Calibration-Free Eye Tracking

by

Daniel Cheng

A thesis submitted to the
School of Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
November 2005

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-15272-0
Our file Notre référence
ISBN: 978-0-494-15272-0

Abstract

Typical eye tracking methods require a calibration process to obtain the geometrical positions of the camera, the visual display, and the eyes. This thesis presents a novel method of eye gaze estimation, one without a need for calibration or knowledge of the aforementioned geometrical properties.

The calibration-free (CF) method exploits the refractive characteristics of the corneal surface as well as the retro-reflective behaviour of the eye.

Infrared light emitting diodes (IR LEDs) are used to mark the perimeter of a visual display. The reflection of these markers on the cornea provide a reference to where the visual display is in relation to the viewer's field of view. Secondly, a viewer's focus of attention is synonymous with the pupil centre. Thus by observing where the pupil centre lies relative to the corneal reflections of the IR markers, it is possible to estimate where the subject is looking. To map the location of the pupil centre to coordinates on the visual display, a novel interpolation routine is presented.

The method is implemented as a prototype CF eye tracker and design considerations are discussed. The CF prototype provides a non-intrusive means for eye gaze estimation. The prototype is easier to use than conventional techniques, and simpler and more economical to construct than currently available commercial systems.

Acknowledgments

First and foremost, I would like to thank my supervisor, Roel Vertegaal; without his wisdom and tutelage this thesis would not have been possible. To Changuk Sohn, our resident hardware guru, a special thank you is deserved for supporting this thesis through diligent and meticulous work on the hardware. As well, I would like to thank my fellow students of the Human Media Lab for the great memories, their guidance, and more importantly, their friendships. Most of all I would like to thank my parents, Allen and Christine, for their unwavering belief in me and constant support throughout the years. To them my deepest love and appreciation.

Finally, I would like to dedicate this thesis in loving memory of my mother. I thank her for teaching me the value of education and the importance of a strong work ethic. Mom, this one is for you - may you be in peace.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

"The eyes are the window to the mind and the mind's window to the scene" John W. Senders

1.1 Motivation

Humans have evolved to be highly cognizant of eye gaze. In interpersonal interactions, eye gaze provides a non-verbal cue that is highly communicative of one's intentions or emotional state [43]. Similarly, a great deal of a user's actions may be contextualized by eye gaze. When compared to manual interfaces such as keyboards and mice, eye gaze provides a much more natural mode for human interactions with computing devices. This is mainly because humans naturally tend to fixate on objects of interest *before* we consciously act upon them; eye gaze provides an observable cue for inferring our intentions. By tracking eye gaze, our innate tendency of fixating on objects of interests may be used as a natural trigger for measuring and initiating interactions. Since eye gaze is reflective of our intentions and thus our mental state, fields exploring these areas have provided eye gaze tracking a plethora of wide-ranging applications, from studying physiological and neurological disorders, to examining various cognitive functions like vision and reading; from augmenting communication with devices, to evaluating the efficacy of visual designs and marketing, with new applications constantly being developed.

Synergistic with the growing number of applications, is the movement to mature gaze tracking technologies past their current niche applications. We are beginning to integrate them into more commonplace systems, such as in the viewfinder of high-end cameras [13], or as gaze-contingent displays for bandwidth-minimized telecommunication [5]

1.2 Problem

For this integration to be realized, several shortcomings of current eye trackers must be resolved. Typically, eye trackers are expensive, impose restrictions on the user's head movements, and often require a calibration routine. Calibration is required to determine the parameters for describing the mapping between the real-world coordinates of the visual scene, and the eye coordinates from the captured camera image. As well, there are systems that can perform accurately given a modicum of head movement; however, these systems require additional knowledge such as the distance between the camera and the eye. More often than not, extra hardware, such as stereoscopic cameras, or ultrasonic sensors must be employed to gather the additional information, adding further complexity to the system.

1.3 Contributions

This thesis proposes a non-intrusive method of eye tracking that does not require calibration and demonstrates it as a prototype. Non-intrusive embodiments are preferable as they are generally more comfortable. Furthermore, the method may be applied remotely, without the need to physically contact the user, facilitating more natural interactions. Infrared illuminators positioned around a visual display produce specific reflections of light on the surface of the user's eye. By observing the location of these reflections, and the location of the centre of the pupil, the subject's point of regard on the visual display may be estimated. To obtain this information, efficient computer vision algorithms were developed to extract the visual features and used in conjunction with supporting hardware.

Chapter 2

Background

"The difficulty seems to be, not so much that we publish unduly in view of the extent and variety of present-day interests, but rather that publication had been extended far beyond our present ability to make real use of the record" *Vannevar Bush* [3]

2.1 Vannevar Bush

As World War II was concluding, war-motivated research and development wound down and allowed scientists to pursue new research avenues. One such scientist was Vannevar Bush, and what he proposed in 1945 helped lay the foundation and guiding philosophy of HCI research. Vannevar Bush hoped to bring recognition to the growing problem of knowledge asymmetry. At the time, there existed a large imbalance in the aggregate time spent writing scholarly works compared to the time spent reading them. Exemplifying this asymmetry, Bush cites how Mendel's break-through laws of genetics were lost to the world for generations because his publications did not reach the experts capable of grasping their significance and furthering his ideas.

"Professionally our methods of transmitting and reviewing the results of research are generations old and by now totally antiquated for their purpose... The summation of human experience is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships" [3]

In his seminal article, Vannevar Bush proposed the memex as a solution to the information storage and retrieval problem. Its purpose was to augment the powers of human memory and association through symbiosis with machines.

2.1.1 Memex

An extension of Vannevar Bush's earlier work with the rapid selector, the Memex [3] was a storage and retrieval device. Bush realized that the efficiency of human memory lay in its ability to recall memories associatively. Similarly, Bush proposed that information retrieval would see similar benefits if the affordances of selection by association could be realized mechanically. Inspired by the associative nature of memory formation in the human mind, the memex allowed links between documents. These links were called associative trails, and were the precursor to modern hypertext. Information such as records, published literature, and written material were stored on microfilm and could be rapidly retrieved and projected onto a desktop viewing panel.

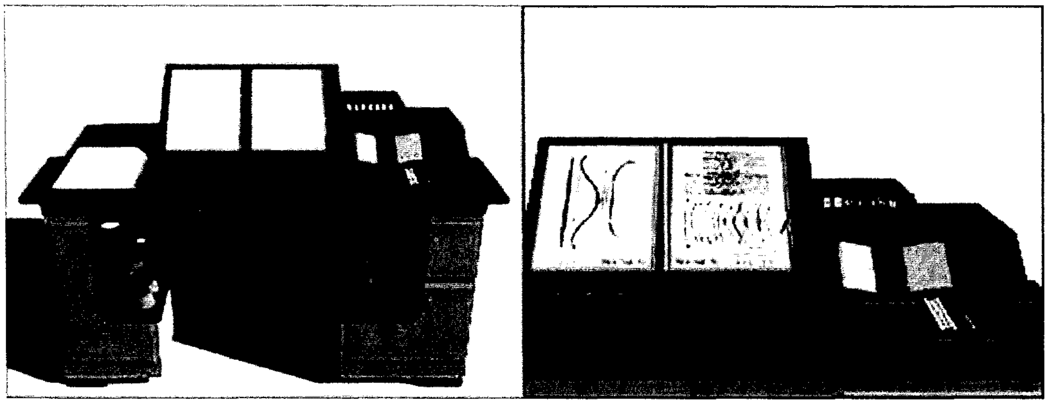

Figure 2.1: Memex System [3]

For example, a user could build a trail by sequentially associating related papers to each other. The interconnections between items allow the items to be reviewed as though they had been physically amalgamated into a new text.

Moreover, any item could be joined with numerous other associative trails, and both items and trails could be shared amongst other researchers in the hopes of providing a more versatile and efficient means of collaborating and researching.

2.2 Human-Machine Symbiosis

symbiosis *- the intimate living together of two dissimilar organisms in a mutually beneficial relationship*

Furthering Vannevar's concept of improving research and collaboration through augmenting humans with computers was J.C.R. Licklider. Licklider coined the term human-computer symbiosis [20] to describe his expectation for the continued development of cooperative interactions between humans and digital computers. Human-computer symbiosis marked a departure from the conventional relationship with machines of that era; machines, typically mechanical systems, were mere extensions of the human body. Human's provided the initiative, direction, and criterion for a problem, while machines simply facilitated the execution of the solution.

The objective of pairing humans with digital computers was twofold. Firstly, it was to enable computers to not only execute the solution of formulated problems, but rather facilitate the formulation of the solution itself. Secondly, it was to allow cooperation between humans and computers in decision making and controlling complex situations without an inflexible reliance on predetermined programming. The computer was more than just a data-processing medium, it could aid the user in thinking, understanding and decision making.

Specifically, Licklider brought attention to the potential benefits that human-computer symbiosis could bring to the domain of communication [21]. With the computer as a vehicle for communication, Licklider described how computing devices would facilitate a more intimate interaction between humans and information.

"[We are] entering a technological age in which we will be able to interact with the richness of living information not merely in the passive way that we are accustomed to with books and libraries, but as active participants in an ongoing process, bringing something to it through our interaction with it, and not simply receiving something from it by our connection to it." [21]

Furthermore, Licklider acknowledged the computer as a medium that would change the nature and value of communication, possible more profoundly than did the printing press or television [20].

Licklider viewed communication as the construction of a mutual external model between different parties. Each party has an internal model of the idea they wish to convey, and through words, diagrams, gestures, etc., this model is externalized. Communication is successful when all parties have reconciled their internal models into a common external model shared by all. From this perspective, the dynamics of communication appeared highly model-centric and from this Licklider concluded that the current methods of two-way communication were inadequate mediums simply because they failed to provide facilities for externalizing models.

"Is it really seeing the expression in the other's eye that makes the face-to-face conference so much more productive than the telephone conference call, or is it being able to create and modify external models?" [21]

Broadening the scope of computer assisted telecommunication to general online communities, Licklider proposed the OLIVER [21] as a method for managing information and requests for user attention. The acronym OLIVER is an homage to Oliver Selfridge who conceived the idea of an "On-Line Interactive Vicarious Expediter and Responder". The OLIVER resided within the online network and was composed of a complex of computer programs that acted on behalf of its user, taking care of routine matters that did not warrant the user's explicit, personal attention. Moreover, the OLIVER provided a means of shielding the user from the demands of the world by mediating attentional requests.

The OLIVER would have knowledge of ones social network, as well as the social standing of each member with that network. Additionally, other value structures could be assimilated, such as what interactions a user may have with others, as well as the priority those interactions carry. Although all users of the OLIVER shared a common initial feature set, the OLIVER's capabilities were extensible via the addition of customized modules, or idiosyncratic learning by means of observing the user's usage behaviour.

Arguably, the OLIVER may be viewed as an archetype for what later developed into the Attentive User Interface paradigm.

2.3 The Mother of All Demos

Another seminal pioneer from the infancy of HCI was Douglas Engelbart. His theories of augmenting human intellect through the use of computers and software gave life to the development of the graphical user interface (GUI) and mouse. Engelbart was deeply inspired by Vannevar Bush's 1945 article "As We May Think," [3] a discussion of the future use of machines as mechanical aids to the human intellect. In that article Vannevar Bush stated that it was without question a catastrophe when the work of a talented individual does not reach the minds of those capable of grasping and extending their work.

Echoing Bush's sentiments, Engelbart realized that the subsequent huge influx of technological innovation and knowledge that followed WWII would quickly outpace the human capacity to manage it. Fortunately for Engelbart, he found a sympathetic ally in a contemporary at the new Information Processing Techniques Office at the Advanced Research Projects Agency (ARPA), J.C.R. Licklider. Licklider was quick to see the analogies between his own efforts with human-computer symbiosis and Engelbart's augmentation of human intellect, and rewarded Engelbart with funding to further his idea. Douglas Engelbart's theory of augmentation was based on his realization that the development of greater human intellect could be facilitated by using machines to handle the routine and mechanical portions of thinking and idea sharing [7]. Engelbart believed that the steady increase in computer usage would eventually drive major sociological changes [7]. Thus a cooperative relationship between humans and computers would be mutually beneficial - if only the proper tools could be developed to allow them to do so.

Drawing from his experience as a Navy radar technician, Engelbart boldly envisioned using the cathode-ray tube as a graphical display for navigating and interacting with information spaces. Additionally, Engelbart was one of the earliest computer scientists to see past the facade of the computer merely being a data-processing device. Instead, Engelbart was interested in harnessing the power of computers for processing documents and images, as well as a tool to enable collaboration and enhance productivity.

"Envision people sitting in front of displays, 'flying around' in an information space where they could formulate and organize their ideas with incredible speed and flexibility" [7]

Engelbart's efforts culminated in 1968, at the Fall Joint Computer Conference in San Francisco, where he gave a ninety minute demonstration about the fruition of his efforts, the oN-Line System (NLS). The NLS was a breakthrough in collaborative groupware and demonstrated convincingly how the computer could be used in everyday tasks. To illustrate this, the majority of the presentation was focused on the scheduling of tasks Engelbart had to do later in the day. The information was presented as a simple hypertext which contained many different methods of organization, each contextual to the task at hand. Navigation and selection was done with relative ease through a combination of mouse and keyboard.

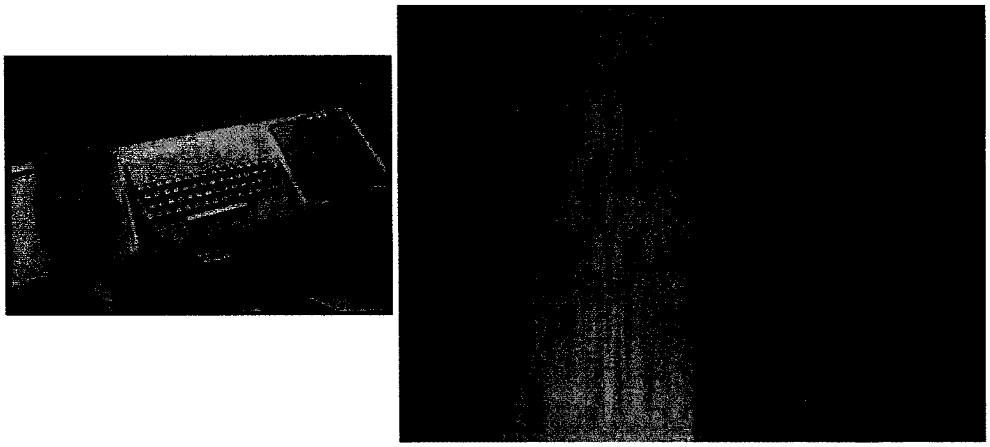

Figure 2.2: NLS Demonstration [7]

NLS utilized a graphical user interface implemented within a windowing environment. Among the features debuted were electronic messaging of other users, a variety of word processing options, as well as on-screen video teleconferencing. Although these technologies have become essentially ubiquitous, in the 1960s it was truly the "mother of all demos".

The work of pioneers such as Vannevar Bush, J.R.C Licklider, Douglas Engelbart, laid the foundations for what would later be formalized as human-computer interaction. The early work of Bush brought to the forefront the possible improvements to human intellect that could arise from augmenting humans with machines. Licklider furthered Bush's call with an argument presenting a clear case for a close interactive relationship between humans and modern computers, illustrating how mutual cooperation would benefit humans through increased productivity and more efficient communication. The core concepts of Bush and Licklider were ultimately consolidated by Engelbart and brought to light with his infamous demonstration of the NLS. Engelbart provided the final push for general acceptance of the great potential gains to be reaped from human-computer interaction. With the 1968 public debut of the NLS, human-computer interaction moved from fancy to reality.

