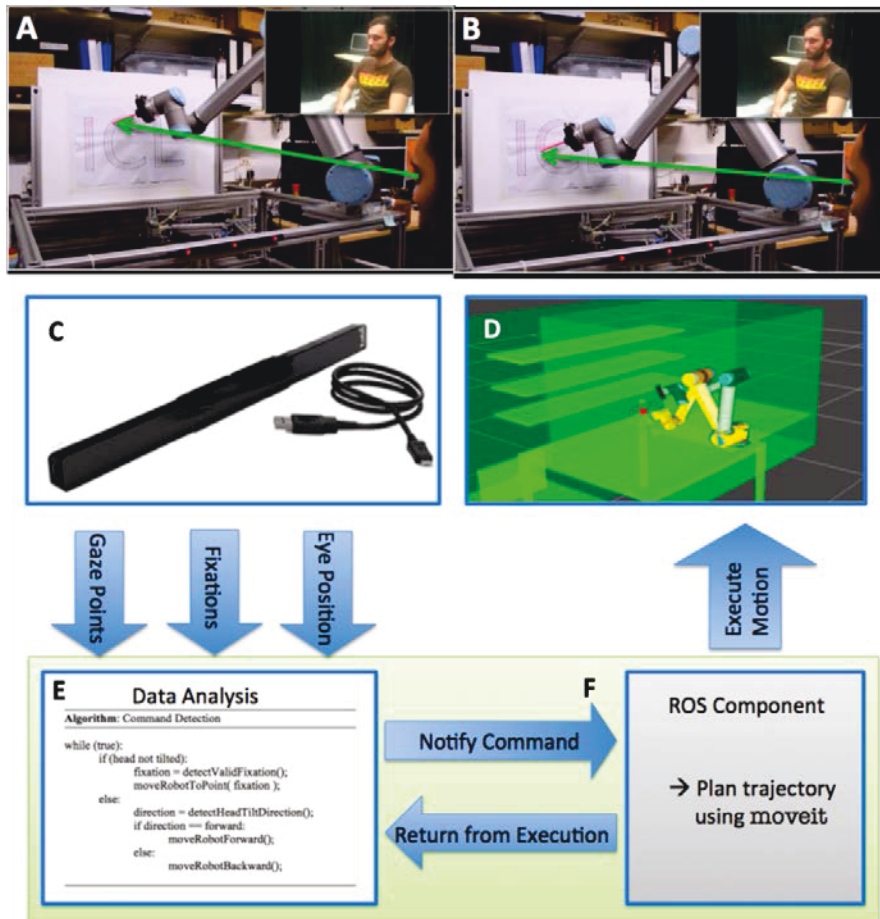


Figure 44: Hardware setup (A,B) and system architecture (C-F). **A.** Hardware setup. The robot arm (UR10 Universal Robotics) and eye-tracker (Tobii EyeX - mounted to the frame - black bar with three red lights, see also C) are mounted in front of the seated user (front view inset). The green arrow represents the gaze vector which controls the robot end-point to draw. **B.** Shows the movement of the robot as the gaze traces the stencil letter. **D.** Virtual model of robot arm and control zone. **E.** Online data analysis pseudocode for gaze control interface. **F.** ROS (Robot Operating System) component to relate gaze commands to plan and execute robot motion.

5.2.2.1 Gaze Calibration:

The Tobii Eye X remote eye-tracking system is a recently released low-cost (£100) system that tracks the eyes and head. It is designed as an additional input for computer control and thus only provides a gaze estimate in the plane of the computer screen. The API does not give access to pupil information or binocular gaze information, outputting a single 2D gaze estimate on the screen. Thus to enable operation with the robot arm a work around solution is devised that calibrates the user to the 2D robot workspace. This is achieved by subjects looking at a static calibration grid placed in the robot drawing plane instead of the computer screen (see Figure 45). Thus by scaling the screen estimate to the drawing plane, the control workspace is greatly enlarged and in the real world.

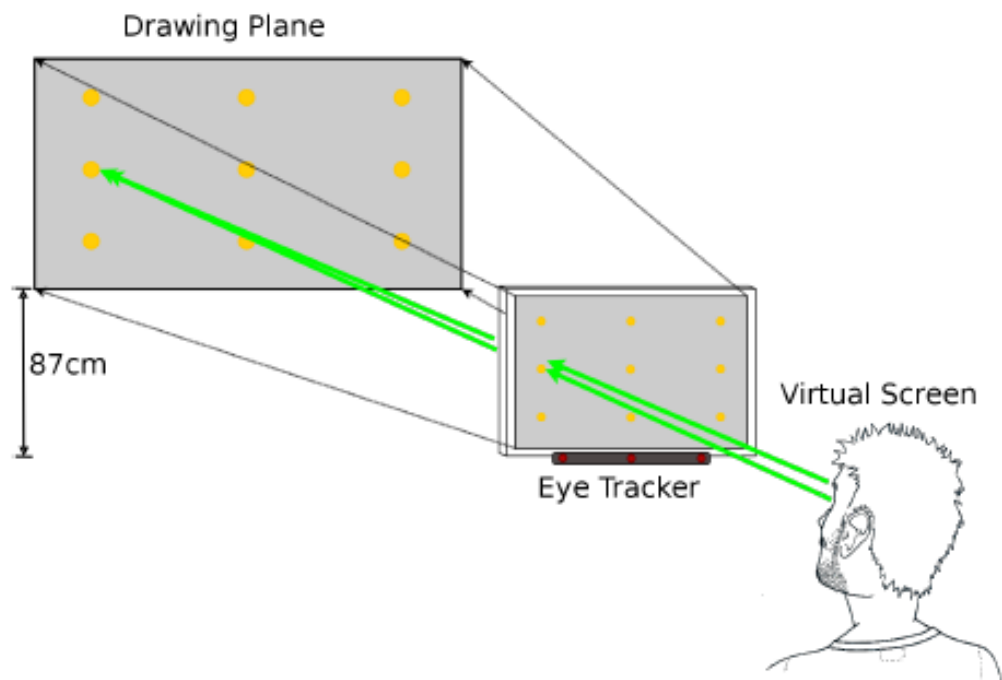


Figure 45: Translating the screen restricted EyeX gaze estimate into the robot drawing plane. As the calibration routine is running on the screen, the subject instead looks at the calibration grid on the robot drawing plane. Based on the geometry of the set-up the gaze estimate in the context of the screen is scaled to the robot drawing plane.

5.2.2.2 3D Robotic control

The robot position in the 2D drawing plane is controlled by gaze, while moving the head forwards and backwards controls the depth dimension, allowing the pen to be moved away from the page when moving from character to character when writing. To solve the Midas touch problem with the gaze input, fixation dwell time is used with a 600ms minimum threshold. The fixations are detected by the EyeX software (the algorithm for this real-time detection is proprietary). Currently the system is demonstrated as a proof of concept with the simple drawing task. The task to trace the stencil letters "ICL" is performed by 8 subjects with no prior experience using their gaze as a control input. The stencil (shown in Figure 46A) contains small circles which the subject must draw through to complete the task. Their challenge is to complete the task as quickly as possible.

5.2.2.3 Results

Figure 46C shows the gaze fixations (red) and robot motion data (green and blue) from a representative trial. For the robot movement the green represents when the pen is in contact with the paper and blue when it has been moved away. The Midas touch solution is effective, distinguishing the large number of fixations from the relatively few required robot commands. The robot clearly marks out the letters, even the more difficult curve of the "C" letter. The task durations for the first and second trial of each subject is shown in Figure 46B, demonstrating the subjects improve significantly, demonstrating the rapid learning of the interface.

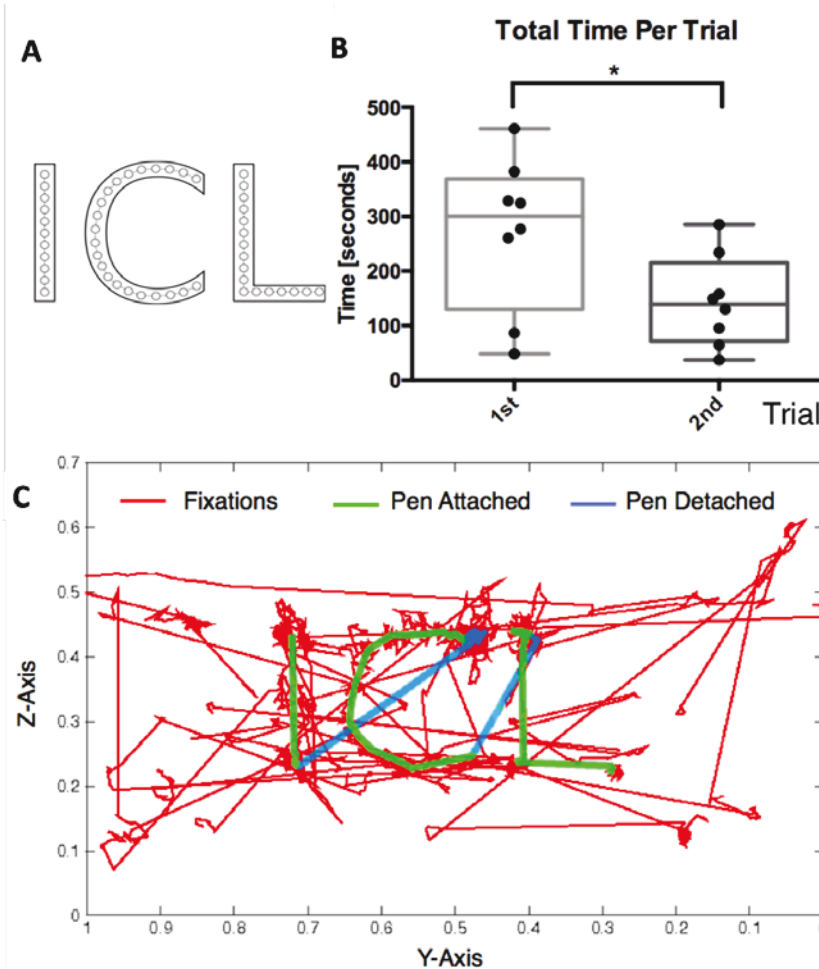


Figure 46 Draw by eye results. **A.** Shows the task stencil to write the three letters ICL. The aim is to write as fast as possible. Subjects must draw through all circles. **B.** The total task duration for the first and second trial ($N=8$, * $p < 0.05$, ** $p < 0.01$). **C** Shows gaze fixations (red) and robot motion data (green and blue) from a representative trial. For the robot movement the green represents when the pen is in contact with the paper and blue when it has been moved away.

5.2.2.4 Discussion

The system framework for gaze control has been developed and tested with a simple drawing task. The EyeX tracker used did not enable 3D gaze estimation and thus head movements were used to control the depth position. This exemplifies how the powerful

gaze input can be used in combination with other input modalities retained by patients. For patients who cannot move their head, this input could be replaced with an alternative input such as electromyography (EMG). It has been argued by reviewers of the paper that accompanied this work (Dziemian et al., 2016) that if head movements are used, why not just use the readily available head joystick. It is important to compare the gaze input with alternative strategies in future work. However it is expected that gaze would prevail as it has the huge benefit of sending a single endpoint positional command rather than a combination of directional movement commands to reach a target point. Thus the intention is communicated in the flick of the eye-rather than perhaps tens of discrete direction commands. Thus this end-point command is leveraged in two of the dimensions while the directional command is used only in depth. Additionally, the head movement was only required because the Eye tracker could not estimate 3D gaze. Thus the drawbacks of using the EyeX system were ameliorated and the power of gaze control is demonstrated without a chin rest or cumbersome head gear. In addition, the integrated sensor suite and real-time architecture is hardware agnostic and thus the GT3D system is to be integrated to improve the control capabilities.

5.2.3 Virtual Wheelchair Platform

A virtual reality wheelchair platform has been developed to allow the safe evaluation of methods for controlling wheelchairs. This allows new approaches to be tested in a safe environment that is fully instrumented to record progress, collisions and user behaviour. Furthermore, it provides a training suite for subjects to safely learn and acclimatise to a new interface before using it in the real world. Using this system, four alternative wheelchair interfaces were tested including two based on previous studies in literature, a novel method proposed here and keyboard input as a control. In addition, eye-movements are recorded during keyboard driving, to measure natural gaze behaviour and thus assess how the different methods detracted from this observed natural behaviour.

5.2.3.1 Methods

The virtual wheelchair system is built using the Unreal Engine, a software suite of integrated tools for game developers to design and build games, simulations and visualisations. This handles all the 3D graphics and interactions, including collision detection. A Virtual model of the department of Bioengineering was built from building plans as shown in Figure 47A,B,C. The system platform is shown in Figure 47E, which has a server client architecture where USARSim (Unified System for Automation and Robot Simulation) enables communication between the Unreal engine and external applications via game bots (virtual robots). Sensors can be added to the game bot (our wheelchair in this case) to track location and interaction with the environment. The front end controller can be any input modality. In the case of the gaze control interface, gaze information is processed and the appropriate command is sent to the simulator to move the virtual wheelchair. The GT3D eye-tracking system is used with 3 alternative gaze control interfaces which are described in the following section.

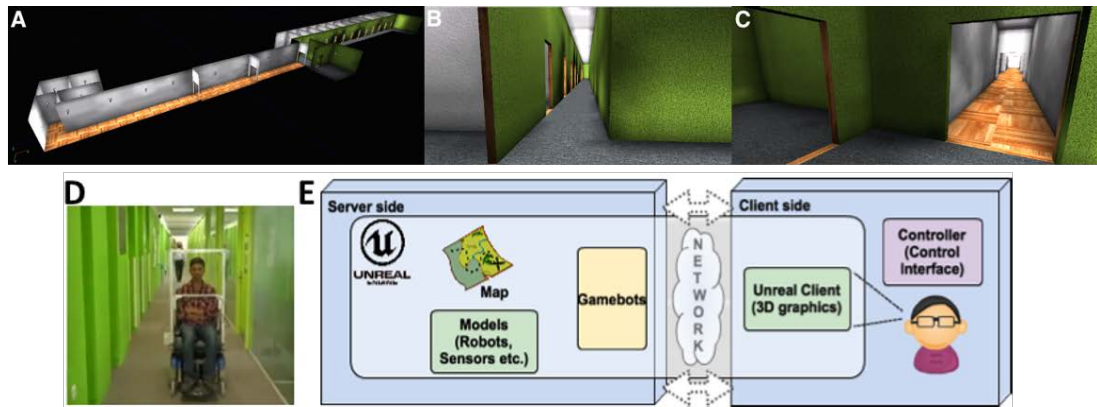


Figure 47. Virtual Wheelchair Overview: A. Virtual model of the department of bioengineering. B,C. Virtual wheelchair view of corridors in virtual map. D. Real wheelchair in department. E. Virtual reality platform architecture including USARSim (Unified System for Automation and Robot Simulation), Unreal Engine and user components.

Gaze interfaces: To translate gaze position into a wheelchair movement, command zones can be mapped to wheelchair heading vectors. This is represented by the vector field diagrams in Figure 48, where the arrows indicate the heading vector that will be sent to move the wheelchair. Two of the approaches are GUI-Free (graphical user interface) which means in real world operation they do not require a screen mounted on the wheelchair displaying a GUI (graphical user interface). This is a preferred approach as it does not take the user's attention away from the natural task of observing the world to guide navigation. Three gaze based approaches are piloted: 1) The Naive GUI-Free Controller, based on (Lin et al., 2006), which divides the pupil position into command zones where looking to the extreme upward position indicates the intention to drive forwards, downwards to reverse and left and right to turn (Figure 48A). In a real-world application, this is a screen-free approach as it does not require the user to look at an onscreen GUI (graphical user interface), however the commands are quite unnatural and thus it is referred to here as the Naive free-view controller. 2) Screen GUI Overlay Controller, is based on (Wästlund et al., 2010) where a GUI with direction arrows is overlaid on the view of the environment (Figure 48B). In the real-world application a real-time video feed of the environment is displayed on a wheelchair mounted screen and the GUI is superimposed on this feed. Here in the virtual set-up the overlay is across the subject's full view of the environment. 3) The Natural GUI-Free Controller proposed here. This controller attempts to create a more natural interface that allows the user to look where they want to go, as they would naturally. This is achieved with a continuous control field as shown in Figure 48C that smoothly transitions between left and right, while looking further into the distance results in an increased heading speed. This works with the natural gaze behaviour, which can be illustrated with the following example. If the subject looks at a desired location, for example to a position ahead and to their right, the wheelchair will rotate to the right while moving forward, if the user maintains fixation on their desired location the eyes will smoothly compensate for the wheelchair motion and move to the centre of the flow field, transitioning smoothly to a forward heading until the desired location is reached. In all interfaces discussed, gazing in the "IDLE" regions indicated in Figure 48 will stop the wheelchair. In other cases a command will be initiated as soon as the gaze position is within a control zone (without minimum dwell).

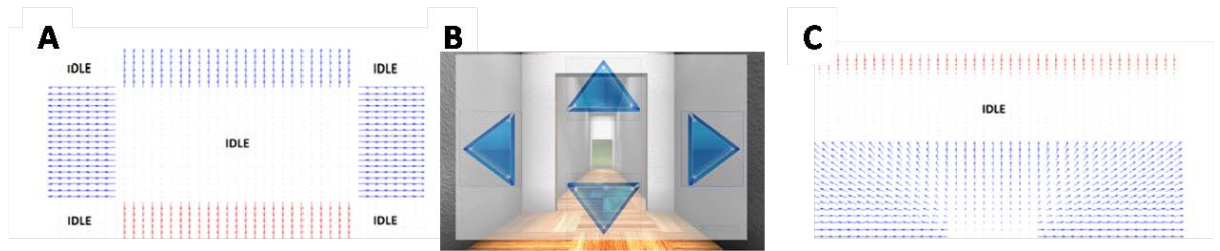


Figure 48: Vector field representation of control interfaces where the arrows indicate the heading vector that will be sent to move the wheelchair. A. The Naive GUI-Free Controller based on (Lin et al., 2006) where looking to the extreme up-down-left and right initiates then respective commands. B. Screen GUI Overlay system, based on (Wästlund et al., 2010) where a GUI (graphical user interface) with direction arrows is overlaid on the view of the environment. C. Natural GUI-Free controller proposed here.

Experimental Protocol: The interfaces described are tested by 6 able bodied subjects (24-27, 2 female and 4 male) who had never used gaze as a control input. A fourth manual keyboard interface is also used as a control during which the free eye-movement behaviour is recorded (arrow keys forwards, backwards, left and right). Subjects were seated in front of a screen (53x30cm) and placed their head on a chin rest. The GT3D system was remotely mounted in front. Each interface involved a short explanation, a calibration routine (standard 3x3 calibration sequence) and then the virtual wheelchair application appeared with a first person view of the environment and the task began. For each test subjects were given 4 four check point locations to navigate to, covering approximately 150m. Subjects were already familiar with the environment as they were students in the department. Following the experiment, the calibration routine appeared again to assess the stability of the gaze tracking.

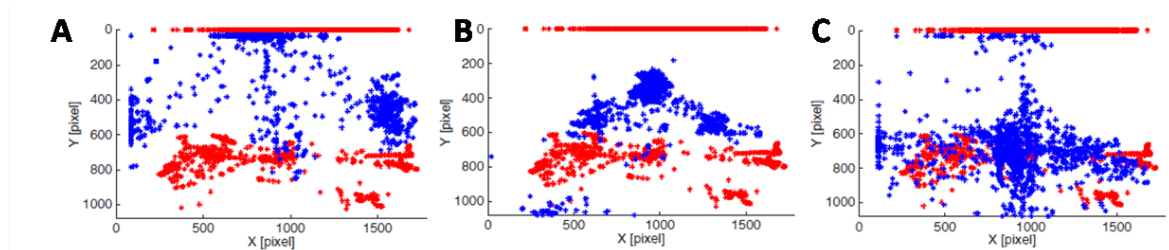


Figure 49: Gaze position distribution comparison between each interface (A - Naive GUI-Free, B - Screen Overlay GUI and C. Natural GUI-Free controller) in blue and natural free viewing during keyboard control in red.

5.2.3.2 Results

The interfaces are compared based on 3 metrics: 1) Gaze similarity to natural gaze behaviour, 2) Task duration and 3) the number of collisions.

1) Gaze similarity to natural gaze behaviour: The gaze behaviour during gaze control is compared to the natural gaze behaviour during keyboard control. This is shown in Figure 49 where the gaze distribution for each control interface (blue markers in Figure 49A - Naive GUI-Free, B - Screen GUI Overlay and C. Natural GUI-Free controller) is shown. This is compared with the natural gaze behaviour recorded during key-board control (red markers Figure 49A). It is clear that the overlap in distributions is much greater for the Natural GUI-Free controller (Figure 49C). This is also numerically reflected when the

Kullback-Leibler (KL) divergence is calculated between the distributions, where the Naive GUI-Free approach yields a KL divergence of 9.1, the Screen GUI Overlay 14.2 and the Natural GUI-Free approach 8.1.

2) *Task duration*: The task duration for all subjects and interfaces is shown in Figure 50. The keyboard interface achieved the fastest completion times, taking on average 301 ± 12 s. Of the gaze based interfaces, the Screen GUI Overlay method achieves the fastest completion of the course, taking on average in 383 ± 80 sec. Among the two GUI-Free approaches, the proposed natural interface outperforms the naive approach, reducing the completion time by 18% (441 ± 161 sec compared to 522 ± 97 sec).

3) *Number of collisions*: The number of collisions per subject and interface is shown in Figure 50B. Again, the keyboard interface achieves the best performance with least collisions (2.3 ± 2.0), while for this metric the Naive GUI-Free interface achieves least collisions out of the gaze based interfaces (7.3 ± 3.4). The Screen GUI Overlay interface makes on average 7.8 ± 3.5 collisions and the proposed Natural GUI-Free approach involves considerably more collisions (22 ± 17.6).

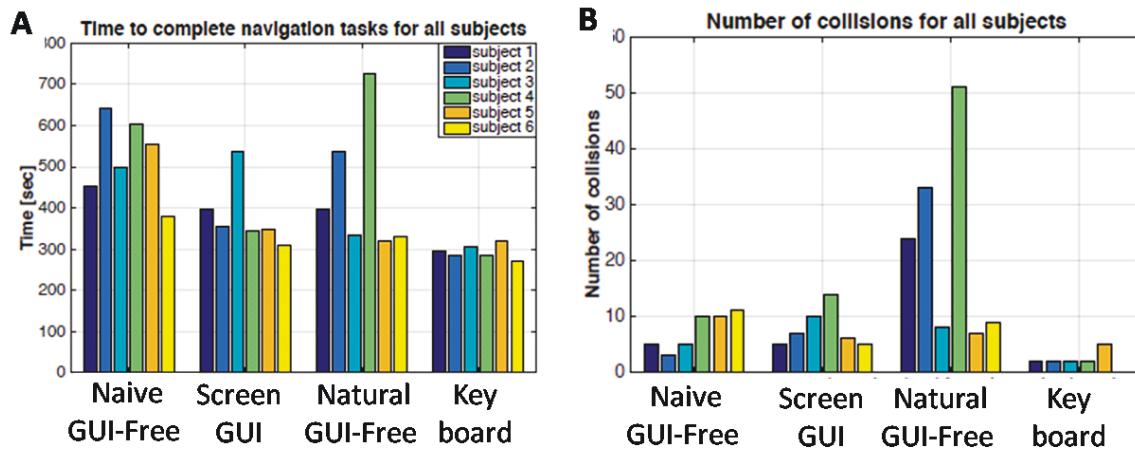


Figure 50. Subject Performance Comparison across interfaces. A. Time to complete navigation task (~150m). B. Number of collisions.

5.2.3.3 Discussion

The VR wheelchair experimental suite enables benchmarking and training of alternative wheelchair control interfaces, a tool that is vital for safely developing new methods. This system has been demonstrated with a pilot study comparing different gaze based control interfaces with able bodied subjects. This benchmark showed firstly that gaze control interfaces achieved performance metrics that were not on average vastly different to the able-bodied keyboard approach. In fact, in some cases the difference was very small indeed. For example, subject 6 took just 9% longer to complete the task with their gaze, despite being completely naive to gaze control. This shows the huge potential of gaze as an input modality for wheelchair control. Looking within the different gaze based interfaces, the method that enabled fastest task completion was the Screen GUI Overlay and the Natural GUI-Free approach was on average faster than the Naive GUI-Free. However, the Natural GUI-Free approach resulted in many more collisions for 3 of the subjects. Though one of these subjects (Subject 4) should be disregarded as the post trial calibration data demonstrated there had been a large head movement during the experiment.

The aim was also to minimise impact on the natural sensory tasks of the eyes. Thus an additional score was based on the similarity of eye-movements while controlling with the gaze versus free eye-movement behaviour while controlling with the keyboard. This showed the proposed natural interface was most natural. Thus to summarise, the Natural GUI-Free interface had a better performance than the Naive GUI-Free interface and had the least impact on gaze behaviour, but its collision rate is strongly affected by gaze inaccuracies. This is something that can be greatly improved with the use of a remote head free tracking system. In addition, the interface with the least collisions (naive GUI free) also had the largest idle zone. This could suggest the idle zones are an inadequate solution to the Midas touch problem which should be improved in future approaches.

It is important to mention the differences between these interfaces and how that may affect the learning rate. Though the Natural GUI-Free interface was shown to have a closer relationship with natural gaze behaviour during keyboard control, it is a more abstract control interface and requires accurate gaze tracking. The other two methods have very clear delineation of control regions and the commands are discrete and thus explaining the control system is easy and it is tolerant to gaze inaccuracies. Thus it is suggested that with a head free-eye-tracker that compensates for head movements and with a little more exposure/training the natural interface could out-perform the other methods. It may take slightly more time for the user to trust the system, but once it is fully understood and tailored to the user it will result in a smoother drive, impinge less on the natural distribution of attention and thus have less impact on cognitive load. In addition, while the screen based GUI system outperformed both free-viewing approaches it is actually not a fair representation of operation in the real world, where the user will have a small screen view of the environment with the GUI overlaid. While in the virtual system, the GUI was as large as their field of view of the environment (Figure 48C). Hence in real-world operation the user will have to constantly share their attention with a small video feed of the world and the real environment, which is very unnatural. This could be better represented in future studies with a smaller feed in one corner of the virtual world displaying a small duplicate feed with the GUI overlaid.

Future work will improve these interfaces and run longer experiments to quantify the learning rates of the alternative methods and to utilise other data metrics available from the simulator. In addition, this is virtual driving and thus real-world driving should be run in parallel to test how well the methods and results translate to real wheelchair driving. An important difference is the absence of ego-motion which will of course affect the eye-movements due to the vestibular ocular reflex (VOR).

5.2.4 Real-world gaze controlled wheelchair

A gaze controlled wheelchair platform has since been developed to allow gaze controlled driving to be tested in the real world. This platform also included instrumentation to record the driver commands (joystick or other input) and to detect object proximity. A full user study is yet to be conducted, thus a brief overview of the system is described and the preliminary results discussed.

5.2.4.1 Methods

The system overview is shown in Figure 51A which shows a diagram of the wheelchair with the eye-tracker mounting frame, data flow diagram and photograph of wheelchair being driven along a corridor. The mounting frame is rather over engineered; however, it was initially used as a chinrest for remotely mounting the GT3D eye-tracker as used for the virtual wheelchair. However, with the release of the first commercial low cost remote eye tracker (EyeTribe), this was used (until it broke) as it allows head free gaze tracking. The Tobii EyeX (low cost eye-tracker used in the second arm study) was subsequently released and this is the current eye-tracker used with the wheelchair. The gaze data from the eye tracker is processed by the gaze interface method (similarly to VR wheelchair) and movement commands are passed to a microcontroller (Arduino) to emulate the joystick. This applies the relevant control voltages to the wheelchair control board and the wheelchair moves. After timing issues with the Arduino board a National Instruments data acquisition (NIDAQ) system was used. This can reliably record and produce voltages and thus enables gaze control of the wheelchair but also records the joystick commands (when in use) and 6 proximity sensors mounted on the wheelchair.

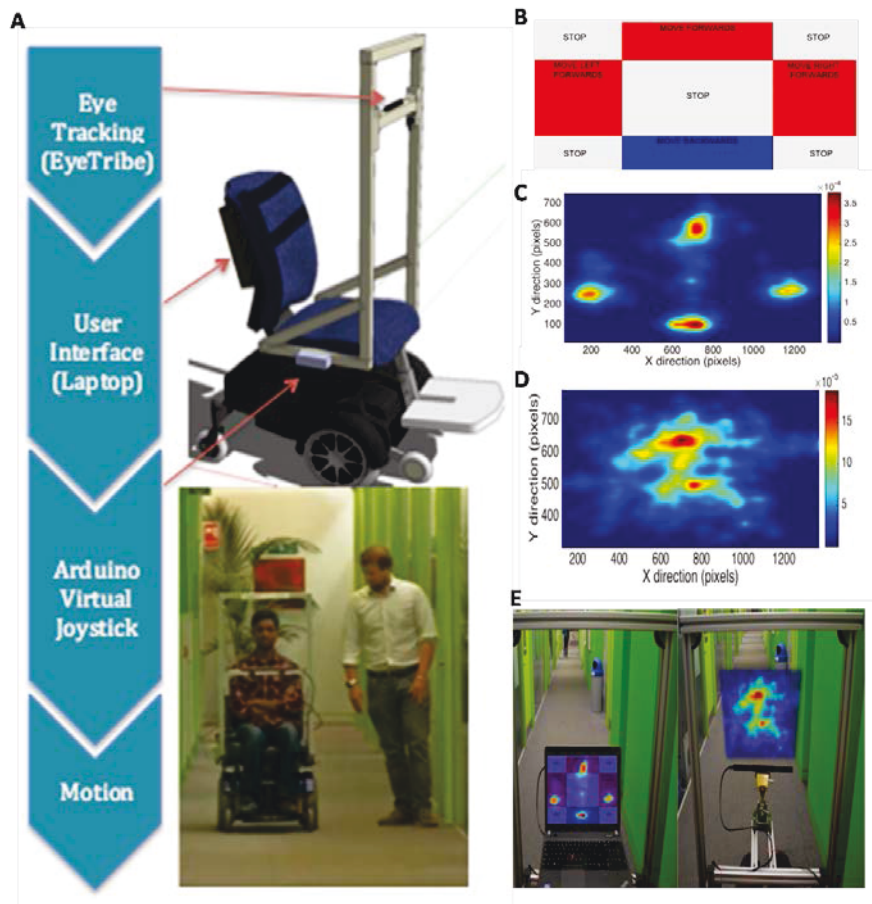


Figure 51: Gaze controlled wheelchair overview. A. Diagram of wheelchair with eye-tracker mounting frame, data flow diagram and photograph of wheelchair being driven along a corridor. B. Screen based GUI interface where gazing in the command regions results in the labelled wheelchair movement. C. Gaze distribution when controlling using the screen based interface. D. Gaze distribution when using the natural screen free interface. E. Respective distributions overlaid on the relevant locations of the visual field. For the screen based this is restricted to the screen, while for the screen free method the gaze is calibrated to a higher portion of the visual field.

5.2.4.2 Results

The wheelchair can be controlled by gaze and an experimental suite has been built around this to enable full user experiments to be conducted with high resolution behavioural and performance data recorded in real-time. This will allow the system to be used in conjunction with the virtual wheelchair simulator to test methods in the real world, while still capturing sufficient data to reliably benchmark control interfaces as they are improved. A number of live demos have been made, however a complete user study has not yet been achieved with reliable data and thus is an ongoing project. Preliminary gaze data has been recorded using the screen based method and the screen free natural interface described for the virtual wheelchair. The distributions are shown for these two methods in Figure 51C and D and also in Figure 51C where they have been superimposed on the first person view from the wheelchair. Here in the real world, the difference between a screen based system and a screen free system becomes even more apparent. The user must spend the majority of their time looking at the computer screen, rather than the environment that they are trying to navigate.

5.2.4.3 Discussion

This preliminary data shows promising results. With the experimental platform built, the current gaze based interfaces can be fully tested in the wild in parallel with the VR training system. Preliminary experiments have begun to link natural gaze behaviour with joystick control; however insufficient data has been collected. Thus a very interesting future study will be to make long term recordings of natural eye-movements as the wheelchair is used with joystick input. This will provide a rich dataset for decoding intentions (joystick commands) from natural eye-movements. This is important to improve solutions to the Midas touch problem (distinguishing control intentions from other gaze behaviours). In addition, future work should integrate an alternative eye-tracker for which the 3D gaze can also be investigated.

5.3 Overview Discussion

Gaze has long been developed as an interface for controlling computers via graphical user interfaces. More recent work has begun to use this approach to control robotics and wheelchairs via computer GUI. This Chapter describes the engineering efforts to enable gaze interaction in the 3D world, without the unnatural proxy of GUI interaction. Current efforts have been mainly bottom up in the sense that platforms have been developed to utilise the gaze signal as a 3D cursor in the real-world. This is in itself a completely novel approach. The design of these interfaces placed a strong emphasis on using gaze in its natural context, taking inspiration from the fact that the allocation of gaze is naturally tightly locked to motor behaviour. The effectiveness of such an approach has been demonstrated, utilising this very high information throughput volumetric cursor for robotic control. The developed novel calibration procedure uses the robotic arm to speed up calibration data collection almost 20 fold while boosting 3D gaze accuracy by >60%. This concept of using one's own prosthesis to calibrate the gaze control input really exemplifies this symbiotic approach. Successful gaze control has also been achieved in wheelchair driving (both virtual reality and reality) where gaze control is achieved with no training, with a performance that is not vastly worse than an able bodied input. In

addition, the proposed natural approach is shown to be related more closely to natural gaze behaviour while driving. These preliminary results need to be further tested and developed both with able bodied subjects and patient groups, while also considering longer durations of operation, to understand the learning rates, optimal achievable performance, and the induced fatigue of alternative methods.

In general, this body of work is predominantly focussed on the development of the technological platforms. The data collected is still limited at this stage but the developed platforms have huge scope for future research and development. The 4 different control interfaces (WidowX arm, UR10 Arm, VR wheelchair and real wheelchair) have all been designed to extract data driven benchmarks and behavioural metrics to support this future work. In addition, this will be leveraged to study gaze behaviour during joy-stick operation of wheelchairs and natural human gaze behaviour during object manipulation in the wild. The following describes some important areas of consideration for taking this forward as well as the justification for such an approach.

Eye tracking: The GT3D system was developed to enable low-cost eye-tracking and to take gaze interaction into the real-world as a non-invasive compliment to BMI control. This has been validated to perform real world 3D gaze control of prosthetics and wheelchairs. The GT3D system was initially developed at a time where the cheapest commercially available eye-tracking system was >£10,000. Since then two low-cost <£100 systems have been released. Thus, these systems are also tested due to a number of benefits they brought, although in the process a number of pitfalls were identified. Both the Tobii EyeX and the EyeTribe are remote eye tracking systems (meaning no equipment must be worn) and in addition they compensate for head movement (avoiding the need for a chin rest). These systems were effective, however the access to raw eye-movement data limited their applicability. As they are designed as an additional input for computer control rather than a scientific tool, extending their use for real-world interaction required workarounds. They could not apply the 3D gaze approach achieved with the GT3D either. While the focus here is not on the hardware used to gather data but how that data is interpreted and applied to control tasks, it has proved a huge asset developing a system in-house to give full access to low level raw data.

An additional important difference is that the GT3D can be head-mounted. For many realms of operation, a head mounted system would be preferable such as for an upper limb amputee. There are currently no low cost wearable eye-trackers and while further development is required to fully integrate head tracking capabilities work towards this has already been done (Sim et al., 2013). Research grade (>£10,000) eye-tracking systems have recently added 3D gaze estimation capabilities (Tobii glasses 2, 2016) and these will be assessed and used in future studies. With all head-mounted eye-tracking system there is inevitable headset slippage which can dramatically degrade the gaze estimation performance. One solution to this in the future is made possible by the improved data driven 3D calibration approach. This probabilistic approach is extendable to incorporate new information as it becomes available. Thus in everyday use, as inaccuracies creep in due to slight shifting of the gaze tracker, a 5 second update to collect 50 points to update the mapping is not a big imposition on the subject. Ideally in the future, this updating can be automated based on the recognition of an error signal. As our understanding of the

interaction between gaze and natural motor behaviour improves perhaps a miss-match response in the user's gaze can be detected when the arm does not quite go where the user expects it to.

Midas-Touch and beyond: grasp, pour, pass, push, place, open. In systems described here, the focus has been mainly on utilising the high information through-put volumetric cursor to interact with the real world. Less emphasis has been placed on solutions to the Midas touch problem, although the novel wink method proposed and used here enables gripper opening and closing. But this is just the beginning, as with interaction in the real world, the range of commands required vastly increases. No longer is a select or click option sufficient, interactions are significantly more complex because having reached the arm to an object, the list of possible actions is vast - e.g. grasp, push, grasp and pour, pass, place, open etc. This is where using gaze as a complement to other non-invasive interfaces such as electromyography (EMG) becomes important. With end-point control achieved, other retained functions such as residual muscle activation can be dedicated to control object interaction. In addition, there is significantly more intention information in the spatial and temporal dynamics of eye-movements, as discussed in Ch2.

User agency and user acceptance: In such an endeavour to infer intentions from natural eye-movements, rather than via a more conventional control interface such as an on-screen GUI, it is important to consider patient acceptance. The eyes are the fastest moving actuator in the body, however we are not consciously aware of this high-frequency shifting of our overt attention due to saccadic suppression (Bridgeman et al., 1975). If a patient's natural eye-movements are used to control the actions of prosthetics, it is important to discuss user agency. In the context of human machine interfaces, this pertains to the user feeling a direct relationship between their intentions and their subsequent enactment on the world. User agency has been shown to be vital for effective interaction in human computer interaction and thus user acceptance of interaction methods (Zeng, 2009). Though there are very little systematic studies, this is a fundamental issue for brain machine interfaces whereby, firstly decoding accuracy can be low and secondly when shared control is used, the system is acting to greater or lesser extent autonomously.

Shared control is where subjects can communicate high level commands which are then enacted by an intelligent system. For example Rebsamen et al. developed a brain controlled wheelchair navigation system whereby a subject used a P300 BMI to select their destination, the wheelchair autonomously navigates to using predefined paths (Rebsamen et al., 2007). This approach is highly promising approach in terms of its ability to restore independence with a very low-throughput signal. However, it is restricted to predefined journeys, significantly limits subject's decision space and ability to update decisions in real-time. This likely to prove significantly frustrating for subjects and detract from their sense of agency. Conversely, However, it has also been shown that the sense of agency can come from high level end-goal related decisions, without lower-level control processes being a requirement for the perception of agency (Kumar and Srinivasan, 2014). However this was involved a motor control task, however in a controlled lab paradigm not directly related to wheelchair or prosthetic control. In a directly relevant task, Kim et al. explicitly probed the effect of autonomy on patient satisfaction when controlling a robotic arm with spinal cord injured subjects (Kim et al., 2012). They found that subjects could see the functional benefit of the robot autonomy (following user input via a touch screen to

select the object to grasp) but they also stressed the satisfaction of being directly in control during the interactive manual mode of operation, despite taking longer to master. Conversely, it is often emphasised that lengthy training times suffered by BMIs are a major challenge for the patient uptake of BMIs (Murphy et al., 2016).

The relationship between agency, shared control, rehabilitative performance and user acceptance is clearly non-trivial but a vital consideration in the design of prosthetic and wheelchair control. More systematic studies are required, with more patient group feedback on the subject (not addressed in recent patient focus groups (Blain-Moraes et al., 2012; Huggins et al., 2011; Schicktanz et al., 2015)). Returning to gaze control, the aim of inferring intentions implicitly from unconsciously elicited eye-movement patterns would seem unlikely to facilitate a sense of agency. However, firstly, many eye-movements are a direct result of a high level intention as discussed at length in Ch2. Thus, providing they can be decoded effectively, the feeling of agency should be achieved when the actuator correctly carries out the intended high-level intention of the subject. In addition, the aim will be to use any additional retained motor function to provide more direct manual control, augmented by the high level intentions decoded from eye-movements. An example of such cross-modal control is demonstrated by Corbett et al. when they significantly improved EMG control of a robotic arm by combining it with target information (2D) from gaze (Corbett et al., 2012). To make such an approach feasible outside of the lab however requires significant developments in the understanding of eye-movement behaviour in the wild, the true context of operation of a prosthetic system. In addition, the vital development of methods to simultaneously quantify eye-movements and the natural body movement repertoire that they are embedded in is required. The next body of work meets this significant challenge by comprehensively recording human behaviour "in the wild".

6 Embodied eye-movement behaviour in the wild

Acknowledgments. The experimental methods and data collection was the output from an equal collaboration with Andreas Thomik in the lab in collecting full body and full eye movement data with subjects. While his work focused on hand movements in relation to Ethomics, my work here is orthogonal to it by focusing on eye movements.

Abstract. The brain is a dynamical system, mapping sensory inputs to motor actions. This relationship has been widely characterised by reductionist lab experiments that highly constrain subject behaviour. However, with the emergence of mobile eye-tracking, increasing emphasis has been placed on the ecological validity of gaze studies, taking them out of the lab and into the “wild”. This aims to make the environmental and task contexts realistic, while moving away from static head-fixed eye-tracking studies that significantly limit behavioural repertoire. However most experiments “in the wild” either focus purely on the bottom up visual drivers of attention or behaviour is annotated by hand. In the latter case, experiments have very short durations due to the time investment required to annotate behaviour. In addition these annotations are subjective and at a symbolic level and therefore do not capture the real-time kinematics of behaviour. This prohibits a data driven approach towards a unified model of eye-movement behaviour, and thus attention, in the wild. A novel experimental paradigm is developed - automatically capturing, rather than constraining, sensory inputs and motor outputs in natural behaviour. Sensory inputs are recorded using a head mounted eye-tracker, scene camera and microphone. Simultaneously recording skeletal motor outputs by motion tracking 51 degrees of freedom in the body and a total of 40 degrees of freedom in the hands. All tracking equipment is marker-less and thus allows unconstrained behavioural monitoring “in the

wild". Experimental durations are hours rather than minutes and involve a very broad range of behaviours from the tasks of daily life. This enables classical relationships to be demonstrated without laborious hand annotation, in ecologically valid environments including 3 daily scenarios: preparing and eating breakfast in the kitchen, evening activities in the home and in-door navigation. Standard aggregated statistics of gaze behaviour are presented and compared to previous pioneering hand coded studies. This highlights a number of differences, caveats, surprising results and, inevitably, a list of future problems to be solved. Primarily, the findings show that the classical categorisation of gaze data into saccades and fixations is insufficient to capture the eye-movement repertoire in natural behaviour. In addition, eye-tracking data is processed post-hoc to give 3D gaze position and for the first time gaze depth statistics in the wild are investigated. The dataset presented moves towards an embodied approach, building a database of human behaviour more complete, extensive and unconstrained than achieved previously.

6.1 Introduction

The eyes are the fastest moving organ in the body and they control the input to a sensory system that occupies almost a third of the human cortex (compared to around three hundredths for hearing). Thus understanding their function and basis is extremely important to furthering the understanding of the human brain. Over the last 50 years eye-movement studies have developed significantly. Vital developments in our understanding of the sub systems that drive the oculomotor system have come from highly controlled psychophysical experiments involving very simple dot stimuli (for extensive review see (Eileen, 2011)). These experiments have been vital in characterising the oculomotor system in isolation - from an engineering perspective the "technical specifications" of eye-movements are physiologically very well understood in terms of their spatial and temporal dynamics (Carpenter, 1988). This approach has lead to the 5 well described subsystems: fixation, image stabilisation, saccadic, smooth pursuit and vergence previously discussed (Chapter 2). This vital work has provided the foundations for understanding active vision, not only in terms of physical repertoire but the neurophysiological mechanisms that produce them and their perceptual implications. However, as will become clear in the following body of work, while these sub systems can be studied in isolation in the lab, natural behaviour in the wild is a complex mixture of these behaviours that is very difficult to categorise so succinctly.

In addition to understanding the repertoire of eye-movements, the way in which these behavioural resources are deployed by the brain, at such high frequency, to gather the required visual information is vital. Extensive electrophysiology studies have demonstrated that the superior colliculus, which is intrinsically involved in eye-movement generation, receives and integrates inputs from visual, auditory, somatosensory, vestibular and proprioceptive systems (Meredith and Stein, 1986). In addition, behavioural evidence over the past century has shown that eye-movements are influenced by visual saliency, auditory, proprioceptive and vestibular stimuli as discussed in Chapter 2. In addition, the integration of memory and vitally visuomotor task demands have been extensively demonstrated (for an overview see (Schneider et al., 2013)).

Studying these influences in isolation has hugely advanced our understanding and provided an important theoretical basis for modelling the allocation of visual attention. However, as with the sub-systems of the eye-movement repertoire, factors influencing allocation of attention do not occur in isolation "in the wild". Thus efforts must be directed towards taking the sub-system understanding and drawing parallels and contrasts to their manifestation in the wild compared to the lab. This will give insights into the cross-modal integration of different drivers of attention, how they are integrated in natural behaviour and their relative importance in different contexts. Further to this, understanding which lab based observations and models depart from what we observe in the wild is a vital development towards a more ecologically valid understanding of eye-movement behaviour.

Over the last 20 years there have been many experimental observations to justify a more integrative holistic approach to understanding eye-movements. The focus of this discussion tends to be on the so called "saliency argument" that has been described at length in Chapter 2 - contrasting bottom up visual saliency models developed based on static lab experiments and the top down task effects that become so clear in natural behaviour, for review see (Tatler et al., 2011). However, the other vital justification to a non-compartmentalised approach comes from lab based experiments that show the embodied nature of eye-movement behaviour which will be discussed. The concept of embodied eye-movement behaviour, proposed in this chapter, has its roots in the idea of embodied cognition which moved to the forefront of thought in computer science and cognitive psychology in the 90s (Gover, 1996). For a review see (Anderson, 2003). It continues to be highly influential and has been adopted in cognitive neuroscience (for a recent review see (Kiverstein and Miller, 2015)) and computational neuroscience (Caligiore and Fischer, 2012). Central to this premise, is that the mind is not just a computing component connected to the body that can be studied in isolation, but the body and the environment influence and shape cognition in the brain. In the context of eye-movements, it is clear from mounting evidence that the study of eye-movements must be embodied, that is not isolated from the body and an ecologically relevant environment. The following section summarises this evidence.

A central question to the concept of embodied eye-movement behaviour is the question of whether eye-movement patterns are shaped by current and planned body dynamics, encoded as an inseparable part of the motor program. Ironically, the experience of being a subject in a static eye-movement experiment gives a firsthand understanding of the difficulty in decoupling eye and body movements. In such a set-up with the head fixed using a bite board, the strain in the neck is prominent when making all but the smallest of eye-movements. In addition, fixing the head in this manner alters the characteristics of the saccadic main sequence, with the peak eye velocity for a given saccadic amplitude being considerably higher in head free conditions (Collewyn et al., 1990). Similar effects are also observed in the presence of limb movements, where eye-movement characteristics are directly affected by the limb dynamics (Donkelaar et al., 2004; Snyder et al., 2002; Tipper et al., 2001). At the simplest level, eye-movement behaviour is different when purely viewing objects compared to when grasping them (Brouwer et al., 2009a). There is a significant difference even when objects are just touched compared to when they are grasped to be moved (Smeets et al., 1996).

