From the Institute for Neuro- and Bioinformatics
of the University of Lübeck
Director: Prof. Dr. rer. nat. Thomas Martinetz

Gaze Guidance for Augmented Vision

Dissertation
for Fulfillment of
Requirements
for the Doctoral Degree
of the University of Lübeck

from the Department of Computer Sciences/Engineering

Submitted by
Laura Pomârjanschi
from Bucharest
Lübeck 2012

First referee: Prof. Dr.-Ing. Erhardt Barth
Second referee: Prof. Dr.-Ing. Andreas Schrader
Chairman: Prof. Dr.-Ing. Achim Schweikard

Date of oral examination: 18.02.2013

Approved for printing. Lübeck, 20.03.2013

Contents

Acknowledgements

I set out into this journey without suspecting how much it would change my life, or how many extraordinary people I would get to meet. I now have the chance to thank a few of those with whom I shared it.

First of all, my sincerest thanks go to my Doktorvater, Erhardt Barth. He is an engineer in the best sense of the word. Every time, he would bring a new, elegant perspective on the most difficult to solve problems.

Unfortunately, official terminology does not have a word for "Doktorbruder", and that is what best describes Michael Dorr: a big brother that constantly offered me advice and support.

Eleonóra Víg was a colleague in science, but also much more than that: a true friend. Thank you for always being there for me!

I am grateful to Thomas Martinetz for giving me the opportunity to pursue my PhD in his institute, and also for his generous sponsorship of my German courses.

I thank Peter Bex for welcoming me in his lab for four months, and to the "Bexlab" for making my stay in Boston memorable.

A warm "Thank you!" goes to Ebba-Maria Dudde, my ever dedicated and enthusiastic German teacher, who patiently bore through my constant abuse of the German grammar.

Christoph Rasche helped me through the beginning of my first independent driving simulator experiment. Maria Schneider and Ronja Möller worked with me during their undergraduate theses.

My colleagues in the INB were always there in moments of frustration and tiredness. I am grateful to them for making this journey easier and more pleasant. I will always remember the coffee and table football breaks.

Finally, my work was financially supported by the Graduate School for Computing in Medicine and Life Sciences funded by Germany's Excellence Initiative [DFG GSC 235/1]. However, the support of the Gradschool was not only financial. It also offered me the chance to work in a multidisciplinary, multicultural environment. I met many friends this way.

Abstract

Although vision is an essential component of most human activities, the brain's capacity to process visual input is limited. In order to cope with the vast amount of stimuli available, visual information is subjected to repeated processes of selection and integration along the visual pathway. A first stage of selection occurs already at the level of the eye, where only about the 2 percent of the visual information that is projected on the central part of the retina is sampled in full detail, while towards the retinal periphery the resolution drops significantly. Because of this, the eyes move frequently (2–3 times per second), while attentional mechanisms select the regions of the surrounding space that need direct processing. However, cases in which this selection is not optimal do exist, and so, unexpected events occurring in the periphery of the visual field can go unnoticed. In safety-critical activities, such as driving, this can prove fatal. In situations of this kind, having an external system to detect safety-critical events, and if needed, to guide the gaze and with it the attention of the observer towards them, could be highly effective.

The primary goal of the present dissertation is to explore the effects of gaze guidance techniques in driving situations, and also to investigate the issues posed by implementing such an augmented vision system as part of a realistic driving assistance system.

Before delving into the actual subject of gaze guidance in driving, we take a look at several issues related to guiding gaze itself. First, we investigate the influence of expertise on eye movement strategies, by examining the eye movement patterns of a novice and an expert group controlling a gaze-operated game. The differences we found between the two groups confirm the fact that eye movement strategies employed by observers are optimized for performing the activity at hand, as well as the fact that these strategies evolve as a function of experience. After this look at the task-related components of gaze allocation, we switch to investigating the influence of the visual input on eye movements. We build a computational framework to predict eye movements on complex stimuli consisting of transparent, time-varying overlaid film clips. We show that eye movements tend to avoid areas of the visual input that are redundant, and also that they can be accurately predicted using low-level features of the visual input.

The main part of our work is focused on gaze guidance in driving, and aims at answering the question whether it can be effective in helping motorists avoid accidents in critical driving situations.

We first set out to investigate whether gaze-contingent cues that highlight pedestrians involved in safety-critical scenarios help subjects drive safer in a desktop driving simulator. The initial results were highly promising: drivers exposed to the gaze-contingent cues caused significantly less accidents with the high-risk pedestrian. However, as it could be argued that pedestrian-centred gaze-contingent markers are unnatural, and at the same time too difficult to implement in a real car, we repeated the experiment using simpler cues. In this case, although as before, the cues were triggered shortly prior to a safety-critical event, they only indicated the general horizontal direction from which

the pedestrian would emerge. The significant reduction in accident rates was confirmed.