2.4 Ubiquitous Computer

"The world has arrived at an age of cheap complex devices of great reliability; and something is bound to come of it." Vannevar Bush [3]

There have been two great epochs in computing. The first epoch began with the production of mainframe computers in the late 1950s. These mainframe computers were bulky, expensive, and mainly operated by experts. The limited accessibility of mainframes made them a scarce resource that was negotiated and shared with others [40]. The relationship of many users sharing one computing resource is definitive of this paradigm of computing.

The dawn of the integrated-circuit era marked the beginnings of the next epoch in computing. This second wave was swept in by the current of Moore's Law and saw the computer mature from a scarce resource that was communally time-shared, to a personal household commodity. Personal computers have become analogous to automobiles as both are special, relatively expensive items that empower the user albeit requiring considerable attention to operate. Similarly, one may own several for different tasks. Personal computing devices have adopted specialized roles that fully engage or occupy the user when in use. This one-to-one relationship is characteristic of the personal computing paradigm.

Currently, we are in a transition phase consisting of a fusion of the previous two computing paradigms. The Internet is essentially client-server computing on a massive scale, with web clients hosted by personal computers, and servers by mainframes, respectively [39]. The Internet has carried us into the age of information and distributed computing. The Internet afforded a massive interconnection of personal, business, and government information. The resulting ubiquity of information has wired our society and established the first inroads for building the next great paradigm, what Alan Kay referred to as the Third Paradigm: *ubiquitous computing.*

The vision of ubiquitous computing is embodied as massive interconnectivity not only amongst people, but with computing devices ranging from the microscopic to the macroscopic. Computing capabilities will be transparently embedded into everything and anything. People will no longer share computers, rather computers will share us. The ideology of ubiquitous computing is to make computing so deeply embedded within society that it becomes invisible [39, 40]. Like other technologies that have become invisible from ubiquity - such as plumbing, electricity, and writing - the proliferation of computers into our surrounding environment necessitates that they are: 1. invisible; and 2. recede into the periphery. For this to be realized, it is necessary to design methods for people to be shared by computers while remaining serene and in control [39].

Table 2.1: Major Trends in Computing

Computing Paradigm	**Relationship**
Main Frame	Many people share a computer
Personal Computer	One computer, one person
Internet - Widespread Distributed Computing	Transitory Phase
Ubiquitous Computing	Many computers sharing many users

2.4.1 Calm Technology

The underlying philosophy of ubiquitous computing is calmness. Information technology is often the antithesis of calmness. Personal digital assistants, cell phones, and email frequently bombard the user with requests for attention [30, 29]. These devices operate under the assumption that they are the sole recipient of the user's attention. Calm technology engages both the centre and the periphery of attention, transitioning back and forth between the two, allowing for a smooth mediating of attention requests.

"[A user's] excursion may be more enjoyable if he can reacquire the privilege of forgetting the manifold things he does not need to have immediately at hand, with some assurance that he can find them again if they prove important" [3]

The periphery can be described as what we are attuned to without attending to explicitly [2]. An example of this is the Cocktail Party effect [38, 36]. The Cocktail Party effect is a colloquialism for describing how one may focus one's listening attention on a single talker among a cacophony of simultaneous conversations and background noise. For example, when you are talking to a colleague at a crowded social gathering, without exerting much mental effort you can focus your listening attention on your colleague and dampen the surrounding auditory noise by pushing it into your attentional periphery. Interestingly, although you may not remember anything about the peripheral conversations, there is still a low-level awareness and monitoring of them. If someone from across the room was to suddenly call your name, you would notice the sound and respond to it immediately - the sudden peripheral stimulus would move from the periphery into your focus of attention.

Similarly, calm technology will easily navigate between our peripheral attention and focus of attention. Peripheral information is informative, but since it resides just off of our primary focus, it avoids being burdensome. By designing for the periphery, users can sense and control what immediately interests them while retaining peripheral awareness of other relevant information that can be brought back into focus at any time. Calm technology enables the user to fully command technology without being dominated by it [39]. To illustrate this, Weiser gives the example of how a video conference may be a calmer interface than a phone conference because with the phone, participants are never quite sure who has entered or left the room at the other end [40].

2.5 Attentive User Interfaces

The Attentive User Interface (AUI) is an exercise in calm technology [40]. AUIs provide a design paradigm for contending with the problems that arise when multiple computers vie for the attention of one or more users. This problem arises because each device behaves as if it were the *sole* recipient of the user's focus of attention.

The proliferation of mobile devices, such as cellular phones, provides a commonplace example for the need of more attentive devices. The inattentiveness of cellular phones stems from three problems [34]: Firstly, cellular phones are unaware of the user's attentive state, and thus often interrupt the user at inopportune times. Secondly, cellular phones have limited notification channels; the phone may have only a handful of alternative notification choices. Lastly, the existing notification channels do not provide any subtlety of expression. The result of this inattentiveness is a familiar example to many: your cellular phone interrupts you at inappropriate times such as during a meeting with colleagues. AUIs turn to human social conventions in an attempt to minimize interruptions and maximize the relevance of the information delivered by basing their interruptions on the user's level of availability.

2.5.1 Human-Human Interactions

Unlike computers, humans are highly attentive beings and abide by a set of conventions for mediating each others attention in social group settings. These conventions allow one to infer the attentive state of another. For example, say you wanted to converse with a colleague in their office, only to find the colleague's office door closed. Typically the closed door is interpreted as a signal of unavailability. If your need to communicate is important, you may knock on the door as a request for attention; if your colleague is available, he/she may answer the door. If there is no answer, it signals unavailability and you know to come back later.

Humans express and perceive attention through nonverbal cues such as inflections in the voice, body gestures, and eye contact [30]. The variety of nonverbal cues allows for considerable subtlety in our requests for attention. As an example, let us revisit the office scenario, only this time you arrive to find the door open, but your colleague is engaged in a conversation with another party. You may position yourself within line-of-sight of your colleague, thus using your proximity and eye contact as a nonverbal request for attention. If your presence is ignored and the information you wish to convey is important, you may move from the initial nonverbal request and wait for a suitable pause in your colleague's conversation before interrupting verbally. Surely your initial request for attention would *not* be to barge into your colleague's office and interrupt the existing conversation. This may seem like an obvious social *faux pas*, yet this is *exactly* what many computing devices do today. Like humans, AUIs address this problem by weighing the importance of the information they convey against the context of the user's current activity.

Of these nonverbal attentional cues, eye contact has been shown to be highly correlated with one's focus of attention [30, 34]. Eye contact provides a peripheral cue for attention without interrupting the central verbal-auditory channels. It has been shown in human-human group conversations that the availability and quantity of eye contact provides a protocol for establishing a communal focus of attention between a speaker and the listener(s). For example, a public speaker seeks out eye contact as a request for the audience's attention. The audience acknowledges the speaker's request by providing sustained eye contact. Upon receiving sufficient eye contact from the audience, the speaker may infer the audience has granted his request for attention and yielded the floor to him. In this manner, both the speaking and listening parties gracefully negotiate a mutual focus of attention.

2.5.2 Human-Device Interactions

In human-to-device verbal interactions, humans exhibit similar nonverbal attentional cues. In a Wizard-of-Oz experiment [23], an attentive office with multiple devices was fabricated and users were instructed to complete a series of tasks with the attentive office machines. As with human-to-human interactions, the relationship between eye contact and focus of attention was maintained; gaze information alone successfully disambiguated the device at the user's focus of attention 98% of the time [23].

Thus the prominence of visual attention in human behaviour communicates a great deal about user activity simply through observation of the user's gaze. By enabling devices to recognize existing nonverbal attentional cues, the need for explicit input from the users may be reduced. As eye gaze is essentially synonymous with a user's interest in a target, it provides a natural trigger for human-to-device interactions. The Attentive TV (ATV) is a device that uses eye gaze in this manner [34, 31, 29]. Through the use of eye contact sensors, the ATV governs the playback of media based on the presence of eye contact.

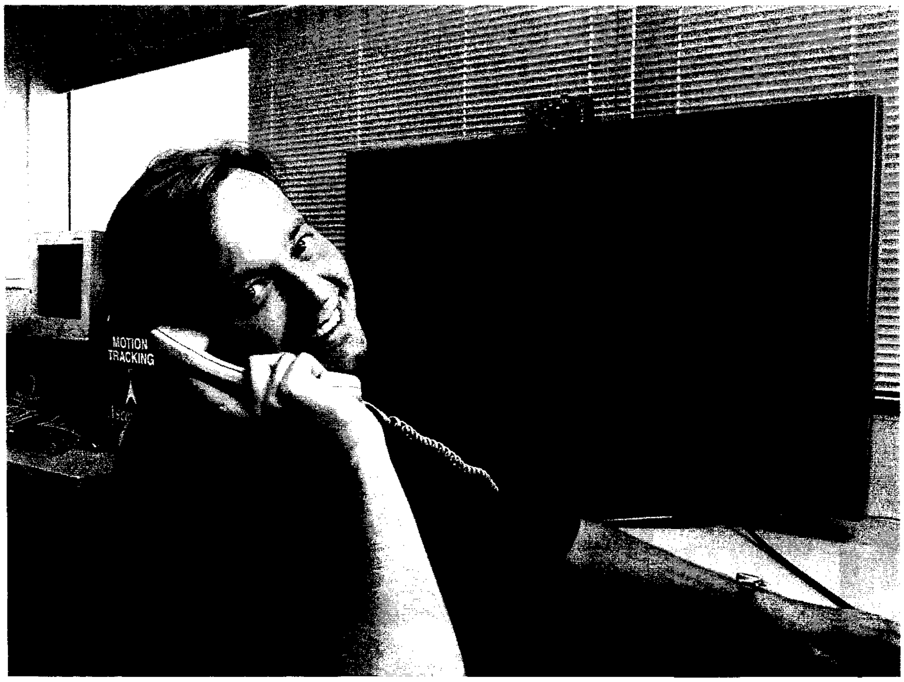

Figure 2.3: Attentive TV [31]

For example, say a user is watching a DVD on the ATV when he is interrupted by a phone call. If the user chooses to attend to the call and turn away from the ATV, the prolonged absence of eye contact signals to the ATV that it is no longer the focus of the user's attention. Consequently, the ATV pauses playback *Figure 2.3*. When the user is done with the call and turns back to look at the ATV, the sustained eye contact indicates the user's interest in the ATV and playback resumes. By using eye contact as an interaction trigger, the ATV is a TV that watches you!

Chapter 3

The Human Visual System

Using a computer system as an analogy, the complexity of the human visual system may be coarsely decomposed into two components. The eye constitutes the lower-level hardware used to sense our surroundings, while the brain constitutes the higher-level software used to make sense of our surroundings. Essentially, the eye is a light sensitive sensor that collects visual information from the surrounding environment and sends it to the brain for analysis. The brain interprets the incoming signals from the eyes and attempts to make sense of it. The brain is responsible for constructing and governing our visual attention, mainly *where* to look, and *what* are we seeing.

3.1 The Human Eye

The structure of the human eye is analogous to that of a camera. The body of the exposed eye is formed by the sclera, a sturdy white opaque membrane that surrounds the cornea. The sclera surface is coarser than the cornea and has a smaller curvature.

The cornea is a transparent membrane that is shaped like a small segment of a larger, imaginary sphere. The cornea acts as the primary lens which performs rudimentary focusing of the incoming light.

Fine adjustment to the eye's lens is governed by a muscle called the zonula, which controls both the shape and positioning of the lens and consequently how the light entering the eye is focused on the retina.

The iris is analogous to a camera's aperture. It is a muscle that when contracted, covers all but a small portion of the lens. This serves to dynamically control the amount of light that enters the eye. In brightly lit environments, the iris will close down to restrict the amount of light that reaches the retina, while in dim environments, the converse is true and the iris opens up. Where the iris does not occlude the lens is a small hole called the pupil.

The human analogy to film is the retina, a photoreceptive layer composed of rod and cone cells. The retina is located at the back of the eye and is where all incoming light is focused and converted into nerve signals. At the centre of the retina is a small depression called the macula lutea. The macula is composed mainly of cones. Central to the macula is the fovea, the part of the retina with the highest density of cones and consequently the highest resolving power. Eye gaze is determined by the line connecting the fovea to the centre of the lens.

3.1.1 The Retina

The retina is composed of several thin layers of specialized cells that line the interior of the eye. Some of these cells serve specialized tasks such as converting incoming light into nerve signals, while other specialized nerve cells transmit these signals to the optic nerve, and ultimately the brain. The arrangement of layers in the retina is counter-intuitive, with light passing through processing and communication layers before falling on the photosensitive layer. In between all these layers, are connection bundles called plexiform or synaptic layers.

Working from the interior of the retina outwards, the outermost layer is the retinal pigment epithelium (RPE) and is primarily a protective layer. It is connected to a blood vessel rich layer called the choroids that provides the blood supply that nourishes the outer two thirds of the retina and supports the function of the photoreceptive layer.

Next is the visual layer, consisting of photoreceptors that convert light energy into electrical impulses (neural signals). These photoreceptors may be functionally generalized into two classes, rods and cones. There are on average 120 million rods, and 6 million cones. Rod cells are very photosensitive and are the primary contributor of vision at low light levels. Cone cells are less photosensitive and operate best at average light levels. Cone cells are the primary contributor of vision in daylight and allow for colour vision. As one approaches the centre of the retina, rod density diminishes while cone density increases. A marked increase in cone cell density is found in the fovea, with density peaking in the fovea centralis. As a result, for an object to be seen at high resolution, the image of the object must align with the fovea centralis. The photoreceptors of the visual layer are interconnected by horizontal cells. Horizontal cells allow lateral signals to be exchanged between adjacent receptors.

Ganglion cells serve as terminators for the nerve fibres connecting to the brain. The ganglion layer is connected to the visual layer via biopolar cells. The biopolar cells are part of the inner nuclear layer, a communication layer where the horizontal and amacrine cells also reside. Bipolar cells serve to funnel the electrical signals from groups of photoreceptors to a single ganglion cell. Throughout most of the retina, biopolar cells gather the signals from several photoreceptors; however, in the fovea there is typically a one-to-one correspondence. Ganglion cells themselves are interconnected by amacrine cells.

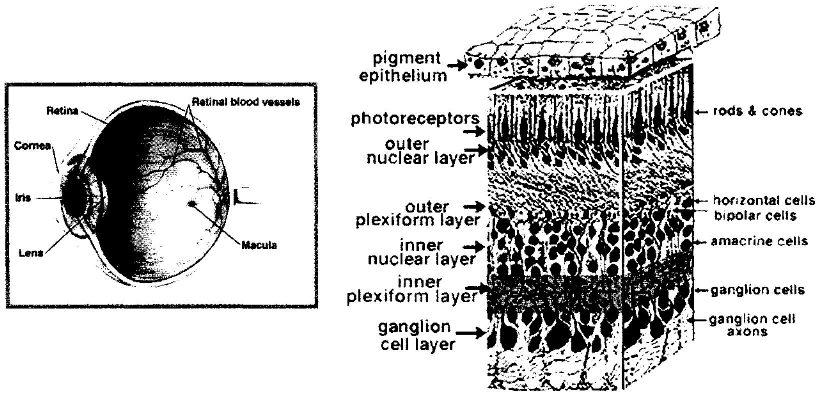

Figure 3.1: Retina Structure [4]

3.2 Human Visual Attention

"Attention to an object is what takes place whenever that object most completely occupies the mind" William James [17]

Attention is the cognitive process of selectively concentrating on one thing while ignoring others. As mentioned, a common example is the cocktail party effect [38, 36], whereby one may listen to what someone is saying while ignoring other noises and conversations in the surrounding environment. Attention may also be distributed, such as when a person is driving a car while simultaneously attending to a cell phone.

Although there are many cognitive processes associated with the human mind, attention is the most closely tied to perception.

Early efforts to study the nature of attention were greatly limited by technology. These studies often consisted of simple empirical observations and speculative philosophical introspection. As technology advanced, the field grew to include several disciplines such as psychophysics, cognitive science, and computer science [4].