Drawing on developmental and motor learning studies also gives additional evidence to support an integrated visuomotor control system - as opposed to a compartmentalised assembly of independent sub-systems. Visuomotor behaviours must be learnt in combination, not as isolated sub-systems, and thus it is likely that the complex behaviours of every-day life are encoded as an integrated control program. Interestingly, it has been shown in developmental studies that eye-movements in infants can be explained predominantly by bottom up effects and then, with increasing age, top down task effects become increasingly important - as the visuomotor repertoire develops (Açık et al., 2010). Further evidence for the unified encoding of visual control and action can be demonstrated in highly habituated motor learning tasks. For example in a cup stacking task, stereo-typical eye-movements are made even in the dark (Foerster et al., 2012). This suggests eye-movements are so tightly embedded in the motor behaviour of the limbs that they endure, even in the absence of visual input. Similar observations have also been made with regard to memory tasks, where subjects make saccades to locations of objects during their recall, even when they are looking at a blank screen. Thus, the tight association between eye-movements, cognitive processes and the body is clear and therefore they cannot be investigated purely as independent systems.

Significant progress has been made towards understanding the role of vision in non-compartmentalised natural behavior. Particularly in the last 25 years, this has been facilitated by the emergence of mobile eye-tracking solutions, which measure gaze in relation to the subject's scene view (2D video feed). For an overview see (Eileen, 2011). In addition, increasing emphasis has been placed on the ecological validity of gaze studies, taking them out of the lab and into the "wild" (Hayhoe & Ballard, 2005; Kingstone et al., 2003; Land & Tatler, 2009). These steps have challenged many of the fundamental results of static eye-tracking experiments and are changing the way we look at eye-movement ethology. As well as the pioneering real-world studies of Land, Hayhoe, Ballard and colleagues in the 90s that disrupted the field of visual saliency (see Chapter 2 and for review see (Tatler et al., 2011)), more recent studies have made direct comparison between lab based experiments and the "wild".

A novel paradigm was developed to elegantly draw direct comparisons between eye-movements in the lab and the real-world by 't Hart et al. (2009). This involved measuring gaze during the free exploration of real-world environments with respect to a head-mounted scene camera view of the world. A direct comparison is made in the lab set-up by re-playing the scene videos to subjects while their eyes were tracked with a head-fixed eye-tracking system. A third condition was compared in which static frames extracted from the video were displayed for 1s intervals. With a similar approach, Foulsham et al. also compared the lab set-up to free behaviour, though in this study there was a more explicit task to navigate across the campus to a specified location (Foulsham et al., 2011). Both studies showed that the central fixation bias (see Chapter2), previously observed in static scenes (Tatler, 2007), endured in natural free-body behaviour. In particular, however 't Hart et al. demonstrated that this effect was the strongest in the static 1s displaying of video frames and in consequence the inter subject consistency of gaze position was highest in this condition. This suggested that the inter subject consistency observed in static lab experiments was predominantly the result of spatial biases. In addition 't Hart et al. found that the eye-movements during video replay were a better predictor of eye-movements in the real-world than those predicted by the static 1s display

condition. This implied the importance of using continuous natural video stimuli when conducting lab experiments to illicit more natural eye-movement behaviors. However the prediction performance of the developed empirical saliency maps, measured by the area under the curve (AUC) of the receiver operator curve (ROC), was still only 60%, where chance level is 50% (for details of this metric see (Hart et al., 2009)). While using a bottom up visual saliency model based on (Itti and Koch, 2000) had a weak, but significant, performance of 53% (free exploration), 55% (continuous video replay) and 53% (static 1s display). In relation to this, moving from low-level visual saliency to quantify what objects were fixated and when, Foulsham et al. showed that there was a larger probability of fixating objects in the near field that were directly related to the task of navigation during the real world experiment than under the lab replay condition. This suggested an enhanced top-down task driven gaze behaviour in the real-world task. Conversely, subjects in the real-world condition were less likely to look at passersby (other people on their path) when they were near compared to far away. They suggest that this may be an effect of the authentic social context (unlike in a video), eliciting the known tendency to avoid the gaze of strangers.

These highly informative studies pave the way for vital cross-domain comparisons between the lab and the wild as our understanding of eye-movement behaviour is translated into ecologically valid settings. In terms of the quest to understand what influences the spatial and temporal allocation of attention, these studies have incorporated spatial biases, visual saliency, object effects (both bottom up and top down) and even social context during free movement "in the wild". However, as the previous discussion demonstrates, there are still many other factors (auditory, somatosensory, vestibular, proprioceptive, motor-plans and memory) that need to be incorporated into a unified model. In addition, little attention has been given to the real-time dynamics of eye-movement behaviour and how this is affected by task and context. In addition, the first study was task free, while the second task was walking and navigating and thus only a limited exploration of human behavioral repertoire has been explored. With more complex tasks such as object manipulation in natural behaviour, the importance of other factors such as the dynamics of the body are likely to become even more vivid.

These aspects of object manipulation in the wild were explored in the pioneering work of Land, Ballard, Hayhoe and others, as discussed in Chapter 2. However, while making no attempt whatsoever to diminish these contributions, it is important to discuss their limitations and to identify new steps forward. Firstly it is important to note the very low number of subjects (N=3 for Land's seminal study (Land et al., 1999)), as well as the very short durations of behaviour recorded (2-4mins). This is because of the huge time commitment required to post process and annotate behaviour frame by frame, firstly coding the eye-movement behaviour into fixations and saccades and secondly to annotate the body movements and identify the objects fixated and interacted with.

To enable more scalable and objective approaches a number of methodological bottle necks and interpretive limitations must be addressed. Firstly, pertaining to the subjectiveness of human annotation of behaviour. In general it is very difficult for human annotators not to impose a certain degree of interpretive bias on the annotations being made, that is to say making inferences and assumptions based on the hypothesis of the experiment or the knowledge of behaviour in the lab (List et al., 2005). In eye-tracking

studies specifically it has been shown that the agreement between experienced coders is sometimes lower than the agreement between experienced coders and algorithmic coding (Munn et al., 2008). Secondly, the experimental duration is strongly limited, for example in this same study by Munn et al., it took in the region of 1 minute of human coding per second of experimental recording. Thus a free recording of 1 hour would take 60 hours to code. The result of this time constraint is that experiment durations are short and capture few repetitions of behaviour and thus cannot represent the true variability. Also similar remarks can be made about the more qualitative descriptions of behaviour that arise from these studies, which have posed very compelling and interesting ideas and certainly enriched the field but still need to be further substantiated with more extensive empirical evidence. Such is the nature of moving away from reductionist approaches, if the complexity of daily life is not reduced and constrained as is the case of the traditional lab experiment, it must be recorded or annotated. This results in a very high dimensional data stream to capture behaviour, with which comes the curse of dimensionality (Bellman, 1960). In this context, it implies that number of observations required to characterise behaviour grows exponentially with the number of variables or dimensions of behaviour and stimulation. In other words, making meaningful deductions from such high dimensional data requires very large quantities of data. Thus objective and scalable (automated) methodologies to capture and quantify unconstrained behaviour are required to make the problem tractable. The following section draws on a number of engineering fields to discuss the state of the art for enabling such methodologies.

The first aspect to discuss is the automated classification of gaze behaviour which although often taken for granted, remains only a partially solved problem, particularly in dynamic scenes and behaviour (Kasneci et al., 2014; Larsson et al., 2014; Pfeiffer and Essig, 2013). The categories of the eye-movement repertoire, and the neurophysiological basis has been extensively characterised in controlled lab environments. In addition there is extensive literature on the automated classification of raw gaze behaviour into some of these categories, with most emphasis on pure fixations and saccades. For a wide range of literature this is sufficient as experimental setups fix the head and present static 2D visual stimuli. In this context the automated classification of gaze behaviour could be considered a solved problem (though ambiguities remain when comparing results from different apparatus, algorithms and settings). However as experiments move from the lab into the natural environment, this classification of just fixations and saccades is no longer sufficient and remains an active area of research (Santini et al., 2015).

There are two main complications that arise for experiments "in the wild". The first is due to limitations of mobile eye-tracking apparatus. When fixing the head and using a desk based eye-tracking system, frame rates in excess of 1000Hz are achieved with a claimed accuracy of <0.1 degrees. While for mobile eye tracking this tends to be at 30Hz with a claimed accuracy of <1 degree (recent releases have brought this to 120Hz). Capturing the dynamics of rapid saccadic eye-movements sufficiently for their classification is difficult with such low frame rates. The second problem is the increased complexity of the eye-movement repertoire during free body movements. In the lab, different components of the oculomotor system can be studied in isolation, with prescribed stimulation of one system (visual / OKR / vestibular / auditory / proprioceptive etc). In natural behaviour, the oculomotor output is a complex interaction of all systems. These interactions are, as yet,

not well characterised. In the following section an overview of the state of the art of event detection will be explored firstly in the lab and secondly "in the wild".

6.1.1 Static fixations and saccades

During head-fixed eye-tracking with static stimuli, the main behaviours elicited are static fixations and saccades. The characteristics used for distinguishing fixations and saccades are generally gaze velocity, acceleration and dispersion, where velocity is taken between data samples, acceleration the second order equivalent and dispersion is the spread of gaze data points as they are sampled (the precise calculation has many alternative definitions). The velocity distribution of gaze does not yield a simple bimodal distribution with a peak at low velocities for fixations and a peak at high velocities for saccades. The peak velocity of a saccade depends on its amplitude (saccade main sequence relationship - amplitudes proportional to peak velocities) and saccades also involve the low velocity eye-movements as the eye accelerates. Thus a common approach to saccade detection is to identify characteristically shaped peaks in the raw data which meet certain criteria. This algorithm known as Velocity threshold identification (I-VT) (Salvucci and Goldberg, 2000) follows the following steps:

- 1. Identify candidate saccade peaks above a defined minimum peak velocity in the raw gaze signal. (thresholds vary considerably but between 30-100 deg/s).*
- 2. For each candidate peak define the full event trajectory by finding the onset and offset of the event by tracing back to where the velocity went over the onset threshold and tracing forward to when it goes below this threshold.*
- 3. With the resulting cropped trajectory, check for symmetry, i.e. that the peak is located within an allowable central percentage of the identified trajectory.*

This method requires 3 parameters, the minimum peak velocity threshold, the onset/offset threshold and the central percentage. The setting of these thresholds will have a significant effect on the algorithm output and there is no clear bimodal distribution for peak velocities either. The peak saccadic velocity is related to the amplitude of the saccade as shown in Figure 52 and therefore peak velocities can be as low as 10 deg/sec for very small saccades. Signal noise and fixational eye-movements (microsaccades) also lie within this region and thus lowering the peak velocity threshold may result in false positives, while raising it may result in false negatives. How the thresholds are chosen will strongly affect the resulting output of the classification algorithm and an awareness of this is vital. The other proviso is that the I-VT method can only be applied when the eye-tracking data has a sampling rate above 200Hz. This is because if a saccade lasts on average 30ms, the Nyquist rate is 66Hz, which would be sufficient to be sure to capture at least 2 samples within the saccade trajectory, but would not necessarily sample close to the velocity maxima. Thus to allow reasonable inference of the saccade peak amplitude requires the generally accepted minimum of 200Hz (Holmqvist et al., 2011). The I-VT algorithm can be made considerably more robust if the acceleration is also used, though taking the second order derivative puts further requirements on the frame rate. Finally, it is important to note that the I-VT is designed for post-hoc classification and cannot be applied online in real-time.

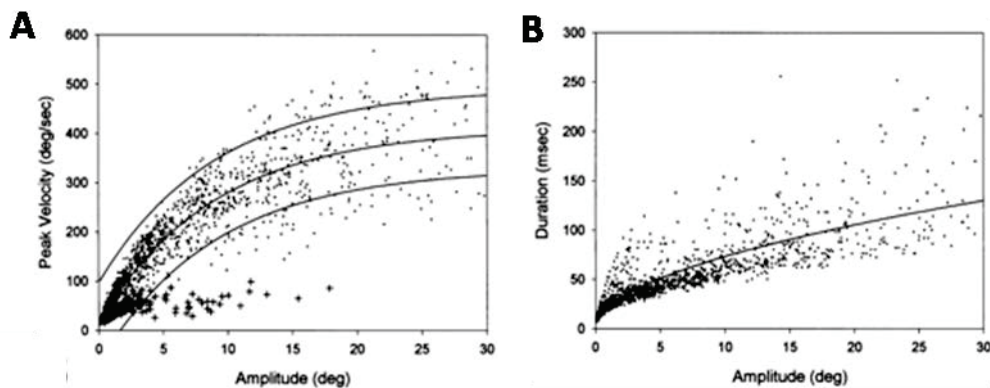


Figure 52 Figure from (Leigh, 2003) Dynamic properties of saccades. (A) Plot of peak velocity versus amplitude of vertical saccades. Data points (dots) are saccades from 10 normal subjects. The data are a fit with an exponential equation of the form: $\text{peak velocity} = V_{\text{max}} \times (1 - e^{-A/C})$, where V_{max} is the asymptotic peak velocity, A is amplitude and C is a constant defining the exponential rise is shown. Also plotted are the 5% and 95% prediction intervals. The + data points indicate vertical saccades from a patient with Niemann-Pick type C disease (not relevant here). (B) Plot of duration versus amplitude.

For eye-tracking frame rates lower than 200Hz, dispersion based methods tend to be used. This approach is based on the premise that either side of a saccade will be a fixation lasting at least 70ms and thus fixations can be unambiguously identified at 30Hz. Thus, by detecting fixations, the saccade initiation and landing positions can be inferred, with the saccade amplitude in between. Fixations can be detected as groups of data points that lie within a certain dispersion threshold. The general procedure for identification by dispersion threshold (I-DT) is as follows:

1. For each data point position take a window of gaze samples the size of the minimum duration
2. While the window dispersion < threshold add new samples to window
3. Move to next position (next position is after window if a fixation was detected or to next data sample if no fixation detected)

The dispersion measurement can take many forms, the simplest of which takes the width and height of a bounding box round the gaze positions. In this case 2 parameters are required, the minimum duration (generally set to 70ms) and the dispersion threshold (generally set to 5 degrees). Again it is important to bear in mind the resulting output from the algorithm may differ considerably with the variation of these parameters (Komogortsev et al., 2010; Salvucci and Goldberg, 2000). Another big impact on the classification will be the context of use and thus these parameters are adjusted to suit the individual viewing behaviour, data quality (e.g. effects from different eye characteristics or lighting conditions), viewing area and user task. For example commercial eye-tracking manuals often recommend task specific setting of the parameters, something referred to by (Holmqvist et al., 2011). For example setting the minimum duration threshold for fixations during reading to 40ms, while for picture viewing a threshold of 200 ms is recommended. While the rational that fixations are typically shorter during reading than during picture viewing is true, applying such a constraint to predetermine expected

results does not seem ideal and may obscure interesting findings or anomalies in such behaviour.

Non-parametric methods have been applied including various clustering methods (Kasneci et al., 2014; Santella and DeCarlo, 2004; Tafaj et al., 2012; Urruty et al., 2007) and hidden markov models (HMM) (Salvucci and Goldberg, 2000). These approaches give a more data driven and robust approach. Kasneci et al. (2014) made a comparison of two probabilistic non-parametric methods - hidden Markov models (HMM) and Bayesian mixture models (BMM), with both performing with precision greater than 96% compared to human annotation (Kasneci et al., 2014). However, the recall of the HMM approach was significantly lower because it did not adapt to the varying distances between fixations, while the BMM adapts in an online fashion. Both implementations require labelled training data (laborious to obtain and highly subjective), make assumptions on the distributions of each behaviour class (fixations and saccades) and finally cluster properties for fixations and saccades show high variability in dynamic scenarios thus requiring the algorithm to adapt online (as BMM does). These requirements are reasonable when it comes to the short 30s data sets on which to perform the evaluation (N=7, t=30s (750 data points, first 300 training, @25Hz head free mobile tracker). However, doing the same in highly dynamic recordings lasting 1hr or more becomes problematic. Added to this, it may be argued that having an adaptive class definition means that data classification is non deterministic. Thus data classified at the beginning of the data stream will have a different class definition applied to that at the end by which point adaptation has occurred. Finally, one further consideration of adaptive approaches is that in adapting the class definition they lose significant information that may be relevant. For example, the fact that behavioural dynamics are so different in different circumstances shouldn't be hidden by the adaptive nature of the algorithm and put in the same class.

6.1.2 Non-static fixations and saccades

Many of the considerations discussed so far are relevant for studies "in the wild". But the addition of free body movement and the dynamic environments of the wild makes the eye-movement repertoire even more complex and difficult to define. Thus, along with frame-rate limitations, the automated classification of behaviours becomes even more challenging. Now the main categories of behaviour (as far as we are currently aware) include static fixations, smooth eye-movements and saccades. This adds another category that has velocity characteristics in between fixations and saccades which are themselves already overlapping. There is no commercially available or standard method for smooth eye-movement detection. Various attempts have been made to classify smooth eye movements such as (Berg et al., 2009) who distinguish smooth eye-movements from fixations based on windowed principal component analysis. Other attempts to consolidate and standardise the detection in high-speed (>1000Hz) eye-tracking data have been made (Larsson et al., 2014). While Rothkopf et al. proposed and implemented a Hidden Markov Model to distinguish saccades, fixations, smooth pursuit and vestibular ocular reflexive eye movements (Rothkopf and Pelz, 2004). The observations used to infer these states were eye-movements and importantly head movements inferred from the scene camera. The same questions arise as discussed previously with regard to training data, how it is obtained, how adaptive should the algorithm be to be robust without shrouding

interesting variability in behaviour etc. What is clear however is the extendibility of such probabilistic models to include new sources of information such as head movement, which are vital to labelling eye-movement data in unconstrained dynamic experiments. Future algorithms should use more and more contextual information, as experimental methodologies make it available, to infer gaze states. Thus where reductionist experiments would target visual / vestibular / auditory / proprioceptive responses individually, methodologies "in the wild" should capture this information simultaneously. The other vital components are thus the recording of subject's body movements and their interactions with objects and other people in the environment. The following subsections will address "in the wild" studies that have achieved this (often manually) and where tools developed from other fields could be applied to automate this.

6.1.3 Head and trunk tracking "in the wild"

Active vision cannot be understood by looking at eye-movements in isolation. It is a combination of body, head and eye-movements that dictate our overt attention in the world. This was observed in natural behaviour by (Land, 2004). To do this the head movements were hand-measured offline based on frame to frame movement of objects in the head mounted scene video. To monitor the trunk position a second tripod mounted scene camera was used. This inference was made by a human annotator who matched the pose of a Barbie doll on a calibrated platform with that of the subject in the video feed. Steps to automate this process were made by Rothkopf and Pelz (2004) by using image processing rather than hand annotation to measure the flow field from the camera (Rothkopf and Pelz, 2004). Another interesting innovation in this area is the "EyeSeeCam" which directly translates eye position to an actuated head mounted camera. This then essentially records the same image being projected on the retina (Schneider et al., 2005). Thus the dynamics of visual stimuli can be explored under natural conditions. In addition, the EyeSeeCam has a static head mounted camera, from which they infer head movements using the flow field, similarly to Rothkopf and Pelz (2004). This was applied by Eisenhower et al (2007) to simultaneously record eye and head movements during free exploration of 3 different environments. The data intensive approach required no human annotation and allowed the authors to apply sophisticated signal processing techniques to give new insights on the dynamics and coordination of head and eye-movements in natural behaviour. They showed the expected smooth compensatory eye-movements to oppose head movements, the synergistic head and eye-movements to make large gaze shifts, but they also showed that synergistic head and eye-movements occurred that were non-saccadic smooth movements. This has two implications, firstly the potential of a new category of eye-movement repertoire existing that adjusts gaze with smooth synergistic head and eye movements. Secondly they claim it has important implications for sensory coding models that currently model vision as a static sequence of frames connected by large saccades during which vision is suppressed. Just like Land observing that VOR is suspended during large head saccades (gaze carried passively with head rotation), these previously unobserved permutations of behaviours will emerge "in the wild", enriching our understanding of behaviour in ecologically valid settings.

6.1.4 Limb Tracking "in the wild"

Beyond tracking the head and trunk, tracking limb movements simultaneously will allow studies "in the wild" to be fully embodied. Capturing the full perception-action cycle will mark a major step towards interpreting gaze behaviour "in the wild" in direct relation to the unfolding motor repertoire of the body. The only known example of this was seen at Vision Sciences Society annual conference in 2015, where the current work was also presented (Abbott et al., 2015; Matthis and Hayhoe, 2015). In this study the Matthis and Hayhoe (2015) used an inertial measurement unit (IMU) body suit and eye-tracker while walking on different terrains "in the wild". The methodology was presented and the integration of the eye-tracker and kinematic data was demonstrated. This high-resolution data driven approach is likely to yield exciting insights for understanding gaze allocation and its association with gait in ecologically valid environments. Though there are no other known studies "in the wild" collecting full body kinematics and eye-movements simultaneously, there is certainly increasing engineering capability for such a methodological shift driven by engineering fields such as ubiquitous computing.

Significant developments are being made in ubiquitous computing and other fields towards activity recognition, which could automate activity annotation. Most approaches concentrate on recognition from body movements or even single IMU sensors alone, but some studies also include eye-tracking in their sensor suite. Velloso et al (2013) developed the activity recognition system AutoBAP (automated coding of Body Action and Posture) which uses the coding scheme (BAP) (Dael et al., 2012) and hard coded rules to classify each body part's state (Velloso, 2013). At this stage the system converts sensor data into low level annotations such as "left shoulder articulation" or "gaze upward". So far the system has been validated with controlled scripted movements in the lab. Another methodology developed by Essig et al. (2012) makes an exciting addition by tracking the gaze, arm position and objects in the environment using optical markers (Essig et al., 2012). The resulting system can annotate the object being fixated on by intersecting the monocular 3D gaze vector with the known object positions and thus associate identity. In addition, it can annotate when an object is being grasped or manipulated. Their preliminary results showed good agreement with human annotations in the lab. One caveat is that it requires every object to have optical markers attached and will require an optical tracking system with sufficient tracking range to allow unconstrained behaviour.

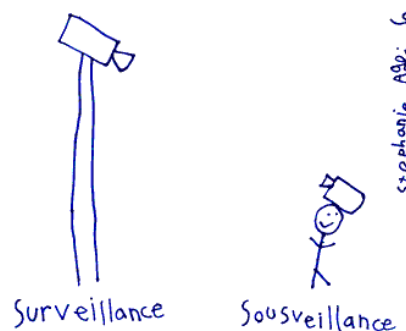


Figure 53: The new surveillance: Sousveillance- Egocentric vision. Figure from(Mann et al., 2014)

Another rapidly emerging field that has exciting implications for behavioural experiments "in the wild" is Egocentric Vision (Mann et al., 2014). With the emergence of high performance portable wearable cameras in consumer products such as GoPro and Google Glass, the capabilities and potential for computer vision and ubiquitous computing are vast, from surveillance (or sousveillance see Figure 53), to life logging as a visual memory aid (Aizawa et al., 2001). An enabling step for these applications is activity recognition, for a review see (Nguyen et al., 2016). Eye-tracking is a logical addition to egovision because it can massively reduce the amount of data being processed in a bioinspired manner. An example of this, with particular relevance here, is (Fathi et al., 2012) who use the egovideo to recognise objects fixated, and detect the hands and associated object manipulations. This approach has allowed them to analyse much larger numbers of subjects (N=17) performing meal preparation tasks (GTEA Gaze dataset), with each recording lasting 4 minutes per subject. Their approach is highly scalable and they subsequently recorded 10 subjects performing 7 meal preparation activities with more than 1hr of data per subject (GTEA Gaze+ dataset). They achieve on average 47% object recognition accuracy (chance=4%), compared to 27% without gaze. They then go on to invert the model predicting gaze based on action recognition, which is then extended further in more recent work to predict gaze based on the features extracted from the ego video (Li et al., 2013). These developments provide promising tools for automating annotation of the scene and object interaction and could be applied to existing scene recordings from other studies "in the wild".

6.1.5 3D gaze behaviour: previous work

A lesser researched aspect of gaze allocation and behaviour is the depth component and associated vergence eye-movements. With the majority of stimuli in lab based experiments coming from a 2D display, monocular eye-tracking and the consideration of 2D gaze position was sufficient. In theory vergence brings both eyes to fixate at the same position to prevent double vision and facilitate depth perception. How this actually manifests in real behaviour however, is as yet unclear. There are only a handful of studies which relate 3D gaze to behaviour in the lab. Steinman reported that vergence was 25-35% beyond the target during the sequential peg task described previously (Malinov et al., 2000). Thus suggesting the visual axes do not intersect perfectly at the visual target. Interestingly the task affected the vergence position with the vergence at 25% beyond the target when the subjects physically tapped the pegs, compared to 35% when they just fixated each one in turn. These preliminary findings add to the increasing body of evidence that suggest that a well characterised behaviour in constrained lab experiments may not be observed in a less constrained task. In this case, although humans have the capability to verge accurately on the target, based on controlled characterisation of the visual system (Carpenter, 1988), this behaviour will not always be present in natural behaviour. To quote Steinman et al. (1998):

"Little is known about how vision and eye movements act under ecologically-valid conditions. What we have learned under analytically pure, impoverished conditions, seems not to apply to the real world"(Steinman et al., 1998)

However, in cognitive neuroscience, similar conclusions have been drawn, where there was a discrepancy between the remarkable cognitive abilities of primates in the lab and

perceived lesser requirements faced "in the wild". Just like Humphrey stated (1976) "We do not expect to find that animals possess abilities which far exceed the calls that natural living makes upon them"(Humphrey, 1976). To relate this to the discussion here, the fact that we possess these abilities in the lab suggest they have a function "in the wild" otherwise they would have been lost through natural selection. The bottom line is thus to discover ecologically valid contexts that elicit such behaviour. Perhaps there are contexts where very precise 3D information (and therefore higher stereo acuity) is required. This is a very interesting area of research but it requires technological development towards a new field of 3D gaze behaviour.

Many technical developments have been made towards making 3D gaze tracking and interaction a reality. Already discussed in this thesis was the developed GT3D system, an ultra-low cost binocular eye-tracker and software for estimating 3D gaze position (Abbott and Faisal, 2012). More recently it has become standard for mobile commercial eye-trackers to be binocular (SMI Eye Tracking Glasses and Tobii Pro Glasses II). Making use of this data is the very exciting "EyeSee3D" system being developed by Pfeiffer et al. Here they intersect the 3D gaze vector with a 3D model of the environment that generated online by instrumentation of the environment with fiducial markers. Previous work of Pfeiffer et al. also explicitly estimated 3D attention volumes when looking at different test objects, developing methods to define 3D attention volumes and scan paths. There has been significant work in this direction, with new tools for visualisation of 3D data (Blascheck et al., 2014; Maurus et al., 2014), improving methods of 3D calibration (Tostado et al., 2016) and human-robot interaction experiments (Renner et al., 2014). However, to the best of my knowledge there have been no studies looking at 3D depth statistics "in the wild". These tools will be integral to future experiments "in the wild", as they capture an informative aspect of behaviour that is generally overlooked.

6.1.6 Specification

Based on the current state of the art and the current limitations, the aim is to develop an experimental methodology that enables efficient and scalable collection and analysis of gaze behaviour in real world scenarios. The aim here is to enable meaningful conclusions to be drawn from freely behaving humans in a scalable manner that does not require painstaking frame by frame annotation of behaviour. Towards this, sensory inputs are recorded using a head mounted eye-tracker, scene camera and microphone. Simultaneously, musculoskeletal motor outputs are recorded by motion tracking 51 degrees of freedom in the body and a total of 40 degrees of freedom in the hands. Recordings are made for three different settings (kitchen, bedroom and indoor navigation) with the total recording for each subject lasting hours rather than minutes. These scenarios involve many different free object manipulations, human-human interactions, navigation and eating covering an extensive compilation of everyday behaviours. Such extensive recordings that enable gaze to be related to full body kinematics in such natural, unconstrained and varied behaviours have not been previously been achieved.

6.2 Methods

6.2.1 Experimental Protocol

The experiment has 3 natural scenarios: breakfast time, evening chores and navigation. During each scenario, high level instructions are given such as “please lay the table for two people” or “please go ahead and have breakfast”. In the kitchen subjects conducted high-level activities such as laying the table for breakfast, making tea, preparing and eating breakfast and clearing up afterwards. While the bedroom scenario involves evening chores and leisure activities including sweeping the floor, making the bed, packing their bag, playing games and socialising. The navigation scenario involves walking to a specified location in the building (involving 4 flights of stairs followed by a lift and multiple corridors) while chatting with the experimenter and for one section carrying a cup of coffee. An overview is shown in Figure 54. It took 2 experimenters to set up all the equipment and run the experiments which last over 3 hours. One experimenter interacted with the subject and guided them through the tasks where absolutely necessary and interacted in tasks requiring a partner e.g. having breakfast together or playing games. The second experimenter observes and performs the online annotation.



- Lay table
- Make tea/coffee
- Prepare breakfast
- Eat Breakfast
- Clean up table



- Tidy room
- Make bed
- Organise book shelf
- Dress/Undress
- Clean away objects



- Play games
- Interact with people
- Read magazine
- Answer phone
- Write a short text

Figure 54: Experimental protocol overview with examples tasks .

6.2.2 Data collection hardware and collection framework

A portable eye-tracker (SMI Eye Tracking Glasses 30Hz, Sensomotoric Instruments, Teltow Germany) is used in combination with a portable full body motion capture suit, measuring 51 degrees of freedom (DOF) from the body with 17 inertial measurement units (IGS-180 Animazoo, Brighton UK) and 22 DOF from the right hand (Cyberglove 1) and 18 from the left (Cyberglove 3, Cyberglove Systems, San Diego California USA). All tracking equipment is marker-less and thus allows extensive behavioural monitoring “in the wild”. The experiments were filmed from a static video camera as well as the integrated scene camera of the eye tracker. In addition, online annotation was made by one of the experimenters to separate tasks and sub tasks.

Eye-Tracking: Eye movement data is collected using SMI Eye tracking glasses which record binocular eye-movements at 30Hz synchronised and aligned to an ego video feed (from the wearer's point of view). Software is written in C++ using the SMI API to communicate with the eye-tracking server. The program gives flexibility to calibrate and repeat calibration when necessary, to collect raw data and to inject time markers for events in the experimental protocol. These events include user calibration, synchronisation points and calibration checkpoints. Calibration checkpoints were made to update calibration periodically, to ameliorate eye-tracker slippage during the long recording duration. This happens at 5 minute intervals at which point a beep sounds and the subject stops what they are doing and looks at a marker on the back of their wrist until a second beep indicates to continue. The data is recorded on a laptop placed in a backpack on the subject.

Body tracking: The movement from the entire body (excluding fingers) is recorded using an IGS-180 motion capture suit (Animazoo UK Ltd, Brighton, UK). The suit is a 2 piece Lycra garment with integrated sensor nodes positioned to track key degrees of freedom in the body. Each sensor is a 9-axis inertial measurement unit (IMU) composed of a 3 axis accelerometer, 3 axis gyroscope and 3 axis magnetometer. The body sensor network is composed of 17 sensor nodes placed on each body segment as illustrated in Figure 56A. Sensor fusion is applied on board and the suit streams 51 degrees of freedom at 60Hz. Calibration of the motion capture suit is performed by a simple routine provided by the manufacturer as part of the control software.

Hand Tracking: The data from the subjects' left and right hand is recorded using a CyberGlove I and III respectively (CyberGlove LLC, San Diego, CA, USA). The CyberGloves are fabric gloves with stretch sensors placed over the joints of the fingers. Note that the right hand glove also measures the movement of the distal interphalangeal (DIP) joints, while the left hand does not. The hand movement is recorded at 140Hz for the left and 90Hz for the right hand, and converted using an 8-bit analogue-to-digital converter (ADC). This results in a claimed resolution of about 0.5 degree. Calibration is performed as a two stage-process: an initial calibration using software provided by the manufacturer and subsequent refinement by manually adjusting sensor gains and offsets while comparing the subject's hand with a visual rendering of the hand. The recorded data is streamed to a laptop on which it is saved along with a time-stamp obtained from the system clock.

Online Annotation: The subjects' behaviour is annotated online by the experimenter using a custom annotation GUI called Dynamic Activity Logging Interface (DALI). This system has been developed to allow an online annotator to provide detailed annotations of the subject's behaviour during the experiment. This is a challenge given the array of possible tasks and actions the subject will perform. The designed system was hierarchical in nature, leveraging the known structure of the tasks to intelligently display most likely behaviours given the setting and sub task. Initially DALI displays the possible settings, upon clicking one setting all possible associated tasks are displayed, upon clicking on a task all possible subtasks are displayed. This hierarchical system is shown in Figure 55. To give an example of the hierarchy : Level 1: "kitchen", Level 2: "make hot drink" , Level 3:"stir". In the event an action is not available, e.g. because the experimenter had not observed a particular behaviour before, a "NEW" action allows the time-stamp of manual annotations to be recorded. This new action can then be easily integrated into the configuration file and are automatically taken into account by DALI.

Scene Recording: The working area of the subject is monitored with a tripod mounted video recorder placed to capture the full scene. In addition the experiment is filmed from the subject's perspective with the integrated scene camera on the head mounted eye-tracker.

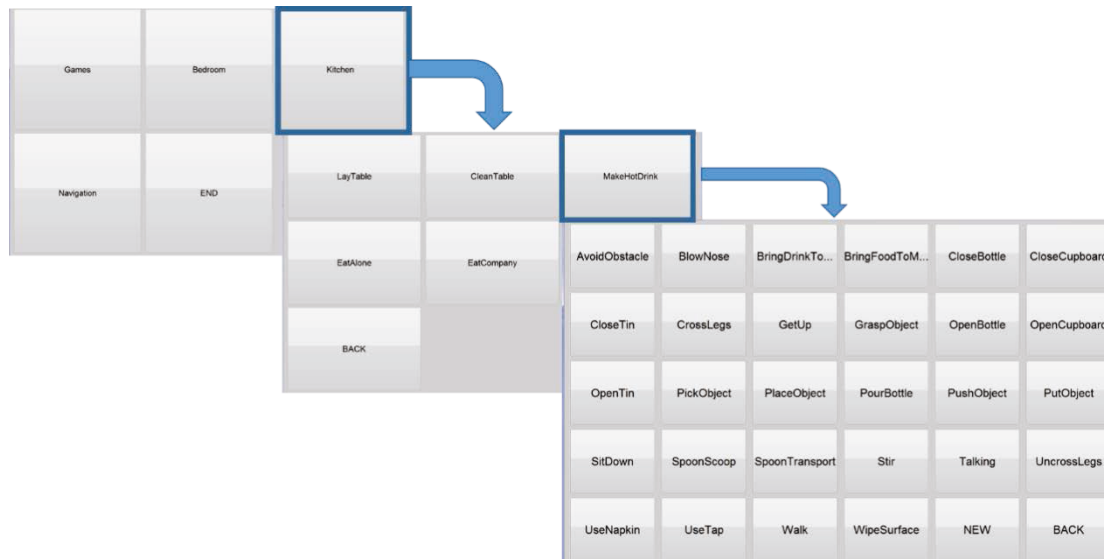


Figure 55: DALI Graphical user interface with multiple views representing an example annotation hierarchy.

6.2.3 Data integration and curation

To collect such rich data, these collection modalities must run on distributed systems simultaneously with the ability to synchronise and integrate data post hoc. This came with it challenges of time synchronisation and data integrity. Time synchronisation was achieved with two approaches, primarily relying on network time protocol (NTP) which allows different systems to poll a universal clock to maintain a common time reference. A secondary backup system involved manual synchronisation via button press injecting timestamp markers into the respective recording systems. Each experiment had as many as 100 files associated with it. To manage this, a database structure has been created to associate and link files to subjects, scenario and recording modality. This database structure is used to extract all data associated with a subject and scenario, to synchronise the modalities and finally to build a data structure amenable to the analysis with requisite pre-processing. Once synchronised, the suit and glove data were sampled at the eye-tracker frame rate (30Hz). This approach was taken to maintain data integrity in eye-movement data which all analysis is relative to here.

Figure 56 shows the overview of the sensor suite and the corresponding coordinate systems. The suit and glove sensor placement is labelled in Figure 56A and C respectively. The gaze tracker coordinate system is shown in Figure 56B, which includes the 3D gaze vectors and the gaze with respect to the scene video. To integrate suit and gaze data a skeleton model is used with the joint angles from the suit to create a 3D model of the body first. Using the head position and absolute rotation the gaze vectors are rotated and translated to be relative to the head Figure 56D. The 3D gaze is then estimated based on

GT3D system. In Figure 56D the blue line protruding from the head represents the head gaze orientation and the red lines represent the gaze vectors for each eye which terminate at their closest point of intersection. The mid-point of these positions is the 3D gaze estimate.

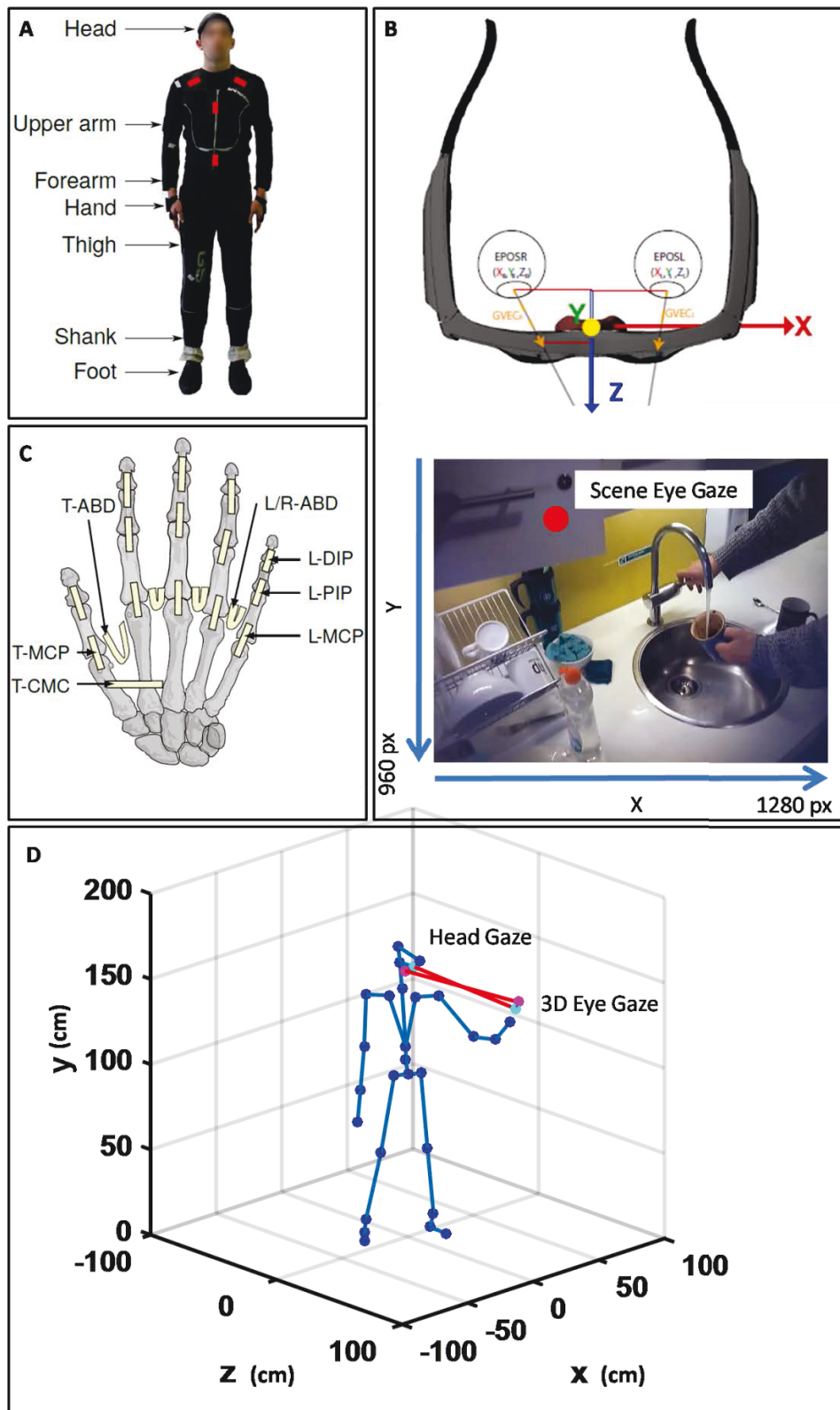


Figure 56: Sensor Suite and Coordinate system. A. Body tracking suit sensor placement. B. Eye-tracker coordinate system in world and in the scene camera view. C. Right Hand Glove sensor placement with respect to the hand joints. D. Integrated suit and gaze data. 3D gaze is estimated based on GT3D system and translated based on head position and orientation. The blue line protruding from the head represents the head gaze orientation and the red lines represent the gaze vectors which terminate at their closest point of intersection. The mid-point of these positions is the 3D gaze estimate.

6.2.4 Data Analytics:

Event detection: The eye tracking system used makes a classification of the behavioural categories of visual intake, saccade, and blink with an additional noise category of 'unknown'. The exact details of how these events are detected are not provided by SMI, but the following describes what information is provided. The unknown category is given when there is an unidentified period of signal such as infeasible acceleration of gaze. Blinks are detected based on sudden change of pupil diameter which suggests pupil occlusion and hence a blink. Visual intake is defined as "fixations and smooth pursuit" in their documentation, however I assume this also encompasses smooth eye-movement in general i.e. also VOR). The algorithm behind this is proprietary and is only alluded to in the documentation: "Saccades are found by evaluating the point of regard velocity and acceleration against adaptively adjusted thresholds. Heuristics are applied to assure minimum amplitude and duration is met for the saccades.". It is unclear how this is achieved given the low data rate. To provide a comparison, and enable classification with the composite head and eye data an alternative algorithm - dispersion threshold identification (I-DT) is used. This algorithm is commonly applied to low frame rate data and is defined in the background section of this chapter (Salvucci and Goldberg, 2000). Here the dispersion is measured as the width and height of the gaze positions' bounding box. The saccades are inferred as the periods that aren't fixations or blinks.

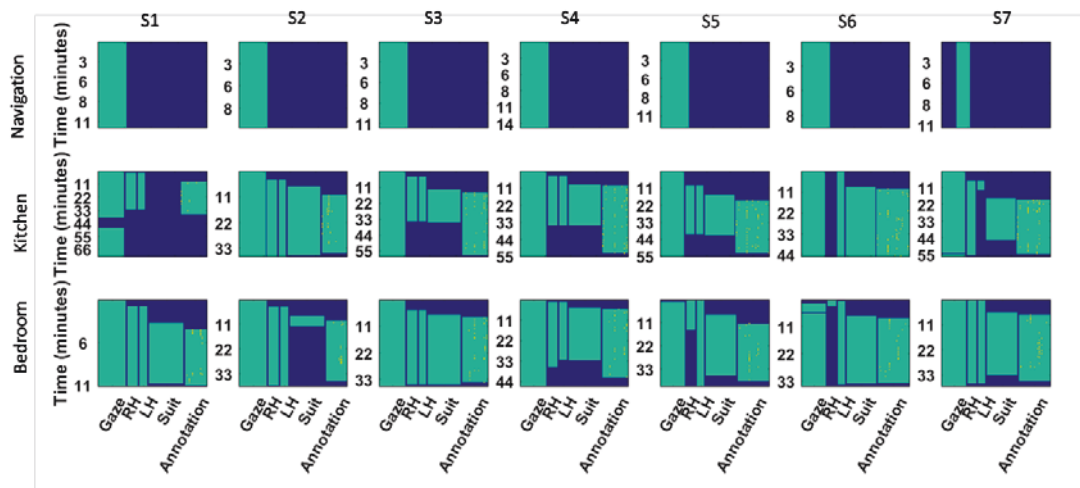


Figure 57: Overview of data collection sets. Rows are samples and columns are sensor modalities. Light blue indicates data exists while dark blue means no data was recorded in this period.

Grasp Annotation refinement: The glove data is used to refine the timings of grasp annotations made by the experimenter. To do this a grasp metric based on the summation of finger velocities is used, where positive corresponds to an opening hand and negative to a closing hand. Grasps take many shapes and forms with different combinations of digits involved and thus this grasp metric enabled a more general grasp to be detected. For each online annotation of a grasp (from DALI - see methods), the nearest hand closure is identified and the annotation time is updated to the end time of this closure. Thus the more crude online annotations were aligned precisely with the nearest high resolution detection of hand closure. The online annotations did not specify which hand had made the grasp and thus this was also identified. If both hands had closing shapes near an

annotation, the grasp was scored based on the amplitude and the distance in time from the annotation and the highest score was chosen.

6.3 Results

6.3.1 Data set overview

Recordings include 7 subjects, each in 3 settings (kitchen, bedroom and navigation), totalling more than 10 hours of data. A complete gaze data record was achieved for all subjects in all settings. Owing to equipment malfunctions the data recorded for the gloves and suit has some gaps. Figure 57 gives an overview of the dataset collected. There was no suit or glove data during navigation due to the suit's wireless range. In total each subject had around 40Gb of data associated with them. Figure 58 gives an overview of a full dataset's recordings. The suit data is converted into endpoint information to allow animations of body movements and 3D gaze to be observed simultaneously with the scene camera. Figure 59 shows a 1 minute section of processed data. This includes the gaze data and the derived event classification (represented with coloured banding (Figure 59A)). This will be shown in more detail in subsequent sections. Figure 59B shows the online annotation. While this does not capture all behaviours it gives rich information that can be refined based on the raw movement data. Figure 59C shows the derived grasp metric (left hand - black, right hand - lilac) , where positive corresponds to hand opening. The lines on this figure represent the online annotations refined with the glove data. The hand endpoint position X (horizontal), Y (vertical) and Z (depth) respectively for the left hand (black), right hand (lilac) and 3D gaze (green) are shown in Figure 59 (D,E,F). There are periods where the tight locking between gaze and hand end-point are clear and interestingly there seems to be more association with the left hand (non-dominant hand) than the right.

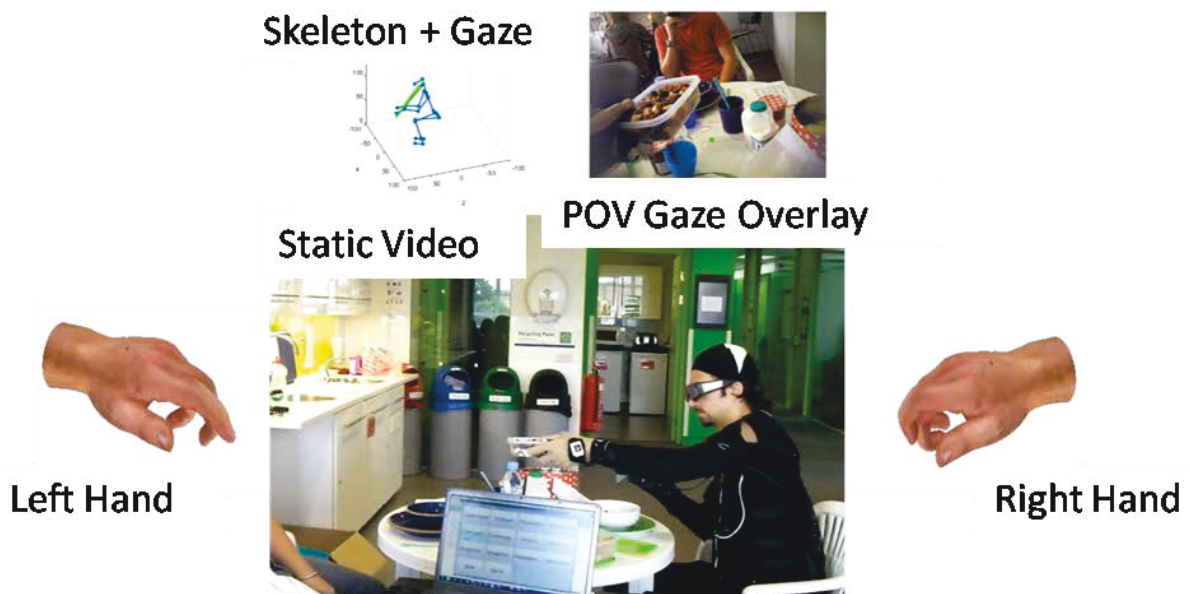


Figure 58: Experimental overview showing all recording modalities.