The final step of our driving research aimed at transposing the previous results in the more realistic setting of a high-fidelity, wide field-of-view driving simulator. We showed that cues implemented with the help of LED arrays that were toggled on and off in a gaze contingent manner are effective in directing the gaze of the driver towards desired locations, without disrupting the driving activity.

Overall, our results reveal gaze guidance to have great potential as an effective tool in future advanced driving assistance systems.

Zusammenfassung

Sehen ist ein wesentlicher Bestandteil fast aller menschlichen Aktivitäten. Da der Informationsverarbeitungskapazität des Gehirns Grenzen gesetzt sind, finden bei der Verarbeitung visueller Stimuli entlang des Sehpfades wiederholt Auswahl- und Integrationsprozesse statt.

Eine erste Stufe der Auswahl lässt sich bereits im Auge finden. Nur ungefähr zwei Prozent der visuellen Information, die auf die Netzhaut projiziert wird, werden mit maximaler Auflösung abgetastet, und die Auflösung nimmt zur Peripherie der Netzhaut hin dramatisch ab. Deswegen müssen sich die Augen zwei- bis 3 mal pro Sekunde bewegen, um die jeweils relevanten Bereiche der visuellen Umgebung seriell abzutasten. Der Mechanismus der Aufmerksamkeit entscheidet dabei, wohin das Auge gerichtet wird. Hierbei kann es jedoch passieren, dass diese Selektion nicht optimal ist, und dass unvorhergesehene Ereignisse, die in die Peripherie des Sehfeldes gelangen, nicht wahrgenommen werden. Bei sicherheitskritischen Aktivitäten, wie z.B. beim Autofahren, kann das schwerwiegende Folgen haben. Unter derartigen Umständen wäre daher ein technisches System wünschenswert, welches potentiell gefährliche Ereignisse detektieren kann, und, wenn nötig, den Blick und damit die Aufmerksamkeit des Fahrers in diese Richtung lenkt.

Diese Dissertation untersucht Aspekte, die auf die Wirksamkeit, aber auch auf die praktische Umsetzung eines solchen erweiterten Sehvermögens ausgerichtet ist, das sich fürs Autofahren eignet.

Bevor wir uns mit dem eigentlichen Thema der Blicksteuerung beim Autofahren beschäftigen, schauen wir uns einige Aspekte an, die grundsätzlich mit Blicklenkung verbunden sind. Zuerst untersuchten wir anhand eines blickgesteuerten Spiels, welchen Einfluss Erfahrung auf die Muster der Augenbewegungen hat. Hierzu verglichen wir die Augenbewegungen einer Gruppe von untrainierten Anfängern und einer Gruppe von erfahrenen Experten.

Die Unterschiede, die wir zwischen beiden Gruppen fanden, bestätigen, dass erfahrene Probanden während dieser Aktivität optimale Augenbewegungsstrategien verwenden, und dass sich diese Strategien für Anfänger in einem Lernprozess verbessern lassen. Nach diesem Blick auf aufgabenbezogene Komponenten der Augenbewegungen gingen wir weiter und untersuchten den Einfluss des visuellen Inputs auf die Blickmuster. Wir implementierten einen Algorithmus zur Vorhersage von Augenbewegungen auf komplexen Stimuli; wir überlagerten zwei oder drei natürliche Videoclips, so dass u.a. transparente Bewegungen resultierten. Wir zeigen, dass Augenbewegungen die aus informationstheoretischer Sicht redundanten Regionen des visuellen Inputs meiden, und dass Augenbewegungen allein durch einfache physikalische Attribute des Stimulus vorhergesagt werden können.

Der Hauptteil der Dissertation konzentrierte sich auf die Blicklenkung beim Autofahren. Unser Hauptziel war die Beantwortung einer grundsätzlichen Frage: Können Unfälle durch Blicklenkung wirksam vermieden werden?

Zu Anfang untersuchten wir in einem PC-basierten Fahrsimulator, ob blickrichtungsabhängige

Hinweise auf Fußgänger, die unerwartet vom Gehweg auf die Strasse treten, die Reaktionszeit der Probanden und damit die Zahl der Unfälle reduzieren. Die Ergebnisse der ersten Versuche waren vielversprechend: Probanden, denen mit blickrichtungsabhängigen Hinweisen geholfen wurde, verursachten statistisch signifikant weniger Unfälle. Eine Einschränkung dieser ersten Versuchsreihe war jedoch, dass die verwendeten blickrichtungsabhängigen Hinweise nicht realistisch genug waren, um sie tatsächlich in Autos einbauen zu können. Daher haben wir das Experiment noch einmal mit einfacheren Hinweisen durchgeführt, die bloss die horizontale Richtung, nicht jedoch die genaue Position des Fußgängers anzeigten. Auch in dieser Versuchsreihe führte Blicksteuerung zu einer signifikant niedrigeren Unfallrate als in der Kontrollgruppe ohne Blicksteuerung.