3.2.1 Hermann von Helmholtz: Where to look?

"We let our eyes roam continually over the visual field, because that is the only way we can see as distinctly as possible all the individual parts of the field in turn" Hermann von Helmholtz

Hermann von Helmholtz's view of attention was primarily concerned with not what object was at the focus of attention, but rather the object's spatial location. Specifically, he was interested in the role of attention in determining what location received our focus.

Typically, eye movements provide explicit cues regarding our visual attention. However, Helmholtz was the first to observe that humans are also capable of consciously focusing their visual attention on a peripheral object *without* moving our eyes [4]. Peripheral vision refers to all visual experiences outside the immediate line of sight.

For example, let us consider a variation of the Titchener illusion in *Figure 3.2.* Fixate briefly on circle A and repeat for circle B. Circles A and B are in fact the same size, but since A is surrounded by smaller circles, we perceive it as being larger than B. In the case of A, although you were fixating on it, you were also aware of the surrounding smaller circles residing in your peripheral vision. You are aware of these smaller circles yet no eye movements were made - your visual attention was shifted briefly from centre into the periphery.

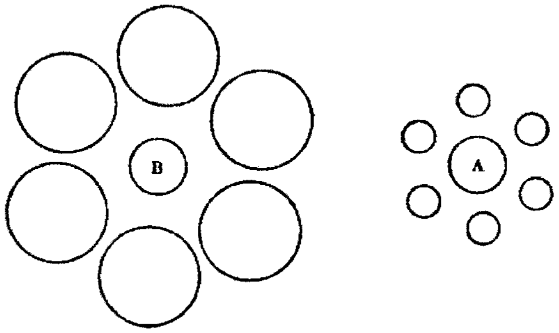

Figure 3.2: Titchener Illusion

Although we may shift the focus of our visual attention into the periphery without moving our eyes, more commonly, we tend to choreograph our eye movements such that the pupil centre aligns with our focus of attention. This is understandable as the most visually acute portion of the eye is aligned with the pupil centre (fovea). Coaxial alignment of the focus of attention with the fovea allows the object under scrutiny to be examined with the full resolving power of the eye. Since the fovea represents a minute portion of field of view, the eye naturally tends to wander over a scene in order to compose a high-resolution impression. Eye movements are an explicit cue for inferring the state of our visual attention: what we are interested in is defined by *where* we look [4].

Helmholtz's interpretation of attention is Gestalt in nature. Gestalt itself is a

German word meaning (approximately) *form*, or *whole figure.* Gestalt psychology describes a method of understanding psychological phenomena not through the summation of their constituent parts, but rather as organized forms or structured wholes.

3.2.2 William James: What are we looking at?

"Everyone knows what attention is. It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, or consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others." William James [17]

One of the earliest investigators of visual attention was William James. James viewed attention, and the process by which we determine what receives our attention, as the unification between incoming stimuli and the expected characteristics (pre-existing impressions) associated with the said stimuli.

James viewed attention, and the process by which humans determined the recipient of its focus, as the unification between incoming stimuli and pre-existing impressions, or expectations, of the characteristics of the aforementioned stimuli [17]. Although humans may maintain several objects in their peripheral attention, "we are limited to focusing on only one such object at any given time" [17]. How an object reaches the focus of attention can be viewed as three stages.

The first stage may be viewed as a precursor to conscious awareness of the object. According to James, before an object enters our conscious attention, it will first trigger premonitory images such as where one should look if the object is to become the primary focus, or what to expect to see. The *what* was James' primary interest, whereas the *where* was more aligned with von Helmholtz's interpretation of attention. The virtual object depicted by the premonitory images prime the same nerve centres of the brain that would respond to the actual object.

As an object enters our attention, it is re-enforced by both the preliminary excitation of the mind's impressions as well as incoming stimuli from the real object. In this intermediate stage, an object generates sufficient excitement for one to become aware of the object; however, one may not care for it as the object may not be of vital importance or relevance to the task at hand. The interest in the object is marginal, and consequently, it remains in the background of our visual attention. At any given moment, there may be several such objects in the periphery of our visual attention, each supplying premonitory images and impressions of their own.

In the last stage, the shift from the periphery to the focus of attention may be viewed as a convergence of the incoming stimuli with the premonitory impressions of an object. Our attention is excited both internally and externally; incoming currents from the periphery arouse it, and collateral currents from the centres of memory already excited by the previous two stages re-enforce it [17]. In this process, James viewed the incoming stimuli as:

> *"In this process the incoming is the newer element; the ideas which re-enforce and sustain it are among the older possessions of the mind. And the maximum [focus] of attention may then be said to be found whenever we have a systematic harmony or unification between the novel and the old."* [17]

3.2.3 Yarbus: How do we look at it?

Helmholtz's *where* and James' *what* provided early insights in the mechanisms of the human visual system. However, *how* humans look at things, or the manner in which we examine the objects at our focus of attention, also conveys non-verbal cues about our interests.

As mentioned earlier, the fovea is the portion of the retina with the highest density of cones per ganglion cell connection. Thus the fovea holds the highest visual acuity in the retina. Only the portion of the visual scene befalling the fovea is perceived with high resolution. When examining an object, the object is typically orders of magnitude larger than the narrow field of view covering the fovea. In order to compose a high resolution impression of an object, it is necessary to sweep our gaze over the object such that all portions of it fall, albeit briefly, over the fovea. These movements are carried out in a series of sudden ballistic movements called *saccades*, interspersed with stationary *fixations*; saccades last on the order of tens of milliseconds (60 msec), while fixations last on the order of hundreds of milliseconds (100-1000 msec). Saccades facilitate rapid relocation from one region of the visual scene to another, while fixations facilitate the examination of the region once it is foveated.

Eye movements allow us to scan the visual field and focus our attention on the parts that represent the most significant information to the task at hand. An early investigator of eye gaze movement and patterns was A. L. Yarbus. Yarbus investigated the nature of eye gaze movement with an apparatus resembling a cap that was affixed directly to the eyeball of his subjects [4]. A small mirror mounted on the cap reflected a light beam onto a sheet of photo-sensitive paper. When developed, the photographic paper would depict a record of the subject's eye movements.

In his most famous study, Yarbus [42] demonstrated how eye movement patterns reveal a great deal about the current task occupying the subject. The subject was asked to inspect a painting several times (*Figure 3.3*), with each iteration requiring the subject to perform a different visual search and note various aspects of the painting. Tasks included a free inspection of the painting without a predefined purpose, taking note of the ages of the figures in the painting, noting the clothes adorning the figures, and observing the relative position of people and objects in the scene. The resulting trace patterns and fixation points were clearly correlated with the goal assigned to the subject. For example, when given the task of determining the ages of the figures, the majority of fixations and eye movements were concentrated on the faces of the figures (*Figure 3.4a*). Similarly, when asked to note the clothing adorning the figures, eye gaze was primarily confined to the upper and lower torso of the figures(*Figure 3.4b*).

The ideas put forth by Helmholtz and James provided the first steps in our understanding of human visual attention. Together, they provided insight on how our attention was distributed as well as the mechanisms that determine the recipient of our attention. Yarbus furthered our knowledge by investigating the methods and techniques humans use to examine the things at the focus of attention. The contributions of these pioneers, and their contemporaries, provided the first steps toward our understanding of attention.

Figure 3.3: *An Unexpected Visitor* (I.E. Repin) [42]

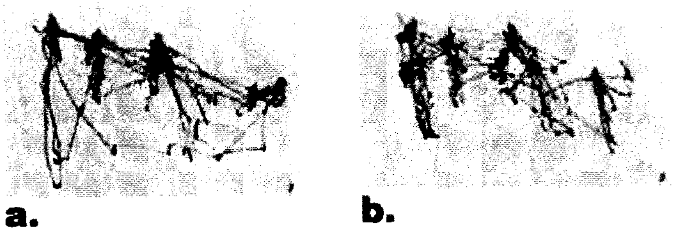

Figure 3.4: Age and Clothing [42]

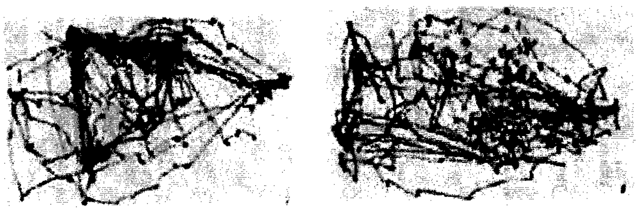

Figure 3.5: Free-examination and Position [42]

Chapter 4

Sensing Visual Attention

4.1 Eye Tracking

Although there are several cues for indicating attention, visual attention is most strongly communicated to other humans through eye contact [10, 4]; humans tend to look at what interests them. Eye contact sensing devices are typically low cost devices that are capable of perceiving this visual, nonverbal attentional cue. It should be made clear that eye *sensing* is not to be confused with eye *tracking*. Eye sensing may be viewed as a limited subset of eye tracking. With eye sensing, typically only a binary awareness is available. For instance, the device may know that eyes are present but not where they are looking, or if eyes are looking at the device or not. Compared to eye sensing, eye tracking technologies afford more detailed examinations of a user's visual attention. Analysis of eye movement behaviour, such as scan patterns and fixation periods, provide valuable clues into the viewer's attentional awareness.

4.1.1 The Bright Pupil Effect

The most commonly used method of sensing visual attention (eye tracking) exploits the retro-reflective characteristics of the eye. Most mammalian eyes are classified as retro-reflectors, meaning that light incident to the retina will be returned along the same path as the incoming source. This physiological phenomenon is commonly seen in flash photography as the "red eye" effect, an example of the *bright pupil* effect. To solicit the bright pupil effect, an illuminator placed along the central axis of the camera lens is directed at the eye. The lens of the eye focuses this incoming light through the pupil and onto a point on the retina. The retro-reflective nature of the retina reflects a large portion of the incident light back through the pupil, along the same path it entered. The resulting effect is that the pupil becomes illuminated and appears to the camera as a bright disk, noticeably contrasted against the surrounding iris. Additionally, there is a second reflection off the cornea surface. However, since the corneal surface lacks the retro-reflective property of the retina, this second reflection is significantly smaller and called a *glint.*

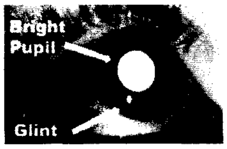

Figure 4.1: Bright-Pupil Effect [24]

4.2 Eye Tracking Overview

The development of eye tracking technology can be classified as three distinct periods. The first generation of eye trackers measured the position of the eye relative to the position of the head. The subject's head had to be fixed in way to ensure that the eye's position (relative to the head) coincided with the point of regard (POR). These early efforts employed invasive technologies such as electro-oculography or scleral contact lens for measuring eye movement, and restrictive measures such as bite bars and chinrests to stabilize the head.

The second generation of eye trackers made the first attempt of tracking the eye independently from the head. Instead of measurements made relative to head position, multiple ocular features were measured in order to disambiguate head movement from eye movement. Photo-oculography and video-oculography analysis are based on frame-by-frame visual inspection of photographic or video content. However, as the analysis is a manual effort, it is time consuming and prone to error, and makes real-time calculations of the POR difficult.

Third generation eye trackers employ analog video-based oculography combined with pupil/corneal reflections, and consolidate the decoupling of the eye from the head. Measurements of the eye in space are rapidly and accurately analyzed with image processing hardware, allowing for the first time real-time POR calculations.

4.2.1 Electro-oculography (EOG)

EOG employs electrodes to measure the electrical potential differences of the muscles in the ocular cavity. Although still in use, the application of EOG has steadily declined, a stark contrast compared to only forty years ago when EOG was the most widely applied technique for eye movement measurement. As EOG measures the position of the eye relative to head position, head movement must be constrained or an additional tracker must be used to obtain the head position. These additional complications of EOG make it generally not well suited from measuring POR.

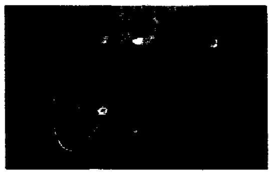

Figure 4.2: Electro-oculography [4]

4.2.2 Scleral Contact Lenses (Search Coil)

This method involves attaching a mechanical or optical reference object to a contact lenses which is then worn directly on the eye. Applications of this technique have been reported as early as circa 1898, where a ring of gesso (plaster of Paris) was affixed atop the cornea and connected by mechanical linkages to pens for recording [4]. In the 1950s, advancements to the technique introduced contact lenses to improve accuracy. Devices such as coils of wires, or small mirrors were attached to the contact lenses. The physical contact of the lenses with the eyeball provides very sensitive measurements of eye movement.

The technique has continued to evolve, with the current embodiment employing modern contact lenses with an attached mounting stalk. The stalk facilitates the attachment of various mechanical or optical devices. Of the possible implements, magneto-optical configurations are the most popular. The underlying principle uses a wire coil that is immersed in an electromagnetic (EM) field. The wire coil acts as a transformer and as it moves within the EM field there is an induced voltage that is proportional to the sum of the horizontal and vertical magnetic fluxes traversing the coil.

As this method provides direct contact with the eyeball surface, it is one of the most precise eye movement measurement techniques, albeit at the cost of comfort. The invasiveness of this method often requires the eye to be anesthetized and trials can only be conducted for about twenty minutes. Furthermore, the coils are delicate and expensive, requiring care and practice for proper insertion. Like the EOG technique, eye measurements are made with respect to head position, thus incurring the same utility limitations as EOG.

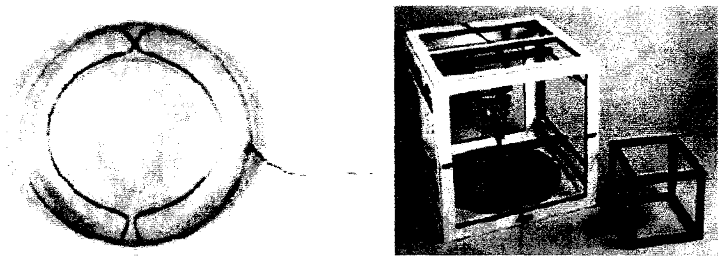

Figure 4.3: Scleral Coil [4]

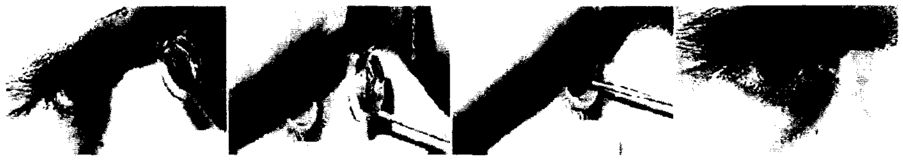

Figure 4.4: Insertion of Scleral Coil [4]

4.2.3 Photo-oculography (POG)/Video-oculography (VOG)

The human eye has several distinct visual features that move systematically in relation to each other. Features such as the shape of the pupil, the position of the limbus (iris-sclera boundary), and the corneal reflections from directed light sources (in close proximity) provide a cues for measuring eye movement independently from head position. Ocular movement traces may be recorded manually or automatically for off-line analysis. Unfortunately, this off-line analysis is more often than not a tedious, frame-by-frame visual inspection of the recorded material. The temporal resolution is limited by the sampling rate of the recording device. As the analysis component of this method is manual labour intensive, not only is it tedious, but is error prone and consequently unsuitable for real-time calculation of the POR [4].

Figure 4.5: Video-Oculography [4]

4.2.4 Video-based Combined Pupil/Corneal Reflection

Video-based combined pupil/corneal reflection is the most widely applied apparatus for measurement of the POR. Its popularity is in a large part due to two factors: firstly, unlike other previous methods, measurements are made remotely and thus are non-invasive; secondly, the POR analysis can be done manually offline, or automatically in real-time.

Utilizing relatively inexpensive cameras, image processing hardware, and the high clock speeds afforded by modern computers, eye tracking systems based on this technique have greatly improved the accuracy with which eye movements can be measured. The apparatus may be stationed on a desktop or portable with a wearable head-mount; both embodiments are essentially identical with the exception of size.