To give more context to the data, skeleton and Scene snapshots corresponding to the shaded periods on the plot are shown in Figure 59(1,2,3,4). The skeleton is shown with the head direction indicated by blue projection from the top of the head and the gaze 3D endpoint indicated by the red line (refer to Figure 56 for more details). On the scene view the gaze position is indicated (red - fixation, blue – blink, green - saccade). In Snapshot 1. the subject offers the box, opens the box and places the lid on the table. The subject fixates on the rim of the lid as they peel it open and then they fixate on the table as they place the lid there. In Snapshot 2. the subject looks at another box, moves their arm towards it to make a grasp, by which point they fit in a blink. As they pick up the box they are already looking to the left hand where they will bring the box to hold while they open the lid. The first grasp has been missed by the online annotator (but can be seen in the grasp metric Figure 59C as a dotted line) while the second is correctly logged in Figure 59B and the identified hand and temporally refined based on the grasp metric in Figure 59C. Snapshot 3 shows another interaction, this time passing the fruit and nuts across the table before grasping another box to serve themselves. This grasp is also correctly logged by the annotator (Figure 59B) and the refined marker is displayed in Figure 59C. During the serving period (snapshot 4) the hand oscillates back and forth scraping the contents into the bowl. This oscillating behaviour is also seen in the gaze position as they look back and forth between the contents of the box and the amount they have served themselves in the bowl. There is very rich information here and the association between the gaze and behaviour is very clear when you can see what is happening in the scene. While this can be seen in the endpoint data, the 3D gaze signal can be noisy, particularly in the y (vertical) and z (depth) dimensions.

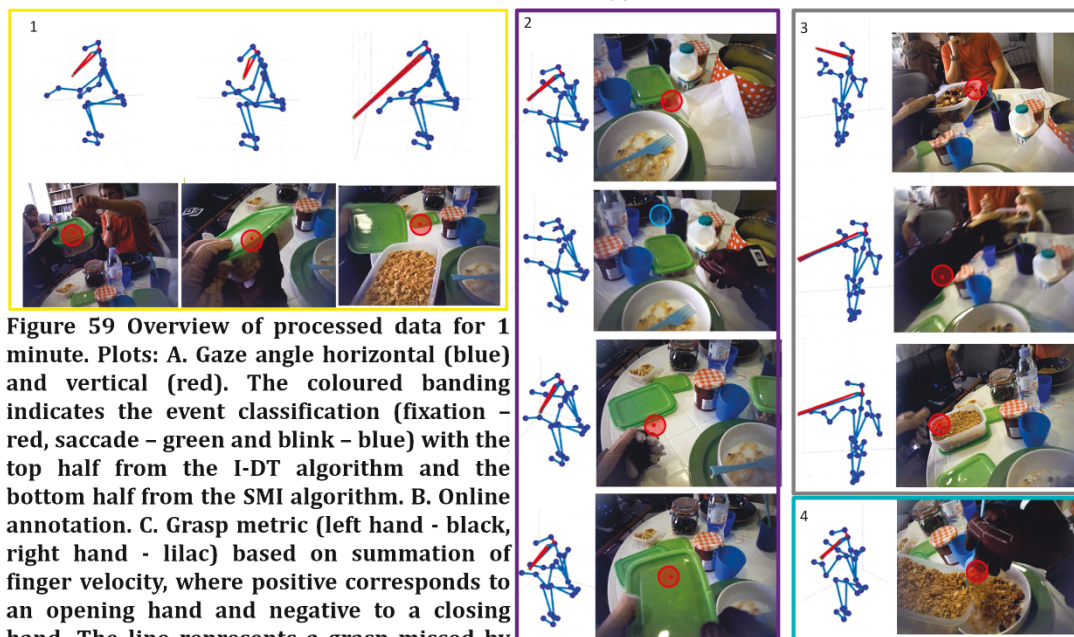
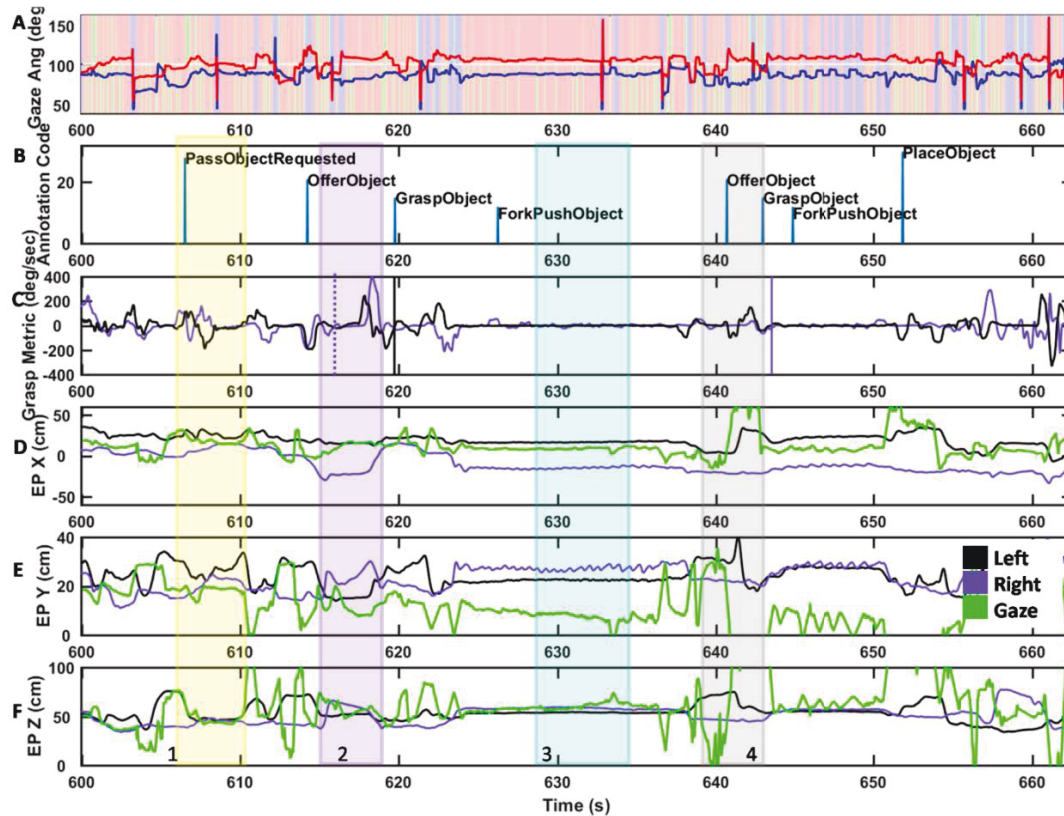


Figure 59 Overview of processed data for 1 minute. Plots: A. Gaze angle horizontal (blue) and vertical (red). The coloured banding indicates the event classification (fixation – red, saccade – green and blink – blue) with the top half from the I-DT algorithm and the bottom half from the SMI algorithm. B. Online annotation. C. Grasp metric (left hand - black, right hand - lilac) based on summation of finger velocity, where positive corresponds to an opening hand and negative to a closing hand. The line represents a grasp missed by the online annotator. D,E,F. End point position X (horizontal), Y (vertical), Z (depth) respectively for the left hand (black), right hand (lilac) and 3D gaze (green). Skeleton and Scene snapshots (corresponding time period shown on plots): The skeleton is shown with the head direction indicated by blue projection from the top of the head and the gaze 3D endpoint indicated by the red line. The scene view and gaze position (red - fixation, blue - blink) Snapshot 1. Subject Offering box, subject opening box and subject putting lid on table. 2. Subject picks up box and brings it to their other hand. 3. Subject offers and passes fruit and nuts and then reaches for the granola. . 4. Subject scrapes granola into bowl (this continues for a long period but view very similar; just oscillations of the hand back and forth – clearly seen in Plot E).

6.3.2 Eye Movement direction and depth Statistics

Raw Gaze direction statistics: Figure 60A shows the gaze direction distribution averaged across subjects for each setting. Firstly the distribution shows significant oblique eye-movements where the peaks for purely vertical or horizontal eye-movements are slightly more likely. In addition, there is a slight difference across settings with subjects looking directly down more often during the kitchen and bedroom settings.

Gaze Depth Statistics: The depth statistics for fixations are shown in Figure 60B for all subjects, grouped by setting. The difference is quite stark with the navigation setting yielding a much broader distribution with significantly more fixations over 1m in depth. The depth distributions for the kitchen and bedroom scenario are more similar with the majority of fixations at depths below 1m, consistent with the fact that in both the kitchen and the bedroom settings involve object interaction and manipulation while navigation is purely walking around a building. Interestingly there are also slight differences between the kitchen and bedroom setting with more close fixations (<20 cm) in the kitchen setting. This may be because a large proportion of behaviour in the kitchen involves eating.

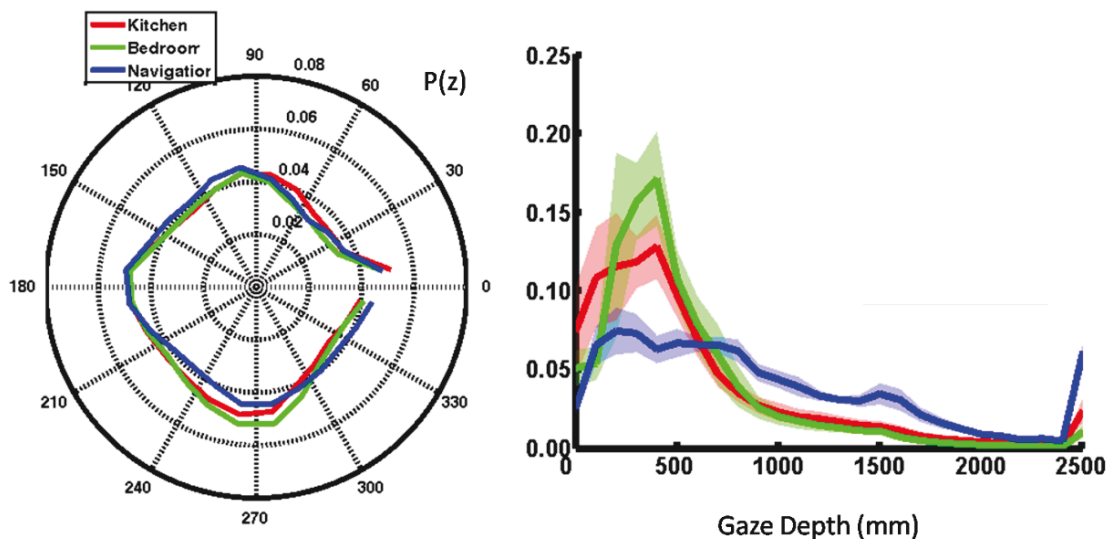


Figure 60: Gaze statistics cross setting aggregated over all subjects. A. Gaze velocity direction probability. B. Gaze Fixation Depth probability.

6.3.3 Gaze Event Segmentation

The dispersion based (I-DT) algorithm is compared to the SMI velocity based algorithm. The event categories are labelled blink, saccade and fixation, though the fixation category for SMI strictly should be labelled visual intake (see methods). Firstly the general event statistics are compared in Figure 61. For each plot (Figure 57A,B & C) the bars represent the mean metric across all settings and subjects with the yellow connecting line representing the 2 tail paired t-test between the two methods taken for every subject and setting. The error bars represent the average and standard deviation when separated by setting (kitchen - red, bedroom - green, and navigation - blue), with the 2 tail paired t-test result within subjects, across setting in black. The event rate, without separating the type of event is considered first in Figure 61A. This general event rate is significantly higher for

SMI at 4.9 events/s compared to I-DT at 4.0 events/s (paired t-test, all subjects and settings, $p < 0.01$). This suggests that the SMI approach makes a finer distinction between gaze events. For the SMI method, the event rate is not significantly different between the different settings, while for I-DT, the event rate is significantly larger during navigation than both bedroom and kitchen scenarios. This implies that there are more event transitions during navigation, suggesting that the gaze dispersion is in general larger.

The relative likelihood of each event type (blink, saccade, fixation) is also shown in Figure 61B. The SMI method detects significantly more fixations than the I-DT method and consequentially significantly less saccades. This suggests that the I-DT classifies many of the SMI fixations as saccades. This can be seen more qualitatively in Figure 62 which compares the event classifications made by the two methods on the same one minute gaze trajectory. For the I-DT method, significantly less fixations are observed during navigation than both kitchen and bedroom scenarios, significantly more saccades during the kitchen task with respect to the bedroom task and no significant difference in blinks. While for the SMI method, there is no significant difference between fixation probability across settings. Instead, a significantly lower saccade probability during navigation compared to the bedroom and a significantly higher blink probability during navigation compared to the bedroom scenario. The number of unknown events (unclassified) is not significantly different across settings. Finally, the blink detection method used is the same prior to each event detection method being run and thus does not differ (probability difference is due to differing total number of fixation/saccade events).

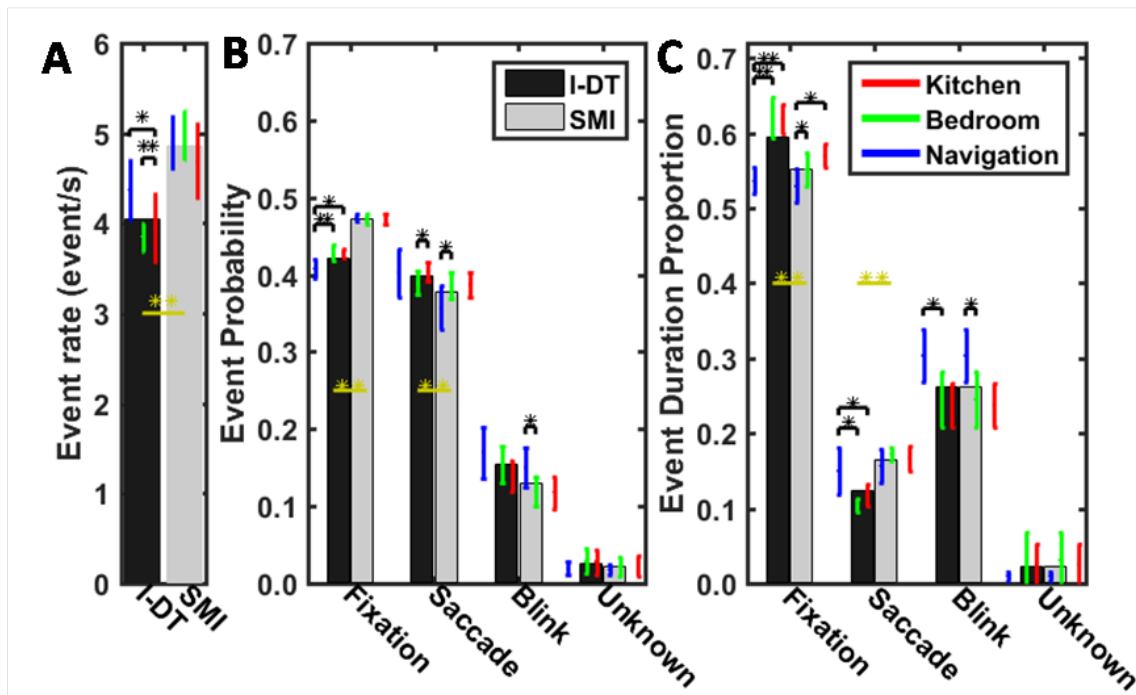


Figure 61: Event Statistics Comparison between SMI and I-DT methods. For each plot the bars represent the mean across all settings and subjects with the yellow connecting line representing the 2 tail paired t test between the two methods for every subject and setting. The error bars represent the average and standard deviation separated by setting (kitchen - red, bedroom - green, and navigation - blue), with the 2 tail paired t test within subjects across setting in black. A. Mean number of events (non-discriminative) per minute. B. Event probability, separated into respective events (fixation/visual intake, saccade, blink and unknown). C. Event duration Proportion, the proportion of total time spent in each event. *is $p < 0.05$, ** $p < 0.01$.

The event probability captures the frequency of events but does not capture the amount of time spent in each event, thus the event duration proportion is presented in Figure 57C. It expresses the proportion of time spent in each event state for the experiment duration. It combines frequency and duration information, an important metric for fixations as it tells us the proportion of time spent taking in information. The I-DT method results in significantly more time spent making fixations and significantly less time making saccades and (the blink duration is identical). Looking at this metric cross setting, the time spent making fixations is significantly less during navigation for both I-DT and SMI, but the effect is much stronger for I-DT. This setting distinction is strengthened further for I-DT saccades which occupy significantly more time during navigation, while no effect is seen for SMI. Finally, for blink duration, significantly more time is spent blinking during navigation. Again, the qualitative differences across settings can be seen in Figure 62 which looks at a 1 minute segment for the same subject in each setting.

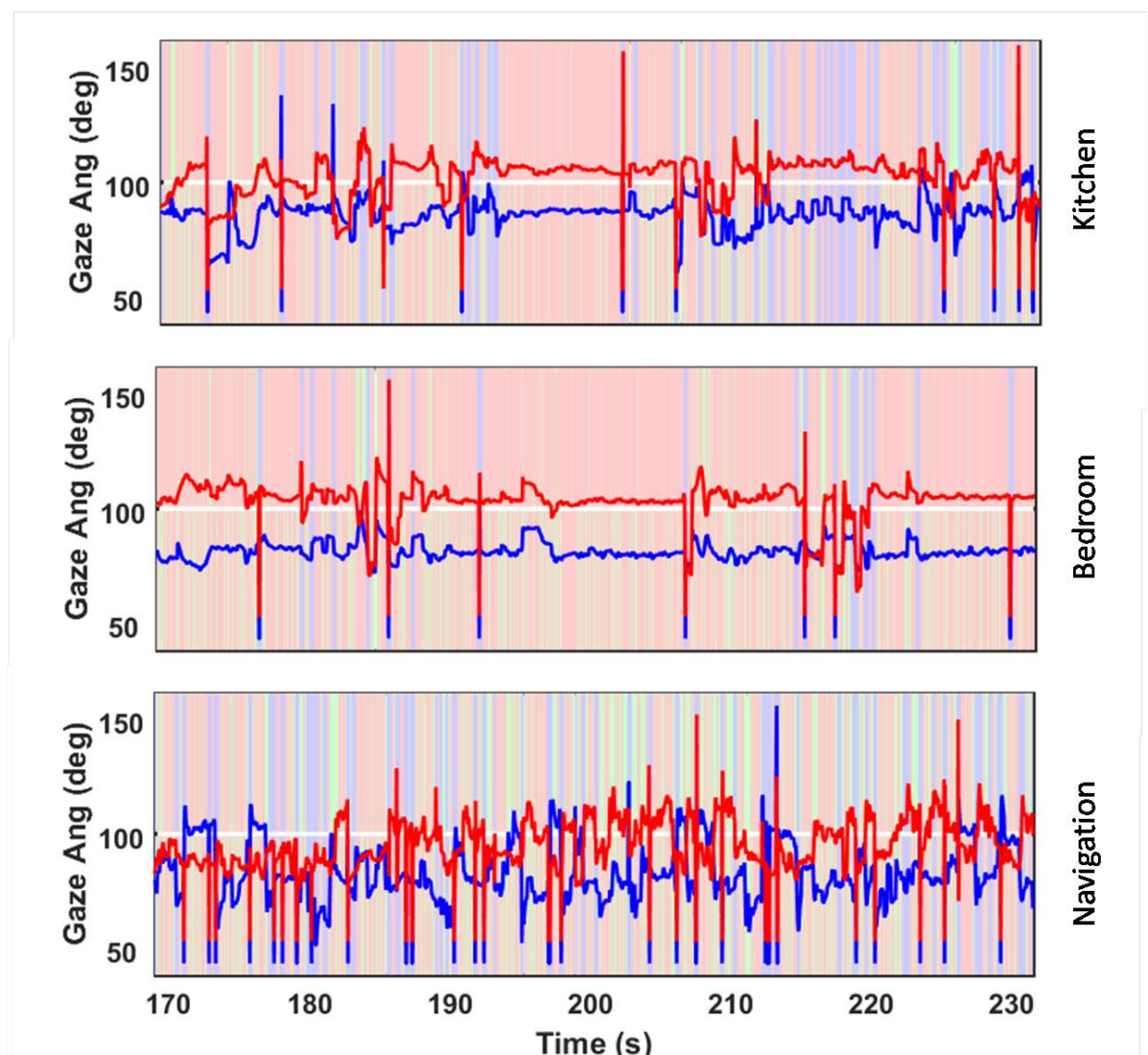


Figure 62: Gaze trajectory and event detection comparison. The gaze angle horizontal (blue) and vertical (red). The coloured banding indicates the event classification (fixation - red, saccade - green and blink - blue) with the top half from the I-DT algorithm and the bottom half from the SMI algorithm.

Figure 63 gives a more qualitative example of the difference between SMI and I-DT classification by comparing the methods on the same gaze trajectory plotted in 2D and with the associated time evolution of position and velocity. It is clear in this example that smooth eye-movements are being classified as fixation by SMI, while I-DT will classify them mostly as saccades, with intermittent fixations depending on the speed of the smooth movement. To further understand the nature of the classified *eye gaze* behaviour, the *gaze in world* behaviour is observed. This is the gaze based on head and eye-movements. The gaze pattern is very different when head movement is also integrated as shown in the right hand panel of Figure 63. The periods of smooth movements are now tightly clustered gaze locations and thus the head movement is equal and opposite to the eye gaze and thus it can be inferred that these smooth eye-movement periods are vestibular ocular reflexive (VOR) movements, stabilising the gaze during a head movement. Applying the I-DT algorithm to the *gaze in world* gives a much more similar classification to SMI, with one remaining difference being that the SMI method captures the full trajectory of the saccade (start and end samples) rather than a single time point.

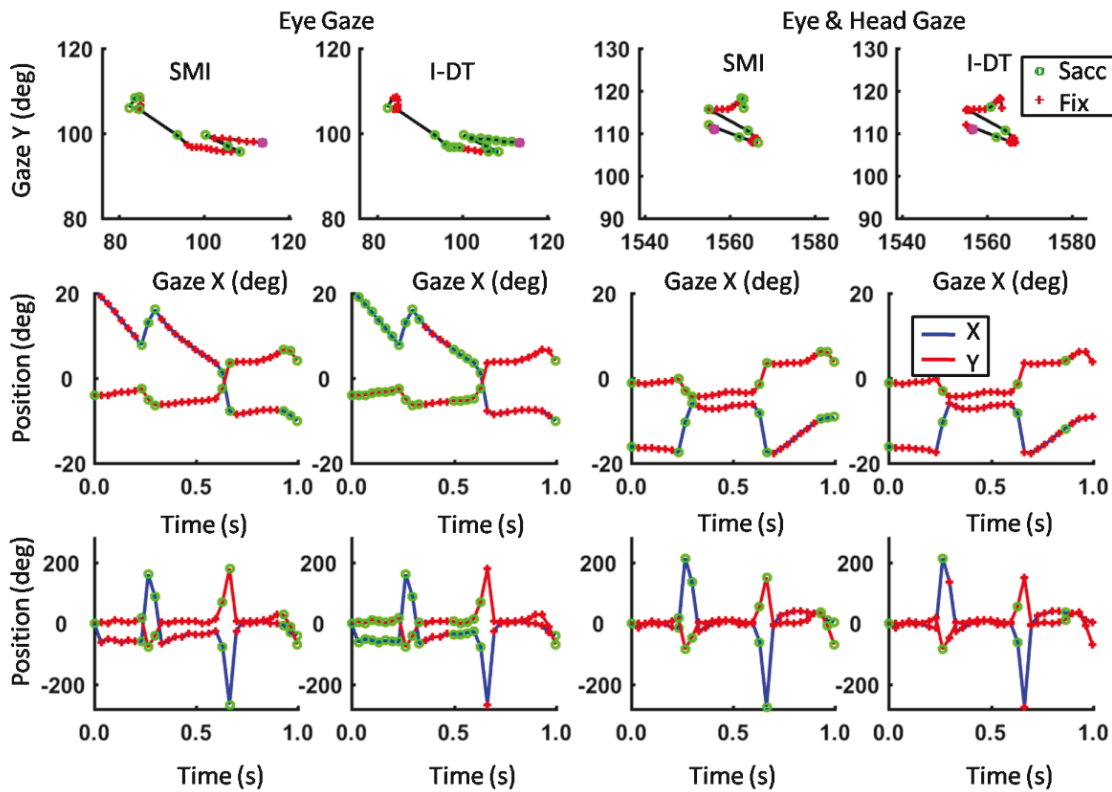


Figure 63: Effect of head on gaze trajectory and subsequent gaze event classification. In the left panel the eye gaze trajectory is shown with the respective gaze event classification for SMI and the dispersion based (I-DT) algorithm. This is then compared in the right hand panel for the same data samples with the eye and head gaze, i.e. the gaze in world as opposed to gaze in head.

6.3.4 Gaze event characteristics: method comparison

Looking at the fixation and saccade characteristics i.e. their duration, amplitude etc, as classified by the two methods further highlights the differences in these methods. The relative frequency distribution over the fixation amplitude, taken as the angle subtended by the first and last data point in the fixation, is shown in Figure 68A. The profiles are very

different with the SMI fixations having a low modal amplitude of 0.5 degrees with a long tail with 15% of fixation amplitudes greater than 5 degrees, compared to I-DT with a modal amplitude of 3.75, with a sharp drop and 0 amplitudes greater than 5 degrees. This cut off is due to the dispersion threshold of 5 degrees. To understand the shape of the fixation trajectories, the direction standard deviation across the fixation gives a measure of how coherent the direction of movement is. A SD of 0 would suggest a straight line. This is shown in Figure 68B, with the I-DT fixations having a higher likelihood of high directional variation suggesting a more random trajectory rather than coherent straight line. So the SMI algorithm has more high amplitude fixations and higher direction coherence, which together are suggestive of smooth eye-movements. The saccade amplitude has a similar profile for both methods (Figure 68C), while the I-DT saccades have a larger proportion of mid range (30-100 degree/s) saccades as shown in Figure 68D. This suggests there are more smooth eye-movements in the saccade category for the I-DT method.

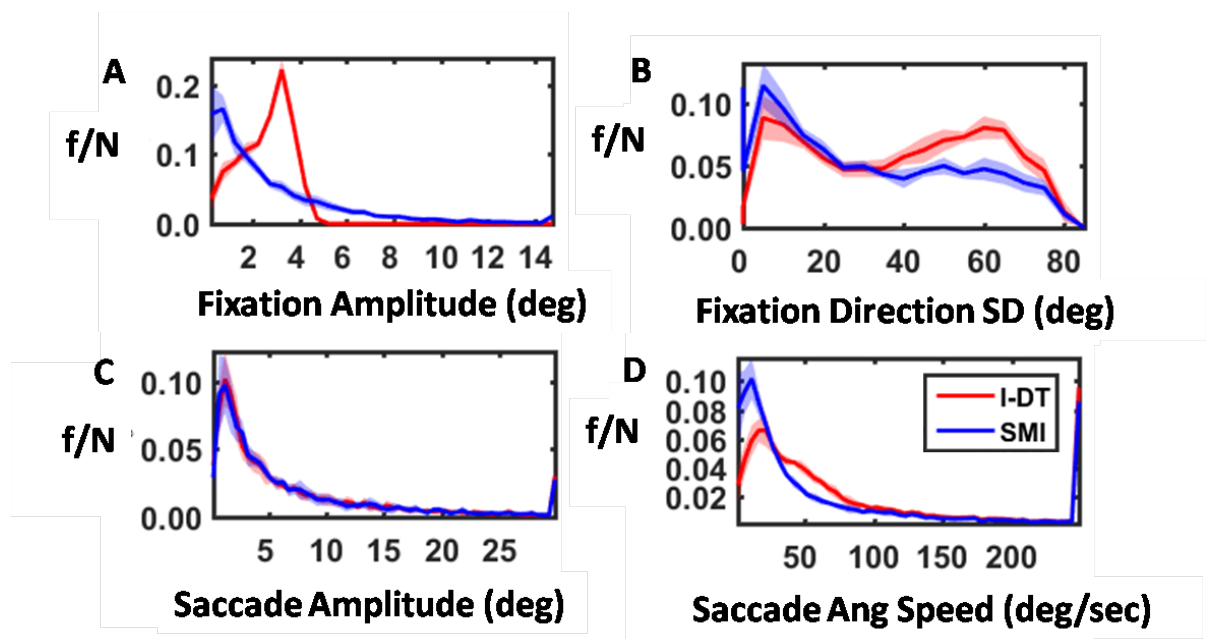


Figure 64: Gaze characteristics relative frequency distributions averaged across all subjects and settings (shaded area is standard error) compared across event detection methods (I-DT - red SMI-blue). A. Fixation amplitude taken as the angle subtended by the first and last data point in the fixation. B. Fixation direction standard deviation taken over the sample to sample gaze direction. C. Saccade amplitude is taken as the angle between the previous the subsequent fixation. D. Saccade Angular Speed is the sample to sample gaze speed (magnitude only).

6.3.5 Gaze event characteristics: setting comparison

To investigate further, the event characteristics across setting were analysed and three characteristics that showed significant differences are shown in Figure 65. The saccade angular speed for I-DT saccades (Figure 65A) shows the bulge observed previously in the mid range of velocities (30-100 deg/sec) is largely due to the navigation setting. Thus as might be expected, more eye-movements in the smooth range are observed in the navigation setting and that the I-DT algorithm classifies them as saccadic eye-movements.

Figure 65B shows the SMI fixation angular speed. Here the same bulge in the distribution is observed for the navigation setting, further suggesting more smooth eye-movements during navigation and that these are classified as fixations for the SMI method. Finally, as the last piece of evidence, Figure 65C shows that the fixation (SMI) direction standard deviation is lower for navigation. Thus again furthering the evidence to support the hypothesis that more smooth eye-movements are made during the navigation setting.

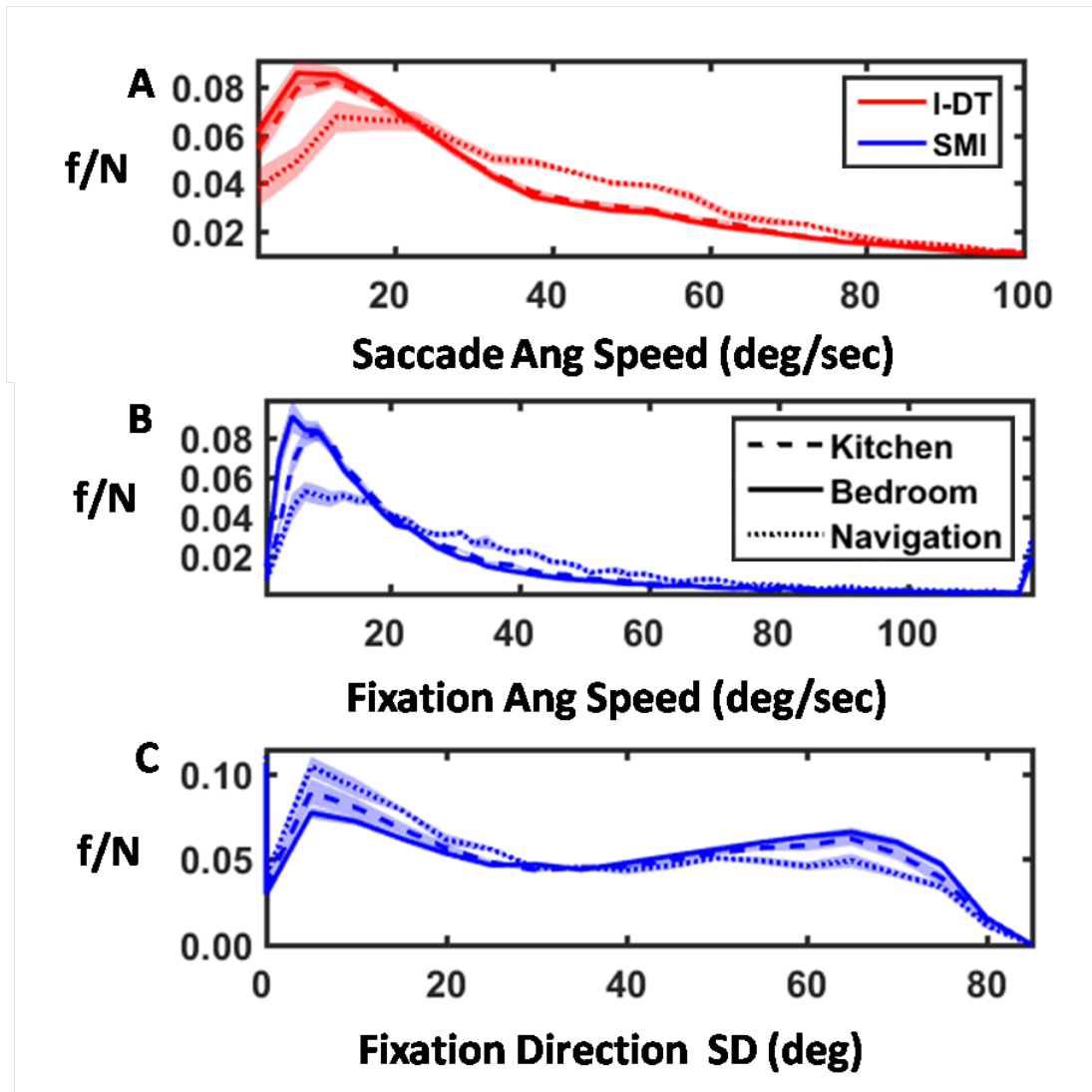


Figure 65: Gaze characteristics relative frequency distributions for each setting, averaged across all subjects (shaded area is standard error, I-DT - red SMI-blue, dashed line -kitchen, solid line bedroom and dotted line navigation). A. Saccade Angular Speed is the sample to sample gaze speed (magnitude only). The range is limited between 0 and 100, above which the distributions are almost identical, both have 24% >100 deg/sec. B. Fixation Angular Speed is the sample to sample gaze speed during fixation (magnitude only). C. Fixation direction standard deviation taken over the sample to sample gaze direction.

6.3.6 Blink Characteristics:

Observing the fixation position prior to and post blink revealed large eye-movements commonly occur during a blink period. Rather than ignoring this data the eye-movements during blinks are interpolated and the resulting characteristics compared with saccades in Figure 66. Over all subjects and settings, the number of blinks per minute was on average

36±12.1 (Figure 66A), compared to saccades at 111±22.3 per minute. Separating this characteristic into each setting there are 46.0 ±12.6 blinks per minute for navigation, 35.2 ±11.1 for the kitchen and 36.2±11.8 for the bedroom, while for saccades (SMI) a reduction is seen for the navigation setting towards 105.1±22.3 compared to 110±19.7 and 116.8±26.3 for kitchen and bedroom respectively. The event duration is significantly longer for blinks, peaking at 264ms compared to 33ms for saccades (Figure 66B). The blink duration distribution is also considerably wider with 20% lasting between 0.5 and 1 second for blinks, with no saccades lasting longer than 0.5s. The eye movements during blinks tend to have larger amplitudes than the majority of saccades with >20% above 20 degrees compared to just 3% for saccades (Figure 66C). The start-finish angular speed profile is calculated based on the amplitude from the end of the previous event to the start of the next event as this is the only information available for blinks (Figure 66C). The distributions are similar when the same metric is applied to saccades, with slightly more lower speed eye-movements associated with blinks (~50% saccades less than 60 deg/s compared to ~75% blinks) with many more saccades reaching speeds greater than 140 deg/s (24% saccades, 2% blinks).

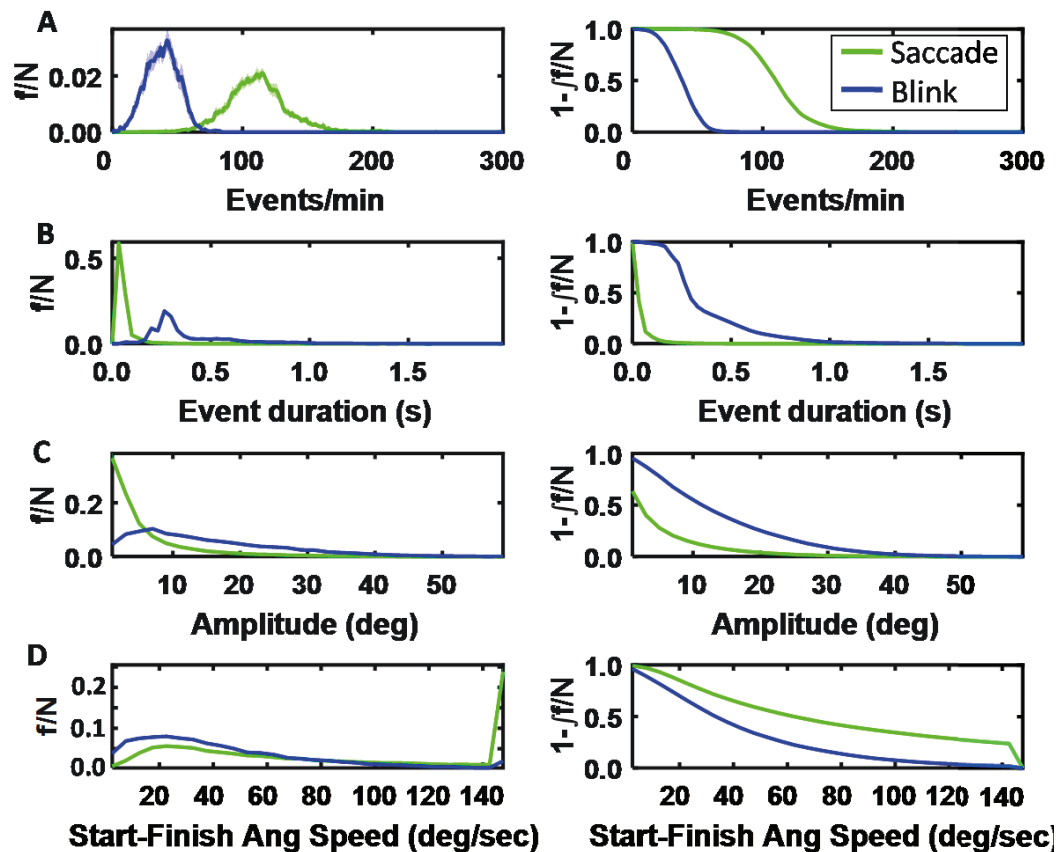


Figure 66: Blink (blue) characteristics compared with saccades (green) Characteristics based on eye-gaze alone. Left hand column: Relative frequency distribution over the following characteristics. Right hand column: The cumulative frequency subtracted from one, expressing the proportion of characteristics above each value on the x-axis. A. Event rate. B. Event duration. C. Gaze Amplitude expresses the angular change from before the event to the after. D. Start Finish Angular speed is the Gaze amplitude from C over the event duration. These statistics are based on all subjects in all settings.

The direction of these blink related eye-movements is also compared with saccades and fixations in Figure 67A. The distribution is similar for fixations and saccades, though there

is a slight bias to the right with 48% saccades in right pointing bins and 42% in leftwards. This bias is similar for fixations at 43% leftwards and 48 rightwards. For blinks the distribution narrows with a smaller vertical component and with a significant bias to the left, opposite in direction to blinks and saccade bias. This significant bias (55% to left and 38% to right) suggests blink related eye-movements are more likely to occur to the left. This effect is also observed when characteristics are separated by setting (kitchen, bedroom, navigation) (Figure 67B) negating a situational effect such as the room layout. In addition this effect is observed in the head gaze alone (Figure 67C), becoming even more pronounced when the eye and head gaze are combined to give the gaze in world (Figure 67D). Given that all subjects were right side dominant, there is a higher probability of gaze evoked blinks when moving gaze to towards the non-dominant side. This bias may imply different informational constraints when looking towards the dominant versus non-dominant sides. Importantly, it shows that this behaviour should not be discarded.

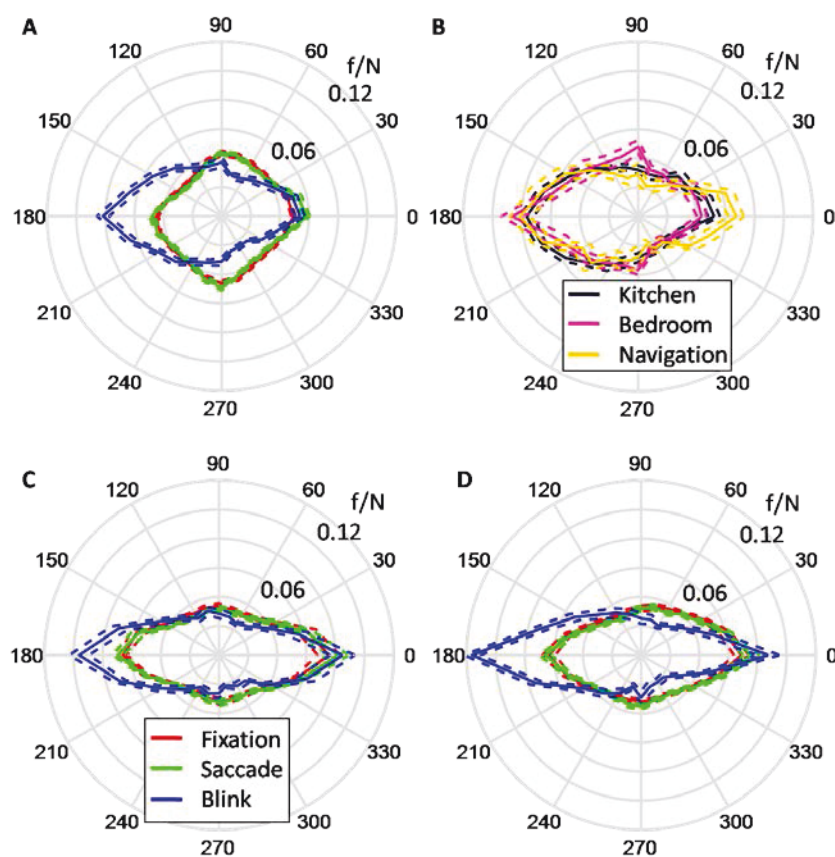


Figure 67: Gaze Velocity Direction Comparison. A. Relative frequency distribution of eye gaze direction averaged over all subjects (N=7) and settings (Kitchen, Bedroom, Navigation) separated for each gaze event - fixation (red), saccade (green) and blink (blue). B. Relative frequency distribution of eye gaze direction during blinks only, averaged over all subjects (N=7) and separated into settings - Kitchen (black), Bedroom (magenta) and Navigation (blue). C. Relative frequency distribution of head gaze direction averaged over all subjects (N=7) and settings (Kitchen, Bedroom, Navigation) separated for each gaze event - fixation (red), saccade (green) and blink (blue). D. Relative frequency distribution of head and eye gaze direction averaged over all subjects (N=7) and settings (Kitchen, Bedroom, Navigation) separated for each gaze event - fixation (red), saccade (green) and blink (blue). Note: Dotted lines indicates one standard error from the mean.

6.3.7 Gaze event characteristics: across subjects and previous studies

In general, previous studies have demonstrated that the fixation duration is affected by task. Figure 68 compares the fixation time across setting for the I-DT method and SMI, it also compares durations taken from previous studies that include both lab based and "in the wild" studies. In general, as has previously been observed, the range of fixation durations is very large in natural behaviour compared to the lab, ranging from 70ms to >2 seconds. Previous studies have shown that fixation duration is much larger during active natural behaviour (tea making mean 457ms and mode 260, sandwich making mean 398ms and mode 120ms) than during sedentary tasks such as reading and picture viewing (reading mean: 224ms and mode: 180ms, picture viewing mean: 295ms and mode: 200ms). In this study, this statistic is very dependent on the method used to classify events. Using the I-DT a similar distribution to previously studies is found, with fixations lasting on average 328ms (mode: 115) in the kitchen and 339 (mode: 115) in the bedroom setting, compared to 457ms (mode 260) in tea making and 398 (mode 120ms) in sandwich making. While during reading, fixations last on average 224ms (mode 180ms) and in picture viewing 295ms (mode 200). In addition, however, during navigation the fixation durations are similarly low, lasting an average 269ms (mode 115ms). As it was found in tea making and sandwich making, for the kitchen task long tailed distributions are observed, with the modal duration being much shorter at 115ms (I-DT) compared to the modal 120ms in sandwich making. For the SMI classification, as observed previously, more events are classified resulting in shorter fixation durations in all settings (mean kitchen 213ms (mode 82ms) , bedroom 199ms (mode 82ms) and 195 (mode 82ms) for navigation). It is interesting to note that the I-DT algorithm gives more similar results to the hand coded gaze events in the previous studies.

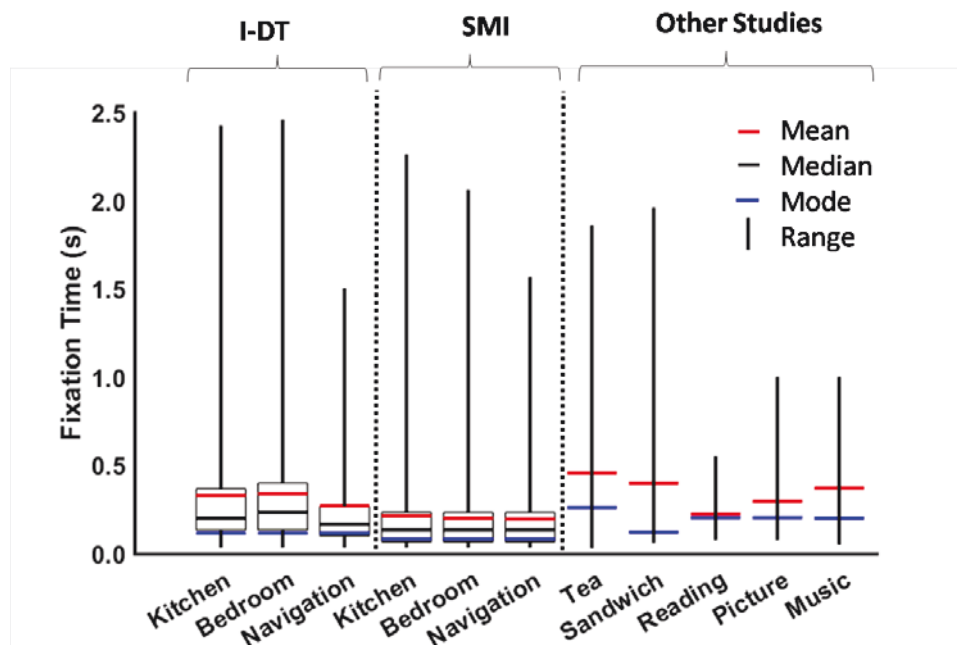


Figure 68: Fixation Duration Comparison. Compares fixation duration box and whiskers for data aggregated across all subjects ($N=7$, $t_{\text{mean}}=55\text{min}$) for this study, with each event detection method (I-DT and SMI) with statistics from other studies that include Tea making (Land et al., 1999) ($N=3$, $t_{\text{mean}}=4\text{min}$) , Sandwich making ($N=4$, $t_{\text{mean}}=2\text{min}$), Silent Reading, Picture viewing ($N=3$) and Sight reading playing piano (M.F. Land & Tatler, 2009). Note the mode is the modal bin centre where bin size is 33ms, the data recording interval.