Der letzte Schritt unserer Forschung zum Fahrverhalten galt der Validierung der PC-Simulator-basierten Ergebnisse in einem wesentlich realistischeren Fahrsimulator. Die blickrichtungsabhängigen Hinweise wurden hier nicht mehr direkt in die simulierte Graphik eingeblendet, sondern wie in einem realen Fahrzeug durch programmierbare LEDs im Rahmen der Windschutzscheibe implementiert. Trotz der größeren Distanz zwischen Hinweis und sicherheitskritischem Objekt konnten wir auch in dieser Versuchsreihe einen signifikanten Effekt der Blicksteuerung nachweisen.

Insgesamt zeigen unsere Ergebnisse ein großes Potenzial der Blicksteuerung für zukünftige Fahrerassistenzsysteme.

1
Introduction

During all our daily activities we are significantly relying on a fact most often taken for granted: we see. We effortlessly interpret and integrate the sensory information received from the visual system into the task currently performed. We are able to recognize objects, people, and situations in the complex world surrounding us, based on what appears to be solely eyesight. And all activities, from simple ones, such as moving through a room, or grasping an object, to complex and highly specialized ones, such as reading, or driving a car, they all need visual information to a certain degree.

With this in mind, does vision need augmenting? Is there any reason, or any situation where we would need help to see better, or more? And should the answer to the previous questions be “Yes”, then how could such a system for augmenting vision be built?

The impression of having access every moment to a high-resolution, high-fidelity representation of the surrounding environment is in fact only an elaborate illusion, the result of intensive brain processing. In reality, every instant, only a small part, approximately the size of about two of the approximately 180 degrees of visual field can be seen with full accuracy. In order to compensate for the lack of detail in the periphery of the visual field, humans move their eyes in average two to three times each second. The full-resolution map of the world that we are under the impression of seeing is built fixation by fixation, through successive jumps from one item of interest to another.

Because of the inhomogeneous accuracy in the visual field, what is consciously perceived from the outside world is conditioned by where one looks, as well as by the sequence in which one has scanned the environment. In addition to that, what we expect to see shapes the way we perceive the world. Because of this, events that occur “outside the fovea”, and that do not fit our expectation, or better said, our model of the world at a certain moment, and that also are not salient enough to attract attention, may easily go unnoticed. From this point of view, it could prove useful to have an automated system that detects events needing to be attended, and then unobtrusively makes its users direct their gaze towards them.

In this thesis we will approach the issue of vision augmentation through gaze guidance. Although the main part of the dissertation will focus on augmenting vision in driving scenarios, we shall also investigate several aspects that relate directly to guiding gaze, namely correlations between expertise and visual strategies, and the prediction of eye movements.

The work we shall present has its premises in, and at the same time continues research carried out in the context of the European project GazeCom. GazeCom, short for "Gaze-based Communication" aimed to "*(i) show that gaze guidance has a high impact on what is perceived and communicated effectively; (ii) advance the level of understanding of the human visual system to the point where gaze guidance becomes feasible, and (iii) build prototype systems that exploit these insights and demonstrate the potential for applications*"[1].

1.1 Outline

The first two chapters of the thesis are designed to give an overview of the basic notions we used throughout our work: Chapter 2 is an introduction into the basics of human vision, while Chapter 3 will introduce the reader to some of the techniques and the theoretical concepts that will be later used in the practical experiments.

After the first two introductory chapters, in the remainder of the thesis, we shall tackle a series of issues related to augmenting vision. First, Chapter 4 develops upon the correlation between a task and the eye movements of the observer that performs it, emphasizing on the influence of expertise on the eye movement strategies adopted by observers. In this context, we present a study that highlights differences existing between the eye movement patterns of novices and of experts that play a gaze-controlled game. Next, as a robust and yet simple framework that can predict eye movements is one of the first requirements for implementing a system that can unobtrusively guide gaze, Chapter 5 will offer a short look into the prediction of eye movements using low-level features of the visual input. This chapter will describe our results on gaze prediction on complex visual stimuli.

In Chapter 6 we will directly approach the issue of augmenting gaze, and we will attempt to build such a system adapted for driving a car. The first part of the chapter presents results obtained in Lübeck, using fairly complex gaze-guiding cues in a simple desktop driving simulator. The final part of the chapter will describe results obtained in a state-of-the-art, wide-field-of-view driving simulator, using simpler gaze-contingent cues, that can be easily implemented in a real car. These final results were obtained following an experiment conducted during a four month research stay in the laboratory of Peter Bex, at the Schepens Eye Research Institute, of Harvard Medical School.

We will conclude the thesis in Chapter 7.