Corneal reflections are known as the Purkinje reflections/images and are due to the physiology of the eye. There are four Purkinje reflections in all. The corneal reflection of a directed light source is measured with respect to the location of the pupil centre. In the case of video-based eye trackers, typically only the first Purkinje image is measured. Given a suitably positioned planar surface on which calibration points are displayed and recorded with appropriately, eye trackers using this technique are capable of measuring the viewer's POR in real-time.

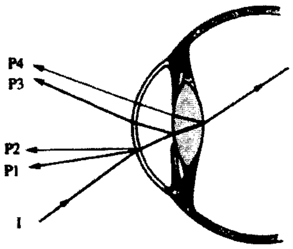

Figure 4.6: Purkinje Images [4]

4.3 Early Corneal-Reflection Eye Tracking Methods

Previous eye tracking systems based on the bright-pupil effect had used an illuminator that was offset from the central camera axis. As the bright-pupil effect relies on alignment of the light source with the central camera axis, a semi-silvered mirror tilted at 45 degrees had to be mounted in front of the lens to redirect light along the camera axis and into the eye. This early method achieved the bright-pupil effect, however, the semi-silvered mirror reflected away half of the reflected light from the eye, effectively reducing the brightness and clarity of the camera image.

4.4 Modern Corneal-Reflection Eye Tracking Methods

In 1986, Hutchinson developed the idea of placing an illuminator, a small IR LED, in the centre of the camera lens [14]. This simplified the equipment configuration and since the LED blocked only a small percentage of the camera aperture, the image of the eye was brighter and clearer. Hutchinson estimated the gaze direction by calculating the corneal reflection/pupil (CR-P) vector.

Given a fixed light source, the position of the corresponding glints remains almost fixed in the image as long as the user's head remains stationary. Conversely, the position of the pupil moves to follow the rotation of the eyeball.

For example, if the subject were to look directly at the camera, the glint would appear in the centre of the pupil image. Now consider if the subject looked above the camera. The pupil centre would move upward against the fixed position of the glint. This phenomenon is irrespective of the dark and bright pupil effects. In both bright pupil and dark pupil conditions, the gaze direction relative to the camera may be estimated based on the relative positions of the pupil and glint centres. The simplified design of Hutchinson's eye tracker provided the archetype for several improved designs. Although these later systems marked an improvement over their predecessors, the techniques employed to estimate the user's POR introduced a necessary evil: calibration. As each person's eyes have unique characteristics (e.g. size, corneal curvature, etc.), with conventional CR-P vector techniques calibration is required to compensate for these interpersonal differences.

4.4.1 The Purpose of Calibration in Eye Tracking

The purpose of calibration is to present the subject with a series of visual points at fairly extreme viewing angles. As the characteristics of each person's eye are unique, each eye reflects light slightly differently, especially at extreme points. Calibrating for each unique user aids the eye tracker in maintaining an adequate level of accuracy while interpolating the subject's POR between extreme points. The general calibration process is as follows [4]:

1. Adjust the eye tracker's pupil and corneal reflection threshold controls
2. Calibrate the eye tracker
3. Reset the eye tracker and run the application
4. Save recorded data
5. Optionally calibrate again

4.4.2 Eye Tracking Methodology

One drawback of Hutchinson's system was that it relied solely on the bright pupil effect. Although the pupil does become illuminated via the bright pupil effect, it does not differ sufficiently in brightness from its surroundings to make detection trivial.

Tomono developed a system involving three CCD cameras (C_1, C_2, C_3) and a pair of infrared illuminators [33]. The two illuminators emitted light of slightly different wavelengths (λ_1, λ_2), with λ_1 being polarized. The unpolarized light source, λ_2, was placed coaxial with the camera lens, while the polarized light source, λ_1, served as the off-axis illuminator. Each of the three cameras were made sensitive to separate attributes of the λ_1 or λ_2. C_1 was only sensitive to λ_2 and captured only the bright pupil images. C_2 and C_3 were sensitive to λ_1, with filters blocking out λ_2. C_2 captured the dark pupil image, while C_3 had an additional polarizing filter that blocked all but diffuse light components. The effect of the polarizing filter on C_3 was to remove specular components, such as the corneal reflections, from the captured image. The pupil was found by subtracting C_1 and C_2, the bright and dark pupil images, respectively. For gaze estimation, the corneal reflection is also required. This was obtained by subtracting C_2 from C_3. Gaze estimation was calculated via a corneal reflection/pupil (CR-P) vector.

Like Tomono, Ebisawa addressed this problem by introducing multiple illuminators [6]. However, Ebisawa only employed one camera and like Hutchinson's design, illuminators placed coaxial with the camera are used to solicit a bright pupil. A key difference was the use of an additional set of non-coaxial (off-axis) illuminators to solicit a dark pupil effect. The difference of the on-axis and off-axis images yields an image where the pupil figures prominently with respect to its surroundings. This image difference provided the input to Ebisawa's pupil detection algorithm. To further aid pupil detection, Ebisawa used a zoom lens with a narrow field of view (FOV) to provide a large image of the pixel while reducing the image search space. Gaze estimation was calculated via a CR-P vector.

Morimoto [24, 25] proposed a similar system (i.e. PupilCam) using a different geometric distribution of the illuminators. Instead of the on-axis illuminators being coaxial with the camera lens, they were distributed around the lens. Although the on-axis illuminators are not coaxial, its close proximity is sufficient to still solicit the bright pupil effect. Furthermore, a lens with a large FOV was used to allow multiple pupils to be detected. Gaze direction was estimated by first calculating the CR-P vector and then correcting for distortions induced by the curvature of the eye by analyzing the results of a calibration pattern. The calibration pattern consisted of a three-by-three point grid that the user sequentially fixated on. The calibration procedure provided a means for corresponding eye movements to scene coordinates. When calibrated well, the system was able to achieve 1^o of spatial resolution, which equates to approximately 1cm at a viewing distance of 50cm. However, the simplicity of the calibration pattern was inadequate to afford acceptable human-computer interaction in a natural environment.

The intolerance to head movement and tedium of calibration pose a large usability obstacle and limits its applications. With the aid of an eyeball model, Ohno [26] reduced the calibration process to two points. The gaze estimation process was composed of three steps. The initial stage examined the captured image and detected the location of the pupil and glint of each illuminator. Next, an eyeball model was generated based on the pupil and corneal reflection locations. This eyeball model provided a three-dimensional gaze coordinate. Finally, the two-point personal calibration was used to correct the eyeball model and accurately estimate gaze direction. With a good calibration, the system was capable of achieving 0.3 - 0.6 degrees of accuracy. A shortcoming of this process was the lack of support for subjects using soft contact lenses. Soft contact lenses produced two glints for each illuminator - one in front of the lens, and another in front of the cornea - making it difficult to correctly identify the correct glint.

4.5 Sampling of Eye Tracking Systems

Although there is a wide diversity of commercial eye tracking technologies available, most can be categorized as either a desktop system, or a mobile wearable system.

Desktop systems are the most prevalent. Typically, they consist of a camera mounted below the visual display, with illuminators adorning the camera body [14], or embedded into the visual display itself [15]. Desktop systems achieve real-time performance, typically communicating the user's POR in the form of screen coordinates. Average accuracy of desktop systems is approximately 1cm at 60cm. However, there are several drawbacks to conventional desktop systems. As the eye tracker is confined to a stationary location, the user must remain within an operational volume for optimal performance. As well, there is usually a small tolerance for head movement, typically averaging 30cm. Tracking accuracy quickly degrades as one approaches the bounds of the operational volume, or the limits of head movement tolerance. Furthermore, the systems are quite expensive, typically in the tens-of-thousands of dollars.

Mobile or wearable eye tracking systems typically employ head-mounted optics that communicate to a wearable computer system [19]. Unlike desktop systems that provide coordinates to real-world objects such as a location on a visual display, head-mounted systems provide coordinates with respect to the user's head or a reference scene image. Thus head-mounted systems cannot detect what object or spatial location is at the user's POR, instead they provide a direction or vector in which the user is looking and further processing is required to identify if an object of interest lies along its path.

The need for calibration inherently makes eye tracking difficult for the mass public to adopt. Coupling that hinderance with the restrictive operational characteristics and high unit cost prevents conventional eye tracking from integrating into more commonplace applications and becoming ubiquitous.

Below is a brief sampling of both desktop and head-mounted eye trackers. These examples were chosen as they represent a good cross-section of the current state-of-the-art.

4.5.1 LC Technologies EyeGaze

In 1988, LC Technologies introduced the first commercial, PC-based eyetracker to use the pupil corneal reflection technique (i.e. bright-pupil effect) [14]. A video camera located below the computer screen observes the eye which is illuminated by a small, low power, infrared LED located at the center of the camera lens. The illuminator generates the corneal reflection and causes the bright pupil effect, which enhances the camera's image of the pupil. Specialized image-processing software identifies and locates the centres of both the pupil and corneal reflection. Subsequent trigonometric calculations project the user's gazepoint based on the positions of the pupil center and the corneal reflection within the video image.

Table 4.1: LC EyeGaze Specifications

Sampling Rate	60Hz
Spatial Resolution	0.45^o - 0.70^o
Effective Accuracy	1^o
Head Movement	3.8cm x 3.0cm x 3.8cm

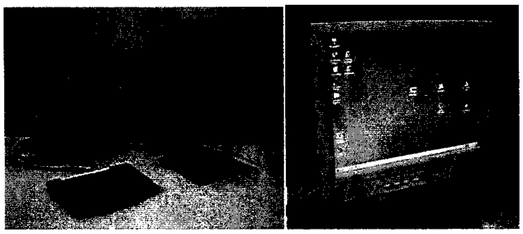

Figure 4.7: LC Technologies EyeGaze Eye Tracker [14]

4.5.2 Applied Science Laboratories H6

In some cases, such as mobile environments, remote eye trackers are inadequate since they are confined to the desktop. Head-mounted eye trackers offer an alternative solution for situations where the subject may wear lightweight, head mounted optics and must have unrestricted freedom of movement. The optics are necessarily lightweight and often mounted on an adjustable headband. An example of typical head mounted eye tracker is the H6 from Applied Science Laboratories (ASL) [19]. The ASL H6 records the scene with a colour camera that is mounted on an adjustable headband. The system measures the user's line of gaze with respect to the head. Unlike remote, realtime eye trackers, the H6 stores the data for later analysis offline. The only realtime indicator available is a broadcast of the camera scene with a displayed cursor, or set of cross hairs superimposed over the scene image of the user's pupil. To determine if a user has looked at a visual target, the target's location in 3-space must be known *a priori*. For further detailed analysis, a magnetic tracker is needed to determine head orientation.

Table 4.2: ASL H6 Specifications

Sampling Rate	50-60Hz (60Hz Default Configuration)
Spatial Accuracy	0.5^o
Spatial Resolution	0.1^o
Head Movement	N/A
Visual Range	50^o (Horizontal) 40^o (Vertical)

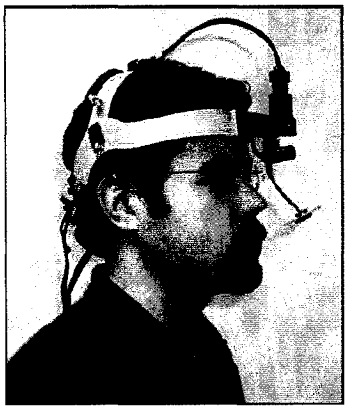

Figure 4.8: Applied Science Laboratories H6 Wearable Eye Tracker [19]

4.5.3 Tobii Eyetracker 1750

The Tobii 1750 is arguably the current state-of-the-art in eye tracking [15]. It employs binocular tracking that observes both eyes simultaneously, automatically determining which eye is left and which is right. Binocular tracking affords greater tolerance to head motion since tracking may continue in the event one eye becomes occluded from the camera's field of view. The scene camera is integrated into an LCD display with clusters of infrared LEDs located along the top and bottom frame of the LCD display.

Table 4.3: Tobii 1750 Specifications

Sampling Rate	50Hz
Spatial Resolution	0.25^o
Effective Accuracy	1^o @ 50cm
Head Movement	30cm x 15cm x 20cm @ 55cm
Camera Field of View	20cm x 15cm x 20cm @ 55cm
Max. Gaze Angle	$\pm 35^o$

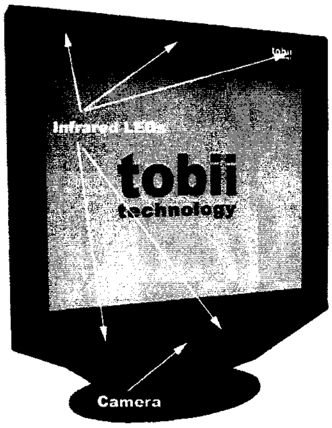

Figure 4.9: Tobii 1750 Eye Tracker [15]

4.6 The Need for Calibration-Free Eye Tracking

Prolonged gazing at a calibration point is both difficult physiologically, and unnatural behaviourally. Physiologically, the eye is rarely at rest. Even when we consciously fixate on an object, the eye exhibits slight jitters called microsaccades. These miniature eye movements prevent retinal adaptation that would result in the visual image gradually disappearing - humans must constantly scan the visual scene to maintain a sufficient level of excitation. Such miniature eye movements may be demonstrated by fixating on any single white dot in *Figure 4.10.*

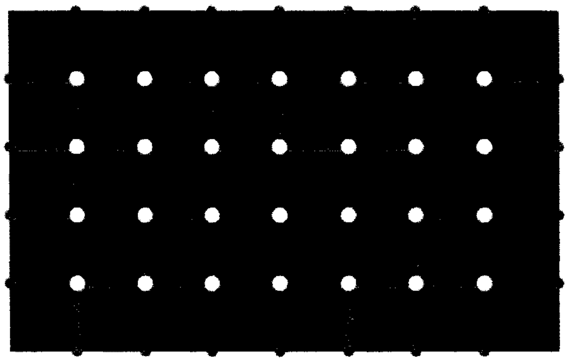

Figure 4.10: Hermann Illusion

Doing so, it appears gray dots appear within the local neighbourhood of the white dot, while black dots appear further away. Any attempt to fixate on a gray/black dot instantly removes it. The illusion of the gray/black dots is in part due to the involuntary miniature movements (e.g. microsaccades) of the eye.

Furthermore, gazing at a point on the screen often causes fatigue. Behaviourally, no other input device, such as keyboards or mice, require static fixation: calibration is an unnatural user interaction [26]. Moreover, the second-order or higher approximations typically used to solve for gaze direction may affect robustness [18]. These approximations are calculated from the calibration patterns; however, since the user may not gaze at the calibration point accurately, error is introduced into the calibrated data. This error may increase further as the calibration process continues, further hindering the accuracy of the gaze estimation.

4.6.1 A Neural Network Approach

Ji and Zhu [18] addressed this need by proposing a calibration-free (CF) system based on combining head pose information with gaze information. The system afforded natural head movements and was unobtrusive. The neural network was trained to recognize and associate specific pupil/glint relationships with a corresponding gaze direction. As these relationships are consistent for all humans, new users to the system could begin using the system immediately without undergoing a personal gaze calibration procedure. Similarly, if a user shifted his/her position while using the system, re-calibration was not necessary. Head pose provided the global gaze direction, while eye gaze provided refined local gaze information. Face pose was determined by examining various geometric and spatial relationships between pairs of pupils. Local gaze determination consisted of a three stage process. The first stage involved the detection and tracking of the pupil and its glints via the bright/dark pupil effects. The second stage was a gaze calibration method involving neural networks. In this stage the spatial relationship between the glints and pupils, as well as the geometric properties of the pupil were parameterized into six variables. These parameters were input into a feed forward, generalized regression neural network (GRNN). GRNNs were used as they exhibit fast training times, the capacity to model non-linear functions, and have demonstrated good performance in noisy environments given sufficient data [18]. Lastly, to demonstrate the ability to use gaze for selection, the gaze direction was quantized into eight regions on the visual display, distributed as a 4x2 grid.