The saccade amplitude is also compared between this study and Land's pioneering study "in the wild" in 1999. For Land's study the amplitude was measured by hand and combined head and eye movements, as such the same has been done here to allow comparison. The saccade detection method used was SMI's as based on previous results presented, the SMI saccade is less polluted by smooth eye-movements than the I-DT method. Figure 69A shows the full relative frequency distribution across all subjects during the kitchen task (same binning (2 degree) and axis as Land's in Figure 69C) while Figure 69B shows the same distribution with a limited range to zoom in on the distribution. Firstly the distributions are very similar across subjects. Secondly, in comparison to the distributions obtained in Land's study, both have a similar shaped long tail distribution, however, the modal and mean amplitude are much lower in this study. The mean amplitude for Land's study was between 18 and 20 degrees while the mean ranged between 8 and 12 degrees here (likely to be because the head information is not used in the classification process). Compared to typical amplitudes of 1-2 degrees when reading (letters ~0.2 degrees). In both our study and Land's the gaze is taken as eye and head movement together. Like Land observed, there are also a number of very large saccades >90 degrees in this study. It is also important to draw attention to the total number of saccades in the distributions, which range from 401-545 saccades for Land's study with recordings lasting around 4 minutes compared to 5584-7827 saccades here, with durations of 37-60 minutes.

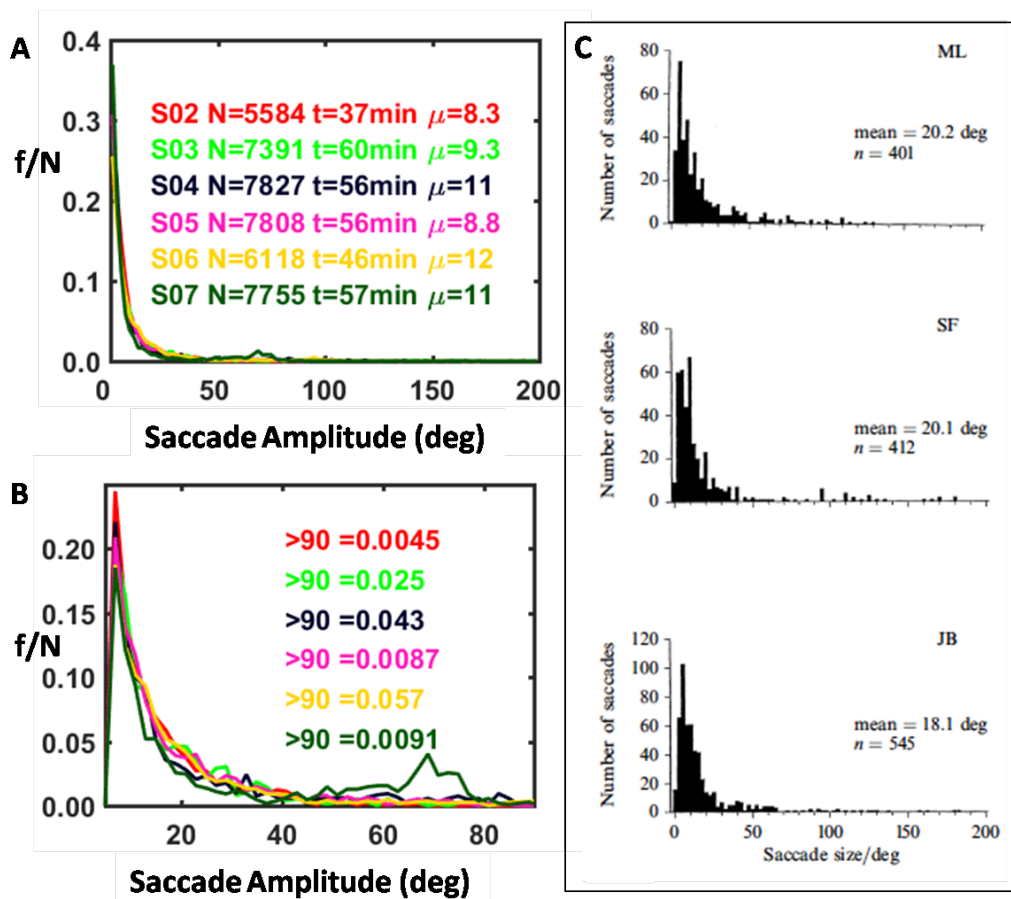


Figure 69: Saccade Amplitude Comparison. A. Gaze Saccade Amplitude (SMI saccades, head and eye amplitude) relative frequency distributions for subjects 2-7 (subject 1 had no suit data) with the number of saccades and the experiment duration and mean amplitude B. Restricted range Gaze Saccade Amplitude (head and eye) relative frequency distributions for subjects 2-7 (subject 1 had no suit data) restricted to between 5 and 90 degrees with the proportion greater than 90 degrees labelled.

C. Comparison study saccade amplitude (from(Land et al., 1999)). Shows the saccade frequency distributions for 3 subjects (each histogram) with the mean and the number of saccades. This was for studies lasting ~4 min.

6.4 Discussion

Developing the experimental framework, collecting the extensive behavioural data and curating and analysing such multimodal data has been a huge task. The result is a completely unique resource to understand human behaviour in ecologically valid environments during unconstrained tasks with uncontrolled dynamics such as human interaction. Previous experiments have tracked the eyes and a scene video from the subjects perspective, using the video to label actions of the hands and head based on a frame-by-frame hand coding of behaviour (Hayhoe et al., 2003; Land et al., 1999). These experiments pioneered the movement from the lab to "the wild" and made huge contribution to the understanding of gaze behaviour in natural behaviour. However the approach is not scalable due to the frame-by-frame hand coding required and thus experiments last just a few minutes. Due to the large number of degrees of freedom in natural behaviour, significantly longer durations of data records are required due to the "curse of dimensionality". Making meaningful deductions from such high dimensional data requires very large quantities of data. Thus objective and scalable (automated) methodologies to capture and quantify unconstrained behaviour are required to make the problem tractable. Work from the fields of computer vision and EgoVision have made great strides to automate this process e.g. (Li et al., 2013). In addition, more data driven approaches have been seen in real-world eye-movement research (Foulsham et al., 2011; Hart et al., 2009), which have shown important differences between the lab and the real-world, however their focus has not been on more purely visual associations of eye-gaze "in the wild". In addition, even though aspects of these studies are more driven by raw rate, significant post processing and hand coding is required to label visual features such as the horizon and objects in the scene. Einhäuser et al. (2007) used more scalable data driven approaches, recording eye and head movements during free exploration "in the wild"(Einhäuser et al., 2007). The current study goes well beyond this, capturing, rather than constraining sensory inputs and motor outputs in natural behaviour. The majority of visual and auditory sensory inputs are recorded using head mounted eye-tracking, scene camera and microphone. Simultaneously, recording the majority of skeletal motor outputs by motion tracking 51 degrees of freedom in the body and a total of 40 degrees of freedom in the hands. Experimental durations are hours rather than minutes and involve a very broad range of behaviours from the tasks of daily life, including making and eating breakfast, daily chores such as tidying the bedroom, making the bed and leisure activities such as playing cards. In addition the task of navigation around a building was performed, however only eye-movements were recorded in this task.

The only other experiment "in the wild" that approaches this level of data comprehensiveness is that of Matthis and Hayhoe (2015) developed to investigate the interaction between gait and gaze on different terrains (Matthis and Hayhoe, 2015). They similarly track the body kinematics with an IMU suit and eye-movements with a mobile eye-tracker. They do not track hand and finger movements, 3D gaze position, involve social interaction or other behavioural situations that involve manual object interaction. However, in a sense their dataset provides a missing component of this study where the

body kinematics are not tracked during navigation due to range limitations of the wireless tracking system. Their results are as yet unpublished (the experimental framework was presented at VSS 2015) so cannot be commented on at this stage, but it would certainly be useful to compare notes in the future. There are many problems and challenges to overcome when dealing with this type of data and thus future collaborations and discussions are vital.

There are a number of important evaluation points to be made about the methodology before the results are discussed. Compared to the previously discussed studies in the wild, the experimental durations were significantly longer, lasting up to 1 hour compared to at most 10 minutes. This also brings some methodological challenges. Firstly the suit used to track the body is marker-less, which has the benefit that it can be used outside the lab without a complex camera set-up as is required by optical tracking methods. The suit uses inertial measurement units (9 axis - accelerometer, gyroscope and magnetometer), which means no issues arise from occlusion (as with optical tracking), however there is significant drift in the marker position estimates over long recording durations (over ten minutes) and thus only velocity measurements can be relied on. This means for example that the 3D hand positions are not accurate over long durations. The velocity measurements provide reliable dynamics of the body which is sufficient for the analysis in the following Chapters, however reliable position information would be useful to answer questions of "gaze-hand span". In addition the magnetometers are strongly affected by external magnetic fields and ferrous materials in the surroundings. For the eye-tracker, over such long recording durations, slippage of the headset is likely to affect calibration. This was ameliorated with calibration updates made at 5 minute intervals, at which point subjects stopped what they were doing and looked at a mark on their wrist. While it is not ideal to interrupt subject behaviour in this way, it is vital to maintain calibration quality. In the future more elegant solutions can be devised.

Aspects towards such improvements were included in the protocol but not yet analysed post hoc. For example, in the current study, fiducial markers were placed on the backs of the subject's hands and key surfaces of interaction, such as the breakfast table. These can be used to give positional information with respect to the head-mounted scene camera, however this information has not been utilised at this stage. Such an approach would be useful to provide positional updates in hand position when they are in the head-mounted scene camera field of view. This information could then be fused with the suit data to minimising the effects of sensor drift. Another caveat is the possible effects of wearing the tracking equipment on the subject's behaviour. This is particularly important for the hands as the subjects wore data gloves to record the hand movements. While the gloves are not bulky, they inevitably have an impact on hand mobility and tactile sensory feedback is reduced. Future control studies should compare eye-movement behaviour for the same situations without the body suit and gloves to ensure consistency. One final draw-back of the set-up is that a slightly different model of cyber glove is used for the left and right hands. This makes direct comparison between hands susceptible to systematic effects due to differences between the gloves.

In this Chapter the basic statistics of eye-movement behaviour are assessed and compared across different detection methods, different behavioural settings and with previous hand annotated "in the wild" studies. The key challenge is how to automate the analysis of such

large data-sets which would have been hand coded traditionally. This has 3 main components: the coding of body behaviour, eye behaviour and environmental context. Tracking motion kinematics already makes a significant step towards this as the body movements no longer need to be hand coded from a video of the subject - a hugely laborious and subjective approach. In addition, to achieve the higher level symbolic representation of behaviour an online annotation system and GUI called DALI - Dynamic Activity Logging Interface is developed. The hierarchical annotation system leveraged the known structure of possible tasks to intelligently display most likely behaviours given the setting and sub task to the annotator. This gives a detailed storyboard of behaviour which can be refined using the high resolution behavioural data to give more precise onset and offset. This was achieved for the grasp annotations as will be discussed. The remaining tasks of eye-movement event classification and interpretation of environmental context will be discussed in detail in the following.

The automation of gaze behaviour classification, moving away from hand coding, has been discussed at length in the background section. Though very well established in the lab, it remains an unsolved problem "in the wild" due to the data recording limitations of mobile eye-trackers and the less bounded eye-movement behaviours exhibited in natural unconstrained behaviour. To investigate this the results of two existing algorithms (dispersion based I-DT and adaptive velocity based SMI) are compared with each other and then with results from hand-coded studies "in the wild". In general it seems the SMI algorithm does a good job of distinguishing saccades and visual intake (both fixations and smooth eye-movements). The fixations (visual intake) have more instances that move in a consistent direction with a higher amplitude -suggesting smooth movement. While the saccades detected by SMI's method are less likely to have intermediate speeds in the range of smooth eye movements (50-100deg/sec). This suggests smooth eye-movements are predominantly present in the visual intake category. For the I-DT method, fixations are more incoherent in direction, while the saccades can have more mid-range velocities (50-100deg/sec). This seems to indicate that the I-DT algorithm tends to shift smooth movements into the saccadic category. This highlights the importance of understanding what the emerging classes actually mean and how this relates to perceptual implications. In the case of SMI's algorithm it is verified here that the visual intake category does indeed capture both static and smooth eye-movements that both imply visual intake. While the I-DT method had smooth eye-movements in the saccadic category thus visual intake was occurring both during saccades and fixations. If head-movement information is included the smooth movements associated with VOR become fixations with the I-DT, which highlights the importance of including additional information.

6.4.1 Cross setting event detection differences

It is still useful to utilise both detection methods due to the different characteristics they capture. For example, knowing the number of pure fixations (I-DT) allows us to more clearly distinguish Navigation from the other settings as there are significantly less pure fixations than in the bedroom, or kitchen setting. This is because there is more smooth pursuit in navigation and thus significantly more saccades and less fixations for the I-DT method (Figure 60). However, this can be seen in the navigation characteristics of SMI fixations which have faster fixations, with less fixations with a peak velocity less than 20 degrees per second (Figure 65). This comparison is important, particularly for low frame

rate eye-trackers where the summary statistics can be misleading if the true nature of the detection is not understood. This clearly becomes more important in natural behaviour where significantly more smooth eye-movements occur. There are less observed differences between the bedroom and kitchen settings as overall the manual nature of the tasks are similar. However, future work can use the annotations to break the setting down into task categories and comparisons made between shorter time segments of behaviour.

6.4.2 Comparing automated coding to hand coding in previous studies

Many findings based on eye-movement behaviour "in the wild" are based on hand coding of eye-movements using the eye image and gaze with respect to the scene video. This is still seen as the gold standard, however it is impractical for large studies of this nature (in the past experiments "in the wild" lasted ~4mins) and is subjective. Thus the results are compared with previous hand-coded studies for which the saccade amplitude and fixation duration were reported. The saccade amplitude of SMI saccades was used as they show characteristics of pure saccades, without smooth eye-movements, as discussed previously. The mean saccade amplitude for Land's tea making study was between 18 and 20 degrees while the mean here ranged between 8 and 12 degrees. It is also important to draw attention to the total number of saccades in the distributions though, which range from N=401-545 for Land's study with recordings lasting around 4 minutes compared to N=5584-7827 saccades here with durations of 37-60 minutes. This also indicates a much higher saccade rate on average for this experiment compared to Land's. Thus it seems in this more extended task, which includes tea making but is by no means restricted to it - including periods of conversation, preparing food, eating, laying the table etc, that the average saccade amplitude is on average lower and the rate higher. The amplitude of saccades can give an idea of targeting accuracy because they represent the metric of gaze adjustment. Thus the smaller the saccade the finer the gaze adjustments being made. Therefore, the observation of many smaller saccades in our task suggests there is significant fine targeting of information, thus providing a higher information signal for inferring intentions. Apart from the larger array of behaviours exhibited in our study, it is hard to tell precisely why these differences occur. It may be that very small saccades within an object are considered as the same fixation by hand coders, thus reducing the number of smaller saccades. Without detailed insight into the hand coder's rules and their level of consistency it is hard to know. This is why more objective metrics are required. Further analysis should explore the variability of this and other metrics across time in relation to the ongoing behaviours (captured by the online annotation).

Another metric considered was the fixation duration, often used to distinguish cognitive tasks (Droll et al., 2005; Hayhoe et al., 1998, 2003; Land et al., 1999). Both SMI visual intake and I-DT pure fixations were compared with previous studies. The I-DT method gives much more similar fixation durations to the hand coded studies showing much larger fixation durations during the kitchen and bedroom settings when compared to sedentary tasks such as reading. For the SMI method, the fixation duration was much lower with a similar magnitude than those usually observed in reading tasks and very little difference between kitchen bedroom and navigation setting. Thus the hand coded fixations are much more similar to the I-DT coded fixations. This is perhaps because the hand coders were used to coding static eye-movement experiments and like the I-DT algorithm (also designed for static experiments) puts smooth eye-movements in the

saccade category. But again it is hard to speculate without knowing the precise rules the hand coders used, notwithstanding the subjectivity of such an approach.

6.4.3 Consistency of behaviour across task and setting

The comparison has previously been made between sedentary tasks such as reading and picture viewing and active tasks such as tea making. Here the additional category of navigation is added. Firstly, the pure fixation (I-DT) duration during navigation is significantly shorter, lasting on average 269ms which is more similar to those made during reading where fixations last on average 224ms compared to 328 in the kitchen task and 339 in the bedroom scenario. In addition there is significantly more smooth eye-movements during navigation based on looking at the visual intake characteristics (SMI) and the saccade characteristics of I-DT as discussed. Other gaze characteristics such as gaze direction distribution have been assessed in different environments (Einhäuser et al., 2007). They found in general that eye-movements were made along the cardinal planes and claim that there was an anisotropy across settings. These results are not observed in our data. There is only a very slightly higher probability of gaze direction along the cardinal axes, compared to the "*remarkable*" anisotropy observed by Einhäuser and colleagues. The difference in results may be attributed to methodological differences. Firstly, the eye-movement direction was extracted from a gaze actuated scene camera (EyeSeeCam) using the 2D cross correlation between adjacent frames. This method of eye-tracking is likely to be affected by the environment as different scene textures and depths will affect the metric. Secondly, different environments were not always viewed by the same subject and thus the anisotropy between apartment and forest was only observed for a single subject. Finally, while subject's behaviour was unconstrained in their study there was no specific task, but only to freely explore the environment for a few minutes, which may also contribute to the difference in observed results.

A spatial descriptor that did show some interesting differences in our study was gaze depth. The fixation depth statistics clearly distinguish gaze behaviour during navigation, with significantly more fixations reaching depths beyond 1m as the subject looks into the distance to gather navigational information rather than within the arm's reach to guide object interaction and manipulation (Figure 60). However owing to the limitations of the suit positional data drift, the 3D end-point relationship between hand and gaze is difficult to study. In addition the 3D gaze measurements were quite noisy and while the aggregate statistics show meaningful differences across settings, the 3D gaze position did not always correspond perfectly to the 3D position of fixation. This observation is based on looking at the subjects gaze in the video and the corresponding 3D estimate. This could either be the natural eye-movement behaviour or the 3D gaze estimation inaccuracies. Since this data was recorded the methods for estimating 3D gaze estimation have been improved by 60% (see Chapter 4) and thus further studies should utilise this methodology. This initial exploration into 3D gaze statistics opens an entirely new avenue of research. Further analysis should investigate the vergence accuracy of gaze behaviour in different situations, perhaps when precise 3D information is required for tasks such as precision grasps the vergence will be more accurate. One final metric that is significantly different between navigation and the kitchen and bedroom situations is the proportion of time spent blinking, with significantly more time spent blinking during navigation. In fact, blinking turns out to be quite an interesting behaviour during natural behaviour in general.

6.4.4 What we miss in the blink of an eye

Blinking is generally seen as an interruption of the gaze tracking signal to be ignored, however blinking occurs 20-30 times per minute and it makes up almost 30% the experimental duration (see Figure 61). Blinks are generally attributed to the function of moistening and protecting the cornea. Thus when large gaze shifts during the blink period were noticed in our data it was intriguing, if a little concerning. In addition, large head and trunk movements were associated with these blink periods, a very surprising result at first. Initially the large head movements were assumed to effect the eye-tracking (due to rapid light level changes), causing blink false positives. However, though not widely discussed, large head-free redirections of gaze are often accompanied with a blink, with the blink probability increasing with the size of the gaze shift (Evinger et al., 1994). It is known that these gaze-evoked blinks are not produced as a reflex to the wind resistance as the eye-makes a rapid movement because muscular innervations of the orbicularis oculi muscle (upper eyelid) are observed even when the eyes are closed (Evinger et al., 1994). There is relatively limited knowledge of gaze evoked blinks in the lab, let alone "in the wild". Generally, as mentioned, blinks are both a protective mechanism and required to moisten the cornea, but they may play another role. One possibility is that the timing of these lubricating blinks is optimised to minimise visual disruption and are thus synchronised to saccades, where the visual input is suppressed anyway. Alternatively, evidence suggests they play an active role in saccadic suppression itself (Williamson, 2004). The latter postulates that blinks take over from saccadic suppression for higher amplitude saccades. Gandhi recently made significant progress to characterise the interaction between gaze direction and blinks for both head-restrained and unrestrained gaze shifts in non-human primates (Gandhi, 2011). It was shown that the average velocity of saccades during gaze evoked blinks is lower and the durations longer when compared with amplitude matched controls.

Here, for the first time, this behaviour has been characterised "in the wild", and observation of some similar and additional characteristics have been made (Figure 66). Firstly there is a higher proportion of blinks associated with high amplitude eye-movements (59% blinks >10 degrees) than there is for saccades (15%>10 degrees) (Figure 66C). Which agrees with the assertion that gaze evoked blinks are increasingly likely for higher amplitude gaze shifts. To understand this relationship further, future analysis should be pool blink and saccade amplitudes and the probability distribution of a gaze evoked blink over the movement amplitude should be observed. On a side note, this may give another reason for the saccades observed in this study having a lower mean amplitude than in Land's (1999) study (Land et al., 1999). Perhaps the annotator inferred these saccades had occurred from the gaze position change, without also recording that that a blink had occurred simultaneously. To continue discussing the eye-movement behaviour during blinks, the inferred angular speed of gaze during blinks is on average lower than pure saccades and the blink duration is longer. This agrees with lab studies, providing strong evidence that what is observed here are the first reported gaze evoked blinks "in the wild". This addition to the taxonomy of behaviour "in the wild" may have important behavioural associations with context and task which will be discussed.

The additional characteristic of a directional bias in gaze velocity during blinks emerges here (see Figure 67), which has not been previously reported. Firstly horizontal saccades

are more likely to evoke blinks than vertical ones (higher amplitude saccades in horizontal direction), and secondly there is a bias for blinks to be evoked when saccades are moving to the left. This effect is also observed for head movements and is seen in all settings, though the effect is less pronounced during navigation. It is unclear why this bias exists, but it is important to remind the reader that all subjects were right side dominant and thus it may be that looking to the non dominant direction may have different informational constraints. This would also suggest why this effect was less prominent during navigation where very few hand-object interactions occur. Evidence of the opposite effect must be observed in left side dominant subjects, but the result as it stands is indeed thought provoking. In general, this ordinarily discarded information associated with blinks has strong implications for future analysis and changes the understanding of blinks in natural behaviour.

The blink rate and characteristics have long been known to be influenced by cognitive processes such as word recognition difficulty (Ohira, 1996), discriminative tasks both auditory and visual (Goldstein, Walrath, Stern, & Strock, 1985), fatigue (Stern, Boyer, & Schroeder, 1994) and task transition (Stern et al., 1984). With the massively expanding field of wearables and ubiquitous computing, information like the blink frequency are being used more and more for fatigue detection (Svensson, 2004), levels of engagement (Ismail et al., 2013), cognitive load (Chen and Epps, 2013) and activity recognition (Ishimaru et al., 2014). Blinks can be detected with very cheap sensor technology such as proximity detectors or simple photodiodes and thus it is an ideal metric for commercial applications. However, given the association between head movements and blinks observed here, there is huge potential to combine blink detection with a cheap IMU sensor to measure head movement and relate this directly to blink frequency. Interestingly, Ishimaru et al. found that combining head movement and blink frequency did greatly improve activity recognition, though they used separate summary statistics that did not utilise the covariance of these behaviours that has been observed here.

Furthering the understanding of blinks and their relationship to eye-movements and information retrieval may provide more descriptive features for inferring ongoing cognitive processes, informational priorities and thus along with other descriptors for intention decoding. For example, asking the question other than amplitude of eye and head movement, what effects the brains decision to make a slower and longer saccade that is accompanied by a blink compared to a faster shorter duration saccade that is not accompanied by a blink? It has often been observed that blinks will be suppressed prior to a stimulus onset or information retrieval. Perhaps therefore saccades that proceed vital information retrieval are less likely to be accompanied by a blink while the following saccade, once information has been retrieved may combine a blink because the information constraints are less demanding. However, the alternative could also be true, i.e. when in a very active state where high frequency information gathering is required to guide a complex set of motor behaviours, the required moistening blinks must be initiated at optimal moments. In this case there could be an increase in gaze-evoked blinks to ensure blinks only occur at the optimal time to minimise impact on the ongoing information retrieval. So far there is some evidence to support the former argument, such as Williamson et al. (2004) who showed that "refixational" saccades at the end of the lab trial were more likely to have blinks associated than the behaviourally relevant saccades during the task (Williamson, 2004). Finding more evidence for this and understanding the

relationship further could provide a vital descriptor for distinguishing fixations that are behaviourally relevant - such as the location of an object to be grasped versus more spontaneous refixations to behaviourally irrelevant locations. Thus given the prevalence of this behaviour observed here "in the wild" and the potential it holds, this should be analysed in more detail in the data collected here and in other studies.

The methodology and eye-movement statistics have been presented and compared to previous studies. This shows that reasonable automatic coding of natural gaze behaviour is possible while showing the importance of understanding the event detection method used and the characteristics of the classes it segments behaviour into. It reiterates the notion that studies in the wild cannot purely consider eye-movement in terms of fixations and saccades. Future methods should integrate body information into the event detection process to give richer descriptors of eye-movement behaviour. Towards this, the following chapter moves analysis beyond eye and head movements utilising the full motor repertoire, leveraging the truly unique embodied aspect of the gaze data set.

As such technological and analysis methods improve, it is also important to consider how such real-world scenarios can be manipulated experimentally to enable hybrid studies that maintain free behaviour while changing conditions in a repeatable manner. An interesting example of this is presented by 't Hart et al. during a walking task where the subjects walked to adjacent paths that had very different terrains but almost identical visual surroundings ('t Hart and Einhäuser, 2012). Thus the experiment is repeatable and well defined from an experimental perspective while allowing the subject to behave as naturally as possible. Another important methodology that should be mentioned as an important tool for furthering our understanding of natural behaviour is virtual reality. This enables controlled lab experiments in which subjects can move freely and the visual stimulus is updated based on their movements. This goes a long way towards making stimuli and interaction more natural compared to a head fixed screen based set-up, while maintaining the investigator's control over the visual stimulus and thus repeatability. In addition, such an approach needs little or no human coding of objects of interest because the virtual world is predefined and known and can thus be inferred directly from the data. Some of the powerful results presented in Ch2 have come from such VR studies e.g. (Rothkopf et al., 2007) and this approach is being adopted more and more widely (Diaz et al., 2013; Kit et al., 2014). It is important to note that although highly immersive and a useful experimental tool, VR set-ups cannot be used as a direct replacement for experiments in the wild. In addition, more systematic studies need to compare the visuomotor behaviour between the lab, VR and natural settings. It cannot be stressed too strongly that these experimental approaches should not be mutually exclusive but must be used in parallel to maximise research effectiveness. This chapter and the following chapter focus on methods to enhance eye-movement studies in the wild, focusing on analysing the embodied gaze data collected in the wild. In the future, this can be compared and enhanced with similar experiments in controlled lab settings as well as virtual or augmented reality in a united front towards the understanding of visuomotor behaviour.

7 Embodied Saliency and gaze descriptors

Abstract. A novel method is presented for analysing and interpreting eye-movements in the wild, by relating them directly to real-time recordings of complete body posture, including hand movements. Experimental context has a huge impact on shaping emerging theories. Static lab studies observing eye-movements in response to purely visual stimuli led to a wealth of literature on bottom-up visual saliency. Now, with the capabilities to capture the full perception action cycle in natural behaviour, a new field of Embodied Saliency can emerge. This works towards a more complete model based not only on bottom up environmental cues and top down high level tasks, but the real-time movement of the body itself. Thus, based on body joint movements, it is shown that eye-movements can be predicted significantly above central fixation bias (that tends to out-perform visual saliency models). In addition the taxonomy of gaze and body behaviour is explored. In the wild, eye-movements are accompanied by a huge repertoire of body movements. In this context the taxonomy of fixation, saccade and blink becomes an impoverished description . Thus to understand the visuomotor repertoire eye-movements are taxonomised by their association with body movements. Firstly, gaze-grasp associations are explored with a new unsupervised data driven approach to extract embodied gaze descriptors. This shows that 42% of saccades are temporally aligned with grasp-like behaviours. The descriptors that emerge show that seconds before the grasp is made there are predictive cues in the eye, head, trunk and arm which could be used to initiate a grasp intention for amputees.

7.1 Introduction

The previous Chapter introduced the novel dataset collected that captures the full perception action cycle. The gaze behaviour alone was analysed and compared to other studies. In this Chapter the relationship between gaze and body movements is addressed. The oculomotor system has been broken down into many sub-systems that have been characterised in the lab, as have been described in Chapter 2. In addition, there are many models to control gaze allocation. However little has been done to unify these systems. In the wild, there is less clear division between all these subsystems where the stimuli affecting eye-movements are multivariate with visual, vestibular, auditory, tactile and proprioceptive stimuli occurring simultaneously with complex task considerations. Thus with the extensive dataset that captures rather than constrains the perception action cycle, eye-movements can begin to be attributed to different combinations of these different subsystems.

Bottom up visual saliency and top down task demands have shown to have predictive power in various contexts (for extensive reviews see (Schütz, Braun, & Gegenfurtner, 2011; Benjamin W. Tatler, Hayhoe, Land, & Ballard, 2011)). However, if the brain is thought of as a dynamical system mapping sensory inputs to motor actions surely the body posture itself holds predictive power. The focus tends to be on visual sensory information but eye movements are also affected by body proprioception (joint positions) and indeed the planned future postures i.e. intentions. Thus the concept of Embodied Saliency is hypothesised: can eye-movements be predicted from the dynamics of the body?

The second line of investigation looks at the embodiment of gaze descriptors. This reflects the idea that gaze events such as saccades, fixations and blinks should be associated with different body-movement behaviours. The taxonomy of fixation, saccade and blink becomes an impoverished description of gaze behaviour when the body is free to move. Thus to understand the visuomotor repertoire eye-movements are taxonomised by their association with body movements. The scope of such an approach is expansive with such a comprehensive dataset. Firstly, grasp movements are explored because of the obvious links with prosthetic control and also in particular these relationships have never been recorded in the wild. As the dataset is so vast and unique, this chapter presents the journey of exploration so far and the discussion realigns these findings with the existing literature.

7.2 Methods

The following gives an overview of some of the more involved analytics used to probe the relationship between eye and body movements. The gaze events used are based on the SMI classification.

7.2.1 Linear Regressive Models

To probe the idea of Embodied Saliency, the first approach is to map from the body joint angles (51 degrees of freedom) to the eye gaze (2 DOF). Three regressive models are used: gaze autoregression, gaze autoregression with body joint angles as exogenous inputs and pure linear regression from body joints to gaze. Figure 70 gives an overview of these models. The number of delays used in the regression is optimised by a parameter sweep.

Figure 70 shows the different models built. The regression is made to training data using the Moore-Penrose pseudo inverse (Moore, 1920). For the autoregressive models, the training is performed in open loop (Figure 72B) with the true current state of the eye position and the body joints being mapped to the next eye position. This is then tested both in open loop, again with known current eye position and in closed loop multistep ahead performance (Figure 72C,D), where the predicted current state of eye-position is used rather than the actual current state.

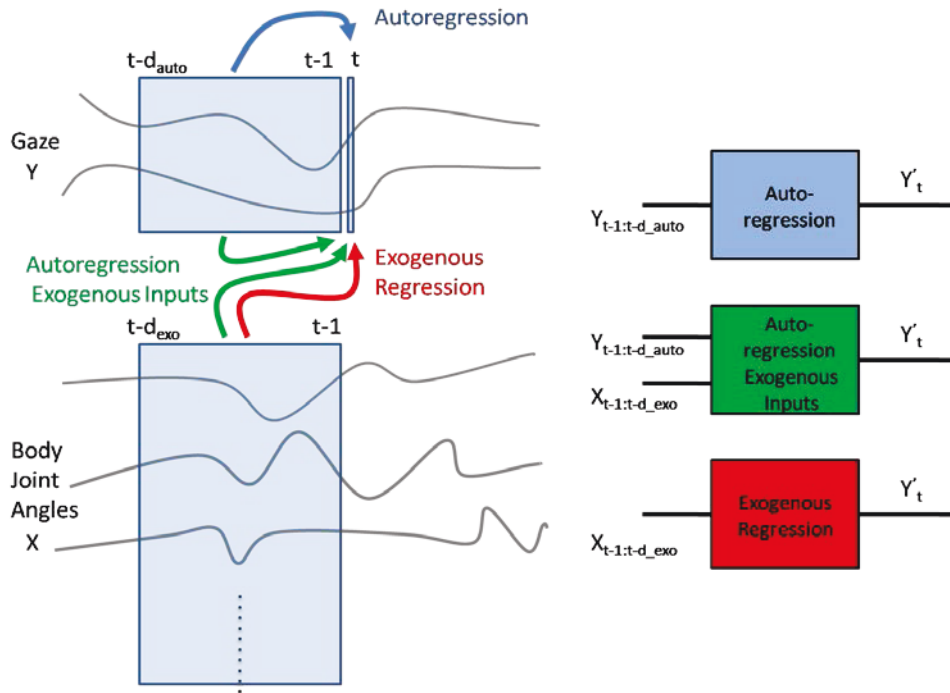


Figure 70: Linear regression models. Y is gaze position (pitch, yaw) and X is the suit joint angles. Three linear models are used: Autoregression (blue), predicting future gaze based on past gaze. Autoregression with Exogenous Inputs (green), predicting future gaze based on past gaze and body joint angles. Exogenous regression (red), predicting future gaze based on past body joint angles. For each model the parameter that must be set is the number of delays (d_{auto} and d_{exo} depending on the model) used to predict gaze. Note: the ":" notation means all integers between, for example 1:5 means 1,2,3,4,5. $Y_{t-1:t-d}$ refers to frame number so $Y_{t-1:t-d}$ is the eye positions for the previous d (delay) frames.

Training and test protocol

When considering how to split between training and test data it is important to consider the auto-correlation function to ensure the training and test data are sufficiently temporally separated. Looking at the auto-correlation function, the full width at half maximum (FWHM) indicates the span of temporal auto-correlation. For the eyes with a vertical FWHM of 780ms and in the horizontal dimension 600ms. This suggests that the data is significantly correlated over intervals of around 1 second. After 5 seconds the correlation function becomes close to zero and thus it is safe to use blocks of training and test data that are at least 5 seconds apart. To be even more conservative, training and test data is separated by 30 seconds, this ensures that there are no direct temporal correlations between test and training data. The data during blinks (detected by SMI BeGaze) was removed from the training data. Cross validation is performed by selecting a period of test data flanked by a minimum of 30s gap and the remaining data is used to train the model (see Figure 72A). For each round of cross validation the training block

(including the flanking gaps) is moved to the next contiguous position at the end of the current block. This is illustrated in Figure 72A. This regime allows the performance to be assessed at different positions of the time course showing where generalisation of the model is successful but also provides a measure of “similarity” or “predictability” of the data chunk.

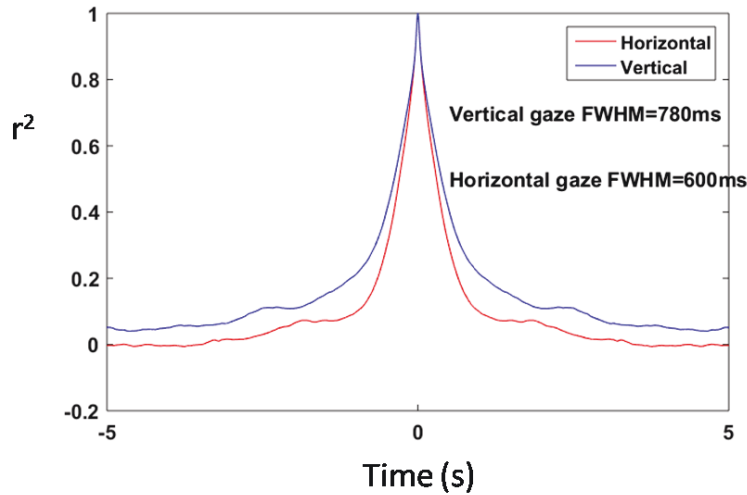


Figure 71. Gaze Auto-correlation for vertical gaze (blue) and horizontal gaze (red), with the full width at half maximum (FWHM) labelled for each dimension.

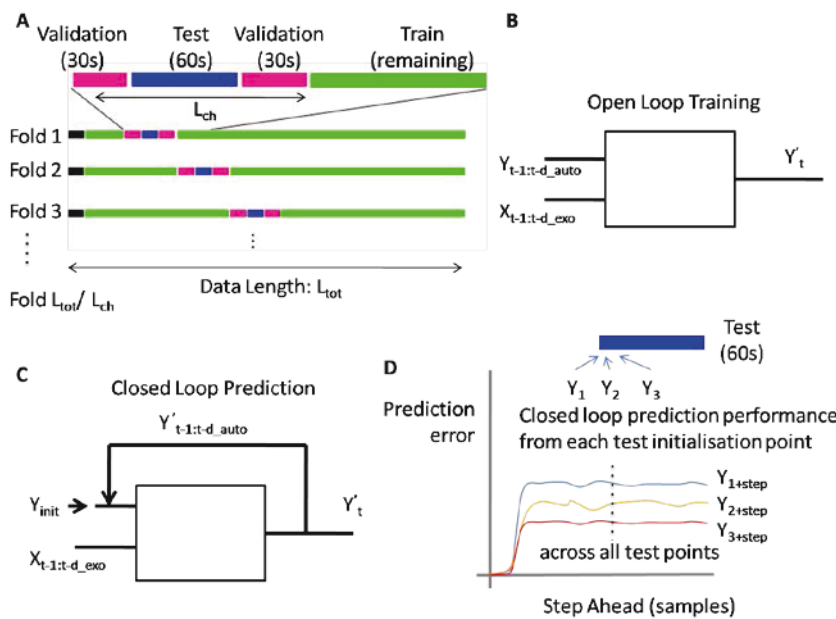


Figure 72: Training regime. A. Cross validation is performed by splitting data into training, validation and test data. Test data is flanked by validation data, providing separation between the test and training data. for each run on the same data set, the test chunk position is stepped through the data until all data has been used to train and test/validate. B. For Autoregressive models, training is performed in open loop meaning that the real past gaze data is available. Open loop performance is then the prediction of the next state given the true current state. C. In closed loop mode the initial gaze position is known (Y_{init}) for the first step prediction, after which the prediction is fed back to be used in subsequent step predictions. D. The prediction error is assessed by looking at the closed loop step ahead performance starting from each test data point location and predicting n steps ahead.

7.2.2 Performance metrics and chance level

The auto-regressive models are run both in open loop (true previous gaze position known) and closed loop (predicted gaze feed-back), while linear regression only has an input out-put mode (see Figure 70 and Figure 72). The open loop performance is calculated for each test data point while for the closed loop performance, the multi-step ahead performance is calculated. This takes each test data point as the current real gaze position (as well as required lags before) and then predicts each subsequent gaze position based on the model operating in closed loop. Thus for subsequent steps the predicted gaze feedback (rather than true gaze) is used to predict the next steps. The multi-step ahead performance is then averaged across each starting position (see Figure 72D). The model performance is assessed based on two metrics, the average (mean) absolute error (AAE) and the squared Pearson correlation coefficient (R^2). The latter assessing the actual error in the prediction and the former capturing the linear correlation between the prediction and the actual gaze data and thus prevents non-stationary effects of drift in the sensors shrouding the performance. This metric also quantifies the similarity of the predicted and true dynamics unlike conventional saliency performance metrics. To provide a baseline comparison or “chance level” two methods are used 1) Random Sampling: the gaze estimate is made by sampling from a uniform distribution across the gaze range. 2) Central Fixation Bias: a normal distribution is fitted to the training data gaze distribution and then sampled from to predict the gaze location. This is based on the observation that the gaze fixation distribution tends to be tightly centred in screen based tasks and that this outperforms many visual saliency based models (Tatler, 2007).

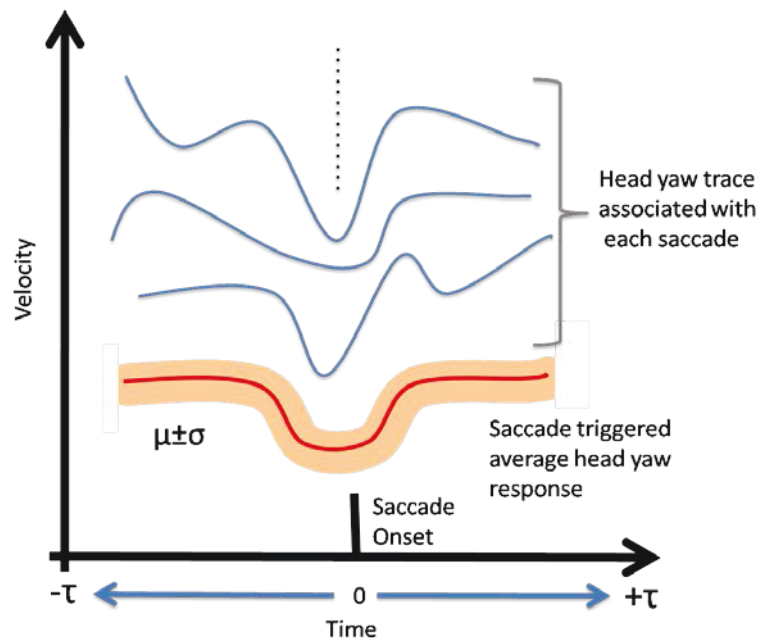


Figure 73: Event triggered average body movement. For all instances of a given gaze event (e.g. saccade), a window spanning τ seconds before and after the event onset is extracted for a given joint angle in the body (e.g. head yaw). The example sketch here shows the head yaw traces associated with saccade onsets. The blue traces represent each saccade associated window of head yaw position. Thus for saccades the average head yaw response can be calculated (red trace). This is repeated separately for all joints in the body.

7.2.3 Event triggered average

The association between gaze events, such as a saccade and body movements is studied locally. To do this, for each instance of a given event, a window of data is cropped and aligned with all other windows associated with the same event type (see Figure 73). This gives a time aligned event response for each joint to a given event. This analysis does not relate the continuous eye-movement behaviour like the cross-correlation, but rather asks the question: given a particular event (fixation/saccade/blink), what is the stereotypical body behaviour associated with it? From this various statistics can be analysed such as the average head movement response to a saccade as depicted in Figure 81. In addition, clustering can be applied to these responses to reveal different taxonomies of behavioural responses.

7.3 Results

7.3.1 Gaze-Body interaction characteristics

Initially the simple correlation structure for the eye gaze and body velocities is analysed. Figure 74A shows the auto-correlation (mean subtracted Pearson correlation coefficient) for gaze pitch and yaw velocity. Important differences are observed in the correlations depending on how the missing data is handled during blinks (Figure 74A, left versus right column). Technically speaking, the two available options include interpolation of eye position across blinks (given the eye must have moved in this period if there is a difference) or to set the eye position to non-numerical values so that no auto-correlation is computed over these periods of missing data. As noted in the previous Chapter, the natural eye-movement literature mostly ignores blinks, despite the associated saccades demonstrated (visually evoked blinks, see previous Chapter). It is even unclear how previous work has handled this issue of blinks during analysis. The full width at half maximum (FWHM) is calculated to give an idea of the span of temporal correlation (interpolation of cross correlation to find width at half maximum). This gives 86ms for gaze yaw and 91ms for gaze pitch, which means the velocities are highly correlated within 50ms but by about 100ms there is no correlation. This corresponds to the timescale of saccade durations which tend to last 30-90ms. When the blink regions are included, which as discussed often shroud slower saccades, the temporal correlation is smeared out. This can be seen in the second column of Figure 74A and the FWHMs increase to 110 and 120 ms for yaw and pitch respectively. Given the longer duration of blinks (mode 250ms) over which the gaze is interpolated this is not surprising. However, it is important to at least be aware of these "invisible" eye-movements.

The cross correlation of the absolute gaze pitch and yaw is shown in Figure 74B with high cross-correlation of 0.5 at 0s lag, suggesting a large proportion of oblique eye-movements (unlike previous shorter studies in the wild (Einhäuser et al., 2007)). The cross correlation is also shown between eye-gaze and the head, spine and hips. These peak cross correlations for each joint angle are summarised in Figure 74C, while the full profile of the cross correlation function is shown in Figure 74D, with the lag at peak correlation summarised in Figure 74E. This is plotted separately for all gaze data (raw gaze data with interpolated blinks), saccades, fixations and blinks alone, with other events set to NaN (not a number) in each case. As it would be expected there is a significant correlation between

eye and head yaw across all eye gaze ($0.31 \{G_{yaw}, Head_{yaw}\}$ and $0.34 \{G_{pitch}, Head_{yaw}\}$). This occurs at a lag of -33ms for pitch and 0 for yaw, which, given the temporal resolution of the eye-tracker, suggests they occur almost simultaneously. A similar result for saccades alone is observed thus suggesting constructive eye and head saccades. For visual intake the correlation is negative for head and gaze yaw (-0.21), reflecting the gaze is in compensatory mode. This correlation occurs at a lag of -230ms suggesting the eye-movement follows the head movement and is thus reactive suggesting VOR. For pitch the result is less clear, there is a smaller negative correlation of -0.1 at 264ms followed by a peak in cross correlation of 0.17 at 0 lag. For blinks there is a large cross correlation between eye and head for corresponding pitch and yaw, reaching 0.56 at 66ms and 0.31 at -33ms for pitch. For blinks constructive interactions are observed between gaze and spine yaw, with correlation of 0.1 at a lag of 165ms, suggesting spine follows gaze and head. There are also other subtle negative correlations between gaze yaw, head roll and head pitch. The roll component of the eye cannot be measured with optical eye-tracking thus it cannot be commented on here. In addition, small correlations were observed between other joints in the body, particularly the shoulders, however further analysis is required to cross validate and to study these at a more local level to ensure significance of these results. The following section looks at what this correlation structure translates to in terms of the predictive power of body moments to gaze movements.

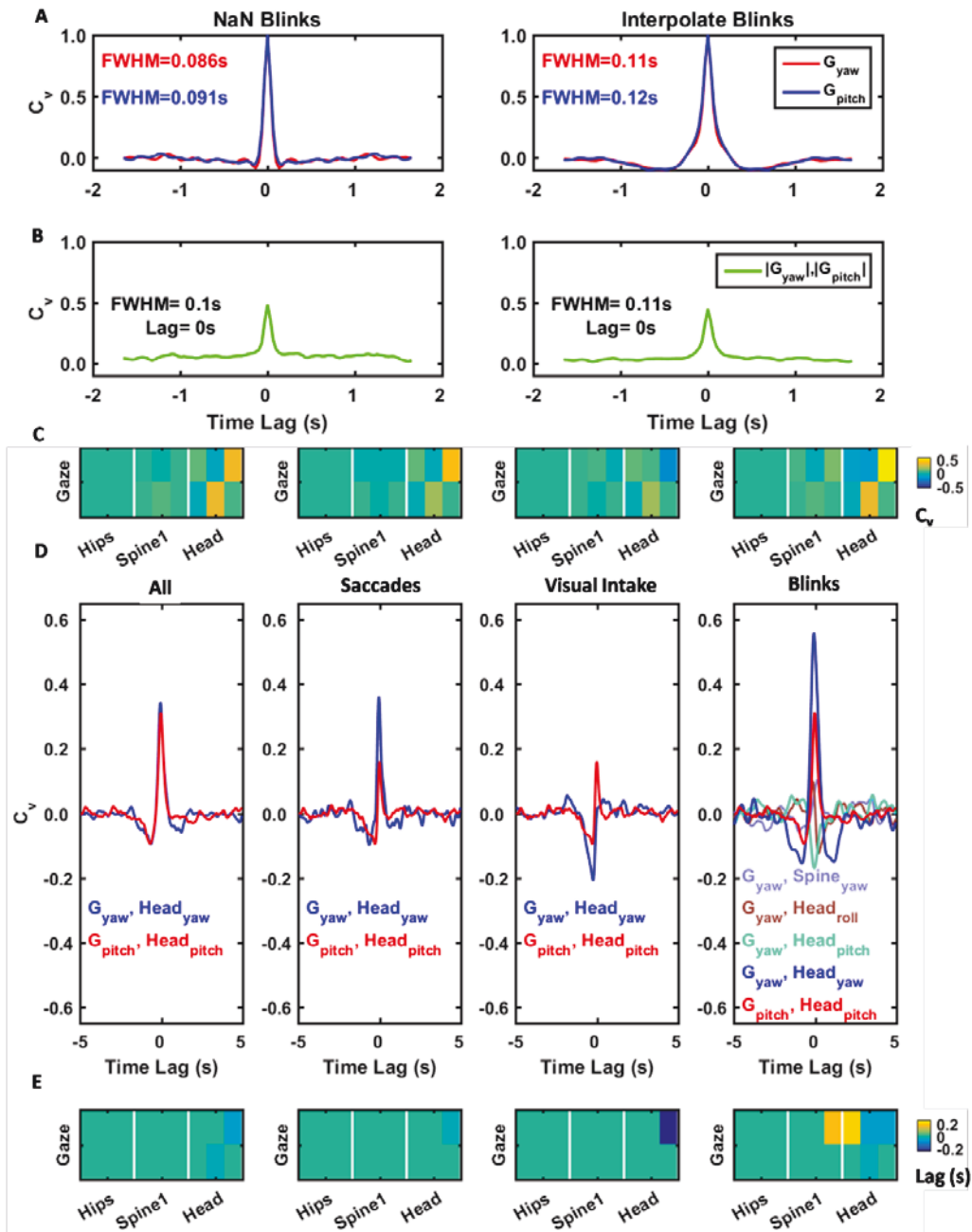


Figure 74: Eye Gaze Auto and Cross-Correlation (Pearson's correlation coefficient - C_v). A. Eye Gaze velocity auto-correlation for horizontal (red- G_x) and vertical (blue- G_y). The Full width at half maximum (FWHM) is reported. The left hand plot is for with blink periods set to NaN and thus not contributing to cross correlation and the right hand plot is where these periods have been interpolated (cubic interpolation) based on the gaze positions either side of the blink event. B. Cross-Correlation between absolute horizontal ($|G_x|$) and vertical ($|G_y|$) eye gaze velocity. C. Cross correlation between eye gaze (first row G_x , second row G_y) and the joint angle of the hips, spine and head, with each joint made up of the euler angles roll, pitch and yaw. D. Displaying the cross correlation of those with a peak correlation coefficient greater than 0.1. Where columns reflect "All" eye movements, those during Saccades, Visual intake (as defined by SMI event detection) and blinks. E. Time lag in seconds at the maximum correlation coefficient for each joint. Those with a correlation coefficient less than 0.1 are set to 0 lag. Each column of 3 plots (C,D and E) corresponds to the event grouped data: All- raw gaze data with interpolated blinks, saccades, fixations and blinks alone with other events set to NaN in each case. Note: the lag is with applied to the head data thus correlation at a positive lag suggests head follows gaze.

7.3.2 Autoregression

The performance of the gaze autoregression model is shown in Figure 75. This shows how well the eye-gaze can be predicted from previous eye-gaze. The open loop performance is separated into training, validation and test performances (Figure 75A). (Note: Open loop prediction is where the current true gaze position is used to predict the next gaze position, thus for each prediction the true previous gaze position is known (see Figure 72)) The number of input delays is optimised as shown in Figure 75C (5 input delays, 1,2,3,4 and 5 frame delays $\sim 165\text{ms}$). The performance is separated into horizontal (X –blue) and vertical (Y- red). As expected, given the auto-correlation function of gaze, the R^2 is very high, approaching 1. The route mean squared error RMSE in the x and y coordinates is 37 and 33 pixels (1.7 and 1.6 degrees) in the horizontal and vertical gaze directions respectively. In closed loop mode, the multi-step ahead performance (Figure 75D) drops off in concert with the auto-correlation, tending towards a zero R^2 after 2 seconds in the vertical direction, and slightly sooner at around 1 second for the horizontal component. The RMSE converges to around 220 pixels (10 degrees) in vertical and horizontal gaze position. The weights matrix that makes this optimum mapping from past gaze to future gaze is shown in Figure 75B. These weights effectively perform a linear extrapolation with the strongest differential banding over the previous two samples. Adding further lag samples does not improve the performance significantly as can be seen in Figure 75C.

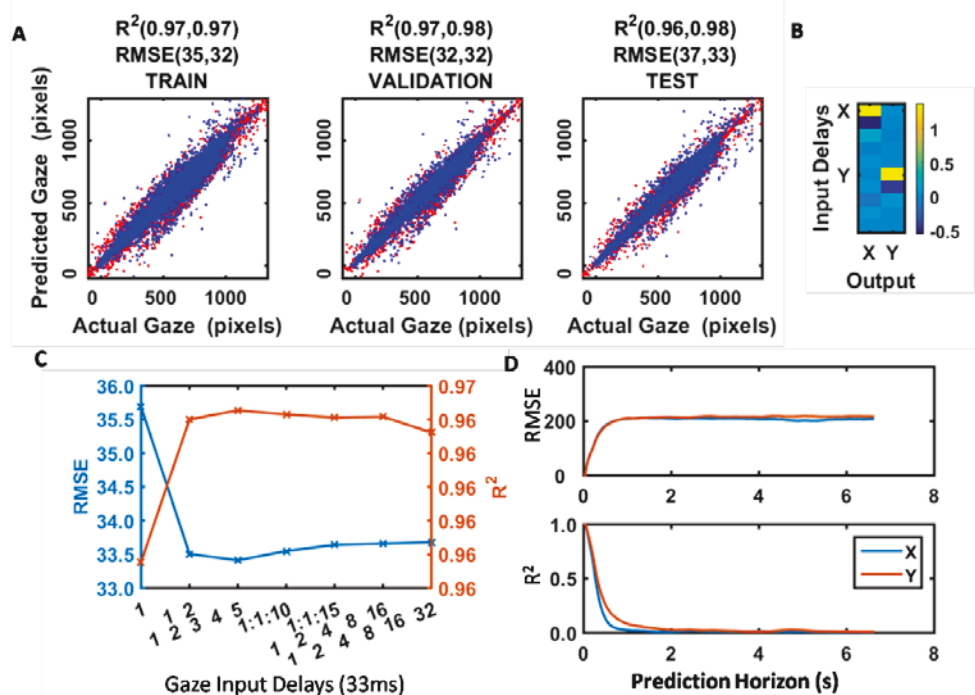


Figure 75: Autoregression Performance. A. Open loop performance showing predicted gaze for vertical y (red) and horizontal x (blue) against actual gaze. Performance separated into train, validation and test data separately. B. Optimal Weight Matrix used to predict gaze. The columns represent the weights used to predict each gaze dimension (x and y) while the rows represent the input gaze and the respective delays (5 delays). C. Delay parameter optimisation in open loop, comparing RMSE and R^2 metrics for different input delays. The x-axis labels indicate the time lags from “1” single frame, the “1 2” the last two frames, frames 1 to 5, frames 1 to 10, frames 1 to 15, as well as logarithmically spaced lags (1,2,4,8,16) and (1,2,4,8,16,32). Each frame is 33 ms apart from the next one, so a 10 lag frame is approx. 1/3 of a second in the past. D. Multi-step-ahead prediction performance (see Figure 72D).

7.3.3 Auto regression with exogenous inputs

To improve prediction significantly the body joint angles were added as exogenous inputs to the auto regression (see Figure 70). This asks the question, given the previous eye (predictions) and body (true) movements, what is the future gaze. The open loop performance is similarly high however, now in closed loop the predictive power remains significant into the prediction horizon (Figure 76C). The multi-step ahead performance no longer drops to a negligible R^2 , but plateaus at around 0.35 R^2 for the horizontal component and at around 0.22 for the vertical component. The RMSE stabilises at 170 pixels (8.0 degrees) in the horizontal and 210 pixels (10.1 degrees) in the vertical. This optimal linear prediction performance is achieved with 2 input delays in both autoregressive feedback and exogenous inputs (see Figure 76B) but does not degrade significantly as the number of delays increases suggesting there is sufficient training data to prevent over fitting.

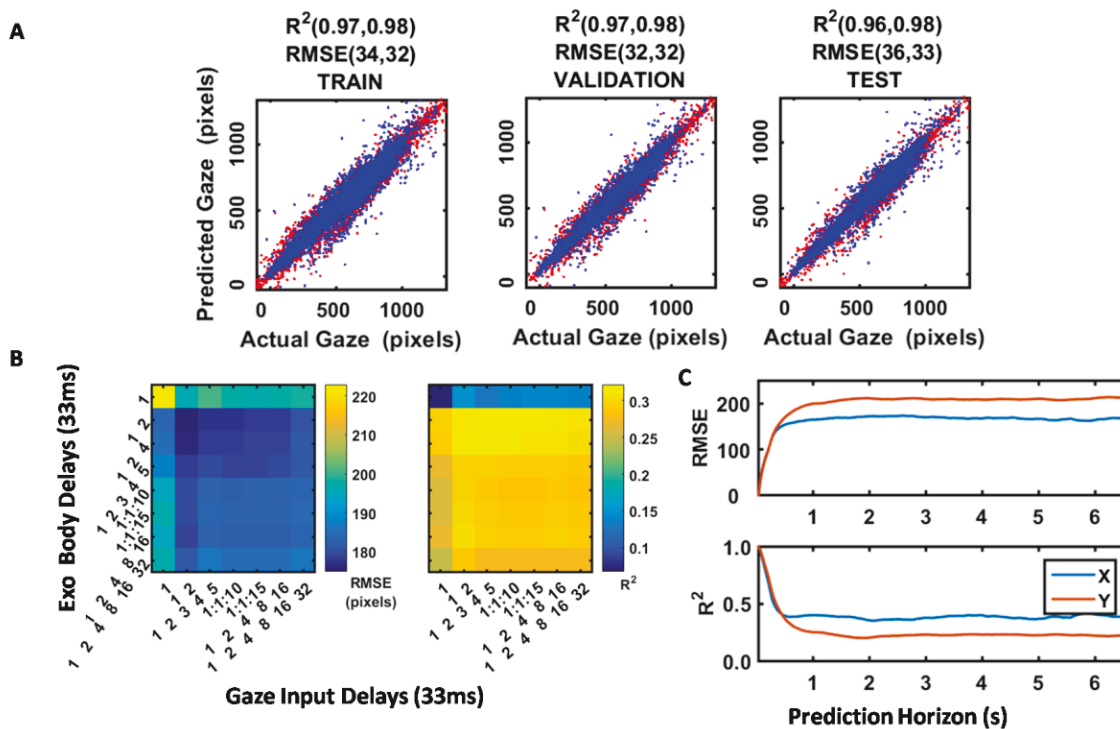


Figure 76: Auto regression with Exogenous inputs: A. Open loop performance showing predicted gaze for vertical y (red) and horizontal x (blue) against actual gaze. Performance separated into train, validation and test data separately. B. Delay parameter optimisation (both input delays and exogenous input delays) in closed loop, showing root mean squared error (RMSE) and the squared Pearson correlation coefficient (R^2) performance for different input delays. Lags as explained the previous figure. D. Multi-step-ahead prediction performance (see Figure 72D).

7.3.4 Exogenous regression

For purely exogenous regression (see Figure 70), i.e. without feedback in the system mapping directly from body to eye gaze, the performance is similar (Figure 77). However to achieve this performance, the optimum number of input delays over the body sensor data increases to 5 input delays spaced with base 2 exponentially increasing delay steps (see Figure 77B). The weights matrix of the exogenous inputs tells us how different joint angles contribute to the predictive power achieved. These weights are shown in Figure 78.

For autoregression with exogenous inputs (Figure 78A), the strongest weights apply to the head sensor information. For the prediction of gaze position in the horizontal, the head yaw has a strong differential banding, in addition slightly less, but significant banding appears in the head roll. This suggests that the head yaw and roll velocities contribute significantly to the prediction of horizontal gaze. This contribution is in the same direction, i.e. turning the head to the left predicts an eye movement to the left and similarly to the right. The same is true for the vertical component but with head pitch and roll. For Linear Exogenous regression i.e. gaze based purely on body positions without autoregressive feedback, a similar pattern for the head weights is observed, over a larger time delay block. However, here it seems slightly more information is used from the spine, hips (rotation and translation) and even right and left shoulders. For the hips and spine rotation, the relationship is similar to the head's i.e. they predict gaze in the same direction. This is not the case for hip translation which seems to show more of a compensatory relationship with banding in the opposite direction. For the shoulders, the weights are strongest in association with shoulder roll (right-hand system, roll is about the Z axis). This suggests reaching movements have some predictive contribution to eye movement direction. The extent of this will become clearer in subsequent analysis which looks at different body parts' regression alone.

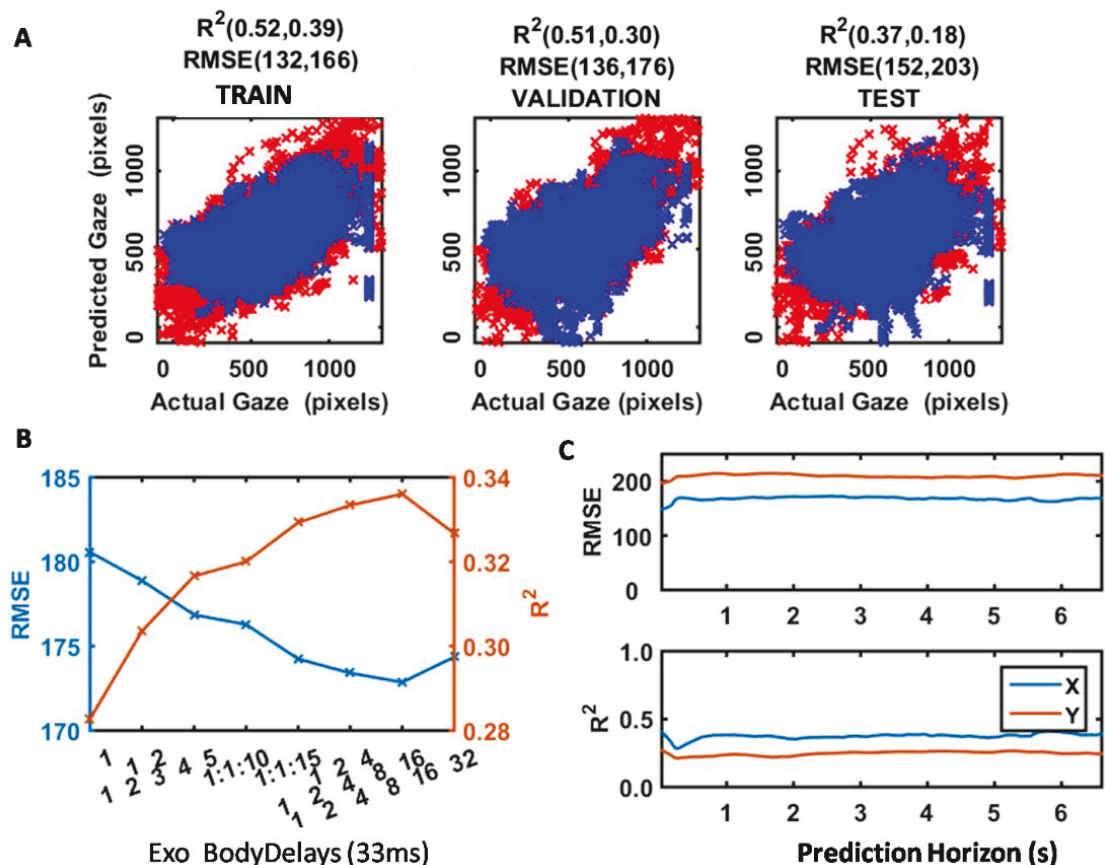


Figure 77. Exogenous regression: A. Prediction performance (neither closed or open loop as there is no autoregressive component) showing predicted gaze for vertical y (red) and horizontal x (blue) against actual gaze. Performance separated into train, validation and test data separately. B. Delay parameter optimisation (both input delays and exogenous input delays) in closed loop, showing root mean squared error (RMSE) and the squared Pearson correlation coefficient (R^2) performance for different input delays. D. Multi-step-ahead prediction performance (see Figure 72D).

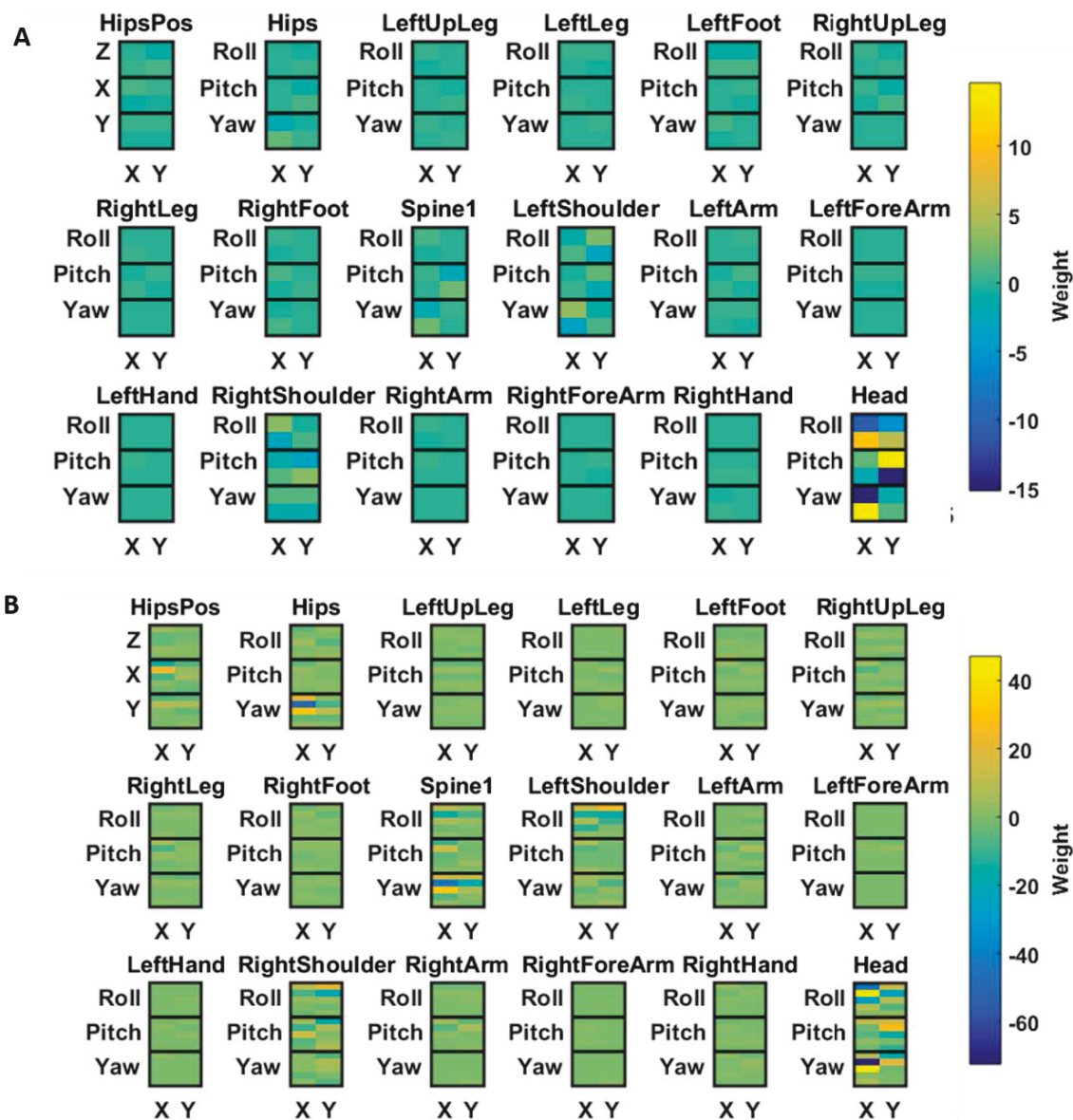


Figure 78 Linear Mapping Weights. A. Weights for Auto regression with exogenous inputs. B. Weights for exogenous regression.

A more qualitative representation of prediction performance for exogenous regression can be seen in Figure 79. This shows a one minute segment of gaze data (actual in red, predicted in -black) with the gaze events highlighted (red is visual intake, green is saccade and blue is blink) as classified by SMI software. The model prediction follows the low frequency trend of the data very well. In addition, some of the large amplitude and high frequency components are captured. Some of these actually tend to occur during blinks, i.e. if an eye movement has been made during a blink (gaze evoked blinks) the model captures the position of the subsequent fixation well. An example of this is shown more clearly in Figure 79, scene shot 1 where the subject has looked up and left from what they are doing to look at the experimenter. The predicted gaze (yellow line with triangular markers) and the actual gaze (white line with circular markers) for the last second has been superimposed on the static image, with recency corresponding to marker size

(largest marker is current data sample). The trajectories are very similar and the resulting fixation position has moved from the bottom centre to the top left - clearly captured by the model. The relevant time series data is indicated with an arrow where a large shift in both horizontal and vertical gaze is observed. Scene shots 2 and 4 also capture this behaviour. The model also captures the central fixation bias (Tatler, 2007), which seems to endure in the natural behaviour - particularly during manipulations. This is exemplified in scene shot 3. Here the subject maintains their focus centrally on the ongoing task of serving muesli, during which only small adjustments of gaze are made. This long period of limited gaze activity can be seen as a flat line on the time series. Interestingly during this period of concentration no blink events are observed.

The close linear relationship between gaze and head movements can be observed in the corresponding head movement velocity traces plotted below gaze in Figure 79. The head movements are constructive i.e. head movements predict eye movements in the same direction. That is to say, the effect of constructive interference between gaze and head movement, i.e. working together to change the point of attention results in a stronger prediction of gaze than the compensatory or destructive interference of eye and head movements during vestibular ocular reflexes. This description of constructive and destructive interference is borrowed from wave theory, where rotational motions can cancel each other when they are antiphasic (head turns left, eyes turn right) or phasic (both head and eyes turn left). It is interesting that the constructive relationship is pronounced during blinks, and relates back to the discussion on large amplitude gaze movements (head + eye) evoking blinks. But to bring the focus back to body movements in general, the following describes the predictive power of different groups of body joints.

7.3.5 Regression Performance Comparison

The performance of the two regression methods, auto regression with exogenous suit inputs and linear regression from suit inputs, is compared in Figure 80A,B. Linear regression from suit to eye gaze has a significantly higher R^2 performance than the auto regression with exogenous suit inputs (two tail paired t-test, kitchen $p=0.03$, bedroom $p=0.02$). The performance difference between settings (bedroom, kitchen and navigation) is not significant however. For the rest of the comparisons the linear regression from suit inputs to gaze (without auto regression) is focussed on. In the kitchen the absolute angular prediction error is 7.1 degrees, which is significantly better than chance metrics (central fixation bias CB - 10.4 degrees and the random uniform sampling performance Rand - 15.1 degrees) (Figure 80C). The difference is even clearer looking at the R^2 metric, where in the kitchen the performance is on average 0.27 compared to less than 0.001 for both chance levels (Figure 80D). In Figure 78B, the weight matrix for the linear regression from body joints to gaze is shown. To analyse the contribution of different joints Figure 80E and F compare the performance using different combinations of joints. The performance is significantly above the chance measures for all combinations of joints, even using joint angles from the legs alone. The R^2 performance is significantly higher for all joint combinations that include the head. However the performance is not improved significantly by combining additional joints when compared to the head alone. This suggests that globally, the current model performance is based on head rotation but there is information in the other joints, albeit seemingly redundant information with the current approach.

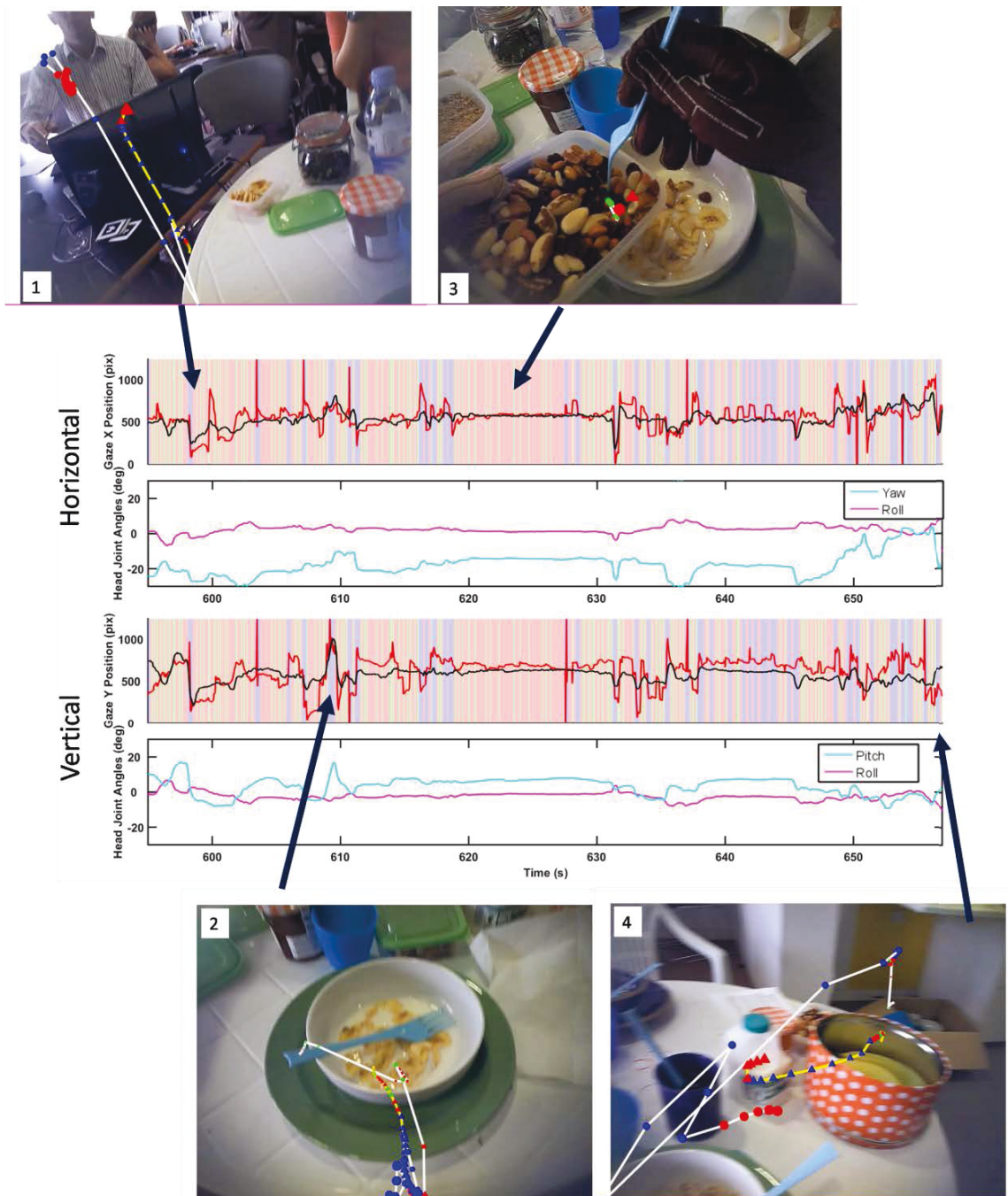


Figure 79: Time series prediction example using Exogenous regression. One-minute segment of gaze data (actual in red, predicted in black) and the highest 2 weight dimensions (yaw and roll for horizontal (x) gaze position and pitch and roll for vertical (y) gaze position). The coloured banding on the gaze plots represents gaze events where “red” is fixation, “green” is saccade and “blue” is blink as classified by SMI software. Frames from the scene camera are shown from various times along the time series and the predicted (yellow with triangle markers) and actual (white with circular markers) gaze trajectories for one second have been overlaid. Similarly to the coloured banding, the marker colour corresponds to the ongoing gaze event and the size represents the recency, i.e. the marker is smaller for samples taken further back in time. Note: The prediction algorithm is completely blind to video. Only previous body joint angles are used to predict current gaze position at each step.

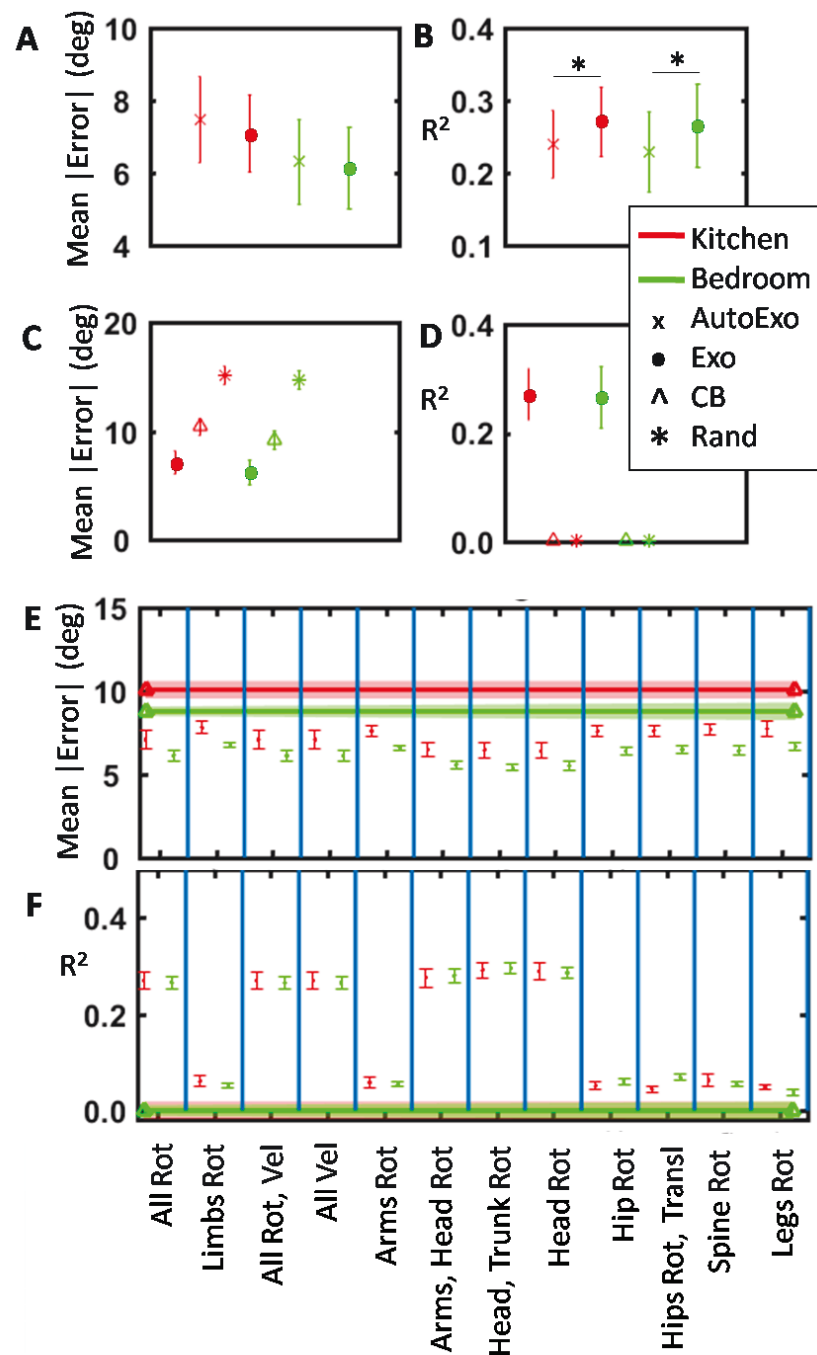


Figure 80: Performance overview. A. Comparing mean absolute error for each regression method to predict eye gaze (cross is auto regression with exogenous suit inputs, dot is linear regression from suit inputs) in each setting (kitchen - red and bedroom -green). All joint angles used, plotting mean and standard error cross validated and across all subjects. B. As for A, but comparing R^2 (Pearson's correlation coefficient squared). Here the linear regression from suit to eye gaze, pair wise, has a significantly higher performance than the auto regression with exogenous suit inputs, though not clear from plot alone (two tail paired t-Test, kitchen $p=0.03$, bedroom $p=0.02$). The difference between settings is not significant. C. Comparison between linear regression (dot) from all suit sensors to gaze and two "chance" metrics. The first metric (CB) is the common saliency benchmark based on the central fixation bias (see methods). The second metric (rand) is based on the performance with gaze estimates sampled from a random uniform distribution. E and F compares the mean absolute error and R^2 for linear regression with different combinations of joint inputs. In this context Rot means joint

rotation angle and Vel is the joint's angular velocity. For example All Rot, Vel uses the rotation angle for all joints concatenated with the rotation velocity for all joints. The straight lines plotted across the whole figure is the central fixation chance level for each setting..

7.3.6 Gaze Event triggered behaviour

To give an alternative perspective of the relationship between eye movement behaviour and body movements, behaviour is conditioned on the instances of the events themselves (fixation, saccade and blink). This analysis does not relate the continuous eye-movement behaviour like the cross-correlation, but rather asks the question: given a particular event (fixation/saccade/blink) what is the stereotypical body behaviour associated with it? The data for all subjects, across all settings is combined, thus analysing the response to 38228 saccades, 47011 fixations and 12113 blinks. Looking first at the head and trunk movement, Figure 81A-C shows the median absolute total rotation (summed over pitch yaw and roll) for the head, spine and hips respectively. The same scale is used for each type of gaze event to enable initial comparison. The absolute is taken because it is the activation rather than the direction that is of interest, as the gaze direction of the event is not considered. With a similar logic, the sum over pitch yaw and roll is taken, also the response is similar for each dimension and thus rather than showing each they are combined (later plots show them separately). For fixations, saccades and blinks there is a clear association with head rotation velocity, with the blinks having the most pronounced effect (Figure 81A). On average, away from an event trigger, there is a total magnitude of 10deg/s (summed across the Euler triplet) head velocity. Both saccades and fixation have a drop in the associated head rotation velocity (around 1-2 degrees/s). This drop has a different temporal associations for fixations and blinks which will be discussed in more detail shortly. Blinks on the other hand have a remarkably high amplitude of head velocity associated with them, rising rapidly by over 5 degrees/s. Finally, these events are compared with a random event, based on uniform-randomly sampled time points without replacement, for the same number of samples as there are saccades (N=27580). With this chance measure there is no observed response (see Rand Figure 81). For the trunk (spine and hips) a similar, albeit less pronounced, response is observed. Again the response is much more pronounced for blinks, making the response to fixations and saccades difficult to observe on the same scale.

Having compared the magnitudes between events, Figure 81D-F enables a closer look at the shape of the temporal response. The same responses are shown, this time purely for absolute pitch velocity alone and the axis ranges are scaled to the data range for each event response so the shape can be seen clearly. The baseline head pitch velocity is around 3 deg/s and falls by around 0.3 deg/s for saccades and 0.15 deg/s for fixations and rises by about 1.5 deg/sec for blinks (Figure 81D). Similarly for the spine and hips there is a reduction in pitch velocity associated with the saccade onset (Figure 81E,F). For fixations there appears to be an increase in hip and spine pitch velocity prior to the fixation onset followed by a decrease in velocity at fixation onset (Figure 81E). This shows there is a generalised locking of head spine and hips to the respective gaze events. Though the association may seem small, this is an average across all events (38228 saccades, 47011 fixation, 12113 blinks), for many of which there will be no associated head or trunk movement. Thus there is clearly a tight association between events and the head and trunk rotation. The temporal association will be examined further.

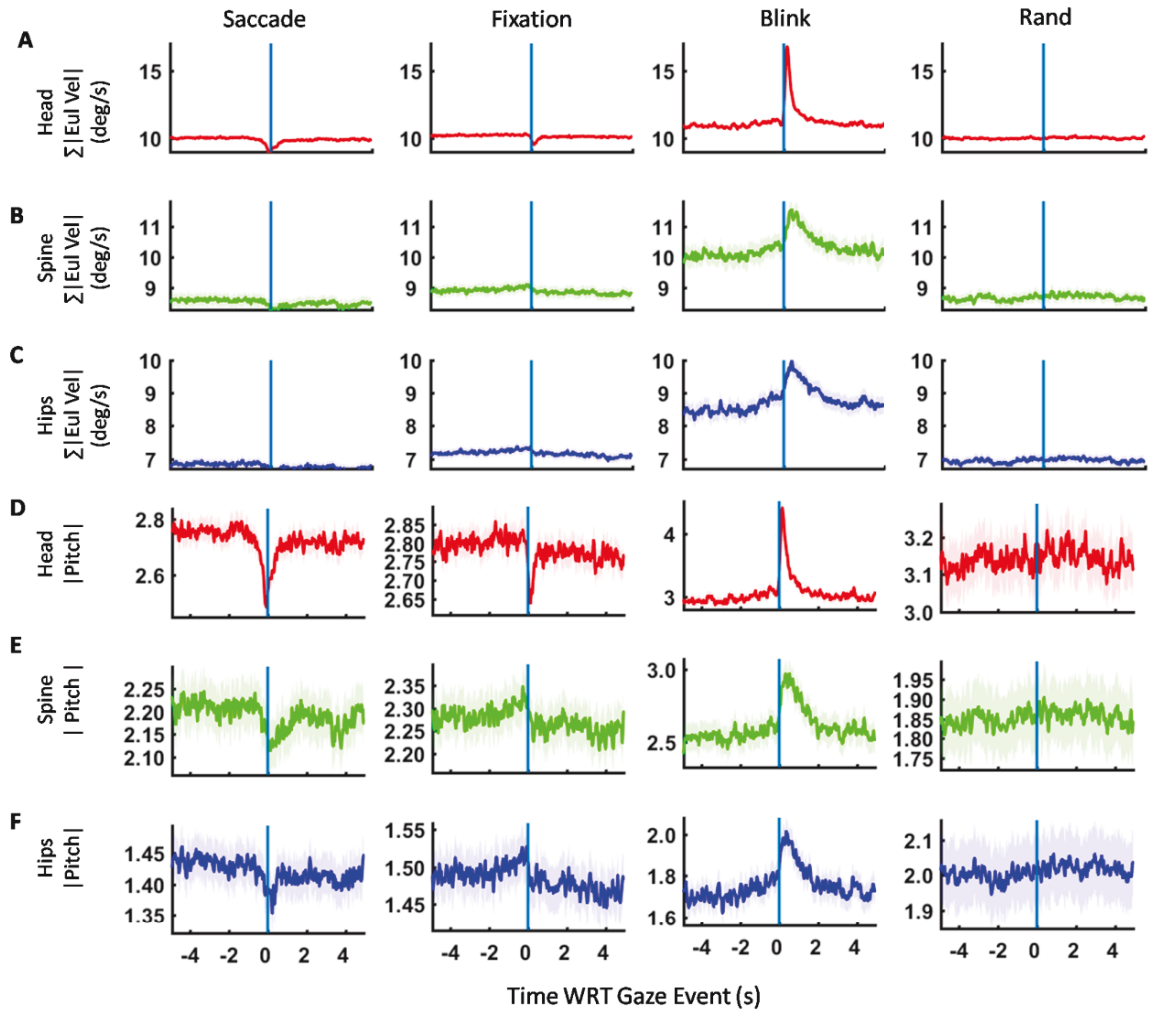


Figure 81: Event triggered averages for trunk and head rotation (median $\pm$ SDE). Data pooled from across all subjects and both kitchen and bedroom scenarios. For each event (saccade, fixation, blink) a window $\pm 5s$ is taken and the average response is calculated across all event instances, for a clear explanation of event triggered averages see Figure 73. A. Median Absolute Total Head Rotation (summed over pitch yaw and roll) triggered on each event type. Note: Random is based on uniformly randomly sampled time points without replacement, with the same number of samples as there are saccades ($N=27580$). B,C as for A but for spine and hips respectively. D. Median Absolute Head Pitch. Same as previous rows but just for pitch and here the axis limits are not fixed, instead scaled to tightly fit the range of the plot. This allows the variability to be observed more clearly. E,F as for D for spine and hips respectively.

To understand the temporal relationship more clearly Figure 82 shows the average head rotation response for each event on the same axis. To allow all event responses to be superimposed, the baseline movement 1.5 s prior to event (2.78 deg/sec fixation, saccade 2.8 deg/sec and blink 3.6 deg/sec) is subtracted. The red line is fixation, green - saccade, blue - blink and black - random trigger. For all events head movement is proactive of the event, i.e. the characteristic response starts prior to the event onset. This suggests characteristic head movements may be used to predict different event onsets. For blinks from 99ms (3 frames) prior to onset there is a rapid increase in velocity which peaks at 2.5deg/s, 165ms after the onset and then reduces rapidly to around 0.6deg/s 500ms after onset, and then less rapidly back towards the baseline around 1500s after. This gives an unambiguous demonstration of large changes of gaze (head driven) evoking blinks. For

saccades the head rotation velocity begins to decrease at $\sim 800\text{ms}$ prior to saccade onset, becoming significantly below the baseline from 330ms , and dropping more rapidly from 264ms prior to onset until a minimum of -0.43deg/s is reached 66ms prior to onset. By 33ms prior to onset the head velocity is increasing, returning to the baseline 693ms after the onset. For fixations the reduction in head velocity starts 66ms before onset with a much more rapid reduction reaching the minimum of -0.28deg/s 132ms after onset and returning to baseline at 363ms after the onset. This suggests that the reduction in head movement is associated with the fixation period not the saccade and the smeared out response of the saccade is due to the fixations of different durations that proceed and follow the saccade. It is a supposition of two fixation responses either side of the saccade. Again, this response is the average response and thus portrays the general relationship. Further analysis will cluster the event triggered trajectories to describe the full taxonomy of behaviours. For all other joint angles the averaged response did not show any trend. Thus this next stage of analysis, to see what is shrouded in the group average, is applied first to hand grasping. This relationship has not been studied in the wild previously and is important in the context of prosthetic control. The following section describes this approach.

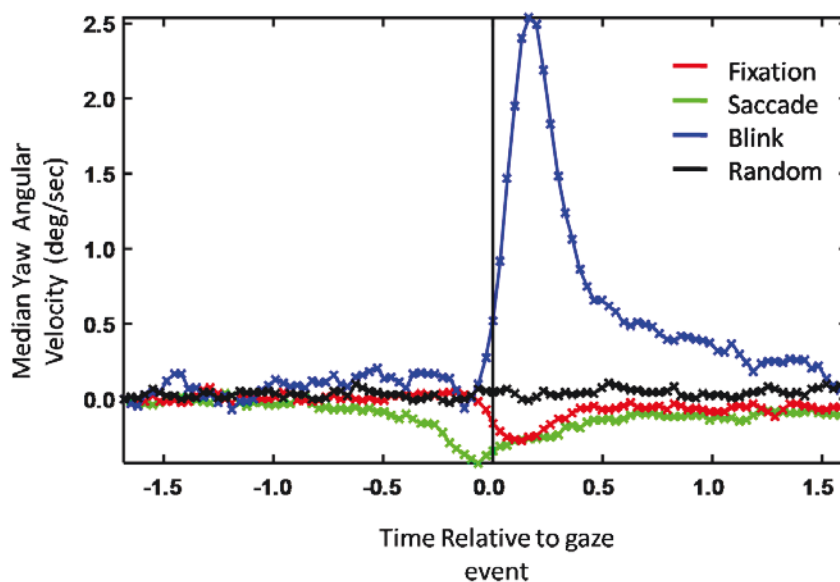


Figure 82: Event triggered head yaw angular velocity with baseline removed. For each event triggered average (median) the baseline movement $>1.5\text{ s}$ prior to event (fixation, saccade 3.2 deg/sec and blink 3.6 deg/sec) is subtracted to allow the different event associated head movement to be compared on the same axis. The red line is fixation triggered, green - saccade, blue - blink and black -random. Note: Random is based on uniform-randomly sampled time points without replacement, with the same number of samples as there are saccades ($N=27580$).

7.3.7 Saccade triggered grasplets

Similarly to head, spine and hip movements the temporal coincidence and dynamics of saccades and hand and finger movements is investigated. Hand movements were simplified to give a grasp velocity sum whereby the appropriate summation of finger and palm joint velocities was made to give a positive value for hand opening and a negative value for hand closing (see Methods previous Chapter). This grasp velocity was used to

temporally refine the online annotation positions by finding the nearest substantial hand closure and adjusting the annotation time to the exact end point of hand closure. The crude online annotation is thus refined to the higher resolution of the behavioural data. As outlined previously, saccade triggered windows are taken, here with a window size of 4.5 seconds. This window size was chosen based on the general findings of Land et al (1999) - that manipulations lasted on average 3 seconds (Land et al., 1999), and thus 4.5s contains this with an additional 50% to allow for different phases or additional behaviour to be captured. The resulting saccade triggered average grasp velocity had little trend, but large variability. This was unsurprising given that we make around 3 saccades per second while the hand may be static for periods much larger than 1 second. Thus to investigate this variability, and indeed answer the question of what proportion of saccades co-occur with grasp like behaviours, the saccade triggered grasp trajectories were clustered. An unsupervised approach was used to cluster the grasp trajectories. Each grasp velocity window is treated as a 141 dimensional descriptor and all descriptors are clustered using an agglomerative hierarchical approach with Ward's criterion used as an objective function (minimum variance of Euclidean distance within clusters). This starts with all data points considered as separate clusters, the clusters are then successively merged to minimise the resulting total within cluster variance at each step up the hierarchy. This approach gives an intuitive description of how the clusters emerge, how they are related to each other and how dense the clusters are. This gives a useful taxonomic representation when it is unknown how many clusters may exist.

Figure 83A shows the resulting hierarchical clustering of right hand grasp velocity. The first trajectory at the top of the hierarchy is the average across all saccade triggered windows where positive x is post saccade and negative is pre-saccade. Moving down the hierarchy new clusters emerge until the average cluster distance between each data point drops by 30%. At this point 8 clusters have emerged. Six of these eight clusters have a very characteristic positive peak followed by a negative peak, representing the hand opening and then closing - a grasp. Hence, in an unsupervised manner, these highly characteristic "grasplets" emerge. Their alignment with the saccade is different for each cluster suggesting saccades are locked to particular time points of the grasp trajectory. The other two clusters that don't show this grasplet shape include a flat trajectory (magenta line in Figure 83B) and a flat line followed by an upward peak alone (yellow line in Figure 83B). The former suggests little grasp activity associated with these saccades (the trajectory also remains flat if the average absolute is calculated). The trajectory which includes an upward peak alone could be one half of a grasp phase or it may be un-grasping i.e. placing an object by opening the hand. In total 27580 saccades have been classified. The proportion of these saccades that fall into each cluster varies significantly. Figure 83C shows the distribution of cluster membership. The fifth cluster (magenta line in Figure 83B) contains the majority of fixations with 14221 (52 %) while the grasp related saccades make up 11537 (42 %) with 1822 (6.6%) in cluster 6, the potential "place object" cluster.

These emerging grasplets can be cross-checked with the grasp annotations to see if they correspond with the labelled grasp behaviours. To do this each saccade is associated with its nearest grasp label and the temporal distance is recorded. The distribution over these temporal distances can then be observed for each cluster, as shown in Figure 84A. The clusters emerge in an unsupervised manner but here, when this is cross referenced with

the grasp annotations clear peaks emerge for each grasplet cluster, close to the grasp annotation. This is not the case for the non-grasplet cluster 5 (magenta line Figure 84A), for which the distribution remains fairly uniform. Cluster 6 (yellow line Figure 84A) also seems to have less association with the grasp annotations, this was the cluster which may be a placing object action rather than a grasp event. Thus it certainly seems that there is an association between these grasplets and the annotated grasps. To further test this, the peak temporal differences from Figure 84A are used to align the associated saccade triggered grasplets with respect to the annotated grasp time (Figure 84B). The phases align for all grasplets except for number 6 which provides further evidence that this may be associated with a different manipulation behaviour such as placing an object. The vertical lines indicate the saccade position associated with each grasplet and therefore the positions along the general grasp trajectory that saccades are likely to occur. This neatly corroborates the grasplets showing their strong relationship with the annotations. Again, to emphasise the point, these grasplets emerge in an unsupervised manner yet they have a very clear relationship with the grasp annotations. Although the peaks of the relative frequency distribution of temporal difference between saccade and grasp annotation only reach around 0.05, the distribution is very broad as the online annotations reflect only a small proportion of the total grasps, as the annotator cannot keep up with the frequency of behaviour.