The quantization of the eye gaze into 4x2 necessarily results in a much lower resolution (about 5^o) than the previous methods discussed. Furthermore, the system employs a neural network that must be trained before the system can be deployed. The performance of the neural network is directly related to the diversity of the training set. For example, if only Caucasian and Asian subjects were included in the training set, system performance with subjects of other ethnicity may not be satisfactory. Thus gathering an adequate training set becomes increasingly difficult as the potential user pool grows. This drawback of neural network poses an obstacle for this method to become ubiquitous.

4.7 Eye Sensing Devices

As discussed earlier, there is a myriad of non-verbal cues expressed by humans for conveying attentional focus and availability(*Chapter 2.5.1*). Humans use these non-verbal cues to govern inter-human interactions. Like humans, devices that are attuned to these non-verbal cues, specifically eye contact, are able to infer attentional availability. Devices with eye sensing capabilities have no knowledge of where the user is looking other than a binary awareness of the user's visual attention: the user is either looking at the device, or not. By augmenting devices with eye contact sensing capabilities and applying AUI design principles (*Chapter 2.5*), we enable them to perceive a user's visual attention and self-govern their behaviour according to whether or not they are the recipient of the user's visual attention. The daunting task of determining a user's focus of attention is simplified to finding the pupil, and observing the centrality of the on-axis glint with respect to the pupil centre.

4.7.1 PupilCam

The PupilCam was designed as a low cost eye tracking system, relating a user's gaze to a coordinate system.

The bright pupil effect is induced through a set of IR LEDs located around the central axis of the camera. A second set of IR LEDs located away from the central axis provides off-axis illumination, or a dark pupil effect. By illuminating the on and off-axis IR LEDs sequentially, bright and dark pupil fields are produced in alternating images. Through subtraction of these two images and analysis by computer vision algorithms, the pupil location is extracted. Besides the bright/dark pupil effect, the on and off-axis IR LEDs also produce respective glints on the corneal surface.

The PupilCam uses the location of the on-axis glint with respect to the pupil centre to map the user's gaze to a coordinate system. For this mapping to be accurate, each user's unique eye attributes had to be compensated for by a calibration process [24, 25].

4.7.2 Eye Contact Sensor

The Eye Contact Sensor (ECS) was developed at the Human Media Lab at Queen's University [30, 34]. It is a light weight eye sensing device that is able to determine if it is the recipient of the user's focus of attention.

Like the PupilCam, the ECS also uses a similar set of IR LEDs to generate alternating bright and dark pupil effects. The method in which the ECS and PupilCam use these glints is what differentiates them. The ECS was designed not to estimate the user's point of regard (POR), but rather to augment other devices and enable them to identify when they were at the centre of the user's visual attention. Noting the discussion on visual attention in the preceding chapter, the pupil centre is often synonymous with the user's focus of attention. Since the on-axis glint is aligned with the central axis of the camera, when the pupil centre becomes co-located with it, it indicates that the user is looking at the ECS. Using this technique, the ECS does away with the calibration process required by the PupilCam, while maintaining its low cost attributes.

The simplicity of the ECS design, its low cost, as well as the lack of a calibration procedure, affords the ECS a wide range of applications in attentive, ubiquitous computing such as: the Attentive TV, a tv that uses eye contact to control playback of media [34, 31, 29]; AuraLamp, a lamp with speech recognition capabilities that only listens for commands when you look at it [30]; and ECSGlasses, a pair of glasses that uses eye contact as a trigger for recording the user's daily interactions with other people [29].

Figure 4.11: Eye Contact Sensor

Chapter 5

Implementation

This chapter describes the individual components of a calibration-free (CF) eye tracker prototype. The layout and function of each hardware component is discussed, albeit briefly as these components are not the focus of this thesis, as well as specification listings when appropriate.

Since the primary contribution of this thesis lies in the computer vision algorithms used in the CF method, a detailed examination is provided. The platform on which the software was developed is explained. Furthermore, extraction of eye features, as well as how these features provide an estimation of gaze direction, are described algorithmically.

5.1 Calibration-free Ocular Mechanics

Before delving into the technical details of the implementation, we will briefly discuss the underlying mechanics of the eye that make the calibration-free possible.

The human eye may be modeled as two overlapping spheres: a larger sphere representing the eye ball; a smaller sphere representing the corneal sphere. The optical axis of the eye is defined as the axis or line segment that intersects the centres of rotation of each of the optical elements of the eye (see *Figure 5.1*) [37]. Interestingly, the region of highest visual acuity, the fovea centralis does not lie on the *optical* axis but on the *visual* axis of the eye. While the optical axis is coaxial with the elements of the eye, the visual axis is coaxial with a subject's fixation point in the real world, and the corresponding point on the image formed on the fovea centralis. In other words, the optical axis is based on physiology, while the visual axis is based on what we are fixating on. Typically, the visual axis and optical axis are not collinear. In practice, it has been shown that the offset between these two axes is not very large [1]. Consequently, the pupil centre may be shifted laterally by the iris muscles such that the optical and visual axes intersect [37]. Thus the optical axis of the eye is synonymous to the gaze vector. Stated differently, the focus of attention may be estimated by the position of the pupil centre. Furthermore, from *Figure 5.1* we can see that the location of the pupil is recessed from the corneal surface. The metric used to measure this distance (along the optical axis) is known as the pseudophakic anterior chamber depth (PACD) [37].

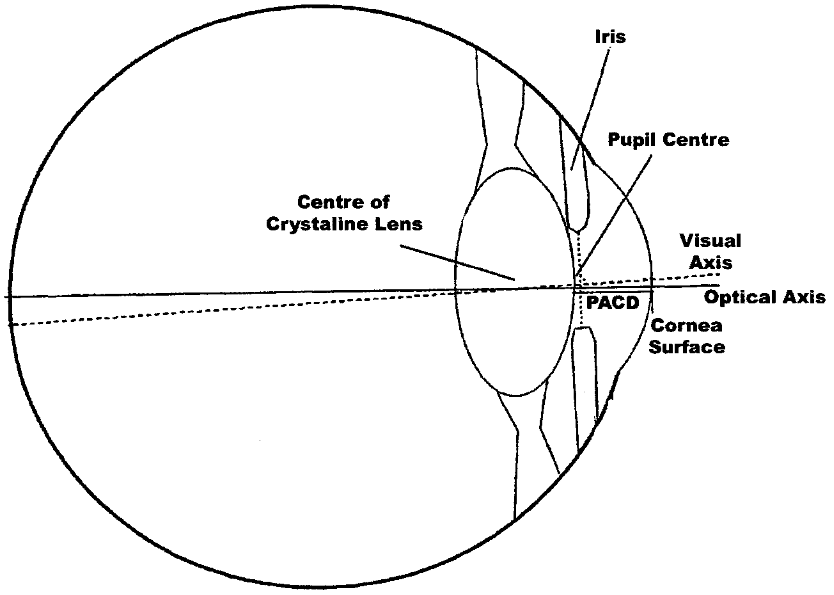

Figure 5.1: Optical Elements of the Eye [37]

For a more in depth analysis of this particular refractive behaviour of the eye, let us refer to *Figure 5.2*. Let M represent the real-world location of the marker the subject is currently fixating on. Let θ represent the angle between the camera's optical axis, and the line segment that connects M with the centre of the corneal sphere. Each off-axis illuminator produces a surface reflection or glint on the cornea surface (see *Figure 5.3*), located at an angle of $\theta/2$ in eye angular coordinates [37]. In *Figure 5.2* we see that the glint G will intersect the optical axis at a distance d from the surface of the cornea. Due to refraction, the projection line of the glint bends when it exits the cornea, intersecting the optical axis of the eye approximately 47% of the distance *(R)* from the centre of the corneal arc towards the corneal surface [11]. As a reference, if one examined the mean pseudophakic anterior chamber depth (PACD) of the general population (i.e. the average location of the pupil), one would find that the mean PACD occurs at approximately 48% of R [27]. As a consequence, although the camera's optical axis may not be aligned with the optical axis of the eye, if a subject fixates on a marker in the real word, the image seen by the camera will have a corresponding glint appear atop the pupil centre. In *Figure 5.4* we see two examples of how the markers align with the centre of the pupil when the subject fixates on it, even though the camera was not coaxial with the optical axis of the eye. In each case, the marker currently being fixated on is circled.

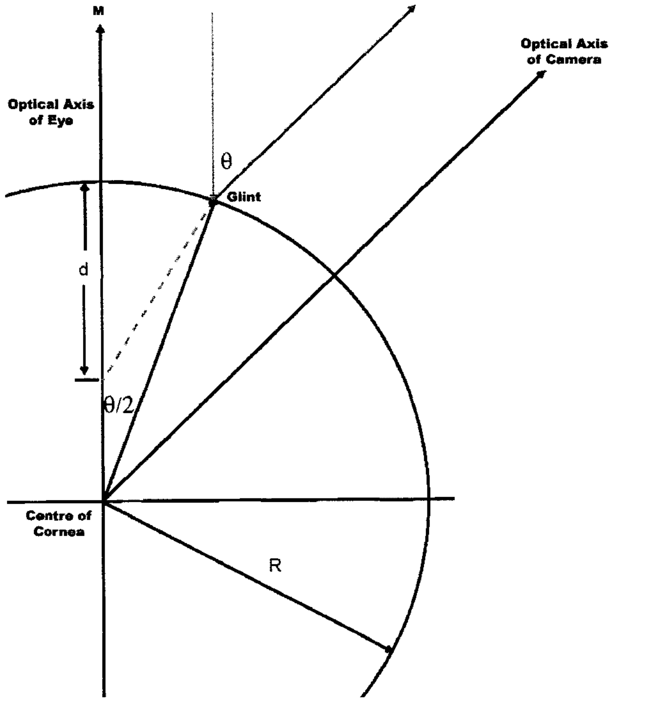

Figure 5.2: Corneal Light Refraction [37]

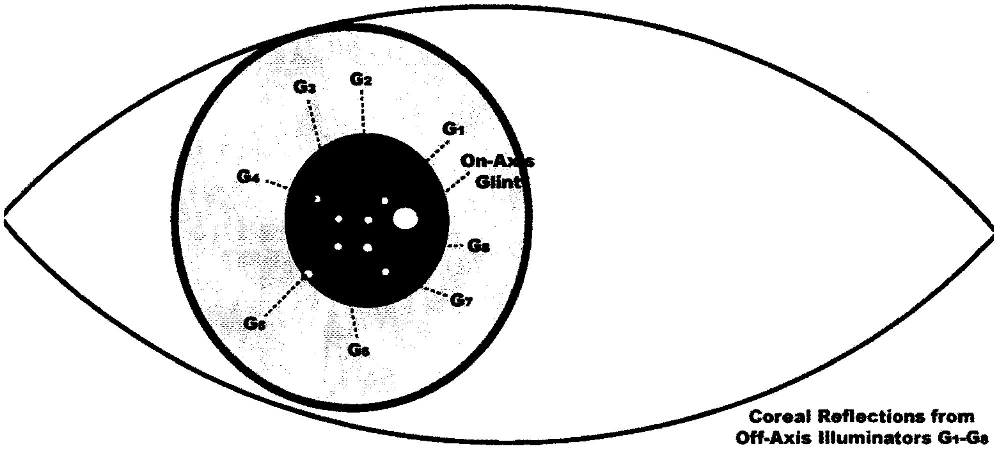

Figure 5.3: Calibration-Free Corneal Glints [37]

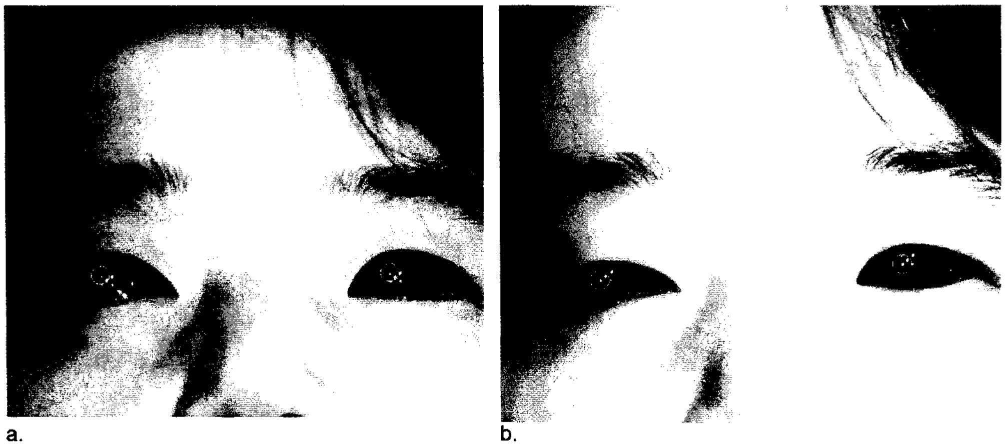

Figure 5.4: Corneal Glints and Pupil Image

To see this relationship visually, let us refer to *Figure 5.5*. In this figure, it is assumed that the mean pupil width is 5mm [9] and that the subject is fixating on marker M(i.e. the location of the marker is coaxial with the optical axis of the eye). Remember that the glint G is the projected image of M on the corneal surface. In

Figure 5.5, we see a ray trace model depicting how the location of the glint G moves (as a percentage of the width of the pupil) with respect to the image of the pupil centre, for each 0.5 standard deviation (SD) of PACD. Note that the glint stays within **10%** of the pupil diameter, at up to 80^o from the camera [37] (curve 805 in *Figure 5.5*). This characteristic of the eye is the underlying mechanism that makes calibration-free eye tracking possible. By observing the position of known glints with respect to the pupil centre, gaze estimation is possible without the need for prior calibration. Even at an extreme SD of 2 on either side, this remains true at up to 40-60^o parallax [37]. Consequently, when a user fixates on a marker, the resulting glint may be identified, by means of computer vision techniques, as being within a threshold distance to the pupil centre. The stability of the projected glint position with respect to the pupil image is the crux of the calibration-free eye tracking method.

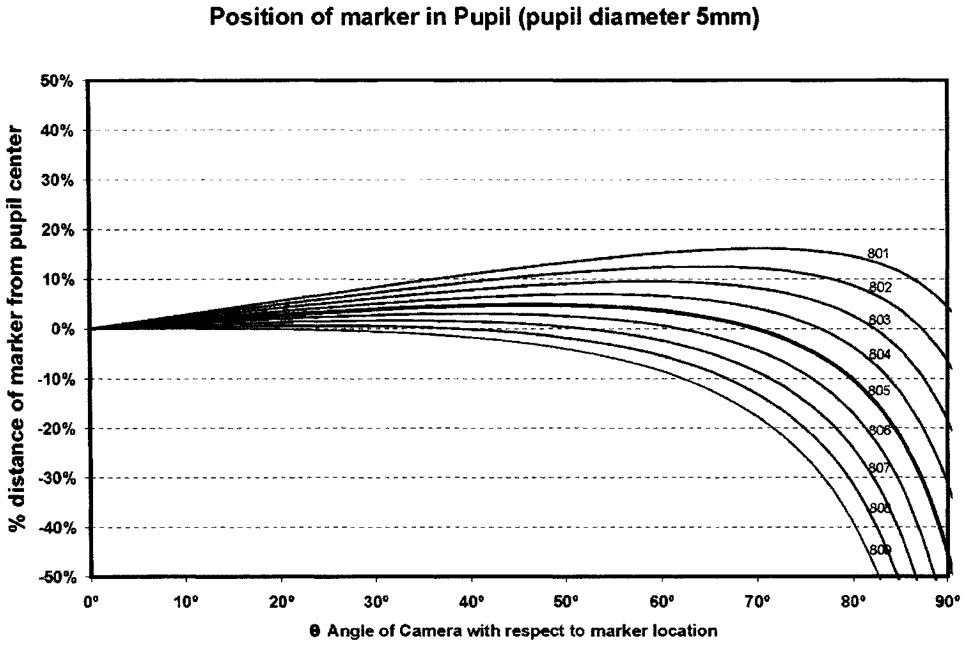

Figure 5.5: Ray Trace of Glint Location [37]

5.2 Calibration-free Algorithmic Overview

The calibration-free algorithm can be broken down functionally into two categories: 1. algorithms concerned with detecting various features of the eye; 2. algorithms for taking those features and estimating the subject's focus of attention.