7.3.8 Grasplets association with body movements

For each cluster assignment, the same temporal labelling can be applied to other data channels and thus the movement of other body parts associated with the grasplets can be explored. The median behaviour is observed for other body parts in Figure 83B, with the right hand activity (endpoint absolute velocity component summed over 3 dimensions), shoulder joint activity (absolute angular velocity summed over pitch yaw and roll) and similarly for head, spine and hip yaw. The median is used to minimise effects of shot noise observed in the velocity data from the suit. Looking at the emerging trajectories the first key observation is that cluster 5, the "non-grasplet", tends to have much less modulation of arm head and trunk associated (magenta line Figure 83B). For the other clusters, there is a strong locking of saccade, arm, head and trunk activities. For the arm, there is an up modulation of movement, reflected both in hand endpoint velocity and shoulder joint angular velocity. This behaviour is associated with all clusters except cluster 5 (magenta line Figure 83B). This modulation is the reaching movement which occurs as the subject opens their hand which peaks as the hand begins to close and then reduces as the grasp finishes. The full phase of this is reflected most clearly in cluster 2 (red line Figure 83B). In terms of the saccade temporal relationship to arm movement, in clusters 1,2 and 7 (blue, red and green lines Figure 83B), the saccade occurs simultaneously, or slightly after the arm begins to move. This may suggest that the fixation that follows this is the guiding fixation.

For cluster 4 the saccade is associated with the time at which the grasp goes from opening to closing, which is also the turning point for the reach velocity i.e. when it starts slowing down. This may be the saccade that finely tunes the fixation, or the saccade that moves on to the next target. For the saccades associated with the end of the grasp, clusters 3 and 8 (green and purple lines Figure 83B), by this point the grasp velocity has returned to the pre grasp baseline value i.e. the reach has finished and the grasp has been made. As for the

head and trunk movements, a similar modulation of activity occurs simultaneously with the reach. Thus it seems that grasp like behaviours tend to be associated with head and trunk movements. There are saccades and trunk movements that are prominently proactive of the object manipulation, with activities starting around 1 second before the grasp starts closing and around 2 seconds before the grasp is complete. Currently these saccades are not associated with one another and multiple saccades will be associated with the same grasp. Thus the next stage will be to look at the probable associations of these saccades within a grasplet and of course the characteristics of these saccades and interleaved fixations to begin to infer cognitive associations. For now this demonstrates the power of such a data driven approach and how, with no hand coding, the taxonomy of gaze and body behaviour during a grasp begins to unfold. Additionally it suggests there is predictive information that could be utilised for grasp intention decoding for prosthetic control.

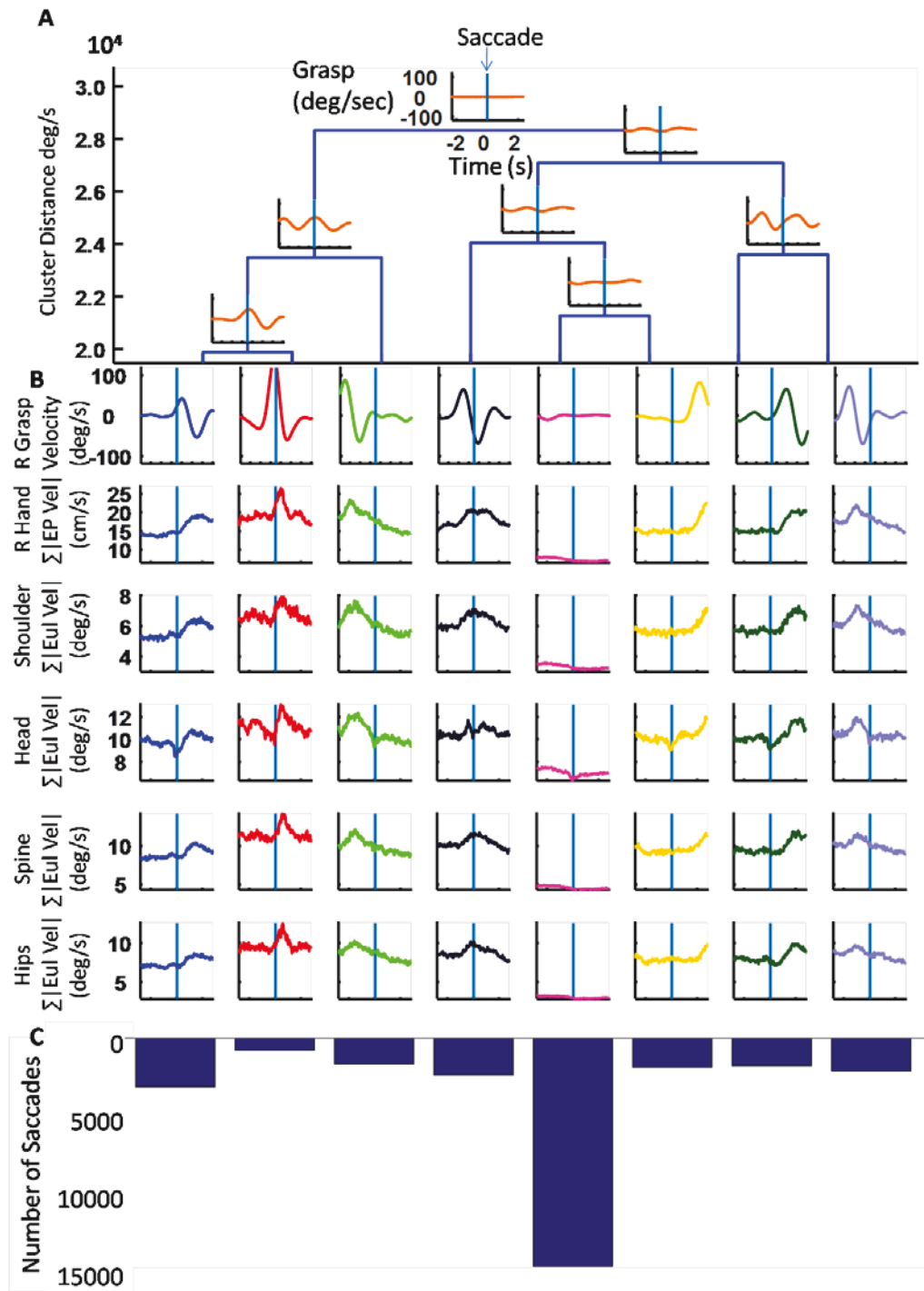


Figure 83: Hierarchical clustering of saccade triggered right hand grasp behaviour and associated body movements. A. Dendrogram representing the hierarchical clustering of right hand grasp velocity (summation over hand joints' velocities where positive is open and negative is closing hand). The first trajectory at the top of the hierarchy is the average across all saccade triggered windows where positive x is post saccade and negative is pre-saccade. Moving down the hierarchy new clusters emerge until the average cluster distance from each data point drops by 30%. At this point 8 clusters have emerged. **B.** The first row are the average grasp velocities of the 8 clusters that have emerged. The subsequent rows use the same cluster assignments but on other movement measurements including: right hand horizontal endpoint absolute velocity component, shoulder roll absolute velocity, head yaw absolute velocity, spine yaw absolute velocity and hip yaw absolute velocity (note yaw is rotation about the vertical y axis). **C.** Distribution of cluster membership.

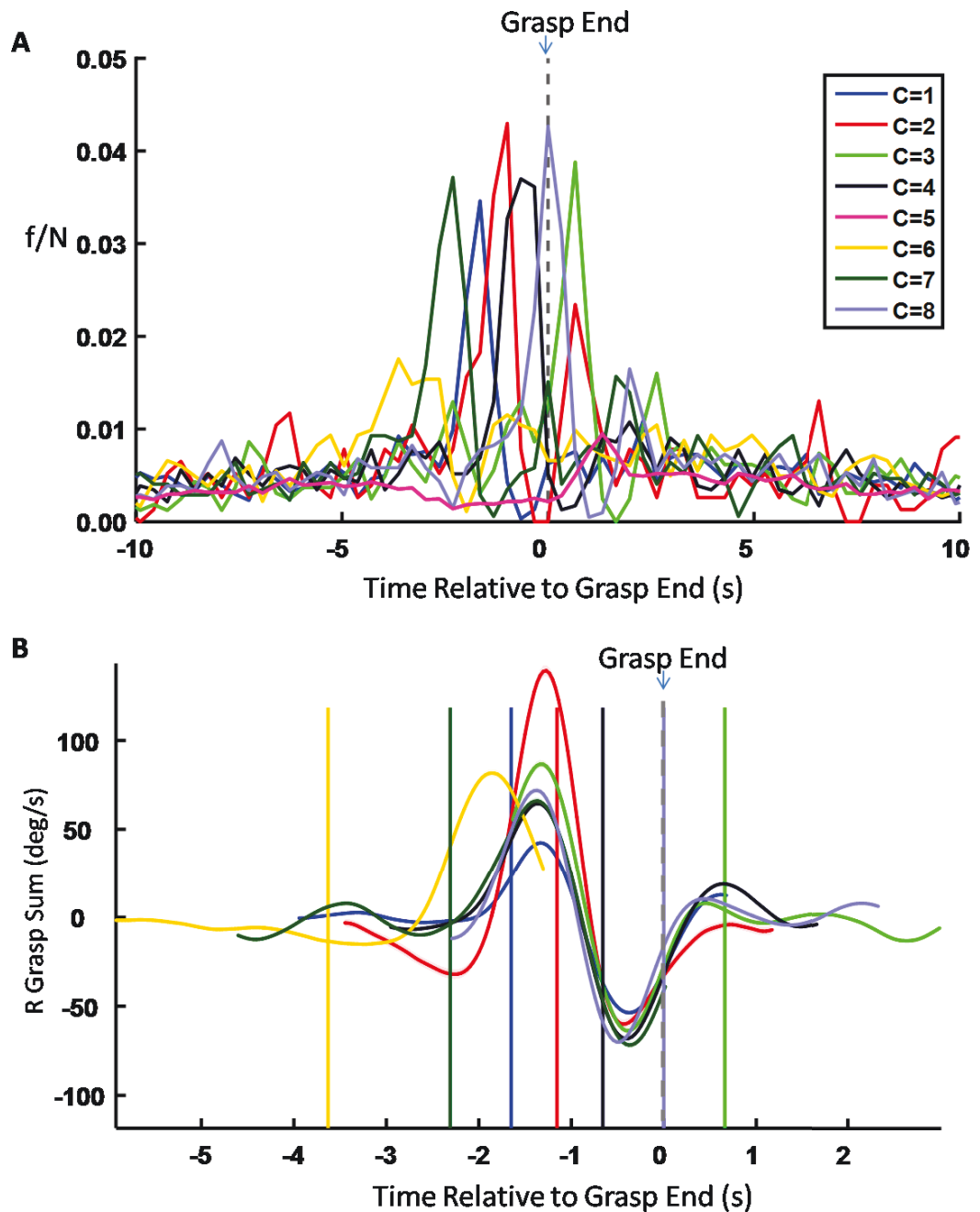


Figure 84: Grasp annotation cross reference: The clusters emerge in an unsupervised manner but here they are cross referenced with the refined online annotations of grasps. First each saccade is associated with its temporal distance from the nearest grasp label. A. The distribution over these temporal distances observed for each cluster. Grasp end refers to the annotated grasp end position. The x axis represents the temporal distance between saccade and nearest grasp annotation with negative time indicating that the saccade occurred before the grasp. B. The peak temporal distance for each cluster saccade-grasp distance distribution from A is used to align the respective cluster trajectory and the phases match up. The vertical lines indicate the saccade position associated with each cluster trajectory.

7.4 Discussion

Classically, allocation of gaze has been attributed to both bottom up scene salience and top down ongoing task demands. A new method is proposed for analysing and interpreting eye-movements in the wild, by relating them directly to body posture. Thus, the interpretation of gaze and attention is integrated, not disembodied, from the kinematics of motor behaviour. Previous data driven methodologies have addressed eye-movement behaviour in the wild, but have focused on natural visual stimuli (Foulsham et al., 2011; Hart et al., 2009). Efforts have also been made to incorporate head movements (Einhäuser et al., 2007) however this is the only known study known to incorporate such comprehensive real-time body movement recordings. The embodied eye-movement methodology described in the previous chapter has led to two major concepts presented here: Embodied saliency - the interaction of body posture and spatial allocation of overt attention and Embodied Eye Event descriptors - the association of eye-movement events such as saccades with different body movements. To elaborate further, Embodied Saliency attempts to capture the concept of interpreting the spatial allocation of gaze behaviour in terms of the conformation of the body. This concept is important when interpreting gaze behaviour in the wild, where gaze can no longer just be thought of as eye-movements in response to visual stimulus. Gaze allocation is instead a coordination of body, head and eye-movements. In addition the actions of the body, which are recorded right down to every joint in the finger, must affect how this attention is allocated. This approach has previously been taken for recording head-movements and eye-movements simultaneously in the wild, for which the results will be compared with those presented here. In addition there is one study that uses head movements in their saliency model to predict head free eye-movements in the lab. Expanding this to the rest of the body, this study is the only known integrated analysis of body and eye movements in the wild. Thus, these novel findings are presented, with the focus on hand movements at this stage, with their implications for prosthetic control.

7.4.1 Head-eye coordination: other data driven approach comparison

Using purely automated methods and no hand coding, the raw gaze direction, gaze events and full body kinematics have been explored. While it only scratches the surface of such an extensive and comprehensive dataset, the following summarises what has been confirmed from other studies, what has been learnt here and the evident power of such a data driven approach. Starting from the top, eye and head coordination has been studied in a large number of studies, some of which in the wild. The only other data driven approach (not hand annotated) to assess the eye-head coordination in the wild was Einhäuser et al. (2007)(Einhäuser et al., 2007). In this study the head and eyes were tracked during natural exploration of various environments. Some of their findings have been discussed previously, but here the head and gaze coordination is discussed. With this data-driven approach the findings here show, similarly to Einhäuser et al. (2007), that eye and head-movements co-occur in a number of different modes(Einhäuser et al., 2007). The first is the well-known smooth head compensating eye-movements (VOR and OKR) and the second is the additive eye and head movements acting together to shift gaze. Within the second category, Einhäuser found that not only saccadic eye-movements were additive or synergistic with head movements to shift gaze, but smooth eye-movements could also be

synergistic with head movements. This could have important implications for modelling the temporal statistics of natural stimuli because it suggests they cannot be treated as a series of static frames interleaved with rapid gaze shifts. There are some important differences between the methodologies which need to be discussed when making comparison.

Einhäuser et al.'s study did not segment gaze behaviour, arguing that the low frame rate of their eye velocity signal (12.5Hz), which was inferred from a 25Hz gaze actuated video camera (EyeSeeCam), was too low for event detection. Unlike Einhäuser et al. in the study presented here, the gaze events were detected based on the 30Hz raw gaze signal and while this is not ideal, as has been discussed at length, there is a clear understanding of what the detected events mean (visual intake category shown to include smooth eye-movements for SMI method, see previous Chapter). Looking at the cross correlation between head and gaze during saccades demonstrates a strong positive correlation (synergistic) at ~ 0 lag for both horizontal and vertical movements. For visual intake (SMI method), there is a negative correlation (compensatory) at a lag of 230ms in the horizontal direction, which could suggest the feedback mechanism of the optokinetic response (OKR, >150 ms latency) has a dominant effect over the feed forward vestibular ocular reflex (VOR, ~ 15 ms latency). For the vertical head-eye correlation there seems to be both a negative peak at a similar lag and a positive peak at zero lag. This supports the observation made by Einhäuser et al. that there are synergistic eye movements during visual intake, as well as compensatory. It is not clear how much of this may be attributed to pursuit, or why this is only observed in the vertical axis at this stage. One critique of the method used here is that cross-correlation is measured globally. Local windowed cross correlation could give a clearer representation and look at how this relationship changes over the course of the experiment.

Looking at the association between gaze events and head movements, using event triggered averages, also showed a strong association with head movements, but also with the hips and spine. This is the first insight into this behaviour in the wild from such a large sample and without hand coding. In general, during fixations a decrease in head, spine and hip rotation velocity is observed, while during saccades the velocities increase. The behaviours are tightly locked in time to each other. Interestingly, the response is far larger for blinks which suggests that head, spine and hip movements are much more likely to be associated with gaze evoked blinks. This analysis is for all subjects, all settings and all events. While interesting responses are observed globally, further analysis is required to break this averaged response into different groups of behaviours. For example, many saccades may have no head movement associated, while others will have large head with spine and hips. To begin to explore this, machine learning methods need to be applied to yield the taxonomy of different behaviours which may have strong implications for interpreting the ongoing task. Currently efforts here focus on the taxonomy of grasp associated behaviour, which will be discussed. However similar techniques should be applied to the head, trunk and other body joint angles to understand the repertoire of gaze control (the combined effect of these movements) in the wild. To look at the implications of these associations, bringing in the whole body, the attempt is made to demonstrate the hypothesis of Embodied Saliency. This is the idea that the conformation and dynamics of body movements can be used to predict eye movements.

7.4.2 Embodied Saliency: the synergies of spatial attention allocation

The factors that influence the allocation of gaze and thus eye-movements is an ongoing active research area. The so called saliency argument, which has been discussed at length in Chapter 5, between bottom up image features and top down task influences continues to be explored. Here an additional model factor is proposed, the idea of Embodied Saliency. This is the idea that eye-movements can be predicted by the conformation and dynamics of the body. Perhaps the body stance we take up and the way we are moving our head and limbs predicts where we are likely to be looking. There is a huge amount of literature that suggests this to be the case, linking eye-movement to action, but as discussed it tends to be qualitative in the wild and while lab studies give exciting justification, the findings need to be demonstrated in a similarly quantitative manner in the complexities of behaviour in the wild. Here by mapping directly from full body kinematics to eye-movements, significant predictive power is shown. The strongest predictive power comes from the head. This predictive power is synergistic i.e. turning the head to the right predicts an eye-movement to the right.

A conceptually related recent study in the lab by Nakashima et al. (2015) showed that head movement was an important input to the saliency model (Nakashima et al., 2015). In some respects these results (including our own) may seem trivial, however. As in essence, claiming that "head information is useful in predicting gaze" is a somewhat misleading statement because head movement is a component of gaze movement in the wild, unless the eyes compensate for it. However it is interesting to understand the extent to which the body is involved in physically shifting gaze location. In addition this relates to work by Tatler et al. demonstrating the central bias of gaze location in screen based experiments (Tatler, 2007). This simple bias that predicts gaze position with an image centred Gaussian probability distribution, outperforms visual based saliency models. This bias was applied as a benchmark in the current study to compare the embodied performance, and it achieved on average 10.4 degrees average absolute error, significantly above chance. Thus this effect also holds for natural behaviour, particularly during object manipulation. This has also been shown in previous studies "in the wild" as discussed in Chapter 6 (Foulsham et al., 2011; Hart et al., 2009). The results presented here build on these findings as they were for navigation/walking tasks, while here the effect is shown during object manipulation. Qualitatively, this effect looks even stronger during object manipulation (see Figure 76, scene shot 3). Future work should directly compare this effect between the scenarios recorded here (navigation, bedroom, kitchen) and sub scenarios within these contexts, e.g. standing versus sitting. In addition, it has been shown here that incorporating head movement also significantly increases performance (average absolute error 7.1 degrees). Foulsham et al. have already suggested that the head movements can be used as a proxy for eye-movements in natural behaviour, due to the enduring effect of central bias in the head centred (Foulsham et al., 2011). Coupling this with the findings here that head movements are even predictive of eye-movements within this head centred reference frame supports this assertion. Further than this, the results enable the dynamics of eye-movements to be predicted from head movements with a highly significant effect.

The fact that this central fixation bias endures in natural behaviour counters arguments that suggest centre bias in static natural image viewing is an effect of photographer bias

(Tseng et al., 2009). The explanation for centre bias enduring in natural behaviour could instead come from the fact that the head and indeed whole body movements are used in combination with the eyes to change gaze heading. While the eye gaze may arrive at the target first, as the head catches up the eyes return to central fixation and thus the amount of time eye-gaze needs to be held at extreme eccentricities is minimised. This would be beneficial for both energy constraints and minimising motor noise. This eccentricity minimisation bias, similar to the concept of orbital reserve referred to here (Tseng et al., 2009), may then endure - even in head restrained screen viewing. This should be investigated further. The primary aim here is to test the hypothesis of embodied saliency, and thus, for the head, it is shown here that it provides a strong predictor of eye-movements, beyond centre bias. But this is just one component of behaviour - the coordination of eye and body movements to synergistically shift gaze.

7.4.3 Embodied Saliency: other body contributions to spatial attention

There are several other aspects that need to be discussed, including compensatory eye-movements and the coordination between hand and eye. The compensatory aspect of eye-movements to body movements cannot be captured simultaneously here due to the limitations of the linear regressive model. The larger synergistic movements dominate the rotation mapping weights for the head and trunk (see Figure 78). The model complexity needs to increase to enable multiple modes for the same joint, i.e. head movements can be both synergistic and compensatory. Perhaps by distinguishing slower head movements, for which the eyes can compensate, from faster head saccades that are likely to be synergistic; these modes can be switched between in the model.

For the hand eye-coordination aspect, the model weights suggest some input from the shoulder, which acts to change the hand position. In addition the contribution from arms alone is significantly better than the central bias benchmark. However when analysed with and without this contribution, the prediction performance does not change significantly. Other recent studies do show a hand contribution to saliency though (Li et al., 2013). This study comes from the Egovision literature (see extensive review in previous Chapter). Here the head motion and manipulation point are inferred from the ego video of free behaviour in the wild. The manipulation point is extracted based on hand recognition, with the location and orientation dictating the manipulation point. Unlike Nakashima et al. (2015) and the study presented here, Li et al.'s results show that head prediction power is actually worse than centre bias alone in terms of absolute angular error, but combining hand information does indeed improve the performance from 10.66 degrees to 8.35 (area under the curve for the receiver operator curve 0.82 to 0.87). This result further supports the hypothesis of embodied saliency; the presence of hands in the egovideo gives a strong predictive bias of where subjects will be looking. Their results raise the question of why the head contribution seemed negligible for their study and also, why the same improvement is not seen with arm information in this study. Firstly, the data sets their results come from are much shorter: 4 min meal preparation tasks. They therefore only represent a subset of behaviour without the requirement of larger gaze shifts around the room that tend to be accompanied with large head and trunk movements (Freedman and Sparks, 1997; Land, 2004). This may account for the limited effect of head movements.

For the hand contribution, their recordings involve an epoch of intensive manipulation of objects in meal preparation, while in our recordings the durations are much longer. Meal preparation is involved, but this is just one of the many aspects of behaviour. For example there are periods of inactivity while talking. Thus the effects observed globally in our experiment are subtle but may be more pronounced in certain subsets of behaviour. Again this is a limitation of the global linear model used to predict eye-movements. The following discusses this and how the model can be extended.

7.4.4 Embodied Saliency: extending the model

Currently body movements can be used to explain up to 60% of eye-movement variance with a simple linear model and thus the next step is to see how the remaining variance can be attributed to other embodied interactions. The current linear model used is dominated by the synergistic effects of head and trunk movements in predicting eye-movements. While subtle effects are seen in other regions of the joint space such as the shoulders and hip position, their contribution do not significantly improve performance across the whole recording. This does not refute the claims of embodied saliency, but demonstrates that the linear model used cannot fully capture the complex and multifaceted nature of these interactions. The model can now be extended and iteratively improved to explain more and more variance of eye-movement behaviour. This is one of the many benefits of capturing such comprehensive behavioural data-sets. It is with this approach that the layers of embodied saliency can be delaminated as more complexity is built into the modelling. To do this the following suggests important aspects that need to be added.

Multiple Time Scales: Eye-movements occur on very different time scales, with rapid saccadic eye-movements and smooth slow eye-movements to completely static fixations. The same is true for different body parts which tend to act over considerably longer time-scales than the eye. For example, for shoulder movements the FWHM for shoulder elevation velocity is 600-800ms (Andreas Thomik, PhD Thesis 2016), compared to ~100ms for the eyes. Thus when a reach is initiated, the gaze movement associated with it is long over by the time the end of the trajectory is reached. This compared to a saccadic head movement, which although slower than ballistic eye-movements, the time courses are much more similar. Therefore, using a constant linear filter to map from arm to eye movement is not ideal, particularly given that by the time the arm reaches the position that was fixated, the gaze is likely to have already moved on. This brings in the question of latencies.

Multiple Latencies: Another limitation of the current linear model is that the number of lags is a single optimisation parameter across all dimension, rather than allowing different lags for different joints. Thus the optimal latency of 5 frames (165ms) used to map from body joints to eye-movements is likely to be suited to leverage the dominant synergistic head movements but not be optimal for other joints. Further to this, the model should also capture proactive inputs, for example the hand position in 1 second may contain information about the current gaze position. It will then be these components of the model that have important implications for prosthetic control, as the model can be inverted to predict future arm movements from the current gaze behaviour. But these proactive planning modes must be separated from the reactive proprioceptive modes, where the gaze is drawn to the current position of the hand, which introduces the next component.

Multiple Modes: As it has been discussed with regard to the head and trunk, the contribution to eye-movements can be both compensatory and synergistic. In the same vein eye movements associated with the hand can be both guiding and following, i.e the eye can lead the hand guiding the action as has already been discussed at length but in other modes the eye can be guided by arm position and thus proprioception (Ren et al., 2006). Thus, the model must enable these different mechanisms to be captured and then to switch between them appropriately. The ideal would be to have multiple controller models, some based on the well characterised subsystems from the lab, and to then fit behaviour locally as a weighted combination of these controllers, yielding interesting results as to how these weightings change over time in different behavioural contexts in the wild. Such an ideal scheme would give a data driven taxonomy to build up descriptors that can predict more and more of the eye movement variance.

7.4.5 Embodied Gaze Descriptors: building blocks of visuomotor behaviour

This endeavour to model the relationship between body and eye-movements and how the different contributions interact in different behavioural contexts is likely to lead to cognitive implications. Embodied saliency should therefore also be thought of, not only as where we look, but how that reallocation of gaze has been achieved (different synergies of trunk, head and eye), and what other body movements are also associated. Einhäuser previously showed the proportion of gaze reallocations that occurred during experiments that were based on eye movement alone (19%), head movement alone (8%), eye and head together (42%), with no reallocation occurring the remaining 31% of the time (Einhäuser et al., 2007). The extent of coupling between eye-movements, head and indeed trunk movements is likely to be determined by the current behavioural requirements as has been purported in other studies (Pelz et al., 2001). The presence of such coupling has particular implications for the temporal requirements, for example increasing the duration of the fixation following a saccade has been shown to increase the probability of an associated head movement (Oommen et al., 2003). As discussed temporal characteristics of eye-movements can be related to ongoing cognitive tasks and thus embodying these descriptors can provide additional information. With the unique data-set collected, these embodied gaze descriptors can include limb and hand movements. As previously mentioned, gaze shifts have been shown to be faster and shorter when physically tapping a set of targets versus just sequentially fixating them (Epelboim et al., 1997). In addition the accuracy of saccades has been shown to be higher in the presence of arm movements (Lünenburger et al., 2000). Finally, the field of embodied attention tells us that the state of the body modulates attention and the level of engagement (Dufour and Touzalin, 2008). Thus, taxonomising gaze events such as saccades based on their association with different body movements enables a step towards investigating these relationships in the wild. These embodied gaze descriptors are the building blocks of visuomotor behaviour and thus their exploration and characterisation has important implications for neuroscience. In addition in the context of prosthetic control, if these associations prove to be as characteristic as they have been described in the lab, they can contribute to the prediction of intended body movements from eye-movements and other retained motor functions. As such the first hand-gaze taxonomy in natural behaviour has been explored here.

Unlike previous methodologies, the kinematics for the entire body have been captured. In general the global relationship between eye-movements and the rest of the body movements (upper and lower limbs) yields no trend for cross-correlation and spike triggered averages (as mentioned this needs to be performed with local windowing in the future). While there were weak contributions from the shoulders in predicting eye-movements from body movements, these were not globally significant. However, Land's hand coding of behaviour shows there is a strong association (Land et al., 1999), but that the relationship is more episodic and less continuous than the head and trunk. For the head and trunk, their movements directly contribute to the control of gaze, while in a sense, gaze contributes to the control of the arm, or rather they are computationally associated rather than physically causal. The arm, moving much less than the eyes, is associated with far fewer gaze events than the head and trunk and thus a global relationship is too heavily diluted to yield significant results. From an intention recognition perspective the association of eye-movements and object interaction is of particular interest here and thus this relationship was mined in more detail. Starting from the saccade triggered average grasp metric (summed over all of hand joints), clusters emerge with grasp-like patterns.

7.4.6 Grasplets: predictive body movement synergies

Grasp like patterns, or "grasplets" as they have been referred to here, emerge in an unsupervised manner from the saccade triggered grasp trajectories (see Methods Figure 73 and results Figure 83). These grasplets are corroborated against the grasp annotations (Figure 84). In addition hand endpoint and shoulder joint rotations precede the grasplets showing the reach component. These grasplets are associated with 42% of saccades. Thus in order to decode grasp intentions, these grasp associated saccades will need to be differentiated from all the remaining 58% of saccades to extract relevant task related gaze metrics such as fixation location in 3D. By associating these grasplets with the head spine and hip movements, a clear pattern of association emerges. Thus it seems that grasp like behaviours tend to be associated with head and trunk movements, as has been shown previously by (Land et al., 1999). The results from their study are shown in Figure 85. This observation that trunk and gaze are tightly locked and proactive of object manipulation gave major evidence to justify the use of gaze and body movements to predict the intended grasp. This result has been reproduced in this study, but on tens of thousands of saccades, rather than on tens of saccades. However unlike Land et al. (1999), trunk movement is not observed to be proactive of the saccade. Instead reach, trunk, head and saccade seem to be initiated almost simultaneously. What is clearly seen, is that there are saccades and trunk movements that are prominently proactive of the object manipulation, with activities starting around 1 second before the grasp starts closing and around 2 seconds before the grasp is complete. Thus to a certain extent the relationship is similar to the observations of Land et al., however there is clearly a larger variability which becomes apparent with such a large sample size.

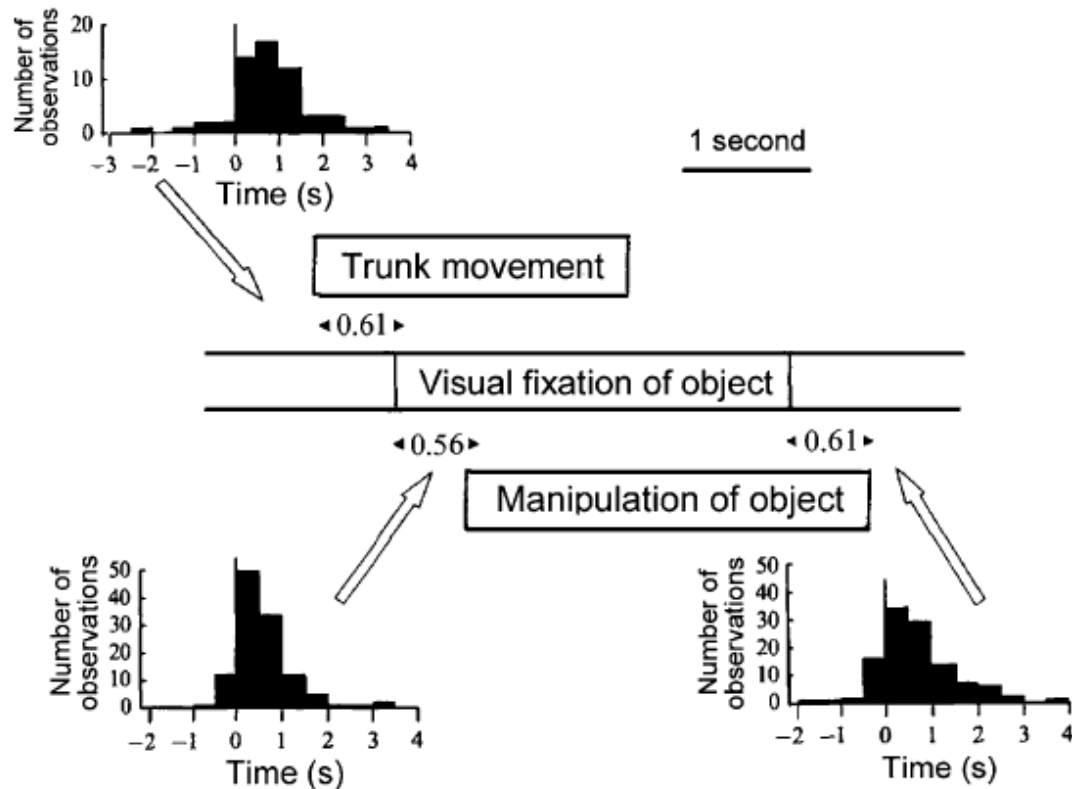


Figure 85. Temporal association of trunk, gaze and object manipulation from(Land et al., 1999). Fixations with long waiting times have been excluded and the data represents 94 object related actions from 3 subjects.

This data driven approach has allowed the gaze behaviour to be associated with characteristic hand movements in a highly efficient and scalable manner. Over 10 hours of data can be automatically analysed with the click of a button. No longer must the data sets be limited to 4 minutes of data due to the extremely laborious hand coding frame by frame. The preliminary results extracted demonstrate gaze and trunk movements are predictive of object manipulation. Furthering the understanding of this association in a data driven approach will have important insights for neuroscience and intention decoding for prosthetic control. The validity of these embodied gaze descriptors should now be explored further in application to intention decoding.

The challenge is to infer the intention to make a grasp and then to pinpoint the time at which to extract end-point information from the highly dynamic gaze. This relates to the Midas touch problem - i.e. the intention information is in the gaze position but when? That is to say, given we move our eyes up to 5 times every second and make grasps much less frequently (at most every 3 seconds), which fixation(s) carry the grasp associated information. The next line of investigation would therefore be to understand how many saccades are associated with each grasp event and the characteristics of these saccades, associated fixations and body movements. These characteristics may identify which part of the information retrieval they are associated with. For example, the first saccade during a grasplet may be a large orienting saccade with head and trunk movement to locate the object in ambient mode; this could then be followed by a smaller saccade to get more specific information (focal mode) from the object handle for example. After this key

information gathering saccade, there may be a saccade away from this object (before grasp is complete) towards the next action landmark (e.g. the bowl to pour milk into). The dataset gathered here is extensive and thus scope for further investigating of such interactions is vast. As the taxonomy of behaviour is expanded, its association with task and cognitive processes will further scientific understanding of ecologically valid behaviour where the brain must control many more degrees of freedom than in the lab.. With this taxonomy will also arise a richer descriptor set. These low level embodied gaze descriptors can be integrated in a hierarchical manner, which in combination with contextual cues from EgoVision and the learnt grammar of action, hold the hope to develop intention decoding systems for the intelligent, intuitive and non-invasive control of prosthetics.

8 Conclusion

This thesis contributes both to behavioural neuroscience, with new ways of interpreting eye-movements in the wild, and to neurotechnology with contributions towards the development of gaze controlled biomimetic prosthetics and other disability aids. Here the major challenge is to derive a practical control signal from eye movements that meets the interface requirements without being intrusive on the natural sensory function of the eyes. To achieve this, interface developments must be intrinsically linked to an understanding of natural human behaviour. Working at the interface between engineering and behavioural neuroscience provides complementary perspectives. Drawing from scientific understanding of behaviour informs the technical development, while developing data driven technological platforms provides tools to further our understanding of science. Figure 86 gives an overview of this approach as presented in the introduction, with some key thumbnails taken from different Chapters in the thesis. This involves developments that span both the entire technical system, as well as to study the behavioural system we are trying to restore: from the eye-image pixels, to 3D gaze estimation (Chapter 2), to robotic end-point control (Chapter 4), while in parallel the perception action loop in natural behaviour (Chapter 6,7). The following summarises and links these developments together. This holistic view enables the current potential of the developed systems to be discussed and the future vision of research and development is suggested.

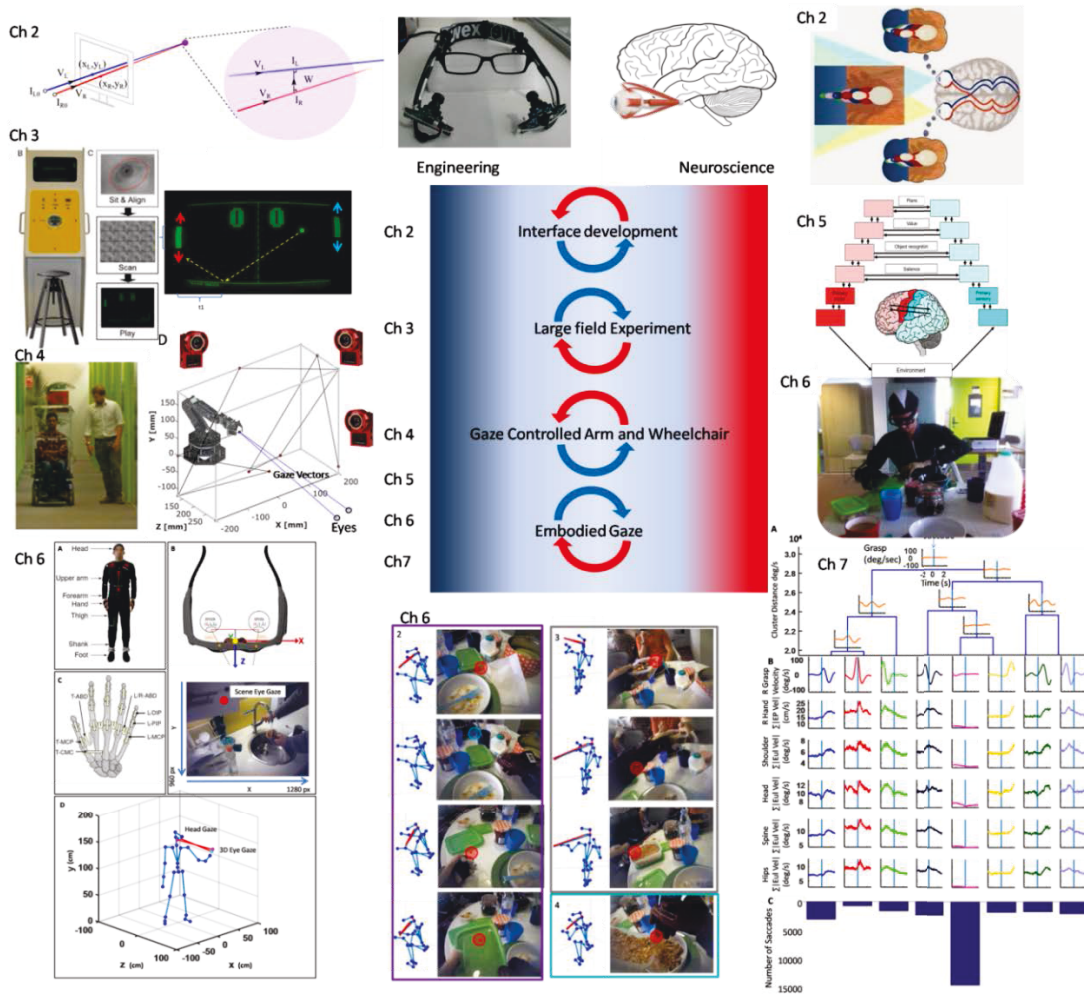


Figure 86: Thesis overview. Shows the four main developments and their relationship at the interface between engineering and neuroscience in combination with key thumbnails of these developments..

8.1 Summary of results

The summary of results is framed by their contribution to the future vision of biomimetic gaze controlled prosthetics. Figure 87 shows a general sketch of the perception action loop. During a task such as making a cup of tea, there is a required sequence of behaviour to achieve the goal. The kettle must be boiled before the water is poured into the cup and the cup must be retrieved from the cupboard before it can be filled. Taking a single sub action "get cup" may involve a search task, memory retrieval as to where the cups are kept or where they are likely to be kept. Once the cups have been recognised, the cup is chosen, the 3D location, orientation and shape may be estimated and a reach followed by a grasp retrieves the cup. If the action component is dysfunctional, whether due to amputation, spinal trauma or disease the aim is to re-interface with and leverage what remains in the loop - user intention, eye-movements and any retained body movements.

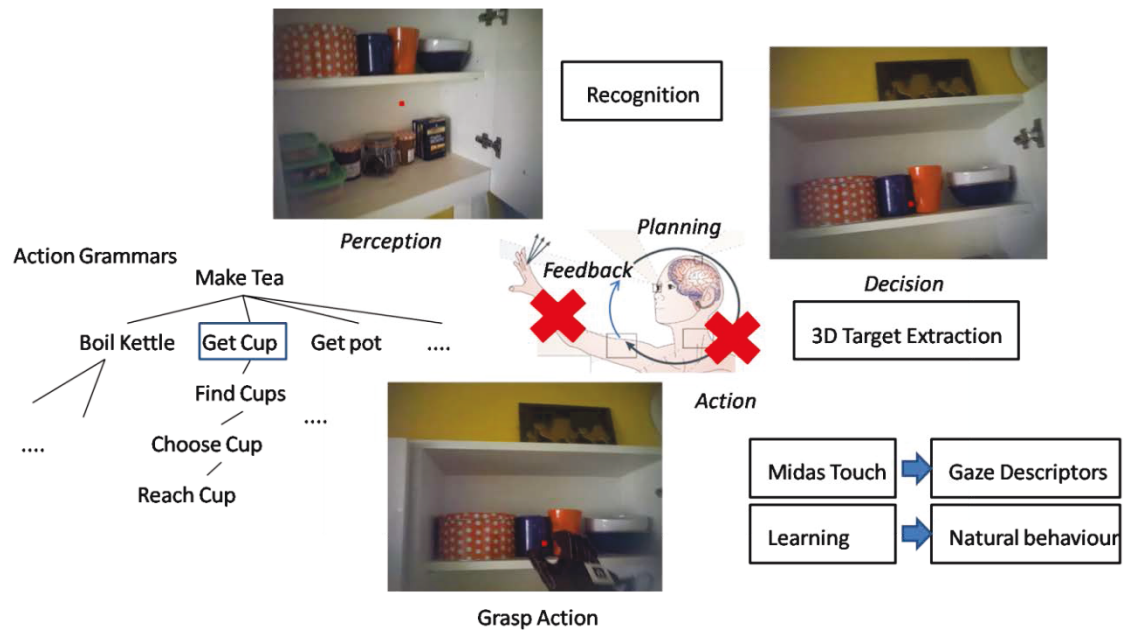


Figure 87: Perception action loop and envisioned components for intention decoding.

8.1.1 Gaze estimation beyond the screen

Vitally, the gaze control input is developed to be in the context of the real world rather than the traditional use in computer control. As the visual system leverages disparity for 3D target extraction by fixating both eyes on the location, the developed GT3D system utilises this process to estimate 3D gaze location in the real world (Chapter 2). Thus rather than communicating a location of intention with a series of movement commands as would be necessary with a joy-stick, in the flick of an eye this high level direct location command can be communicated. Presented in the context of Brain Machine Interfaces, the system has a theoretical information throughput of 43 bits/second, an order of magnitude higher than other non-invasive and invasive brain machine interfaces (BMI). This engagement with the BMI literature highlighted another key benefit of using the eyes as an input modality - the low signal latency. To refer back to Figure 86, this adds the vital link of feedback to close the perception action loop. The world is dynamic and thus commands must be updated even as they are actuated. Thus a new continuous control benchmark is proposed: to play the video game Pong. While this simple computer tennis game seems trivial in comparison to prosthetic control, the fundamental requirement for players to communicate movement intention in real time demonstrates the closing of the perception action loop.

8.1.2 Natural gaze for instant learning

The question of how easily naive subjects could learn to control an actuator with their eyes outside of the confines of the lab is asked. The gaze control interface is benchmarked with Pong in a large field experiment involving over two thousand people (Chapter 3). The experiment found that within 30 seconds most subjects were able to play Pong and in some cases even win. The approach achieved considerably lower set-up times, steeper learning rates and information throughput performance compared to other large scale

benchmarks of this nature. In addition, behavioural analysis showed that controlling the game with eye-movements leveraged similar strategies reported in literature during natural behaviour in ball games. This suggested perhaps the reason subjects could play the game almost instantly was because it didn't require any new behaviour to be learnt. Just by making normal eye-movements, as if manually controlling the game, the control information was present.

8.1.3 Drive and grasp by eye

Wheelchair driving and arm control platforms begin to test this natural approach in Chapter 5. There are two key innovations that make significant steps towards the feasibility of everyday use of gaze for control in the real-world. The first is to use gaze estimations directly in the environment rather than controlling devices via a graphical user interface. To enable this, 3D gaze estimation was achieved in Chapter 3 (Abbott and Faisal, 2012). The second important step was to overcome the challenge of gathering sufficient 3D calibration data to give a more reliable 3D gaze estimate. Thus in Chapter 5, the robot arm set-up enables a novel 3D calibration method to be developed, gathering 600 calibration points per minute (compared to 16 in 30s with conventional screen calibration). This facilitates the more data intensive Gaussian Processes (GP) regression to calibrate 3D gaze, boosting performance by 70%. The method can estimate gaze within <2cm Euclidean error. With these developments, direct 3D robotic arm control is achieved for the first time in Chapter 5. To achieve this, following calibration, users simply dwell on an object they wish to reach for and wink to operate the gripper. For driving wheelchairs, users look where they want to go and the wheelchair moves there. The wheelchair trajectory is updated smoothly as the eyes respond to the wheelchair movement - hence the perception action loop is complete. Preliminary experiments shown in Chapter 5 demonstrate that the impact on natural eye-movement behaviour is much greater when using a graphical user interface rather than our natural approach of looking where you want to go (Ktena et al., 2015).

Further work is necessary to fully test the systems in a full user study. As with the Pong benchmark, these interfaces are built as experimental platforms, recording behavioural and performance metrics for benchmarking and analysing strategies. Preliminary experiments have begun to link natural gaze behaviour with joystick control, however insufficient data has been collected. Thus a very interesting future study will be to make long term recordings of natural eye-movements as the wheelchair is used with a joystick input. This will provide a rich dataset for decoding intentions (joystick commands) from natural eye-movements. This is important to improve solutions to the Midas touch problem (distinguishing control intentions from other gaze behaviours). This is also required for the robot arm where currently only two input commands are possible - to move the robot arm end-point and to open and close the gripper. Although the current system has limitations, it is a very promising start because it shows the potential for the approach. Even in this current stage of development, the level of control is well beyond other non-invasive interfaces and even many of the invasive approaches. Extending the control approach to match the requirements of a fully actuated hand is the challenge for future work. This is a justification for the large behavioural experiment presented in

Chapter 6 - towards a quantitative understanding of what other predictive information is available for prosthetic control.