Using the bright pupil effect, the detection process began by first locating and segmenting the pupil from the captured image. For glint detection, the dark pupil effect was used as it provided strong contrast between the pupil and the corneal reflections. From the captured image of the dark pupil effect, the glint detection algorithm defined a search window localized around the pupil centre and identified all glints within.

Given the location of the pupil centre, and the respective locations of the glints, a validation routine was then run to ensure only valid glints were retained. In the event of missing glints, an approximation routine was run to estimate their locations. Next the glints were registered with their real-world counterparts. For example, imagine a visual display with four IR markers on it, one on each corner. Let us denote the top leftmost marker as M_1. Let us further assume that all four glints were detected successfully. Depending on the order in which the four glints were detected, marker M_1 may correspond to glint G_3. Thus it was necessary to register the location of the detected glints with the positions of the actual markers. Finally, once registration was done, the location of the pupil centre with respect to the glints was used to interpolate the subject's gaze point on the visual display.

5.3 Hardware Setup

5.3.1 Camera Hardware

In selecting a suitable camera, two criteria had to be met. Firstly, as the image of the pupil is small, and images of the corneal reflections even smaller, it was necessary to use a camera with a fairly high resolution and an adequate zoom lens. Secondly, to ensure a tolerance to head movement, a high frame rate was required to minimize the differences resulting from motion in between frames. These two requirements are conflicting, as a high resolution necessarily results in a lower frame rate due to the shear volume of data captured. A fair compromise was found in a high-end commercial digital camera [22]. The camera uses a CMOS detector to capture images at a native resolution of 1600x1200, with a frame rate of 10 frames per second. The camera utilized a USB2.0 interface to provide rapid transfer of the captured image from the camera to the computer. As well, it featured a programmable I/O port for strict synchronization of external devices, such as illuminators, with the camera's timing clock. A 25mm lens was used to provide a zoomed image of the eye while maintaining a sufficiently large field of view to tolerate head movements.

5.3.2 Illuminator Configuration

The CF eye tracker uses five sets of infrared (IR) LED illuminators. The first four sets each consists of a square 3x3 array of IR LEDs that are distributed about the corners of the visual display. These illuminators strobe synchronously and are used for off-axis illumination. The remaining set of illuminators is a ring of IR LEDs that fits around the camera lens. This set provides the on-axis illumination. The on-axis and off-axis illuminators were activated alternatively to create a strobe effect.

5.3.3 Illuminator Synchronization

To regulate the strobe of the on and off-axis illuminators, a flip-flop circuit was used as a toggle. The circuit was connected to the camera's I/O port and synchronized with the camera's timing clock. The on-axis illuminators were set active while the off-axis illuminators were turned off with each positive edge of the clock pulse, and vice versa with the each negative edge of the clock pulse. This produced the successive on and off-axis images required for pupil and glint detection.

5.3.4 Image Capture Synchronization

Since each frame is captured at a very high resolution, rapid transferring of the data from the camera to the computer becomes a potential bottleneck. In the event that images are being captured faster than they can be transferred and/or processed by the host computer, frame dropping occurs. This is because the frame buffer of the camera is capable of holding a maximum of two frames: one is the frame currently being transferred from the camera to the system; the second being the most recently captured frame. The latter frame is held in a circular buffer, meaning if transmission of the captured frame is not initiated before the next frame is captured, the frame is overwritten and permanently lost. Should the processing overhead of the image processing layer become too demanding, the system will not be able to obtain every frame before it is overwritten and frames are dropped. Dropped frames pose a synchronization problem due to the alternating illumination of the on/off-axis illuminators. The system expects to see a sequential stream of alternating images - an on-axis image (bright pupil effect), repeatedly followed by an off-axis image (dark pupil effect). If a frame were dropped, say the loss of an off-axis image, the system would receive a second, consecutive on-axis image when it expected to see an off-axis image. Although additional software checksums may be used to disambiguate on-axis images from off-axis images, these validation routines themselves incur a high computational overhead and, in extreme scenarios, may only compound the initial problem. A hardware solution to this was to have the camera mark each captured frame as either on-axis or off-axis. This was done by toggling the least significant bit (LSB) in the image header for each captured frame. This had minimal impact on the overall image as the LSB only affected the first pixel of each image. The toggling of the LSB results in at most a change of +/- 1 in the intensity of the first pixel; visually speaking, it is effectively unnoticeable.

5.4 Software Setup

The software was designed using a tree-like architecture, consisting of a root, and two child branches or layers. The advanced programming interface (API) [22] provided by the camera manufacturer served as the central root from which all the image processing and communication functionalities stemmed. Placing the API at the core of the software architecture affords low level control of the camera hardware. This served to minimize the bottleneck incurred in transferring the large volume of data produced by each captured frame. The high resolution of the captured images necessitated direct access of the image data from the camera's memory buffer.

Comprising one child layer were the image processing algorithms used to extract the visual features from the eye for gaze estimation. These were implemented using the OpenCV image processing library. This library was chosen for its wide range of capabilities as well as its platform-independent performance. The image processing layer was divided into three major sections. The first section involved the detection and segmentation of the user's pupils from the captured image. The next section dealt with the detection and identification of the glints of the off-axis illuminators. Since the user was assumed to be looking at the visual display, the glints would all fall within the bounds of the pupil. Thus, the results of the initial pupil finding section were used to seed the glint search, limiting the scope of the search to only regions where pupils were found. Finally, in the last section, the centre point of the pupil relative to the off-axis glints was interpolated to give an equivalent point (the user's

POR) in screen coordinates of the visual display.

The remaining child layer was devoted to communication. It facilitated a simple server application that outputted gaze information using a standard TCP/IP communication protocol. This allows other applications, such as AUIs, to easily access the eye gaze data.

5.5 Communication Layer

A communication layer was incorporated into the prototype with the intent of allowing the eye tracker to act as a server and relay its information to client devices listening on a TCP/IP port. Data transmission followed a standard protocol designed for the exchange of relevant eye gaze information [ECS protocol]. To lessen the impact of the communication layer on the image processing layer, the communication layer was implemented as an individual thread. Semaphores were used to maintain proper synchronization with the main application thread.

5.6 Image Processing

To estimate the point of regard of the viewer, the accurate extraction of the locations of both the pupil centre and off-axis corneal reflections are of paramount importance. The feature extraction method must be exacting as well as timely so that the system may operate in real time. In this section an intelligent and rapid feature extraction method is described.

5.6.1 Rolling Subtraction

Through strobe illumination of the on and off-axis IR LEDs, sequential bright, then dark pupil images may be captured. This presents an ideal situation to use image subtraction.

Examination of the bright and dark pupil images will show that (ideally) the images differ visually only in the area where the pupils are located. Through a technique called image subtraction, the bulk of inter-image commonality can be removed and the pupils isolated. Image subtraction retains only the *differences* between images. With evenly distributed illumination, and a sufficiently high frame rate, the only difference between successive on/off-axis frames would be the presence of the bright pupil effect in one frame, and the dark pupil effect in the other. The resulting image difference would yield only brightly illuminated pupils, with small black dots within the pupil due to the presence of corneal glints. In practice, one seldom is privy to such conditions and additional measures are required to sharpen the post-subtraction image. Typically, the illumination characteristics of the on-axis illuminators is more akin to a spotlight, while the characteristics of the off-axis illuminators resembles that of diffused, ambient lighting. This results in uneven illumination between the dark and bright pupil images, the effects of which may be seen in the subtracted image as faint, low intensity regions throughout the image. A common post-subtraction operation is to impose an intensity threshold to the image, with all pixels failing to exceed the threshold being discarded. Typically, the threshold operation is sufficient in removing the undesired, low intensity regions, while leaving the brightly illuminated pupils unaffected.

To increase temporal resolution, a rolling subtraction method is used [29]. For example, consider the alternating sequence B_1, D_1, B_2, D_2, where B_n denotes a bright pupil frame, while D_n denotes a dark pupil frame. These frames would be subtracted as follows: $|B_1-D_1|$, $|D_1-B_2|$, $|B_2-D_2|$, and so on. Note the use of absolute values. Absolute values are used to ensure that all results are non-negative. In this way, rolling subtraction incurs no loss of temporal resolution, except for a fixed delay of one frame - the initial frame B_1 must wait for D_1 to arrive before rolling subtraction may commence.

5.6.2 Pupil Detection Algorithm

Detection of the pupil is more difficult than glint detection since the increase in intensity from the bright pupil effect is significantly less pronounced than that of corneal reflections.

For accurate interpolation, the image resolution must necessarily be quite high; however, the resulting volume of data places a significant demand on computing resources. Although the rolling subtraction yields images with less noise, the computational overhead incurred by the subtraction operation is exacerbated by the copious volume of data that must be processed. The result is a drop in frame rate as the cameras buffer is overwritten with a new image before the existing one has been processed. To alleviate this problem, a new pupil extraction method was developed that did *not* employ image subtraction.

Although theoretically sound, hardware limitations prevented image subtraction from being used in practice. Instead, an approach based on dynamic search windows was used. Automatic, real time adjustments to the scope of the search was accomplished by observing the results of the past search history and by using that information to determine the current window size. The search scope ranges from a global examination to several highly localized regions of interest (ROI).

To initialize the system, a global search window is used. If a pupil candidate is found a ROI that completely encloses the pupil is defined and recorded. In the next search iteration, the search scope is reduced to examine only the ROIs where pupils were previously found. For the localized search, the ROI is defined slightly larger than the pupil to provide a tolerance for any movement that may have occurred in between frame captures. With a sufficiently high frame rate, this movement can be kept quite minimal. The locality of ROI search greatly reduces the computational overhead. Statistical averaging is used to smooth out the window size adjustments. Furthermore, each ROI is initialized with a decaying lifespan, say three search iterations. For each examination of a ROI that returns negative, the lifespan is decreased; conversely, for each examination that returns a positive, the lifespan is increased, up to a maximum value. If the lifespan reaches zero, it is deemed that pupils are no longer present in that ROI, and the ROI is discarded.

Due to the bright pupil effect, the pixels of the pupil image are raised above the mean intensity. To segment the pupil pixels from the image, a threshold intensity value is calculated as follows:

$$\tau_n = \mu + w\sigma \tag{5.1}$$

where τ_n is the threshold value for search iteration n, μ is the mean intensity of the image, w is a scaling factor, and σ is the image standard deviation. All pixels below the intensity of τ_n are discarded. By basing the threshold on the image mean and standard deviation, the threshold value has an inherent rudimentary robustness to ambient illumination changes to the surrounding environment. The post-threshold image provides a coarse discrimination of possible pupil candidates. Although the intensity of the pupil pixels is higher than the image mean, it may not necessary be sufficient to preclude a non-trivial quantity of noise. To reduce the noise further, several additional image processing techniques are applied.

The threshold operation itself introduces noise into the image, typically in the form of either pixilation. A morphological closing operation is performed to first erode orphaned pixels, and secondly close gaps that may have been introduced by the initial threshold or erosion operations. An active contour extraction algorithm then clusters pixels together based on locality. Contours with areas that are either too large or too small to be legitimate pupil candidates are discarded. Since the pupils are fairly circular and thus symmetrical by nature, one further discriminator is to calculate the ratio between the width and the height of the contour. This is approximated by first calculating the bounding rectangle that just encloses the contour completely. The width and height of this bounding box provides the values for the ratio calculation, which should be close to unity.

Locating the Pupil Centre

The pupil centre is indicative with the viewer's POR. Since the quadrilateral formed by the corneal reflections is quite small, it is of paramount importance to determine the location of the pupil centre accurately. One pixel in image space may correspond to several on the visual display, thus even a small error when determining the pupil centre location will result in a large error when the POR is mapped onto the visual display. Occlusion by eyelids or eye lashes, suboptimal threshold values, as well as noise will shift the centre of mass of the segmented pupil away from the actual pupil centre. To compensate for these problems, the boundary of the pupil is determined by ellipse fitting. The shape of the pupil exhibits a low eccentricity through a wide range of viewpoints; the pupil may be modeled as an ellipse and ellipse fitting applied to lessen the effects of noise and improve detection accuracy.

The Random Sample Consensus [8], RANSAC, algorithm was chosen for calculating the ellipse of best fit. RANSAC is used to estimate the parameters of a mathematical model based on a data set that may contain many outliers. The RANSAC algorithm randomly selects a subset of data points to generate a hypothetical model. The hypothesis is tested against the remaining points in the data set to see how well the model conforms to the data. Given an error tolerance, all the points that fit the hypothesis while remaining below the tolerance are noted. If the number of points that fit the hypothesis is sufficient to meet the desired level of compatibility, a smoothing technique, such as least squares, is used to compute an improved estimate of the model parameters from those data points that fit the hypothesis.

The input to the RANSAC algorithm is a set of data values, a model to judge the agreement between data, and some confidence parameters (see table 6.1)

Table 5.1: RANSAC Parameters

Parameter	**Purpose**
n	Minimum number of data values required to fit the model
k	Number of iterations required
e	Error tolerance used to determine whether a data point is compatible with the model
d	Number of compatible data values required to imply a model is correct

The parameter values of e and d may be estimated from empirical data. Given that a good data point occurs with probability w, the parameter k can be calculated from the probability, z, of only bad data values occurring.

Let

$$z = (1 - w^n)^k \tag{5.2}$$

Then the value of k is defined as

$$k = \frac{log(z)}{log(1 - w^n)} \tag{5.3}$$

To gain additional confidence, the standard deviation or multiples thereof can be added to k. The standard deviation of k is defined as

$$SD(k) = \frac{\sqrt{1 - w^n}}{w^n} \tag{5.4}$$

Equation of an Ellipse

The comparison model was defined as the equation for an ellipse. The equation of an ellipse can be defined with knowledge of five parameters: the first four being any four points along the arc of the ellipse; the last being the inclination of the semi-major axis with respect to one of the coordinate axes, say the x-axis. Ignoring the inclination, the equation of an ellipse is defines as follows:

$$\frac{(x-p)^2}{a^2} + \frac{(y-q)^2}{b^2} = 1 \tag{5.5}$$

where p and q are the x-offset and y-offset from the origin, respectively.

5.6.3 Glint Detection Algorithm

For glint detection, the off-axis image is used as it provides the most contrast between the darkened pupil and high intensity glints (see *Figure 5.6*). It is a reasonable assumption to require the glints to be in close proximity to the pupil location, therefore all pixels outside the pupil ROI are ignored. Glints arising from corneal reflections are typically quite high in intensity, lending to relatively easier detection. A relatively high threshold value is used to remove all but the brightest pixels. The value of this threshold is calculated similar to Eq (5.1) but with the appropriate mean, standard deviation, and scale factor.

Glint Registration

Once the corneal reflections have been detected, they must be registered so that the computer can associate each glint with the appropriate off-axis illuminator on the monitor. The off-axis illuminators are arranged in a structured grid-like layout. In the current implementation, the grid is a symmetrical pattern, however, it is not required to be so. A repeating pattern, asymmetrical pattern, etc. are equally suitable. Due to the relatively moderate surface area of the monitor and close viewing distance of the user, the four corner distribution chosen for this implementation provided a fair compromise between performance and simplicity. The simplicity and symmetry of the chosen illuminator configuration affords a simple and rapid binary encoding registration process.

Figure 5.6: Off-Axis Glints

Each glint is assigned a two digit code, with the least significant bit representing the column location, while the most significant bit denotes the row location. Assuming an ideal case of successful detection of all four off-axis glints, the registration process begins by arbitrarily selecting one of the four glints. This glint serves as a reference point for the registration process. An encoding scheme is used to identify the position of each glint with respect to the configuration of the off-axis illuminators. The first step assigns a code to each glint as follows:

- If a glint is to the right of the reference glint, assign a column code of 1
- If a glint is to the left of the reference glint, assign a column code of -1
- If a glint is below the reference glint, assign a row code of 1
- If a glint is above of the reference glint, assign a row code of -1

If at the end of the encoding phase there are negative row codes for any glints, they are removed by adding 1 to all row elements. Similarly, if there are negative column codes, they are removed through the addition of 1. The result is that each glint now has a two digit binary code that indicates its location in the grid.

For example, consider *Figure 5.7a.* The bottom-left point (0, 10) has been arbitrarily chosen as the reference point. The remaining three glints are then encoded according to the coding scheme aforementioned. Since the encoding process generated negative row elements, one was added to all row elements (*Figure 5.7b*). The resulting binary code identifies the correct position of the glints with respect to the off-axis illuminator configuration (*Figure 5.7c*).

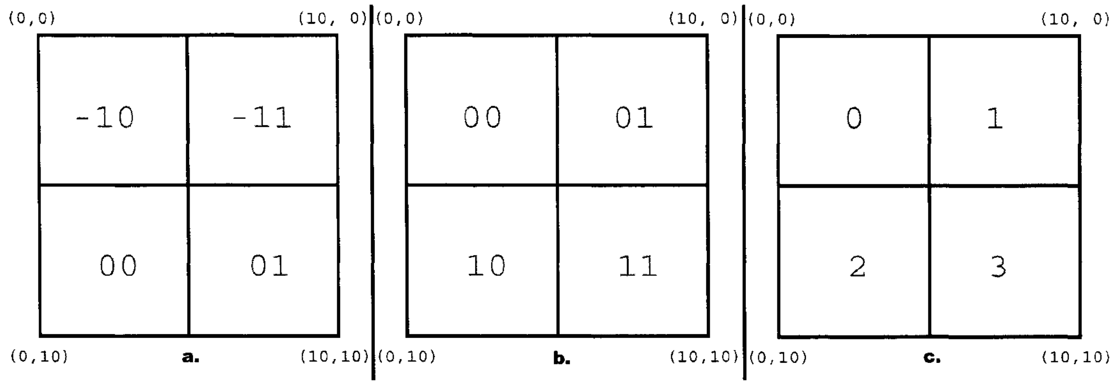

Figure 5.7: Glint Encoding

Error Correction

In the event that only three of the four glints were detected, the encoding process allows for an estimation of the missing glints locations. By summing the value of the codes returned, an error code is generated that identifies the missing glint location.

For example, consider the result of the encoding step would be as shown in *Figure 5.8*.

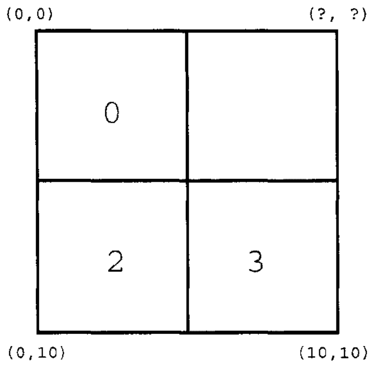

Figure 5.8: Glint Error Correction

Here the top-right glint was not detected. Summation of the binary codes yields an error code of 5. Given an error code generated by three glints, there can only be one combination that would result in an error code of 5 (see *Table 5.2*). Thus the location of the missing glint is known, as well as a means for estimating its location with the existing glint

Table 5.2: Error Codes for 3 Glint Scenario

Error Code	Glint Missing
6	0
5	1
4	2
3	3

5.7 Projective Geometry

The polygon formed by the corneal reflections may be slightly warped, predominantly by projective distortions, as well as minor distortions due to the camera lens, and the curvature of the cornea surface. With the exception of projective distortions, the remaining distortive effects are assumed negligible and may be ignored. To achieve a more accurate estimation of the viewer's point of regard, a transform is required to map the pixels enclosed by the image polygon, back to the physical space of the visual display.

To compensate for any distortion in the polygon, homography, specifically projective geometry, is used. Projective geometry is well suited for modeling the imaging process of a camera because it encompasses a large set of transformations, including Euclidean, affine, and perspective transformations amongst others. A drawback to projective geometry is that geometric relationships such as parallelism, the distance between points, and the angles between lines, are *not* maintained. Under projective transformations, only four attributes are maintained:

1. Points remain points.
2. Lines remain lines.
3. Point collinearity is preserved.
4. A measure known as the cross ratio.

5.7.1 The Cross Ratio

The last attribute, the cross ratio, is of particular interest. Under projective transformations, the cross ratio is invariant and provides a means for correcting the projective distortions. Using the cross ratio, it is possible to transform the distorted polygon formed by the glints, back to the original rectangle formed by the infrared markers placed about the visual display.

The cross ratio is a ratio composed of distance ratios. Furthermore, there is a duality in the definition of the cross ratio, as it may be described by a set of collinear points $\boldsymbol{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3, \boldsymbol{p}_4\}$, *or* a set of lines $\boldsymbol{L} = \{\overline{p_1O}, \overline{p_2O}, \overline{p_3O}, \overline{p_4O}\}$, with a common intersection point $\boldsymbol{O}$ (see *Figure 5.9*).

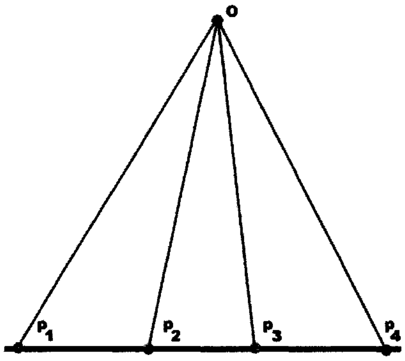

Figure 5.9: Cross Ratio Duality

Let the Euclidean distance between two points, $\boldsymbol{p}_i$, and $\boldsymbol{p}_j$, be denoted by Δ_{ij}. The cross ratio of $\boldsymbol{P}$ is then defined as follows:

$$CR_P = \frac{\Delta_{13}\Delta_{24}}{\Delta_{14}\Delta_{23}} \tag{5.6}$$

Let us consider a visual example. In *Figure 5.10a* a quadrilateral is defined with specific geometric relationships - in this case each side is perpendicular to its adjacent neighbors and the quadrilateral is a rectangle. The rectangle then undergoes a two-dimensional projective transformation and is distorted into a quadrilateral whose sides may no longer hold the same defining geometric relationships as the original figure (see *Figure 5.10b*).

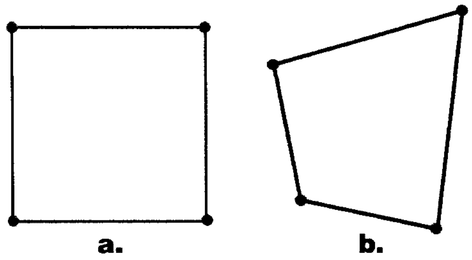

Figure 5.10: Example of Polygon Distortion on Corneal Surface

Since the value of the cross ratio of the distorted quadrilateral is unchanged from that of the original rectangle, we may use it to correct the distorted polygon back to its original form. Extending this idea to eye tracking, an analogous transformation applies. The original, undistorted rectangle is replaced by the screen of the visual display, and the distorted polygon is formed by the four corneal reflections from the off-axis illuminators.

To begin, we identify a set of reference points (see *Figure 5.11*):

1. The four corners of the distorted polygon, $P,\ Q,\ R,\ S$.

2. The pupil centre g.

3. Point c, the point of intersection of lines $\overline{PR}$ and $\overline{QS}$.

Next we extend lines $\overline{PQ}$ and $\overline{RS}$ toward infinity such that they intersect at point m. From m, we extend lines through points c and g, respectively. This results in four collinear points along $\overline{QR}$ which form the set $P' = \{Q,\ g_{QR},\ c_{QR},\ R\}$ (see *Figure 5.11*).

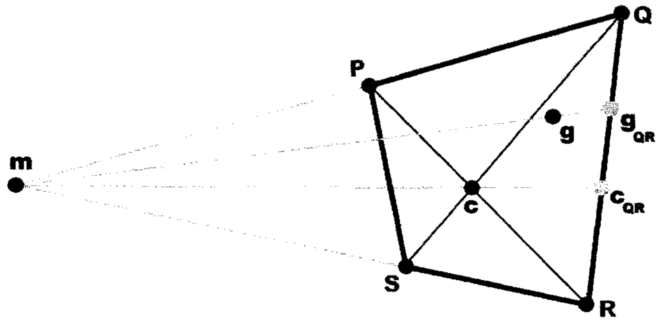

Figure 5.11: Example of Projected Transformation

The user's POR, point ***G(x,y)*** in *Figure 5.12*, is calculated in a piecemeal fashion. As an example, the y-coordinate of the users POR is calculated using cross ratios as follows:

1. Using the locations of the glints, calculate the point set $P' = \{Q,\ g_{QR},\ c_{QR},\ R\}$, as shown in *Figure 5.11*

2. Using the points of P', find the cross ratio of the image y-coordinate, $CR_{y-image}$, as follows:

$$CR_{y-image} = \frac{\Delta_{Qc_{QR}}\Delta_{g_{QR}R}}{\Delta_{QR}\Delta_{g_{QR}c_{QR}}} \tag{5.7}$$

3. Since the cross ratio is invariant, the image cross ratio, $\mathrm{CR}_{y-image}$, is equal to the cross ratio of the original screen, $\mathrm{CR}_{y-screen}$(see *Figure 5.12*):

$$CR_{y-screen} = \frac{\Delta_C\Delta_{G_yR}}{\Delta_R\Delta_{G_yC}} \tag{5.8}$$

where $\boldsymbol{G_y}$ is location of the y-component of the user's POR on the screen. The calculation of $\boldsymbol{CR}_{y-screen}$ is simple since one of the four collinear points is the origin, another is the midpoint of the screen, and a third is the height of the rectangle (see *Figure 5.12*). The only unknown is the screen coordinate for the user's POR, $\boldsymbol{G(x,y)}$.

4. By setting $\boldsymbol{CR}_{y-image} = \boldsymbol{CR}_{y-screen}$ and solving for $\boldsymbol{G}_y$ we may obtain the y-component of the user's POR.

5. Similarly, to find the x-component of the user's POR, repeat steps 1-4 but substitute $\boldsymbol{m}$ with the intersection of lines $\overline{PS}$ and $\overline{QR}$. Then find the corresponding four collinear points along $\overline{PQ}$ and calculate $\boldsymbol{CR}_{x-image}$ and $\boldsymbol{CR}_{x-screen}$, solving for $\boldsymbol{G}_x$.

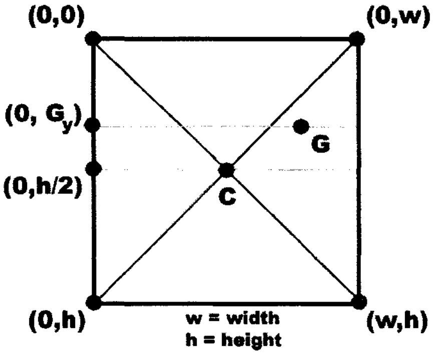

Figure 5.12: Coordinates of Visual Display

It should be noted that the polygon on the corneal surface, defined by the image glints, is significantly smaller in area than the source of the glints, the original visual display. Consequently, there is a drop in resolution as a pixel within the image polygon may correspond to a region defined by several pixels on the original visual display. By means of cross ratios we may account for projective transformations and provide a mapping between the distorted image space and the physical screen space of the visual display.

Chapter 6

Scenarios and Applications

Currently, the bulk of eye tracking applications are input driven. For instance, eye gaze provides a means for navigation, or selection of a target on a visual display. The display adopts a passive role, and interactions are limited to the display reacting to a user's momentary point of regard. Essentially, eye gaze behaves as a pointer.

More interesting is the notion of devices adopting more active roles, whereby content is intelligently displayed based on interest [12, 35], or further still, devices themselves initiating interactions.

Yarbus demonstrated that human eye gaze patterns are reflective of our current interests [42], supporting the notion that eye movements and fixations provide cues for inferring observer interest. Noting that, let us examine an analogy from retailing. In retail sales, the observation of non-verbal cues is the hallmark of a skilled salesperson. The ability to observe the attentional cues of the consumer provides an insight into the consumer's interests. In particular, the good salesperson identifies which items have received the most eye gaze, and infers that they hold the most interest to the consumer. Similarly, a computer system capable of observing gaze patterns may infer what information currently being displayed holds the most interest to the user. The user's gaze patterns would provide cues for determining what content the system would elaborate on visually and/or auditory.

6.1 Self-Disclosing Devices

A system embodying the aforementioned attentive design criteria was proposed by Stark and Bolt [32]. The content displayed by the system was built around Antoine de Saint Exupery's children's tale, *The Little Prince.* Three dimensional graphics representing the prince's world were presented to the subject while an eye tracker analyzed the subsequent gaze patterns. Items of interest were determined by aggregating the locus of eye fixations over the very recent past. For the item(s) of interest, the system would zoom in and a synthesized voice would further contextual information regarding the item(s). The system was self-disclosing in that it communicated its repository of information according to the interests exhibited by user eye movements. The generality or specificity of the system's narration was a function of the scope and focus of the user's attention, as inferred from the user's eye gaze patterns. As feedback to the user, when an item was determined to be of interest, it would "blush" or flash momentarily by modulating its opacity on the display [32].

There are two drawbacks to Bolt's system. Firstly, there was the need to calibrate the eye tracker. Secondly, a camera with a zoom lens was used to provide a high resolution image of the eye, resulting in a narrow FOV that limited the amount of user head movement. At high magnification levels, even modest head movements may move the eye completely out of the FOV. As a consequence, a chinrest was required to restrict the user's head movements and steady the user's head, thus preventing truly natural interactions.

6.1.1 A Calibration-Free Approach

The CF method proposed in this thesis addresses those drawbacks in the following manner. Firstly, since the system does not need a calibration procedure, users may walk up to the system and immediately begin interacting. Users are free to move about naturally when interacting with the system - an ability that is essential if eye tracking is to be applied to more commonplace scenarios.

Secondly, the CF eye tracking method is capable of handling a moderate amount of head movement without the need for restrictive measures such as chinrests. This characteristic is due to the fact that our eyes move independently of our head orientation. For example, one may fixated on a point while rotating one's head from left to right. To maintain focus on the point, our eyes move in an equal and opposite direction than then head. Similarly, if the point were a directed light source, the position of the resulting glint would remain stationary relative to the position of the pupil. This is true as long as the viewing angle is not extremely oblique to the user - a reasonable assumption as the view would not be clearly visible from such extreme viewpoints in the first place.

By augmenting devices with CF eye tracking capabilities, not only can devices respond to user initiated interactions, but the interactions may be extended to *device* initiated dialogues. For example, a large-screen display augmented with CF eye tracking capabilities would not only be able to respond to user attention, but also solicit it as well. Through the application of AUI design principles, the system would abide by the same progressive turn-taking protocols found in human-to-human interactions.

One may envision a large-screen display as a medium for attentive information - information that remains in the periphery of the user's attention until called for through attentional cues, such as sustained eye contact.

6.2 Applications

Calibration-free (CF) eye tracking substantially increases the utility of eye tracking systems by reducing both the intrusiveness and obtrusiveness of conventional eye tracking methods. Users are no longer confined by restrictive operational environments, nor adorn cumbersome equipment that hinder interactions from being fully natural. CF eye tracking affords many more natural interactions, as users may literally walk-up to the system and begin interacting immediately. The ability to perceive a user's visual cues allows the system to anticipate the user's intention(s) before any overt physical gesture is made. The net effect is a seemingly intelligent system that provides a smoother, more intuitive interface for human-computer interaction. Below are two examples of how CF eye tracking technology may be applied to AUI design.