8.1.4 Natural behaviour: embodied gaze methodology

Natural behavioural is captured rather than constrained in the novel methodology used to embody eye-movement recordings in the wild (Chapter 6). It follows on from pioneering real-world experiments by authors such as Michael Land, Dana Ballard and Mary Hayhoe over the last 2 decades (for extensive review see (Land and Tatler, 2009)). Vially, unlike previous methods, the approach is sufficiently scalable to collect the volumes of data necessary to combat the curse of dimensionality of free behaviour. That is to say, making meaningful deductions from such high dimensional data requires very large quantities of data. Previous studies which provided important insights into eye and full body behaviour in natural tasks did so through manual annotation of the video data, thus limiting recordings to a few minutes. Here we build on this by automatically capturing sensory inputs and motor outputs in natural behaviour. Visual and auditory inputs are recorded using a head mounted eye-tracker, scene camera and microphone. Simultaneously, recording musculoskeletal motor outputs by motion tracking 51 degrees of freedom in the body and a total of 40 degrees of freedom in the hands. All tracking equipment is marker-less, thus allowing unconstrained behavioural monitoring “in the wild”. Consequentially analysis can be performed with the click of a button rather than hundreds of man hours of painstaking manual annotation. To achieve this many technical and algorithmic challenges are met to measure, synchronise and analyse the large volumes of data required to capture the full repertoire of human behaviour. The result is a unique and fascinating database, with implications well beyond the scope of this thesis. The interpretation of gaze and thus attention is integrated, not disembodied, from the kinematics of motor behaviour.

Collecting data in this way requires new protocols for analysis and care must be taken when applying techniques developed in the lab. The challenge to understand the perception action loop requires the raw gaze and body data to be related. Three main aspects of this data-driven approach are investigated here: 1) Existing gaze event (saccade fixation, blink and more?) detection methods with new insights in the wild (Chapter 6), 2) Extending gaze event detection to the novel Embodied Gaze Descriptors (Chapter7) and 3) The novel concept of Embodied Saliency (Chapter 7).

8.1.5 Gaze behaviour in the wild

Standard aggregated statistics of gaze behaviour are presented and compared to previous pioneering studies where everything was annotated from video by hand. This highlights a number of differences, caveats, surprising results and of course a long list of future problems to be solved. Gaze event detection helps to interpret raw gaze data into meaningful events of fixations, saccades blinks and potentially other events. The characteristics of these events is often used to distinguish the nature of the visual task, for example fixations are much shorter during sedentary tasks such as reading or picture viewing compared to active tasks such as making a cup of tea (M. Land, Mennie, & Rusted, 1999). However, it is considerably more challenging to classify events in unconstrained free behaviour than in head-fixed desk based experiments. The eye-movement repertoire is no longer sufficiently categorised by traditional fixation/saccade detection - with at

least an additional smooth eye movement category and its sub classes. In addition mobile eye-tracking equipment records lower quality data to base this categorisation on (spatial and temporal resolution). Two existing methods are compared to demonstrate how much the algorithm used affects the results. This is discussed at length, however the main conclusions can be summarised as: Firstly if traditional methods are used it is important to present the resulting event characteristics so it is clear what behaviours they capture, with key metrics including fixation and saccade amplitudes, fixation direction variability and saccade speed. Secondly, the question being asked affects the choice of algorithm. For example the distinction in gaze behaviour between navigation and manipulation task was more significant as quantified by the dispersion based method (I-DT) as it could classify pure fixations. While the velocity based method (SMI) correctly identified visual intake by combining smooth and fixation eye-movements. Thus if the question asked is what periods of the experiment is information sampled from the scene, this would be the preferred metric. The final conclusion is therefore new algorithms are required to enhance the descriptive nature of event detection and some of the promising approaches in development are discussed.

One other aspect of event detection that was explored in more depth than previous studies in the wild is the often overlooked blink. Considering this can occupy 20-30% of our waking life, and that 60% of blinks have eye movements larger than 10 degrees associated with them, or put another way 50% of large eye-movements (>10 degrees) are discarded if gaze evoked blinks are not considered. These gaze evoked blinks - confirmed in laboratory studies (Evinger et al., 1994), are clearly an important component of behaviour. They are essentially hidden saccades and thus represent a large hole in saccadic behaviour if they are discarded. Just as the characteristics of saccades and fixations are linked with cognitive processes, there is more information associated with blinks than just their frequency. A discussion is made regarding their implications for information retrieval and therefore ongoing cognitive processes. The important conclusion in the context of intention decoding is that they provide an additional informative descriptor for interpreting gaze behaviour.

8.1.6 Embodied gaze descriptors for prosthetic control

In the wild eye-movements are accompanied by a huge repertoire of body movements. The taxonomy of fixations saccade and blink becomes an impoverished description of gaze behaviour when the body is free to move. Thus to understand the visuomotor repertoire eye-movements are taxonomised by their association with body movements. The scope of such an approach is expansive with such a comprehensive dataset. Firstly, grasp movements are explored in Chapter 7 because of the obvious links with prosthetic control and because these relationships have never been recorded in the wild. Thus saccades are clustered according to the associated hand movements. Interestingly grasp like movements emerge for 42% of saccades. When these clusters are associated with arm, trunk and head movements there is a clear relationship. Synchronous saccade, body turn head turn, arm reach and then grasp movements are observed. Thus seconds before the grasp is made there are predictive cues in the eye, head, trunk and arm which could be used to initiate a grasp intention. This is very promising for the context of trans-radial amputees (below elbow), where all of this information is available to control a hand

prosthesis. Currently the grasp association is made with an aggregated hand joint velocity but this can be extended to separate digits to associate different types of grasp shaping (eg, precision index-thumb pinch grip, whole hand grasp etc).

8.1.7 Embodied Saliency

The long term aim is to predict intended body movements from eye-movements, thus as an intermediate step the counter question is asked: can body movements predict eye-movements? Bottom up visual saliency and top down task demands have shown to have predictive power in various contexts (for extensive reviews see (Schütz, Braun, & Gegenfurtner, 2011; Benjamin W. Tatler, Hayhoe, Land, & Ballard, 2011). However, if the brain is thought of as a dynamical system mapping sensory inputs to motor actions surely the body posture itself holds predictive power. The focus tends to be on visual sensory information but eye movements are also affected by body proprioception (joint positions) and indeed the planned future postures i.e. intentions. Thus with a simple linear model mapping purely from the body (51 degrees of freedom) to 2 dimensional eye-movements it is shown in Chapter 7 that body movements alone predict eye-movements within 7.1 degrees average absolute error, significantly above chance and central fixation bias. While the majority of this predictive power comes from head movements, running the model with all body joints except the head still has a performance that is significantly above chance and central fixation bias. Thus the data driven approach allows the new concept of embodied saliency to be demonstrated.

8.1.8 Future perspectives

Experimental context has a huge impact on the shaping of emerging theories. Static lab studies observing eye-movements in response to purely visual stimuli led to a wealth of literature on bottom-up visual saliency. Now, with the capabilities to capture the full perception action cycle in natural behaviour, a new field of Embodied Saliency can emerge. As the taxonomy of natural behaviour is furthered, its association with task and cognitive processes will inform scientific understanding of ecologically valid behaviour where the brain simultaneously controls many more degrees of freedom than in the lab. The main control loops that have been demonstrated to be involved in the allocation of overt attention (eye-movements) include motor plans, a value function that is tailored to the ongoing task, object recognition and Visual Saliency (Schütz et al., 2011), and now here Embodied Saliency. Such an integration of top down and bottom up signals has been proposed to occur at higher levels in the visual hierarchy in so called priority maps for which neurophysiological studies are beginning to provide evidence (Schütz et al., 2011). Furthermore distinguishing eye-movement characteristics have been associated with the ventral and dorsal visual streams in lab based experiments (Unema et al., 2005). With the Embodied approach presented here the interaction of all these different sub components of eye-movement control can be modelled for natural behaviour. The collected database includes visual, vestibular, auditory, proprioceptive and future body configuration (motor plans). The object identities and ongoing tasks can be extracted from the scene video using applications from EgoVision fields. Thus a more complete model can be derived that fully embodies the perception action loop by capturing, not constraining, natural behaviour.

With this rich taxonomy will also arise a richer descriptor set for predicting intentions. With the semantic labelling of objects and actions from egovision to train action grammars (likely sequences of actions in natural behaviour - see Figure 87) to provide a prior on the solution space for intention decoding. The hope is that this wealth of information, combined with embodied gaze descriptors, will be sufficient to train intention decoding systems for the intelligent, intuitive and non-invasive control of prosthetics. The future vision is for embodied prosthetic control frameworks, customised to the individual patient based on their individual abilities and needs. Initially, specific disabilities can be simulated with the dataset by removing data channel configurations to represent different levels of amputation (similarly to (Thomik et al., 2013)). Attempts to predict missing dimensions from the retained behavioural data can then be applied. If successful, the future vision would be a data driven model customised to each individual based on the natural behaviour database and the details of their particular movement deficits. Thus by capturing rather than constraining natural behaviour, it is hoped that future biomimetic prosthetics, principled on the full perception action loop, can liberate patients from the constraints of disability.

9 Bibliography

Abbott, W.W., and Faisal, A.A. (2012). Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain–machine interfaces. *J. Neural Eng.* *9*, 046016.

Abbott, W.W., Zucconi, A., and Faisal, A.A. (2013). Large-field study of ultra low-cost, non-invasive task level bmi. In *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, (IEEE), pp. 97–100.

Abbott, W.W., Thomik, A.A., and Faisal, A.A. (2015). Embodied salience for gaze analysis in ecologically valid environments. *J. Vis.* *15*, 366.

Açık, A., Sarwary, A., Schultze-Kraft, R., Onat, S., and König, P. (2010). Developmental Changes in Natural Viewing Behavior: Bottom-Up and Top-Down Differences between Children, Young Adults and Older Adults. *Front. Psychol.* *1*, 207.

Aizawa, K., Ishijima, K., and Shiina, M. (2001). Summarizing wearable video. In *2001 International Conference on Image Processing, 2001. Proceedings*, pp. 398–401 vol.3.

Allison, B., Luth, T., Valbuena, D., Teymourian, A., Volosyak, I., and Graser, A. (2010). BCI Demographics: How Many (and What Kinds of) People Can Use an SSVEP BCI? *IEEE Trans. Neural Syst. Rehabil. Eng.* *18*, 107–116.

Allopenna, P.D., Magnuson, J.S., and Tanenhaus, M.K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *J. Mem. Lang.* *38*, 419–439.

Anderson, M.L. (2003). Embodied Cognition: A field guide. *Artif. Intell.* *149*, 91–130.

Antfolk, C., D’Alonzo, M., Rosén, B., Lundborg, G., Sebelius, F., and Cipriani, C. (2013). Sensory feedback in upper limb prosthetics. *Expert Rev. Med. Devices* *10*, 45–54.

Arai, K., and Mardiyanto, R. (2011). A Prototype of Electric Wheelchair Controlled by Eye-Only for Paralyzed User. *J. Robot. Mechatron.* *23*.

- Awh, E., Belopolsky, A.V., and Theeuwes, J. (2012). Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends Cogn. Sci.* 16, 437–443.
- Babiloni, F., Cincotti, F., Marciani, M., Salinari, S., Astolfi, L., Tocci, A., Aloise, F., Fallani, F.D.V., Bufalari, S., and Mattia, D. (2007). The Estimation of Cortical Activity for Brain-Computer Interface: Applications in a Domotic Context. *Comput. Intell. Neurosci.* 2007, 1–7.
- Baldauf, D., and Deubel, H. (2010). Attentional landscapes in reaching and grasping. *Vision Res.* 50, 999–1013.
- Ballard, D.H., Hayhoe, M.M., Li, F., Whitehead, S.D., Frisby, J.P., Taylor, J.G., and Fisher, R.B. (1992). Hand-eye coordination during sequential tasks. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 337, 331–339.
- Ballard, D.H., Hayhoe, M.M., and Pelz, J.B. (1995). Memory Representations in Natural Tasks. *J. Cogn. Neurosci.* 7, 66–80.
- Barea, R., Boquete, L., Bergasa, L.M., López, E., and Mazo, M. (2003). Electro-Oculographic Guidance of a Wheelchair Using Eye Movements Codification. *Int. J. Robot. Res.* 22, 641–652.
- Barr, C.C., Schultheis, L.W., and Robinson, D.A. (1976). Voluntary, non-visual control of the human vestibulo-ocular reflex. *Acta Otolaryngol. (Stockh.)* 81, 365–375.
- Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* 91, 276–292.
- Beatty, J., and Lucero-Wagoner, B. (2000). The pupillary system. In *Handbook of Psychophysiology* (2nd Ed.), (New York, NY, US: Cambridge University Press), pp. 142–162.
- Bellman, R. (1960). Adaptive control processes: A guided tour.
- Berg, D.J., Boehnke, S.E., Marino, R.A., Munoz, D.P., and Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *J. Vis.* 9, 19.
- Betz, T., Kietzmann, T.C., Wilming, N., and König, P. (2010). Investigating task-dependent top-down effects on overt visual attention. *J. Vis.* 10.
- Blain-Moraes, S., Schaff, R., Gruis, K.L., Huggins, J.E., and Wren, P.A. (2012). Barriers to and mediators of brain–computer interface user acceptance: focus group findings. *Ergonomics* 55, 516–525.
- Blankertz, B., Dornhege, G., Krauledat, M., Schröder, M., Williamson, J., Murray-Smith, R., and Müller, K.-R. (2006). The Berlin Brain-Computer Interface presents the novel mental typewriter Hex-o-Spell.
- Blascheck, T., Kurzhals, K., Raschke, M., Burch, M., Weiskopf, D., and Ertl, T. (2014). State-of-the-art of visualization for eye tracking data. In *Proceedings of EuroVis*, p.
- Bobrov, P., Frolov, A., Cantor, C., Fedulova, I., Bakhnyan, M., and Zhavoronkov, A. (2011). Brain-Computer Interface Based on Generation of Visual Images. *PLoS ONE* 6, e20674.
- Brabender, V. (1977). Cognition and Reality. *Thought* 52, 457–460.

- Brennan, S.E., Chen, X., Dickinson, C.A., Neider, M.B., and Zelinsky, G.J. (2008). Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition* 106, 1465–1477.
- Bridgeman, B., Hendry, D., and Stark, L. (1975). Failure to detect displacement of the visual world during saccadic eye movements. *Vision Res.* 15, 719–722.
- Brolly, X.L.C., and Mulligan, J.B. (2004). Implicit Calibration of a Remote Gaze Tracker. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 8 - Volume 08*, (Washington, DC, USA: IEEE Computer Society), p. 134–.
- Brouwer, A.-M., Franz, V.H., and Gegenfurtner, K.R. (2009a). Differences in Fixations Between Grasping and Viewing Objects. *J. Vis.* 9.
- Brouwer, A.M., Franz, V.H., and Gegenfurtner, K.R. (2009b). Differences in fixations between grasping and viewing objects. *J. Vis.* 9, 18–18.
- Brunner, C., Blankertz, B., Cincotti, F., Kübler, A., Mattia, D., Miralles, F., Nijholt, A., Otal, B., Salomon, P., and Müller-Putz, G.R. (2014). BNCI horizon 2020—towards a roadmap for brain/neural computer interaction. In *Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access*, (Springer), pp. 475–486.
- Brunner, P., Joshi, S., Briskin, S., Wolpaw, J.R., Bischof, H., and Schalk, G. (2010). Does the “P300” speller depend on eye gaze? *J. Neural Eng.* 7, 056013.
- Buckley, M., Vaidyanathan, R., and Mayol-Cuevas, W. (2011). Sensor suites for assistive arm prosthetics. In *Computer-Based Medical Systems (CBMS), 2011 24th International Symposium on*, pp. 1–6.
- Buswell, G.T. (1935). How people look at pictures: a study of the psychology and perception in art.
- Caligiore, D., and Fischer, M.H. (2012). Vision, action and language unified through embodiment. *Psychol. Res.* 77, 1–6.
- Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., and Rust, N.C. (2005). Do we know what the early visual system does? *J. Neurosci.* 25, 10577–10597.
- Card, S.K., English, W.K., and Burr, B.J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. *Ergonomics* 21, 601–613.
- Carpenter, R.H.S. (1988). *Movements of the eyes* (Pion).
- Chen, S., and Epps, J. (2013). Blinking: Toward Wearable Computing that Understands your Current Task. *IEEE Pervasive Comput.* 12, 56–65.
- Collewijn, H., Steinman, R.M., Erkelens, C.J., Pizlo, Z., and Van der Steen, J. (1990). The effect of freeing the head on eye movement characteristics during 3-D shifts of gaze and tracking. *Head-Neck Sens. Mot. Syst.* 412–418.

- Collinger, J.L., Wodlinger, B., Downey, J.E., Wang, W., Tyler-Kabara, E.C., Weber, D.J., McMorland, A.J., Velliste, M., Boninger, M.L., and Schwartz, A.B. (2013). High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet* **381**, 557–564.
- Corbett, E.A., Kording, K.P., and Perreault, E.J. (2012). Real-time fusion of gaze and EMG for a reaching neuroprosthesis. *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Conf. 2012*, 739–742.
- Corbett, E.A., Kording, K.P., and Perreault, E.J. (2013). Real-Time Evaluation of a Noninvasive Neuroprosthetic Interface for Control of Reach. *IEEE Trans. Neural Syst. Rehabil. Eng. Early Access Online*.
- Dael, N., Mortillaro, M., and Scherer, K.R. (2012). The Body Action and Posture Coding System (BAP): Development and Reliability. *J. Nonverbal Behav.* **36**, 97–121.
- Daunys, G., and Ramanauskas, N. (2006). Nonlinear Mapping of Pupil Centre Coordinates from Image Sensor to Screen for Gaze Control Systems. In *Computers Helping People with Special Needs*, K. Miesenberger, J. Klaus, W.L. Zagler, and A.I. Karshmer, eds. (Springer Berlin Heidelberg), pp. 962–965.
- Diaz, G., Cooper, J., Rothkopf, C., and Hayhoe, M. (2013). Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. *J. Vis.* **13**.
- Diefendorf, A.R., and Dodge, R. (1908). An experimental study of the ocular reactions of the insane from photographic records. *Brain* **31**, 451–489.
- Dionisio, D.P., Granholm, E., Hillix, W.A., and Perrine, W.F. (2001). Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology* **38**, 205–211.
- Donkelaar, P. van, Siu, K.-C., and Walterschied, J. (2004). Saccadic output is influenced by limb kinetics during eye—hand coordination. *J. Mot. Behav.* **36**, 245–252.
- Drewes, H., and Schmidt, A. (2007). Interacting with the computer using gaze gestures. In *Proceedings of the 11th IFIP TC 13 International Conference on Human-Computer Interaction-Volume Part II*, pp. 475–488.
- Droll, J.A., Hayhoe, M.M., Triesch, J., and Sullivan, B.T. (2005). Task Demands Control Acquisition and Storage of Visual Information. *J. Exp. Psychol. Hum. Percept. Perform.* **31**, 1416–1438.
- Duchowski, A.T. (2007). *Eye Tracking Methodology: Theory and Practice* (Springer).
- Dufour, A., and Touzalin, P. (2008). Improved visual sensitivity in the perihand space. *Exp. Brain Res.* **190**, 91–98.
- Dziemian, S., Abbott, W.W., and Faisal, A.A. (2016). Gaze-based teleprosthetic enables intuitive continuous control of complex robot arm use: writing & drawing. *Proc. 6th IEEE RASEMBS Int. Conf. Biomed. Robot. Biomechatronics BioRob2016 June 26-29 2016 Singap.*
- Ebenholtz, S.M. (2001). *Oculomotor Systems and Perception* (Cambridge University Press).

- Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C., and Tanenhaus, M.K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *J. Psycholinguist. Res.* **24**, 409–436.
- Eileen, K. (2011). Eye movements: The past 25 years. *Vision Res.* **51**, 1457–1483.
- Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., and König, P. (2007). Human eye-head co-ordination in natural exploration. *Netw. Comput. Neural Syst.* **18**, 267–297.
- Einhäuser, W., Rutishauser, U., and Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *J. Vis.* **8**, 2–2.
- Epelboim, J. (1998). Gaze and retinal-image-stability in two kinds of sequential looking tasks. *Vision Res.* **38**, 3773–3784.
- Epelboim, J., Steinman, R.M., Kowler, E., Edwards, M., Pizlo, Z., Erkelens, C.J., and Collewyn, H. (1995). The function of visual search and memory in sequential looking tasks. *Vision Res.* **35**, 3401–3422.
- Epelboim, J., Steinman, R.M., Kowler, E., Pizlo, Z., Erkelens, C.J., and Collewyn, H. (1997). Gaze-shift dynamics in two kinds of sequential looking tasks. *Vision Res.* **37**, 2597–2607.
- Evinger, C., Manning, K.A., Pellegrini, J.J., Basso, M.A., Powers, A.S., and Sibony, P.A. (1994). Not looking while leaping: the linkage of blinking and saccadic gaze shifts. *Exp. Brain Res.* **100**, 337–344.
- Faisal, A.A. (2009). Stochastic Simulation of Neurons, Axons, and Action Potentials. In *Stochastic Methods in Neuroscience*, C. Laing, and G.J. Lord, eds. (Oxford University Press), pp. 297–343.
- Faisal, A.A., Laughlin, S.B., and White, J.A. (2002). How reliable is the connectivity in cortical neural networks? In *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on*, pp. 1661–1666.
- Faisal, A.A., White, J.A., and Laughlin, S.B. (2005). Ion-Channel Noise Places Limits on the Miniaturization of the Brain's Wiring. *Curr. Biol.* **15**, 1143–1149.
- Fathi, A., Li, Y., and Rehg, J.M. (2012). Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012*, (Springer), pp. 314–327.
- Fenn, W.O., and Hursh, J.B. (1936). Movements of the Eyes When the Lids Are Closed. *Am. J. Physiol. -- Leg. Content* **118**, 8–14.
- Foerster, R.M., Carbone, E., Koesling, H., and Schneider, W.X. (2012). Saccadic Eye Movements in the Dark While Performing an Automatized Sequential High-Speed Sensorimotor Task. *J. Vis.* **12**.
- Follet, B., Le Meur, O., and Baccino, T. (2011). New insights into ambient and focal visual fixations using an automatic classification algorithm. *-Percept.* **2**, 592–610.

Foulsham, T., and Underwood, G. (2008). What Can Saliency Models Predict About Eye Movements? Spatial and Sequential Aspects of Fixations During Encoding and Recognition. *J. Vis.* 8.

Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Res.* 51, 1920–1931.

Freedman, E.G., and Sparks, D.L. (1997). Eye-Head Coordination During Head-Unrestrained Gaze Shifts in Rhesus Monkeys. *J. Neurophysiol.* 77, 2328–2348.

Friedburg, C., Allen, C.P., Mason, P.J., and Lamb, T.D. (2004). Contribution of cone photoreceptors and post-receptoral mechanisms to the human photopic electroretinogram. *J. Physiol.* 556, 819–834.

Frisoli, A., Loconsole, C., Leonardis, D., Banno, F., Barsotti, M., Chisari, C., and Bergamasco, M. (2012). A New Gaze-BCI-Driven Control of an Upper Limb Exoskeleton for Rehabilitation in Real-World Tasks. *Syst. Man Cybern. Part C Appl. Rev. IEEE Trans. On* 42, 1169–1179.

Gandhi, N.J. (2011). Interactions between gaze-evoked blinks and gaze shifts in monkeys. *Exp. Brain Res.* 216, 321–339.

Ganguly, K., and Carmena, J.M. (2009). Emergence of a Stable Cortical Map for Neuroprosthetic Control. *PLoS Biol* 7, e1000153.

Gilja, V., Nuyujukian, P., Chestek, C.A., Cunningham, J.P., Yu, B.M., Fan, J.M., Churchland, M.M., Kaufman, M.T., Kao, J.C., Ryu, S.I., et al. (2012). A high-performance neural prosthesis enabled by control algorithm design. *Nat. Neurosci.* 15, 1752–1757.

Gips, J. (1997). Preliminary investigation of a semi-autonomous robotic wheelchair directed through electrodes. In *Proc. Rehabilitation Engineering Society of North America 1997 Annual Conference*, pp. 414–416.

Goebel, R., Sorger, B., Kaiser, J., Birbaumer, N., and Weiskopf, N. (2004). BOLD brain pong: Self regulation of local brain activity during synchronously scanned, interacting subjects. In *34th Annual Meeting of the Society for Neuroscience*, p.

Goodale, M.A., and Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25.

Gover, M.R. (1996). *The Embodied Mind: Cognitive Science and Human Experience* (Book). *Mind Cult. Act.* 3, 295–299.

Guger, C., Edlinger, G., Harkam, W., Niedermayer, I., and Pfurtscheller, G. (2003). How many people are able to operate an EEG-based brain-computer interface (BCI)? *IEEE Trans. Neural Syst. Rehabil. Eng.* 11, 145–147.

Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., Gramatica, F., and Edlinger, G. (2009). How many people are able to control a P300-based brain-computer interface (BCI)? *Neurosci. Lett.* 462, 94–98.

Hart, B.M. 't, Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P., and Einhäuser, W. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Vis. Cogn.* 17, 1132–1158.

- 't Hart, B.M., and Einhäuser, W. (2012). Mind the step: complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Exp. Brain Res.* 223, 233–249.
- Hayhoe, M., and Ballard, D. (2005). Eye movements in natural behavior. *Trends Cogn. Sci.* 9, 188–194.
- Hayhoe, M., Mennie, N., Sullivan, B., and Gorgos, K. (2005). The role of internal models and prediction in catching balls. In *Proceedings of the American Association for Artificial Intelligence*, p.
- Hayhoe, M.M., Bensinger, D.G., and Ballard, D.H. (1998). Task constraints in visual working memory. *Vision Res.* 38, 125–137.
- Hayhoe, M.M., Shrivastava, A., Mruczek, R., and Pelz, J.B. (2003). Visual memory and motor planning in a natural task. *J. Vis.* 3.
- Henderson, J.M., Brockmole, J.R., Castelano, M.S., and Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye Mov. Window Mind Brain* 537–562.
- Hennessey, C., and Lawrence, P. (2009a). Noncontact Binocular Eye-Gaze Tracking for Point-of-Gaze Estimation in Three Dimensions. *IEEE Trans. Biomed. Eng.* 56, 790–799.
- Hennessey, C.A., and Lawrence, P.D. (2009b). Improving the Accuracy and Reliability of Remote System-Calibration-Free Eye-Gaze Tracking. *IEEE Trans. Biomed. Eng.* 56, 1891–1900.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Weijer, J. van de (2011). *Eye Tracking: A comprehensive guide to methods and measures* (OUP Oxford).
- Hubbard, E.M., Piazza, M., Pinel, P., and Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nat. Rev. Neurosci.* 6, 435–448.
- Huey, E.B. (1898). Preliminary experiments in the physiology and psychology of reading. *Am. J. Psychol.* 9, 575–586.
- Huey, E.B. (1900). On the psychology and physiology of reading. I. *Am. J. Psychol.* 11, 283–302.
- Huggins, J.E., Wren, P.A., and Gruis, K.L. (2011). What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Off. Publ. World Fed. Neurol. Res. Group Mot. Neuron Dis.* 12, 318–324.
- Humphrey, N.K. (1976). The social function of intellect. *Grow. Points Ethol.* 303–317.
- Hyrskykari, A., Istance, H., and Vickers, S. (2012). Gaze gestures or dwell-based interaction? In *Proceedings of the Symposium on Eye Tracking Research and Applications*, (New York, NY, USA: ACM), pp. 229–232.
- Inhoff, A., and Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Atten. Percept. Psychophys.* 40, 431–439.

- Ishimaru, S., Kunze, K., Kise, K., Weppner, J., Dengel, A., Lukowicz, P., and Bulling, A. (2014). In the Blink of an Eye: Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass. In *Proceedings of the 5th Augmented Human International Conference*, (New York, NY, USA: ACM), p. 15:1–15:4.
- Ismail, E., Ghazali, A.S., and Sidek, S.N. (2013). Physiological Signal — Based Engagement Level Analysis under Fuzzy Framework. *Appl. Mech. Mater.* 373–375, 1768–1775.
- Istance, H., Bates, R., Hyrskykari, A., and Vickers, S. (2008). Snap clutch, a moded approach to solving the Midas touch problem. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*, pp. 221–228.
- Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40, 1489–1506.
- Itti, L., and Koch, C. (2001). Computational modeling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203.
- Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259.
- Jacob, R.J.K. (1990). What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Empowering People*, (New York, NY, USA: ACM), pp. 11–18.
- Johansson, R.S., Westling, G., Bäckström, A., and Flanagan, J.R. (2001). Eye-hand coordination in object manipulation. *J. Neurosci. Off. J. Soc. Neurosci.* 21, 6917–6932.
- Jordansen, I., Boedeker, S., Donegan, M., Oosthuizen, L., Girolamo, M., and Hansen, J.P. (2005). Report on a market study and demographics of user population (cogain) ist-2003-511598: Deliverable 7.2).
- Judd, C., Mcallister, C., Cloyd, N., and Steele, W. (1905). General introduction to a series of studies of eye movements by means of kinesiographic photographs. *Psychol. Monogr.*
- Kamavuako, E.N., Scheme, E.J., and Englehart, K.B. (2014). On the usability of intramuscular EMG for prosthetic control: a Fitts' Law approach. *J. Electromyogr. Kinesiol. Off. J. Int. Soc. Electrophysiol. Kinesiol.* 24, 770–777.
- Kaminski, H.J., Al-Hakim, M., Leigh, R.J., Bashar, M.K., and Ruff, R.L. (1992). Extraocular muscles are spared in advanced duchenne dystrophy. *Ann. Neurol.* 32, 586–588.
- Kaminski, H.J., Richmonds, C.R., Kusner, L.L., and Mitsumoto, H. (2002). Differential Susceptibility of the Ocular Motor System to Disease. *Ann. N. Y. Acad. Sci.* 956, 42–54.
- Kasneci, E., Kasneci, G., Kübler, T.C., and Rosenstiel, W. (2014). The Applicability of Probabilistic Methods to the Online Recognition of Fixations and Saccades in Dynamic Scenes. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, (New York, NY, USA: ACM), pp. 323–326.
- Kim, D.-J., Hazlett-Knudsen, R., Culver-Godfrey, H., Rucks, G., Cunningham, T., Portee, D., Bricout, J., Wang, Z., and Behal, A. (2012). How Autonomy Impacts Performance and

Satisfaction: Results From a Study With Spinal Cord Injured Subjects Using an Assistive Robot. Syst. Man Cybern. Part Syst. Hum. IEEE Trans. On 42, 2–14.

Kim, J., Park, H., Bruce, J., Sutton, E., Rowles, D., Pucci, D., Holbrook, J., Minocha, J., Nardone, B., West, D., et al. (2013). The Tongue Enables Computer and Wheelchair Control for People with Spinal Cord Injury. Sci. Transl. Med. 5, 213ra166-213ra166.

Kim, S.-P., Simeral, J.D., Hochberg, L.R., Donoghue, J.P., and Black, M.J. (2008). Neural control of computer cursor velocity by decoding motor cortical spiking activity in humans with tetraplegia. J. Neural Eng. 5, 455–476.

Kingstone, A., Smilek, D., and Eastwood, J.D. (2008). Cognitive Ethology: A new approach for studying human cognition. Br. J. Psychol. 99, 317–340.

Kit, D., Katz, L., Sullivan, B., Snyder, K., Ballard, D., and Hayhoe, M. (2014). Eye Movements, Visual Search and Scene Memory, in an Immersive Virtual Environment. PLOS ONE 9, e94362.

Kiverstein, J., and Miller, M. (2015). The embodied brain: towards a radical embodied cognitive neuroscience. Front. Hum. Neurosci. 9.

Knapp, M.L., Hall, J.A., and Horgan, T.G. (2013). Nonverbal communication in human interaction (Cengage Learning).

Knops, A., Thirion, B., Hubbard, E.M., Michel, V., and Dehaene, S. (2009). Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. Science 324, 1583–1585.

Koch, C., and Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. Hum. Neurobiol. 4, 219–227.

Komogortsev, O.V., Jayarathna, S., Koh, D.H., and Gowda, S.M. (2010). Qualitative and quantitative scoring and evaluation of the eye movement classification algorithms. In Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, (New York, NY, USA: ACM), pp. 65–68.

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., and Kopacz, A. (2016). Discerning Ambient/Focal Attention with Coefficient K. ACM Trans. Appl. Percept. 13, 1–20.

Krepki, R. (2004). Brain-Computer Interfaces. Universitätsbibliothek.

Ktena, S.I., Abbott, W.W., and Faisal, A.A. (2015). A virtual reality platform for safe evaluation and training of natural gaze-based wheelchair driving. In 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 236–239.

Kuiken TA, Li G, Lock BA, and et al (2009). Targeted muscle reinnervation for real-time myoelectric control of multifunction artificial arms. JAMA 301, 619–628.

Lamare, M. (1892). Des mouvements des yeux dans la lecture. Bull. Mém. Soc. Fr.-Aise D'Ophthalmologie 10, 354–364.

Land, M.F. (1993). Eye-head coordination during driving. In Proceedings of IEEE Systems Man and Cybernetics Conference - SMC, pp. 490–494 vol.3.

- Land, M.F. (2004). The coordination of rotations of the eyes, head and trunk in saccadic turns produced in natural situations. *Exp. Brain Res.* 159, 151–160.
- Land, M.F., and Furneaux, S. (1997). The Knowledge Base of the Oculomotor System. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 352, 1231–1239.
- Land, M.F., and Lee, D.N. (1994). Where we look when we steer. *Nature* 369, 742–744.
- Land, M.F., and Tatler, B.W. (2009). *Looking and acting: vision and eye movements in natural behaviour* (Oxford University Press, USA).
- Land, M., Mennie, N., and Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 1311–1328.
- Larsson, L., Nyström, M., and Stridh, M. (2014). Discrimination of fixations and smooth pursuit movements in high-speed eye-tracking data. In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3797–3800.
- Lee, E.C., and Park, K.R. (2008). A robust eye gaze tracking method based on a virtual eyeball model. *Mach. Vis. Appl.* 20, 319–337.
- Leigh, R.J. (2003). Using saccades as a research tool in the clinical neurosciences. *Brain* 127, 460–477.
- Lemahieu, W., and Wyns, B. (2011). Low cost eye tracking for human-machine interfacing. *J. Eye Track. Vis. Cogn. Emot.* 0.
- Lewis, A., Garcia, R., and Zhaoping, L. (2003). The distribution of visual objects on the retina: connecting eye movements and cone distributions. *J. Vis.* 3.
- Li, D., Winfield, D., and Parkhurst, D.J. (2005). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, p. 79.
- Li, D., Babcock, J., and Parkhurst, D.J. (2006). openEyes: a low-cost head-mounted eye-tracking solution. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, (New York, NY, USA: ACM), pp. 95–100.
- Li, Y., Fathi, A., and Rehg, J. (2013). Learning to predict gaze in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3216–3223.
- Lin, C., Ho, C., Chen, W., Chiu, C., and Yeh, M. (2006). Powered wheelchair controlled by eye-tracking system. *Opt. Appl.* 36, 401.
- List, T., Bins, J., Vazquez, J., and Fisher, R.B. (2005). Performance evaluating the evaluator. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 129–136.
- Loconsole, C., Bartalucci, R., Frisoli, A., and Bergamasco, M. (2011). A new gaze-tracking guidance mode for upper limb robot-aided neurorehabilitation. In *2011 IEEE World Haptics Conference (WHC)*, (IEEE), pp. 185–190.

- Loetscher, T., Bockisch, C.J., and Brugger, P. (2008). Looking for the answer: The mind's eye in number space. *Neuroscience* 151, 725–729.
- Lünenburger, L., Kutz, D.F., and Hoffmann, K.P. (2000). Influence of arm movements on saccades in humans. *Eur. J. Neurosci.* 12, 4107–4116.
- Mace, M., Abdullah-al-Mamun, K., Naeem, A.A., Gupta, L., Wang, S., and Vaidyanathan, R. (2013). A heterogeneous framework for real-time decoding of bioacoustic signals: Applications to assistive interfaces and prosthesis control. *Expert Syst. Appl.* 40, 5049–5060.
- MacKenzie, I.S. (1989). A note on the information-theoretic basis for Fitts' law. *J. Mot. Behav.* 21, 323–330.
- Majaranta, P., and Rähä, K.-J. (2002). Twenty years of eye typing: systems and design issues. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, (New York, NY, USA: ACM), pp. 15–22.
- Malinov, I.V., Epelboim, J., Herst, A.N., and Steinman, R.M. (2000). Characteristics of saccades and vergence in two kinds of sequential looking tasks. *Vision Res.* 40, 2083–2090.
- Mann, S., Kitani, K.M., Lee, Y.J., Ryoo, M.S., and Fathi, A. (2014). An Introduction to the 3rd Workshop on Egocentric (First-Person) Vision. (IEEE), pp. 827–832.
- Martinez-Conde, S., Macknik, S.L., and Hubel, D.H. (2004). The role of fixational eye movements in visual perception. *Nat. Rev. Neurosci.* 5, 229–240.
- Matthis, J., and Hayhoe, M. (2015). Eye, head, and foot tracking during locomotion over real-world complex terrain. *J. Vis.* 15, 1322–1322.
- Maurus, M., Hammer, J.H., and Beyerer, J. (2014). Realistic Heatmap Visualization for Interactive Analysis of 3D Gaze Data. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, (New York, NY, USA: ACM), pp. 295–298.
- Mavridis, N. (2015). A Review of Verbal and Non-verbal Human-robot Interactive Communication. *Robot Auton Syst* 63, 22–35.
- McFarland, D.J., Krusienski, D.J., and Wolpaw, J.R. (2006). Brain–computer interface signal processing at the Wadsworth Center: mu and sensorimotor beta rhythms. *Prog. Brain Res.* 159, 411–419.
- McFarland, D.J., Sarnacki, W.A., and Wolpaw, J.R. (2010). Electroencephalographic (EEG) control of three-dimensional movement. *J. Neural Eng.* 7, 036007.
- McIntosh, E., Gray, A., and Aziz, T. (2003). Estimating the costs of surgical innovations: the case for subthalamic nucleus stimulation in the treatment of advanced Parkinson's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* 18, 993–999.
- McMullen, D., Hotson, G., Katyal, K., Wester, B., Fifer, M., McGee, T., Harris, A., Johannes, M., Vogelstein, R.J., Ravitz, A., et al. (2014). Demonstration of a Semi-Autonomous Hybrid Brain-Machine Interface using Human Intracranial EEG, Eye Tracking, and Computer Vision to Control a Robotic Upper Limb Prosthetic. *IEEE Trans. Neural Syst. Rehabil. Eng.* *Early Access Online*.

Meredith, M.A., and Stein, B.E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J. Neurophysiol.* 56, 640–662.

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., and Dodd, M.D. (2011). Examining the influence of task set on eye movements and fixations. *J. Vis.* 11.

Moore, E.H. (1920). On the reciprocal of the general algebraic matrix, Abstract. *Bull Amer Math Soc* 26, 394–395.

Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press).

Morimoto, C.H., and Mimica, M.R.. (2005). Eye gaze tracking techniques for interactive applications. *Comput. Vis. Image Underst.* 98, 4–24.

Munn, S.M., Stefano, L., and Pelz, J.B. (2008). Fixation-identification in Dynamic Scenes: Comparing an Automated Algorithm to Manual Coding. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, (New York, NY, USA: ACM), pp. 33–42.

Murphy, M.D., Guggenmos, D.J., Bundy, D.T., and Nudo, R.J. (2016). Current Challenges Facing the Translation of Brain Computer Interfaces from Preclinical Trials to Use in Human Patients. *Front. Cell. Neurosci.* 9.

Nakashima, R., Fang, Y., Hatori, Y., Hiratani, A., Matsumiya, K., Kuriki, I., and Shioiri, S. (2015). Saliency-based gaze prediction based on head direction. *Vision Res.* 117, 59–66.

Nguyen, T.-H.-C., Nebel, J.-C., and Florez-Revuelta, F. (2016). Recognition of Activities of Daily Living with Egocentric Vision: A Review. *Sensors* 16, 72.

Onose, G., Grozea, C., Anghelescu, A., Daia, C., Sinescu, C.J., Ciurea, A.V., Spircu, T., Mirea, A., Andone, I., Spănu, A., et al. (2012). On the feasibility of using motor imagery EEG-based brain–computer interface in chronic tetraplegics for assistive robotic arm control: a clinical test and long-term post-trial follow-up. *Spinal Cord* 50, 599–608.

Oommen, B.S., Smith, R.M., and Stahl, J.S. (2003). The influence of future gaze orientation upon eye-head coupling during saccades. *Exp. Brain Res.* 155, 9–18.

OpenStax, C. (2013). Illustration from Anatomy & Physiology, Connexions Web site. <http://cnx.org/content/col11496/1.6/>, Jun 19, 2013.

Patla, A.E., and Vickers, J.N. (1997). Where and when do we look as we approach and step over an obstacle in the travel path? *Neuroreport* 8, 3661.

Pelz, J.B., and Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. *Vision Res.* 41, 3587–3596.

Pelz, J., Hayhoe, M., and Loeber, R. (2001). The coordination of eye, head, and hand movements in a natural task. *Exp. Brain Res. Exp. Hirnforsch. Expérimentation Cérébrale* 139, 266–277.

Pelz, J.B., Canosa, R.L., Kucharczyk, D., Babcock, J.S., Silver, A., and Konno, D. (2000). Portable eyetracking: A study of natural eye movements. In *PROCEEDINGS-SPIE THE INTERNATIONAL SOCIETY FOR OPTICAL ENGINEERING*, pp. 566–583.

- Penkar, A.M., Lutteroth, C., and Weber, G. (2012). Designing for the Eye: Design Parameters for Dwell in Gaze Interaction. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, (New York, NY, USA: ACM), pp. 479–488.
- Pfeiffer, T., and Essig, K. (2013). *SAGA 2013: Proceedings of the First International Workshop on Solutions for Automatic Gaze-Data Analysis 2013 (SAGA 2013)*; Bielefeld, 24.-25. Oktober 2013 (Universitätsbibliothek Bielefeld, Bielefeld eCollections (BieColl)).
- Phillips, B., and Zhao, H. (1993). Predictors of assistive technology abandonment. *Assist. Technol. Off. J. RESNA* 5, 36–45.
- Phillips, A.T., Wellman, H.M., and Spelke, E.S. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition* 85, 53–78.
- Porterfield, W. (1737). An essay concerning the motions of our eyes. Part I. Of their external motions. *Edinb. Med. Essays Obs.* 3, 160–263.
- Rantanen, V., Verho, J., Lekkala, J., Tuisku, O., Surakka, V., and Vanhala, T. (2012). The effect of clicking by smiling on the accuracy of head-mounted gaze tracking. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, (New York, NY, USA: ACM), pp. 345–348.
- Rasmussen, C.E., and Williams, C.K.I. (2006). *Gaussian processes for machine learning* (Cambridge, Mass: MIT Press).
- Ray, A., and Bowyer, S.M. (2010). Clinical applications of magnetoencephalography in epilepsy. *Ann. Indian Acad. Neurol.* 13, 14–22.
- Rebsamen, B., Burdet, E., Guan, C., Zhang, H., Teo, C.L., Zeng, Q., Laugier, C., and Jr, M.H.A. (2007). Controlling a Wheelchair Indoors Using Thought. *IEEE Intell. Syst.* 22, 18–24.
- Ren, L., Khan, A.Z., Blohm, G., Henriques, D.Y.P., Sergio, L.E., and Crawford, J.D. (2006). Proprioceptive Guidance of Saccades in Eye–Hand Coordination. *J. Neurophysiol.* 96, 1464–1477.
- Renner, P., Pfeiffer, T., and Wachsmuth, I. (2014). Spatial References with Gaze and Pointing in Shared Space of Humans and Robots. In *Spatial Cognition IX*, C. Freksa, B. Nebel, M. Hegarty, and T. Barkowsky, eds. (Springer International Publishing), pp. 121–136.
- Richer, F., and Beatty, J. (1985). Pupillary Dilations in Movement Preparation and Execution. *Psychophysiology* 22, 204–207.
- Robinson, D.A. (1963). A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Trans. Bio-Med. Electron.* 10, 137–145.
- Rothkopf, C.A., and Pelz, J.B. (2004). Head Movement Estimation for Wearable Eye Tracker. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, (New York, NY, USA: ACM), pp. 123–130.
- Rothkopf, C.A., Ballard, D.H., and Hayhoe, M.M. (2007). Task and context determine where you look. *J. Vis.* 7, 16.

- Rozado, D., Rodriguez, F.B., and Varona, P. (2012a). Low cost remote gaze gesture recognition in real time. *Appl. Soft Comput.* **12**, 2072–2084.
- Rozado, D., Agustin, J.S., Rodriguez, F.B., and Varona, P. (2012b). Gliding and saccadic gaze gesture recognition in real time. *ACM Trans Interact Intell Syst* **1**, 10:1–10:27.
- Sacrey, L.-A.R., and Whishaw, I.Q. (2011). Subsystems of sensory attention for skilled reaching: Vision for transport and pre-shaping and somatosensation for grasping, withdrawal and release. *Behav. Brain Res.*
- Salvucci, D.D., and Goldberg, J.H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, pp. 71–78.
- San Agustin, J., Skovsgaard, H., Hansen, J.P., and Hansen, D.W. (2009). Low-cost gaze interaction: ready to deliver the promises. In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems*, (New York, NY, USA: ACM), pp. 4453–4458.
- Santella, A., and DeCarlo, D. (2004). Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, (ACM), pp. 27–34.
- Santini, T., Fuhl, W., Kübler, T., and Kasneci, E. (2015). Bayesian Identification of Fixations, Saccades, and Smooth Pursuits. *ArXiv151107732 Cs*.
- Schicktan, S., Amelung, T., and Rieger, J.W. (2015). Qualitative assessment of patients' attitudes and expectations toward BCIs and implications for future technology development. *Front. Syst. Neurosci.* **9**.
- Schneider, E., Dera, T., Bard, K., Bardins, S., Boening, G., and Brand, T. (2005). Eye movement driven head-mounted camera: it looks where the eyes look. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, p. 2437–2442 Vol. 3.
- Schneider, N., Bex, P., Barth, E., and Dorr, M. (2011). An open-source low-cost eye-tracking system for portable real-time and offline tracking. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications*, p. 8.
- Schneider, W.X., Einhauser, W., and Horstmann, G. (2013). Attentional selection in visual perception, memory and action: a quest for cross-domain integration. *Philos. Trans. R. Soc. B Biol. Sci.* **368**, 20130053–20130053.
- Schütz, A.C., Braun, D.I., and Gegenfurtner, K.R. (2011). Eye movements and perception: A selective review. *J. Vis.* **11**.
- Shannon, C. (1993). *Prediction and Entropy of Printed English*. 1951 (Shannon: Collected Papers. IEEE Press).
- Sheng-Wen Shih, and Jin Liu (2004). A novel approach to 3-D gaze tracking using stereo cameras. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **34**, 234–245.

- Sheng-Wen Shih, Yu-Te Wu, and Jin Liu (2000). A calibration-free gaze tracking technique. In 15th International Conference on Pattern Recognition, 2000. Proceedings, (IEEE), pp. 201–204 vol.4.
- Sim, N., Gavriel, C., Abbott, W.W., and Faisal, A.A. (2013). The head mouse; Head gaze estimation; In-the-Wild; with low-cost inertial sensors for BMI use. In 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 735–738.
- Sliney, D.H., and Freasier, B.C. (1973). Evaluation of optical radiation hazards. *Appl. Opt.* **12**, 1–24.
- Sliney, D., Aron-Rosa, D., DeLori, F., Fankhauser, F., Landry, R., Mainster, M., Marshall, J., Rassow, B., Stuck, B., Trokel, S., et al. (2005). Adjustment of guidelines for exposure of the eye to optical radiation from ocular instruments: statement from a task group of the International Commission on Non-Ionizing Radiation Protection (ICNIRP). *Appl. Opt.* **44**.
- Smeets, J.B.J., Hayhoe, M.M., and Ballard, D.H. (1996). Goal-directed arm movements change eye-head coordination. *Exp. Brain Res.* **109**, 434–440.
- Snyder, L.H., Calton, J.L., Dickinson, A.R., and Lawrence, B.M. (2002). Eye-Hand Coordination: Saccades Are Faster When Accompanied by a Coordinated Arm Movement. *J. Neurophysiol.* **87**, 2279–2286.
- Spering, M., Schutz, A.C., Braun, D.I., and Gegenfurtner, K.R. (2011). Keep your eyes on the ball: smooth pursuit eye movements enhance prediction of visual motion. *J. Neurophysiol.* **105**, 1756–1767.
- Steinman, R.M., Epelboim, J., Forofonova, T.I., and Bogacz, S. (1998). Gaze-control when humans do what they do best. In 6th International Conference on Ophthalmic Vision Science. **39**, S457.
- Stern, J.A., Walrath, L.C., and Goldstein, R. (1984). The endogenous eyeblink. *Psychophysiology* **21**, 22–33.
- Stevenson, I.H., and Kording, K.P. (2011). How advances in neural recording affect data analysis. *Nat. Neurosci.* **14**, 139–142.
- Svensson, U. (2004). Blink behaviour based drowsiness detection: method development and validation.
- Tafaj, E., Kasneci, G., Rosenstiel, W., and Bogdan, M. (2012). Bayesian online clustering of eye movement data. In Proceedings of the Symposium on Eye Tracking Research and Applications, (New York, NY, USA: ACM), pp. 285–288.
- Tatler, B.W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.* **7**, 4–4.
- Tatler, B.W., and Vincent, B.T. (2009). The prominence of behavioural biases in eye guidance. *Vis. Cogn.* **17**, 1029–1054.
- Tatler, B.W., Hayhoe, M.M., Land, M.F., and Ballard, D.H. (2011). Eye guidance in natural vision: Reinterpreting salience. *J. Vis.* **11**.

- Taylor, D.M., Tillery, S.I.H., and Schwartz, A.B. (2002). Direct Cortical Control of 3D Neuroprosthetic Devices. *Science* 296, 1829–1832.
- Tehovnik, E.J., Woods, L.C., and Slocum, W.M. (2013). Transfer of information by BMI. *Neuroscience* 255, 134–146.
- Thomik, A., Haber, D., and Faisal, A. (2013). Real-time movement prediction for improved control of neuroprosthetic devices. In 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 625–628.
- Tipper, S.P., Howard, L.A., and Paul, M.A. (2001). Reaching affects saccade trajectories. *Exp. Brain Res.* 136, 241–249.
- Tonet, O., Marinelli, M., Citi, L., Rossini, P.M., Rossini, L., Megali, G., and Dario, P. (2008). Defining brain-machine interface applications by matching interface performance with device requirements. *J. Neurosci. Methods* 167, 91–104.
- Tostado, P., Abbott, W.W., and Faisal, A.A. (2016). 3D Gaze Cursor: Continuous Calibration and End-Point Grasp Control of Robotic Actuators. *IEEE Int. Conf. Robot. Autom.*
- Treisman, A.M., and Gelade, G. (1980). A feature-integration theory of attention. *Cognit. Psychol.* 12, 97–136.
- Tseng, P.-H., Carmi, R., Cameron, I.G.M., Munoz, D.P., and Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *J. Vis.* 9, 4.
- Unema, P.J.A., Pannasch, S., Joos, M., and Velichkovsky, B.M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Vis. Cogn.* 12, 473–494.
- Urruty, T., Lew, S., Ihadaddene, N., and Simovici, D.A. (2007). Detecting eye fixations by projection clustering. *ACM Trans Multimed. Comput Commun Appl* 3, 5:1–5:20.
- Velichkovsky, B.B., Rumyantsev, M.A., and Morozov, M.A. (2014). New Solution to the Midas Touch Problem: Identification of Visual Commands Via Extraction of Focal Fixations. *Procedia Comput. Sci.* 39, 75–82.
- Velichkovsky, B.M., Joos, M., Helmert, J.R., and Pannasch, S. (2005). Two visual systems and their eye movements: Evidence from static and dynamic scene perception. In *Proceedings of the XXVII Conference of the Cognitive Science Society, (Citeseer)*, pp. 2283–2288.
- Velloso, E. (2013). AutoBAP: Automatic Coding of Body Action and Posture Units from Wearable Sensors.
- Wade, N., and Tatler, B.W. (2005). *The moving tablet of the eye: The origins of modern eye movement research* (Oxford University Press, USA).
- Wade, N.J., Tatler, B.W., and Heller, D. (2003). Dodge-ing the issue: Dodge, Javal, Hering, and the measurement of saccades in eye-movement research. *Perception* 32, 793–804.
- Wästlund, E., Sponseller, K., and Pettersson, O. (2010). What you see is where you go: testing a gaze-driven power wheelchair for individuals with severe multiple disabilities. In

Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, (New York, NY, USA: ACM), pp. 133–136.

Wells, W.C. (1792). An essay upon single vision with two eyes : together with experiments and observations on several other subjects in optics (London : Printed for T. Cadell, in the Strand).

Werner, J.S., and Chalupa, L.M. (2004). The Visual Neurosciences (MIT Press).

Westheimer, G., and McKee, S.P. (1975). Visual acuity in the presence of retinal-image motion. *J. Opt. Soc. Am.* 65, 847–850.

Williamson, S.S. (2004). Modulation of Gaze-Evoked Blinks Depends Primarily on Extraretinal Factors. *J. Neurophysiol.* 93, 627–632.

Wolpaw, J., and McFarland, D. (2004). Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proc. Natl. Acad. Sci. U. S. A.* 101, 17849–17854.

Wolpaw, J., Ramoser, H., McFarland, D., and Pfurtscheller, G. (1998a). EEG-based communication: improved accuracy by response verification. *IEEE Trans. Rehabil. Eng.* 6, 326–333.

Wolpaw, J.R., Ramoser, H., McFarland, D.J., and Pfurtscheller, G. (1998b). EEG-based communication: improved accuracy by response verification. *Rehabil. Eng. IEEE Trans. On* 6, 326–333.

Yarbus, A.L. (1967). Eye movements and vision. 1967. N. Y.

Zeng, L. (2009). Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th Edition) by B. Shneiderman and C. Plaisant: Pearson Addison-Wesley, 2009. xviii + 606 pages. ISBN: 978-0-321-53735-5. *Int. J. Hum.-Comput. Interact.* 25, 707–708.

Zhao, X. (Amy), Guestrin, E.D., Sayenko, D., Simpson, T., Gauthier, M., and Popovic, M.R. (2012). Typing with eye-gaze and tooth-clicks. In Proceedings of the Symposium on Eye Tracking Research and Applications, (New York, NY, USA: ACM), pp. 341–344.

10 Appendices

GT3D Image processing improvements

10.1 Image processing improvement and calibration automation:

The eye tracking system extracts pupil coordinates using binary image processing to segment pupil pixels (Figure 89(b)), these pixels are grouped into blobregions using connected component labelling. The blobs are then classified as noise or pupil blobs using expected pupil size and shape (Figure 89(c)). The contour pixel locations of the classified pupil blob are then used to fit an ellipse using least squares regression. However this methodology is very prone to specular reflections which partially occlude the pupil contour resulting in the extracted contour deviating round the occlusion, as shown in Figure 89(d). Though least squares regression is robust to outliers, when the incorrect contour makes up such a large proportion of the data points, they are not outliers. As can be seen, this results in a poor fit to the actual pupil, giving a noisy pupil location measurement (seeFigure 89(f)). To remedy this, an alternative approach was taken to extract the contour from the binary image. Rather than using all the edge pixels of the blob to make up the contour, the convex hull of the binary region is found. This has the result that the non-convex internal contour created by the specular reflection is not selected, leaving just the true pupil contour pixels (see Figure 89(e)). This allows an ellipse to be fitted that is highly representative of the pupil giving a much more accurate and precise pupil position (see Figure 89(g)).

The full image pipeline in GT3D system requires 6 parameters be set. This included the intensity threshold for each eye feed, the size of the structuring element used in morphological operations and the pupil classification parameters of maximum and minimum areaand axis ratio. The GUI used to manually set up these parameters is shown inFigure 88. Since then, an additional region of interest (ROI) option has been added to the pipeline to mask out possible noise sources and misclassifications due to shadows in the image periphery (see Figure 89(b)). Initially, this was a user set elliptical ROI requiring 5

parameters to be adjusted for each eye - a laborious process. So efforts have been made to design a system to automate the setting of these 15 processing parameters.

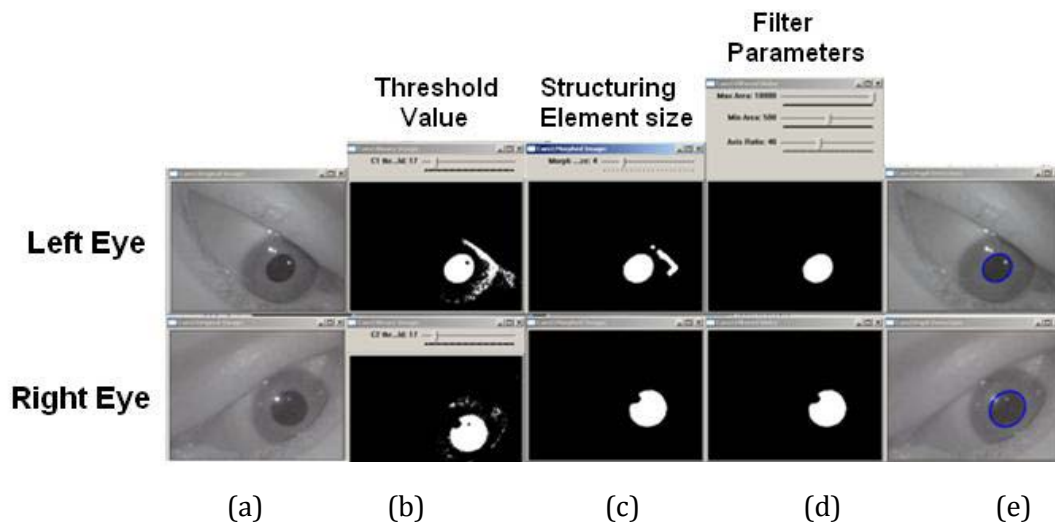


Figure 88.Real-time GUI used to customise image-processing parameters to the user. Each row represents the processing stages for each eye. Column (a): Greyscale image. (b): Binary image with a slider to control the intensity threshold value used to segment pupil pixels. (c): Effect of Morphological operations with a slider corresponding to the size of the structuring element used. (d): Filtered binary image with a slider to control each of the 3 filter parameters of maximum area, minimum area and axis ratio. (e): Resulting ellipse of best fit superimposed on the original greyscale image.

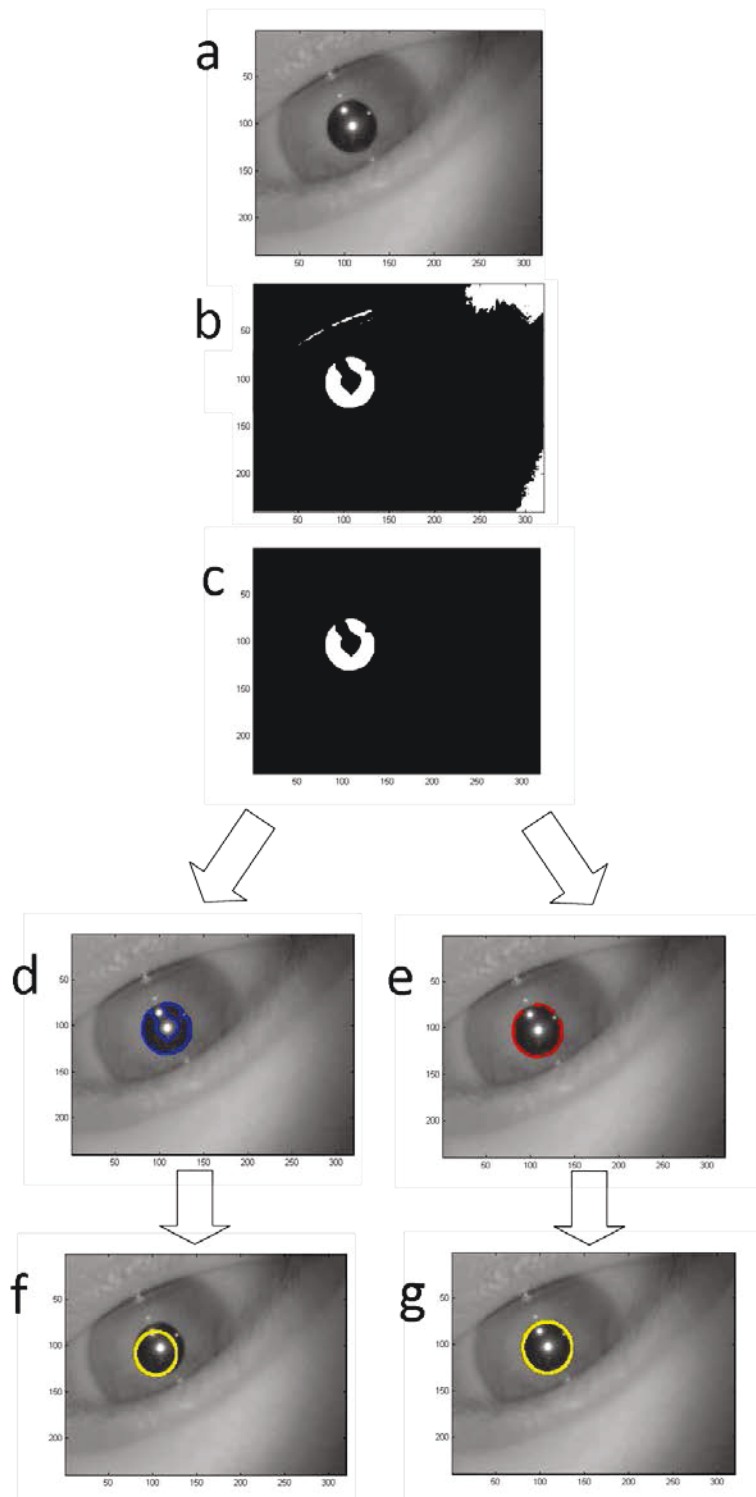


Figure 89.Example of noise introduced by specular reflections partially occluding the pupil contour and the comparison between the old direct contour method and the new convex hull method used to minimise the effect of this noise. (a) Raw greyscale image. (b) Binary image of segmented pixels (pupil/non-pupil). (c) Classified pupil blob. (d) Contour extracted directly (old method). (e) Contour extracted by finding the convex hull (new method). (f) Resulting ellipse fit for the directly extracted contour (old method). (g) Resulting ellipse fit for the convex hull contour (new method).

A system to calibrate the gaze and automate processing parameters has been developed and an overview is shown in Figure 90. The first step is to obtain a set of training images of the eye when looking at different locations on the screen, used to calibrate the gaze location. These images are also used to generate a ROI mask automatically. During calibration the user looks at each point in an evenly spaced calibration grid covering the extent of the screen and thus the full range of eye positions are recorded (see Figure 91(a)). The eye region has movement while the surroundings remain constant and thus background subtraction is used to locate the eye region. Background subtraction is applied to each of these images by subtracting the mean of calibration image stack from each image (see Figure 91b). This removes the background pixels, i.e. the pixels that remain constant through all the images, highlighting just the pixels in which there is significant difference from the mean. The resulting images are then summed into one image (see Figure 91(c)). The eye region can then be segmented using a simple threshold that is set based on the mean intensity of this image minus one standard deviation. This gives the binary image shown in Figure 91(c), from which a polynomial mask is created by obtaining the convex hull (shown here by the blue line). This polynomial is used to create a binary mask which is applied to each eye frame image. The effect of this is demonstrated on the calibration images from Figure 91(a) in Figure 91(e).

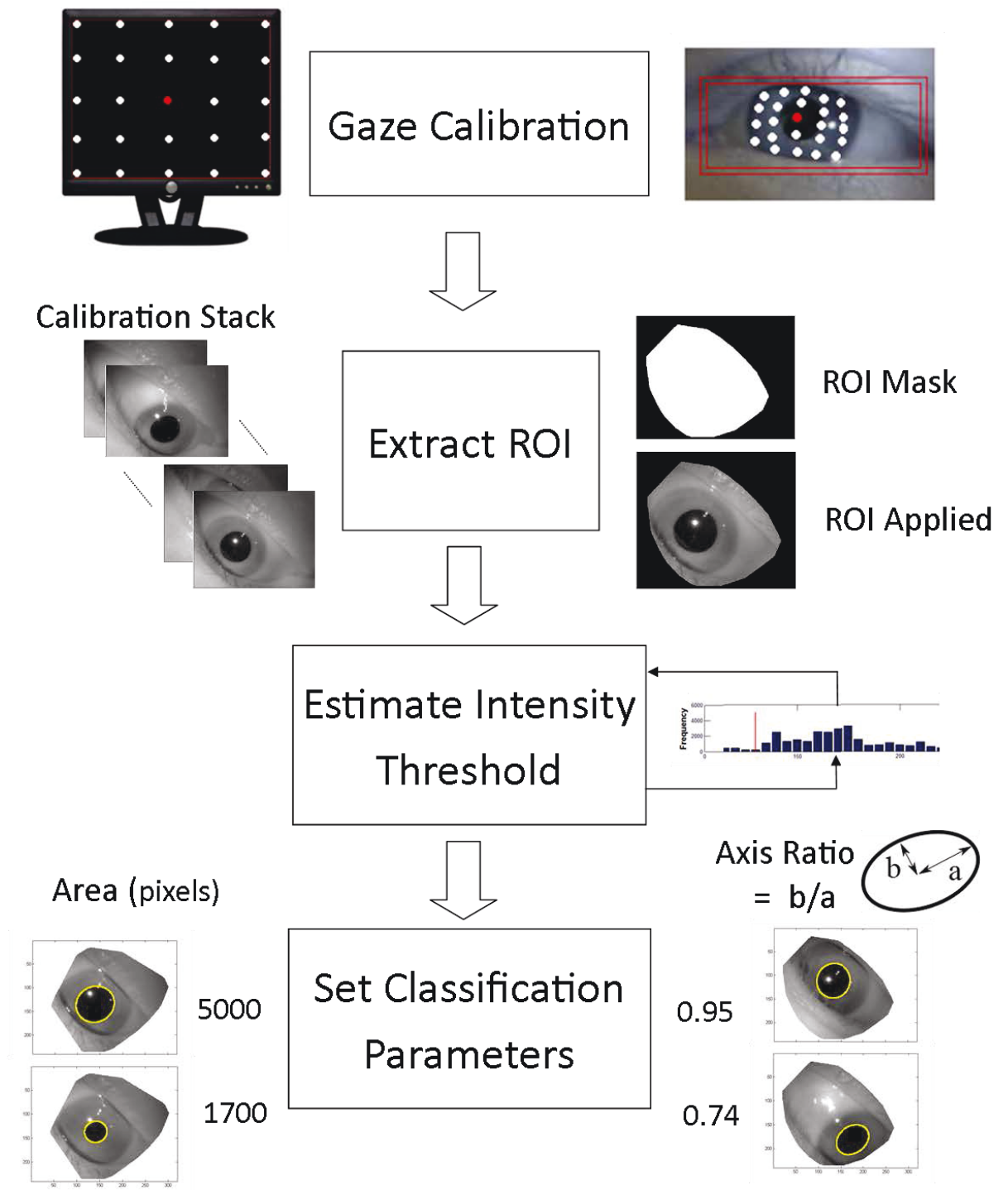


Figure 90. Calibration overview with additional automated parameter estimation required to create a ROI, estimate intensity threshold for pupil segmentation and set pupil size and shape parameters for pupil classification.

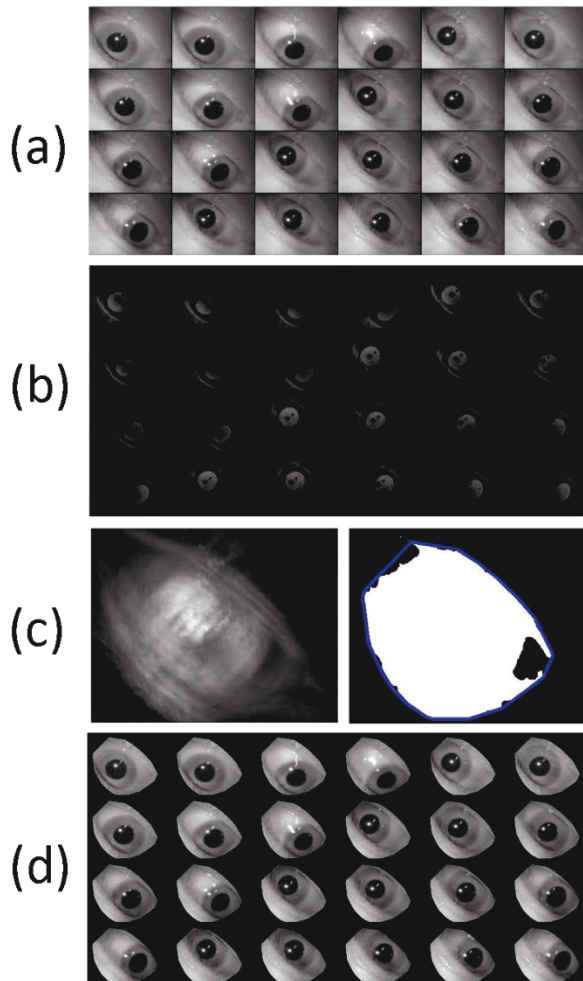


Figure 91.Auto-ROI generation steps. (a) Eye position at each calibration point. (b) Background subtraction applied by subtracting the mean of the calibration images stack from each individual image. (c) Summing over all the background subtracted images in (b) giving a silhouette of the regions of the image in which motion occurs and thus the eye region. A binary image of this motion silhouette is created with an intensity threshold set to the mean intensity minus the standard deviation. The convex hull is found and a binary image mask is created using this polynomial (blue line). (e) ROI mask applied to the calibration images in (a).

As mentioned, the pre-setting of the intensity threshold (Figure 88(b)) made the system very light level dependant which can be controlled in the lab but it is not practical for real world applications. A system was developed to customise the threshold on a frame-by-frame basis, allowing the system to be robust to light level. The intensity histogram of the eye image was found to have fairly distinct peaks for each region of the image. This is shown in Figure 92 where the image pixels are segmented using the intensity level at each minima in the intensity histogram. As can be seen, the pupil pixels to the first peak in the histogram as this is the darkest image region. Thus an algorithm was written to efficiently locate the first minima in the image histogram. To do this the turning points are found from the discrete histogram values by finding the zero crossing points. The process is illustrated graphically in Figure 93 and Figure 94. Figure 93 shows the threshold assignment process for a very bright image, taken immediately after a lamp with a conventional light bulb is turned on, giving high IR illumination. The same process is shown in Figure 94 for a dark image, taken just after the light has been turned off, before the camera automated gain and exposure control has adapted to the new lighting

conditions. As can be seen the histogram shape changes and its position shifts dramatically between the two extreme conditions but the threshold assigned (130 for bright and 30 for dark) gives good pupil segmentation and an accurate pupil ellipse is fitted in both cases as shown in Figure 93f and Figure 94f.

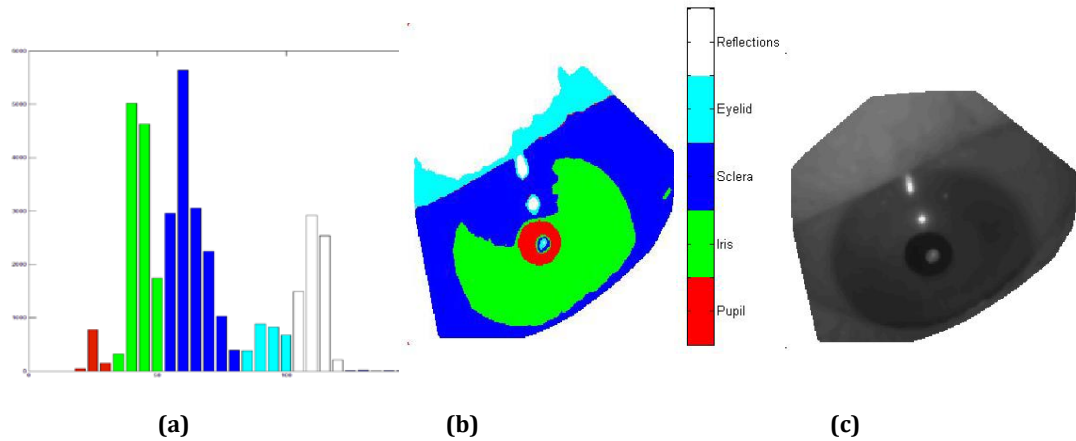


Figure 92. Segmentation of eye regions. (a) Histogram of pixel intensities for a particularly dark frame shown in (c). (b) Segmentation achieved using histogram minima intensities as thresholds.

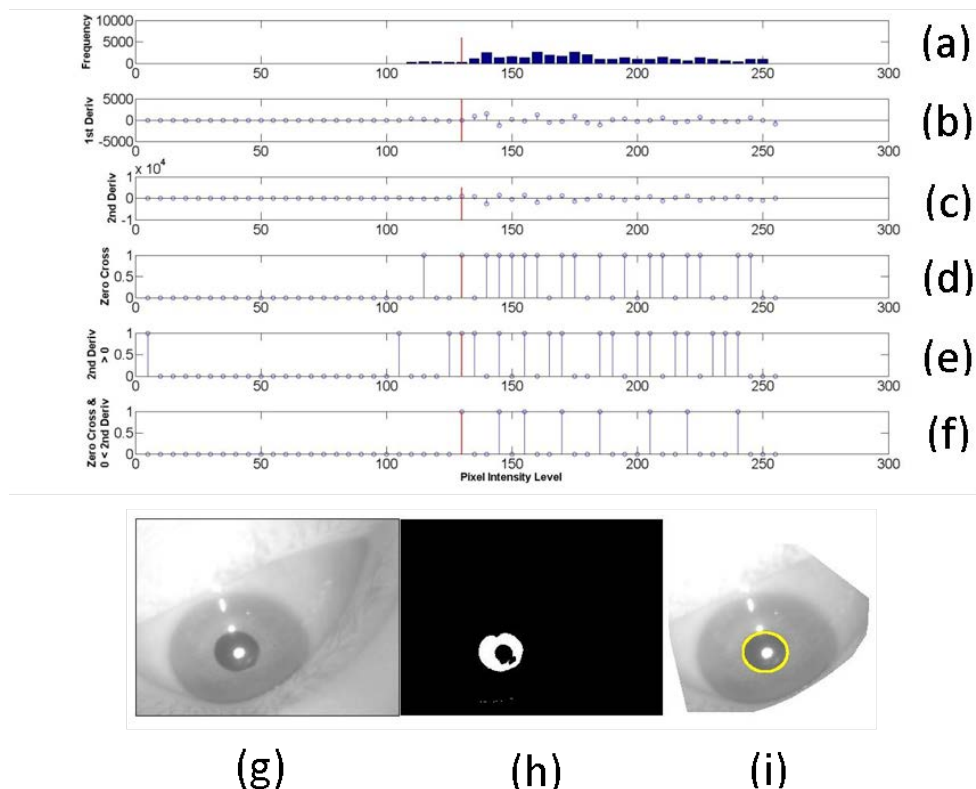


Figure 93. Graphical illustration of automatic intensity threshold calculation when a bright lamp is turned on, before the built in camera auto gain and exposure have reacted. (a) Incoming image intensity histogram. (b) First derivative (c) Second derivative. (d) Zero crossing points. (e) Positive elements of the second derivative. (f) Locations of each minima based on the criteria of zero crossing with a positive second derivative. The threshold is then taken as the intensity level at the location of

the first minima – represented by the red line. (g) Incoming image. (h) Binary image extracted using the custom threshold. (i) Resulting classification and ellipse fitting of the pupil segmentation made using the automatically assigned threshold.

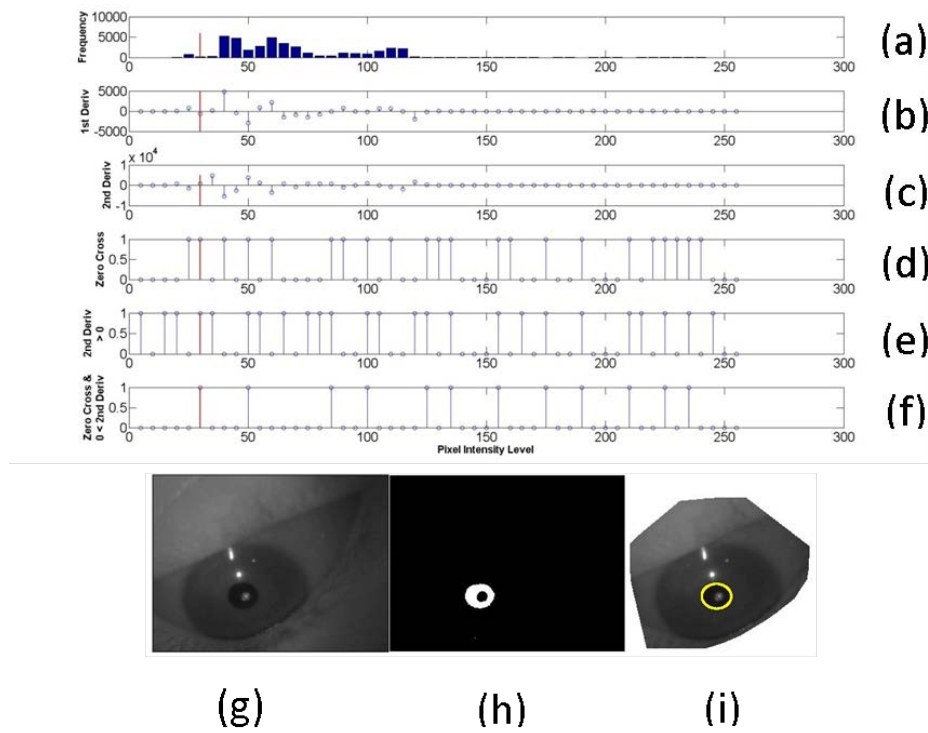


Figure 94. Graphical illustration of automatic intensity threshold calculation for a dark image, following a bright lamp being turned off, before the built in camera auto gain and exposure has reacted. (seeFigure 93 for description of each part).

The final settings required for the pupil extraction are the expected pupil size and shape parameters used to classify the segmented blobs as pupil or non-pupil as shown in Figure 89

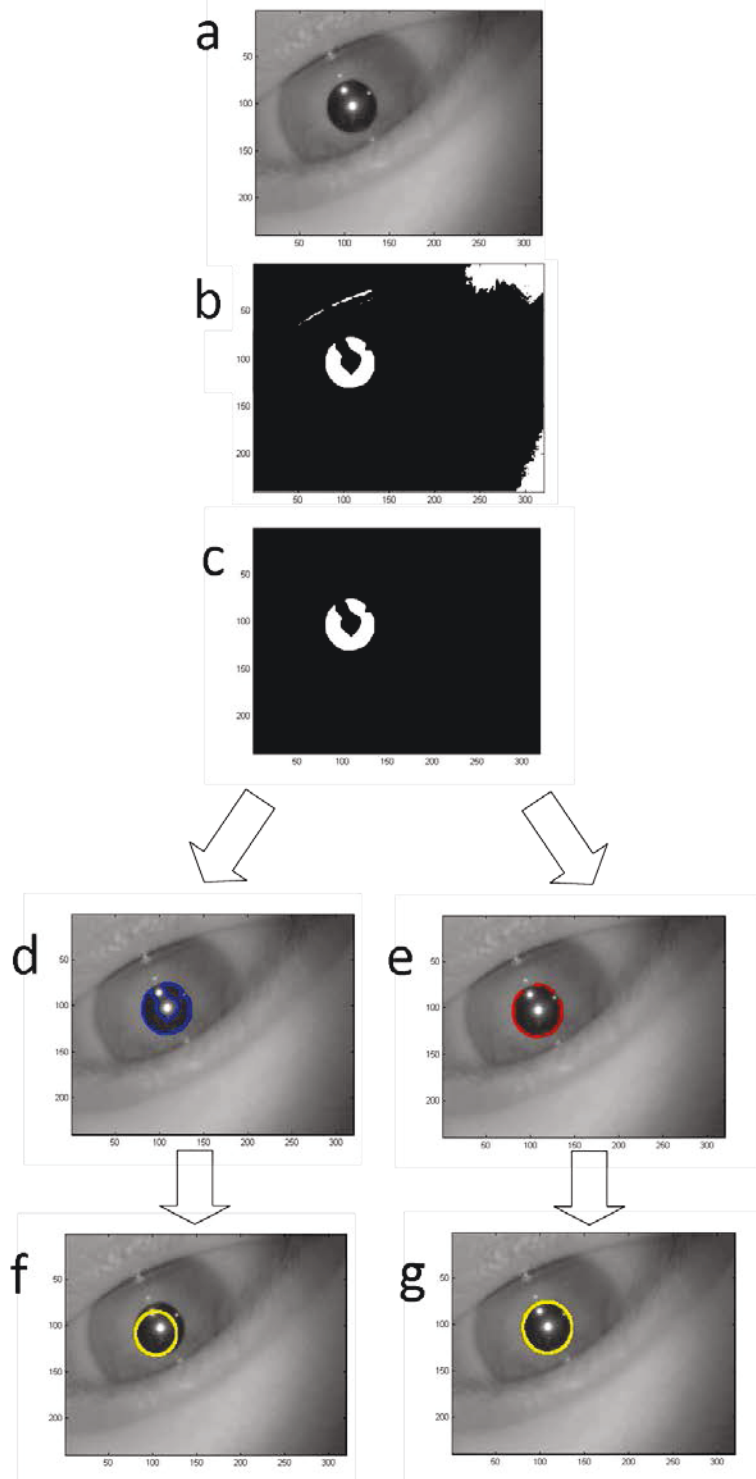


Figure 89. Example of noise introduced by specular reflections partially occluding the pupil contour and the comparison between the old direct contour method and the new convex hull method used to minimise the effect of this noise. (a) Raw greyscale image. (b) Binary image of segmented pixels (pupil/non-pupil). (c) Classified pupil blob. (d) Contour extracted directly (old method). (e) Contour extracted by finding the convex hull (new method). (f) Resulting ellipse fit for the directly extracted contour (old method). (g) Resulting ellipse fit for the convex hull contour (new method).

(d). Thus the possible range of pupil sizes and eccentricities in the image must be estimated to correctly classify the pupil blob while avoiding false positives. This range can vary between people and for different adjustments of the eye-tracking camera positions. As such with an additional 10 second stimulus added to the current calibration screen, suitable parameters can be estimated automatically. To obtain the pupil area maximum and minimum, the subject is presented with a blank screen for 5 seconds, leading to pupil dilation (Figure 95a) followed by a white screen for 5 seconds, leading to pupil constriction (Figure 95b). The pupil area is recorded during this stimulus; an example recording is shown in

(a)
(b)

Figure 97a. The range of pupil areas is used to set the pupil area maximum and minimum. To set the range of possible eccentricities, the pupil eccentricities during the gaze calibration step are analysed. An example trace is shown in Figure 97b, as can be seen the axis ratio varies as each calibration point is fixated and thus the range of eccentricities is found.

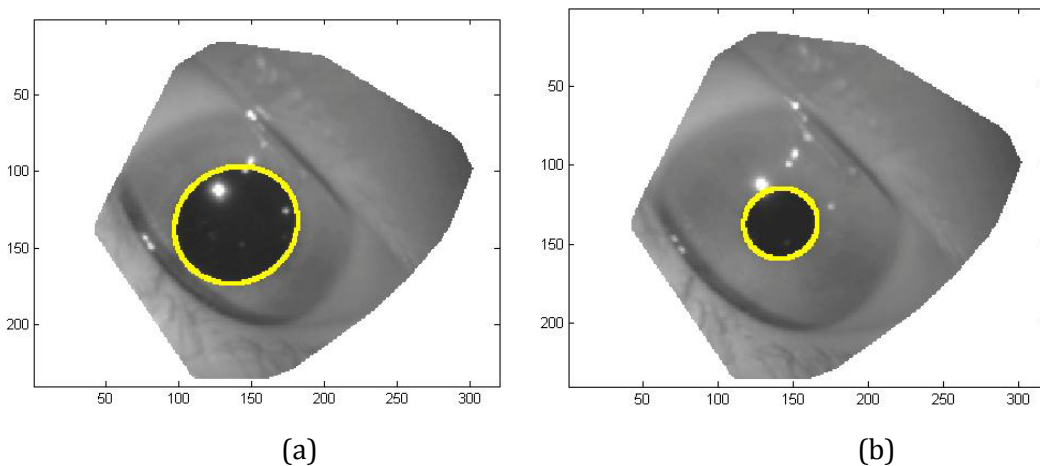


Figure 95. Pupil area variation. (a) Pupil dilated due to dark screen- area =5000 pixels. (b) Pupil contracted due to white screen- area =1700 pixels.

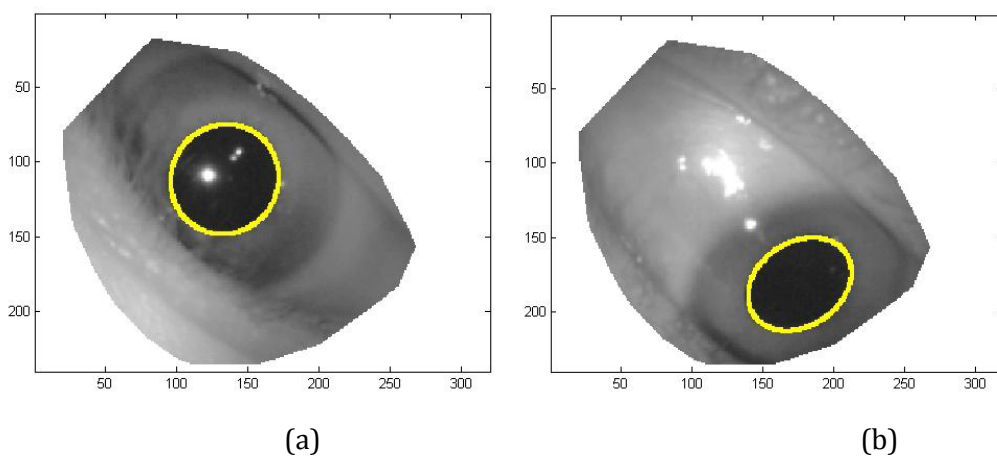


Figure 96. Pupil Axis ratio variation. (a) Looking straight ahead gives a high axis ratio of 0.95. (b) When the eye moves to higher eccentricities, the axis ratio decreases as little as 0.74.

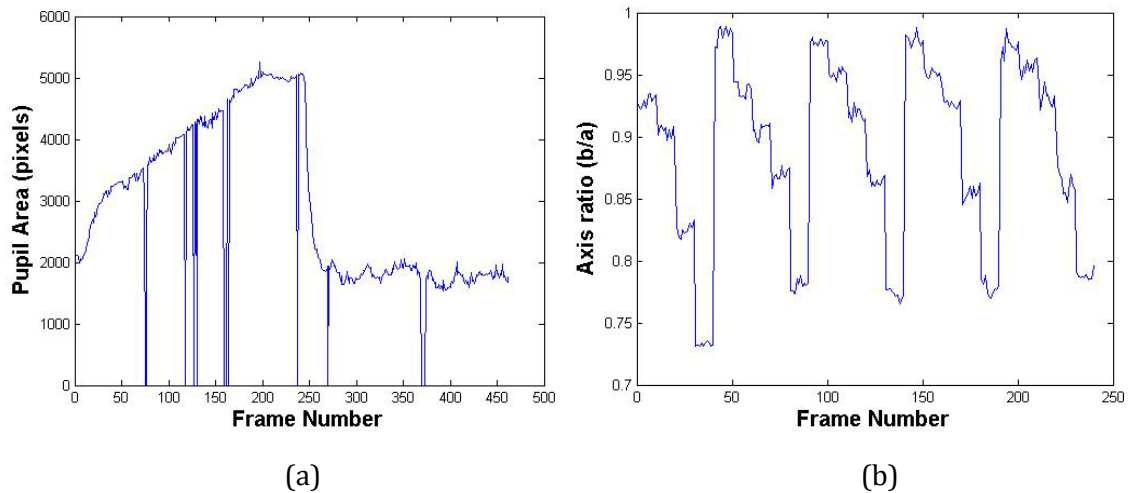


Figure 97. Pupil size and axis ratio measurements for automatically setting the classification boundaries. (a) Pupil area is recorded when presented with a black screen followed by a white screen to stimulate the pupillary response. (b) Pupil eccentricity (axis ratio) during the gaze calibration process as the user looks around the screen producing different pupil eccentricities.

10.2 Image processing Results.

As a preliminary assessment of the improvements of the dynamic threshold and convex hull contour extraction were tested. To give a practical measure, the system was calibrated to the screen and following calibration, 50 test points randomly distributed on the computer screen were displayed. The user had to look at each in turn and press space bar. The performance of the system was then based on the mean Euclidean error of the gaze estimates over all the test points. During the experiment, a bright light facing the subject was switched on and off to simulate varying light conditions. The eye-images were saved and post processed using the different methods to allow direct comparison. Figure 98 shows a short trace of the auto threshold in action during a period when the light is turned on and then off again. The threshold adapts to this and the pupil locations are found effectively while the static threshold does not locate the ellipses for the more extreme lighting levels. To isolate the effects of the different improvements on gaze accuracy, first the old method was applied - a fixed threshold with ellipse fitting applied directly to the full contour of the pupil blob (first row Table 4). Then the automatic threshold was used with the full contour and subsequently, the convex hull method was used. The first method was strongly affected by the varying light levels because of the hard coded intensity threshold and the pupil was often not located at all - resulting in the very high mean Euclidean error of 36.6 mm for the left eye and 60.9 mm for the right eye as shown in Table 4.

Table 4. Comparison between fixed threshold and dynamic threshold during varying light levels. Contour pupil extraction and using the convex hull contour. The mean Euclidean error between 50 test points and the gaze estimates and the standard deviation is shown.

Method	Left Eye Error (mm)	Right Eye Error (mm)
Fixed threshold with full contour	36.6±61.6	60.9±250
Full Contour with dynamic threshold	10.7±8.7	7.1±6.9
Convex Hull with dynamic threshold	9.4±7.1	6.32±4.63

When the eye is not located the estimate is set to zero which explains the very high standard deviations seen in Table 4. Though it would be less naïve to set the estimate to say the centre of the screen, this approach penalises the method for not locating the pupil. When the dynamic threshold is applied, the results improve significantly with the mean Euclidean error dropping to less than 1cm, with standard deviations of a similar order. This 1cm error corresponds to approximately 1 degree of visual angle at 50cm, the distance between the screen and the user. In the third method, the convex hull was used to extract the pupil contour. This had a surprisingly small effect on the error decreasing the overall mean very slightly –by around 1mm. This is perhaps because the methods results will be equivalent except for the cases when the pupil is severely occluded by specular reflections.

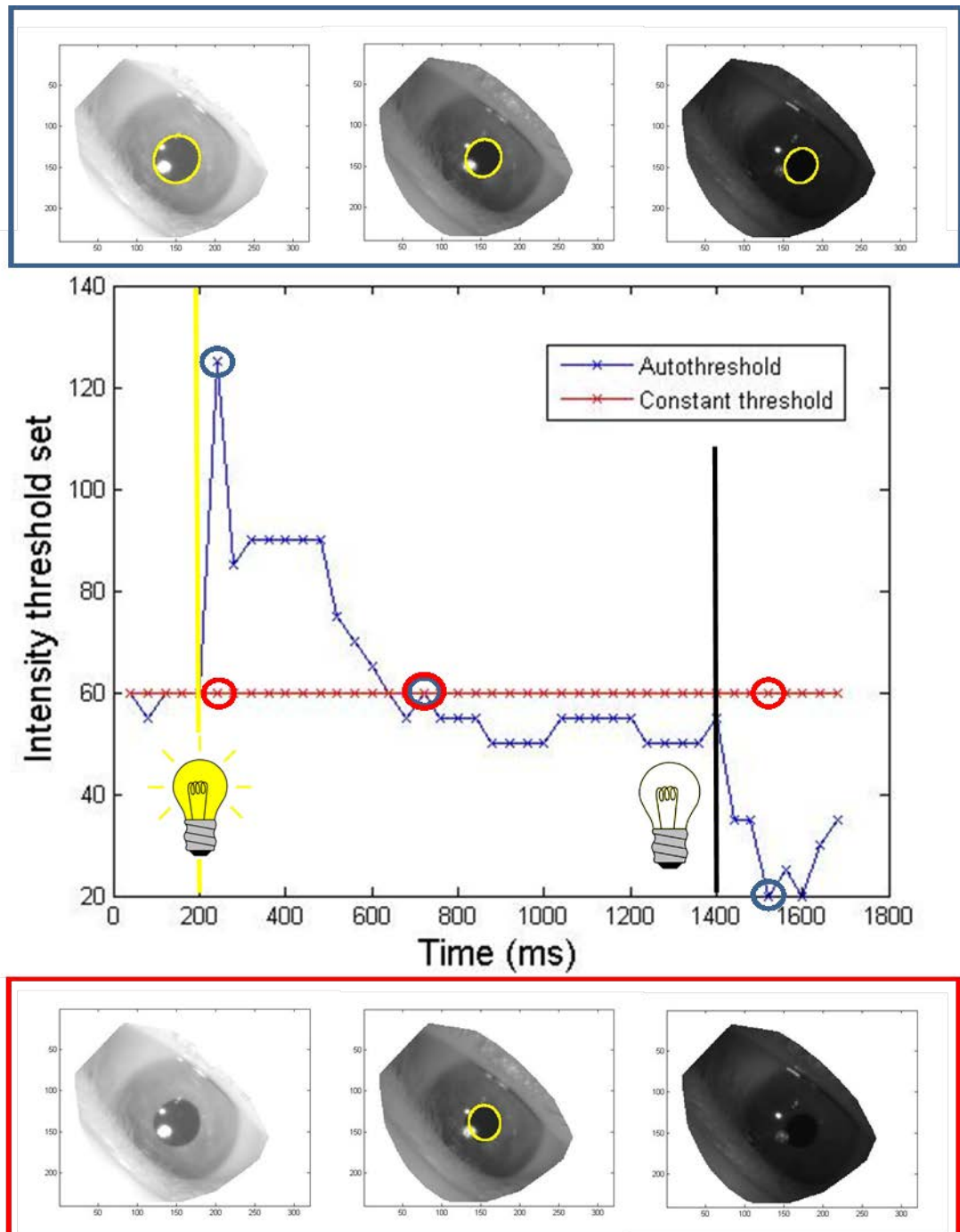


Figure 98. Demonstration of automatic threshold in action. During the experiment the user looked at different test locations on the screen while a bright light was turned and off. The plot above shows the evolution of the threshold for a short sequence in which the light is turned on (yellow line) and then off (black line). As can be seen, the threshold increases to compensate for the light being turned on and then rapidly drops again when it is turned off. The eye images are frame samples of the respective algorithm outputs at the times indicated on the graph by the circles. In the blue box the results of the automatic threshold and the convex hull contour extraction are shown to find and fit the ellipse in all cases. Images in the red box correspond to the static threshold method with full contour extraction and show that the pupil can only be found in the case when the lighting levels are similar to when the threshold was set. In this case you can also see that the ellipse is slightly distorted by the specular reflection.