6.2.1 EyeDisplay

The EyeDisplay is an example of a Media EyePliance [35]. Essentially, a Media EyePliance was a normal home theatre appliance that was augmented with eye tracking capabilities to facilitate perception of the attentional cues exhibited by humans. The EyeDisplay consisted of a 23" LCD visual display, augmented with the proposed CF eye tracker. A series of off-axis illuminators were distributed about the perimeter of the display. This quantized the visual display into regions called media tiles. The relative position of the six off-axis corneal reflections with respect to the pupil centre provided the gaze location. For example, when the user looked at the top-right corner of the screen, the corresponding infrared marker for that corner appeared centred in the pupil. If the subject's point of regard fell between markers, interpolation was used to estimate the gaze location on the display. Each of the tiles was assigned a media file, in this case a video stream. Users selected a media file for playback by looking at the corresponding tile and pressing the play button on a remote. Upon activation, the selected tile would begin playback and magnify its contents to fill the screen. Similarly, by hitting the stop button on the remote, the tile would shrink and revert back to its minimized state. The EyeDisplay was a display that watched the user and mediated its content based on the user's focus of attention. It determined when it was being watched, and when not. When the user's gaze was lost, the EyeDisplay inferred that it was no longer the recipient of the user's focus of attention and consequently playback was paused. Conversely, when the user resumed watching, the presence of sustained eye gaze signified to the EyeDisplay that it was once again the focus of attention and playback resumed.

Using the eyes as an input channel means the user does not require an additional input device, such as a remote control, for initiating interactions. Since the eyes have the fastest muscles in the human body, they are able to move quicker than any other body part [35]. Furthermore, researchers have reported that during target acquisition, users tend to look at a target *before* initiating manual action [16]. When tracked effectively, eye gaze may provide one of the fastest possible input methods. Through the use of calibration-free eye tracking and AUI design principles, the user experience traditionally associated with visual displays is made more intuitive and consequently, more natural.

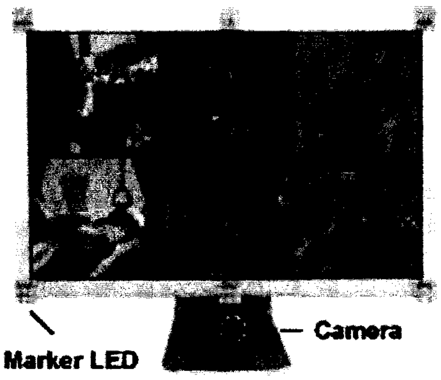

Figure 6.1: Media EyePliance - EyeDisplay [35]

6.2.2 Attentive Art

Attentive art [12] is a media art piece that interacts with viewers by modulating the visibility of the artwork based on quantity of user attention received. A large plasma display provided the canvas for this art piece. The surface of the visual display was quantized into several regions via strategic placement of IR illuminators on the display surface. The art piece measured the audience's interest in any given region of the artwork by tallying the quantity of eye gaze that fell upon each region within a given time interval. This value, representative of the audience's attention for that region, was reflected visually by highlighting regions the audience found interesting, while darkening regions that received little attention. Whereas Wooding [41] expressed attention with static fixation maps, Attentive Art does so dynamically.

The motivation behind Attentive Art was twofold. Firstly, large screens may contain more information and consequently more visual clutter. Secondly, critical information may be missed due to the large display area relative to the smaller user field of view. Attentive Art demonstrated that users gaze information could be applied as a filter. Unattended regions were removed or abstracted into the periphery while regions of interest were highlighted and made apparent - the displayed information was contingent upon the users' visual focus of attention. Furthermore, areas that were of interest to the user may be used to trigger specific sounds, motions, or other meaningful responses.

While the ideas behind Attentive Art were artistic, its underlying principles may be applied more generally to managing content on, and designing interactions for large visual displays. The unobtrusiveness of the CF method allows the user to interact more naturally, without the hinderance and constraints of traditional eye trackers. By tracking the focus of the user's visual attention, the cognitive load associated with large display visualizations may be managed more effectively [12].

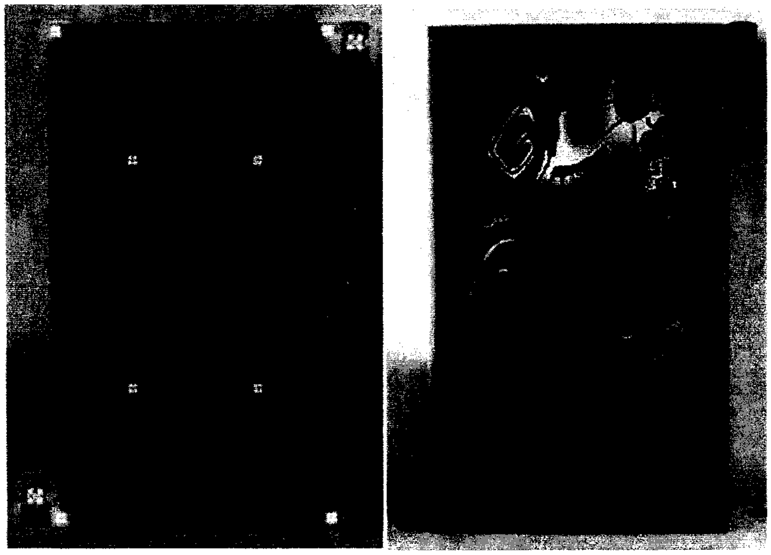

Figure 6.2: Attentive Art [12]

Chapter 7

Future Work

Currently, we are examining methods in which to improve the prototype CF eye tracker. By improving the pupil and glint detection method, a greater tolerance to movement is possible. As well, additional hardware modifications and supporting software will extend the capabilities of CF eye tracking to include wider, more ubiquitous applications.

7.1 Frame Rate

By increasing the frame rate, the responsiveness of the system correspondingly also increases. Currently, the system averages approximately 9 frames-per-second (fps), without the use of any bright/dark pupil subtraction techniques. Subtraction-based approaches require a sufficiently high fps value since the higher frame rates yield lower temporal differences between successive images. The low temporal difference between captured images results in better difference images. We are examining methods of improving the frame rate through increased efficiency of the computer vision algorithms, as well as modifications to the hardware to provide better illumination, both on and off-axis.

7.2 Nyquest-Shannon Illuminator Encoding

The Nyquist-Shannon sampling theorem states a transmitted signal may be perfectly reconstructed if the receivers sampling rate is *greater* than twice that of the transmission rate. Applying this theorem opens the possibility of encode information with the IR illuminators. For instance, pulse-width modulation allows each illuminator to be assigned a unique identification. The benefits of embedding such information can be demonstrated by considering a retail environment [37]. Within this retail environment, a clothes rack is augmented with a one or more eye tracking cameras. For this rack, each piece of clothing (or he accompanying hangar) is fitted with a tag consisting of illuminators that strobe with a pulse-width modulated identification pattern. These illuminators, although invisible to the naked eye, are detected by the eye tracking cameras and determine which items are of interest to consumers as reflected by their gaze patterns. If a consumer's fixation on an object exceeds a threshold value, the system infers interest and begins an iterative process of self-disclosure, and progressive turn-taking to establish a mutual focus of attention. Furthermore, as the encoded illuminators are not confined to the plane of a visual display, this technique allows eye tracking to occur in a three-dimensional space.

Chapter 8

Summary

This thesis examined a novel method of eye tracking that was free of the calibration process found in conventional eye trackers. We discussed a prototype system to demonstrate how calibration-free eye tracking could be used in the context of improving human-computer interactions. The system provided a more natural way to interact with computing devices in less restrictive environments than afforded by conventional eye trackers.

The prototype was built around the API supplied by the camera manufacturer. This was necessary as low level control of the hardware was crucial to maintain real-time performance; the alternation of the on and off-axis illuminators had to be tightly synchronized with the delivery of captured frames. The image processing component was developed with the OpenCV computer vision library.

As a demonstration, the calibration-free technique was incorporated into the design of a media art piece, Attentive Art [12], and an attentive media appliance, the EyeDisplay [35].

Bibliography

[1] A. Bradley and L. Thibos. Modeling off-axis vision: the optical effects of decentering visual targets or the eye's entrance pupil. Technical report, School of Optometry Indiana University, 2003.

[2] John Seely Brown and Terry Winograd. *Bringing Design to Software.* ACM Press, New York, NY, USA, 1996.

[3] Vannevar Bush. As we may think. *Atlantic Monthly*, pages 101–108, July 1945.

[4] Andrew T. Duchowski. *Eye Tracking Methodology: Theory and Practice.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[5] Andrew T. Duchowski, Nathan Cournia, and Hunter Murphy. Gaze-contingent displays: Review and current trends. *CyberPsychology and Behavior*, 7(6), 2004.

[6] Y. Ebisawa. Improved video-based eye-gaze detection method. *IEEE Transactions on Instruments and Measurement*, 47(4):948–955, August 1998.

[7] Douglas Engelbart. Augmenting human intellect: A conceptual framework. Technical Report AFOSR-3233, Stanford Research Institute, Octoboer 1962. http://www.bootstrap.org.

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[9] J. Forrester, A. Dick, P. McMenamin, and W. Lee. *The Eye. Basic Sciences in Practice.* W.B. Saunders, London, England, UK, 1996.

[10] R L Gregory. Perceptual illusions and brain models. pages 179–296, 1968.

[11] R. G. L. van der Heijde, M. Dubbelman, and H. A. Weeber. The shape of the back surface of the cornea. *S. Afr. Optom.*, 62(3):132, 2003.

[12] David Holman, Roel Vertegaal, Changuk Sohn, and Daniel Cheng. Attentive display: paintings as attentive user interfaces. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 1127–1130, New York, NY, USA, 2004. ACM Press.

[13] Canon Inc. Canon eos-5. September 2005. http://www.canon.com.

[14] LC Technologies Inc. Eyegaze system. September 2005. http://www.eyegaze.com.

[15] Tobii Technologies Inc. Tobii 1750. September 2005. http://www.eyegaze.com.

[16] R.J.K Jacob. The use of eye movemments in human-computer interaction techniques. *ACM Transactions on Information Systems.*

[17] William James. Interest and attention. *Atlantic Monthly*, 83(498):510–518, 1889.

[18] Qiang Ji and Zhiwei Zhu. Eye and gaze tracking for interactive graphic display. In *SMARTGRAPH '02: Proceedings of the 2nd international symposium on Smart graphics*, pages 79–85, New York, NY, USA, 2002. ACM Press.

[19] Applied Science Laboratories. H6. September 2005. http://www.a-s-l.com.

[20] J. C. R. Licklider. Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, pages 4–11, July 1960.

[21] J. C. R. Licklider. The computer as a communication device. *Science and Techonlogy*, April 1968.

[22] Lumenera. Lu205. September 2005. http://www.lumenera.com.

[23] Paul P. Maglio, Teenie Matlock, Christopher S. Campbell, Shumin Zhai, and Barton A. Smith. Gaze and speech in attentive user interfaces. In *ICMI '00: Proceedings of the Third International Conference on Advances in Multimodal Interfaces*, pages 1–7, London, UK, 2000. Springer-Verlag.

[24] Carlos Morimoto, David Koons, Arnon Amir, and Myron Flickner. Pupil detection and tracking using multiple light sources. Technical Report RJ-10117, IBM Almaden Research Center, 1998.

[25] Carlos Morimoto, David Koons, Arnon Amir, and Shumin Zhai. Keeping an eye for hci. In *SIBGRAPI '99: Proceedings of the XII Brazilian Symposium on Computer Graphics and Image Processing*, pages 171–176, Washington, DC, USA, 1999. IEEE Computer Society.

[26] Takehiko Ohno, Naoki Mukawa, and Atsushi Yoshikawa. Freegaze: a gaze tracking system for everyday gaze interaction. In *ETRA '02: Proceedings of the*

symposium on Eye tracking research & applications, pages 125–132, New York, NY, USA, 2002. ACM Press.

[27] T. M. Rabsilber, K. A. Becker, I. B. Frisch, and G. U. Auffarth. Anterior chamber depth in relation to refractive status measured with the orbscan ii topography system. *J. Cataract Refract. Surg.*, Nov 29(11):2115–2121, 2003.

[28] Ted Selker, Andrea Lockerd, and Jorge Martinez. Eye-r, a glasses-mounted eye motion detection interface. In *CHI '01: CHI '01 extended abstracts on Human factors in computing systems*, pages 179–180, New York, NY, USA, 2001. ACM Press.

[29] Jeffrey S. Shell, Roel Vertegaal, Daniel Cheng, Alexander W. Skaburskis, Changuk Sohn, A. James Stewart, Omar Aoudeh, and Connor Dickie. Ecsglasses and eyepliances: using attention to open sociable windows of interaction. In *ETRA'2004: Proceedings of the Eye tracking research & applications symposium on Eye tracking research & applications*, pages 93–100, New York, NY, USA, 2004. ACM Press.

[30] Jeffrey S Shell, Roel Vertegaal, Aadil Mamuji, Thanh Pham, Changuk Sohn, and Alexander W. Skaburskis. Eyepliances and eyereason: Using attention to drive interactions with ubiquitous appliances. In *Extended Abstrats of UIST 2003*, pages 93–100, New York, NY, USA, 2003. ACM Press.

[31] Jeffrey S Shell, Roel Vertegaal, and Alexander W Skaburskis. Eyepliances: attention-seeking devices that respond to visual attention. In *CHI '03: CHI '03 extended abstracts on Human factors in computing systems*, pages 770–771, New York, NY, USA, 2003. ACM Press.

[32] I. Starker and R. A. Bolt. A gaze-responsive self-disclosing display. In *CHI '90: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3–10, New York, NY, USA, 1990. ACM Press.

[33] A. Tomono, M. Iida, and Y. Kobayashi. A tv camera system which extracts feature points for non-contact eye movement detection. In *Proceedings of the SPIE Optics, Illumination, and Image Sensing for Machine Vision IV*, volume 1194, pages 2–12, 1989.

[34] Roel Vertegaal, Connor Dickie, Changuk Sohn, and Myron Flickner. Designing attentive cell phone using wearable eyecontact sensors. In *CHI '02: CHI '02 extended abstracts on Human factors in computing systems*, pages 646–647, New York, NY, USA, 2002. ACM Press.

[35] Roel Vertegaal, Aadil Mamuji, Changuk Sohn, and Daniel Cheng. Media eyepliances: using eye tracking for remote control focus selection of appliances. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1861–1864, New York, NY, USA, 2005. ACM Press.

[36] Roel Vertegaal, Robert Slagter, Gerrit van der Veer, and Anton Nijholt. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 301–308, New York, NY, USA, 2001. ACM Press.

[37] Roel Vertegaal, Changuk Sohn, Daniel Cheng, and Victor MacFarlane. Method and apparatus for calibration-free eye tracking. Technical Report WO 2005/046465, Queen's University, June 2005. International Patent Application No. PCT/CA2004/001965.

[38] Roel Vertegaal, Gerrit van der Veer, and Harro Vons. Effects of gaze on multiparty mediated communication. In *In Proceedings of Graphics Interface*, pages 95–102, Montreal, Canada, 2000. Morgan Kaufmann Publishers.

[39] Mark Weiser. Open house. *Review*, March 1996. http://www.itp.tsoa.nyu.edu/ review/.

[40] Mark Weiser and John Seely Brown. The coming age of calm technology. Technical report, Xerox PARC, October.

[41] David S. Wooding. Fixation maps: quantifying eye-movement traces. In *ETRA '02: Proceedings of the symposium on Eye tracking research & applications*, pages 31–36, New York, NY, USA, 2002. ACM Press.

[42] A. L. Yarbus. *Eye Movements and Vision.* Plenum Press, New York, NY, USA, 1967.

[43] Shumin Zhai. What's in the eyes for attentive input. *Commun. ACM*, 46(3):34–39, 2003.