[1] www.gazecom.eu

As stated above, the work we will present in the following was at times based on results and methods developed by others. For efficient video data representation and processing, as well as for the recording and analysis of eye movement data we have used the complex software framework developed by Michael Dorr. Also, an important part of the eye movement prediction section uses the framework developed by Eleonora Vig.

Not last, the pilot study presented in Chapter 6 was run by Maria Schneider as part of her bachelor thesis work.

1.2 Overview of main contributions

Several directions can be distinguished when summarizing the contributions of our research.

Our major contributions are concerned with the investigation of the effects of gaze guidance in the context of driving. Effects both on eye movements and driving behaviour were explored, first using the basic settings of a desktop driving simulator. Later, the first steps towards a generalization of these results were taken, by extending both the gaze guiding cues, and the driving environment to more realistic settings. One of the most important achievements of this series of experiments was showing that gaze-contingent cues are effective in significantly reducing the number of accidents with pedestrians in a driving simulator environment.

Second, we extended a general framework that used low-level features of the visual input to predict eye movements on time-varying stimuli, to the more general case of transparent overlays of multiple motions. We have shown that eye movements on such complex stimuli can be predicted with a high accuracy using only the geometry of the visual input.

Finally, we studied the influence of expertise on eye movements in a gaze-controlled environment.

Our results on the effect of gaze guidance in driving, and on eye movement prediction on complex stimuli have been made public in two journal articles [1, 2], as well as in several conference proceedings [3, 4] and presentations [5, 6, 7].

Our conclusions on expertise and eye movement strategies can be found in [8].

2

A Short Introduction to Human Vision

Before delving into aspects of augmenting vision, a few basic notions about the human visual system should to be understood. The way in which the human brain interprets visual information is of particular interest, since it is what justifies the need for an augmented vision system.

There is a complex path between the photons reflected by objects around us and the final visual perception that reaches consciousness. An impressive part of our brain is at all times occupied with selecting, processing, and analysing visual information from the surrounding environment.

This chapter is meant as an attempt to a short introduction into the basic biology of vision. The first section is conceived as a summary of both anatomical and physiological aspects of the human visual system. The second section is focused on eye movement types, while the last part of the chapter will give a very short overview of visual attention.

2.1 The Human Visual System

On a very simplistic level, seeing is sometimes compared to photography. Although the eye could indeed be likened to a camera – light rays projected from objects in the world pass through an aperture of adjustable size, then through a lens which focuses them on a light sensitive surface, the retina – most similarities between vision and photography end here. The retina is far more than a photosensitive film, it is a part of the central nervous system, and it is where visual information goes through a first stage of processing. The snapshot being projected onto the retina is not transmitted forward as is: from the retina onward, it is subjected to multiple stages of information extraction and analysis before it reaches the visual cortex. Only then, after undergoing further processing and integration, visual information reaches awareness.

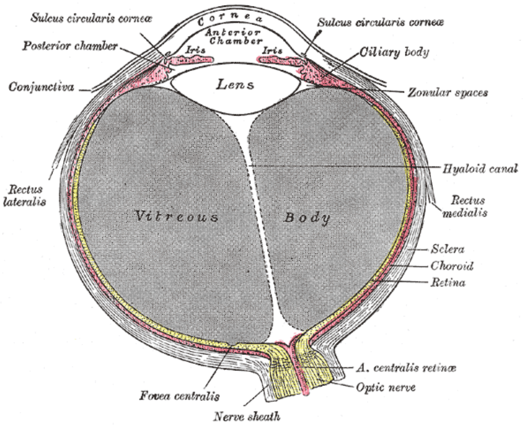


Figure 2.1: A cross-section through the human eye. Light rays entering the eye pass through the transparent anterior chamber and vitreous body, and are bundled on the light sensitive retina. The retinal area with maximal visual acuity is the pit-like structure called the fovea. Optical information, converted into neural signal by cells in the retina leaves the eye via the optic nerve. (Illustration reproduced from *Gray's Anatomy*).

2.1.1 The Eye

Light rays reflected by objects in the world enter the eye via the pupil and then travel through a refractive medium that focuses them back into a clear image on the light-sensitive retina (Figure 2.1).

The first, and the most significant refraction occurs as light meets the curved surface of the *cornea*. Light then travels through the anterior chamber of the eye and enters the internal chamber through an opening in the *iris*, called the *pupil*. The pupil contracts or dilates depending on the luminosity of the environment and thus controls how much light is allowed to enter the eye. A secondary, much smaller refraction of the light rays occurs on the *lens*. The main role of the lens is to bring objects nearby into focus, a role which it accomplishes through the fact that it can alter its curvature, a characteristic known as accommodation [9].

From the lens, the light rays are projected onto the *retina*, where the photoreceptors transform the incident photons into neural signals.

When viewed in cross-section, it can be observed that the retina has a laminar organization (Figure 2.2). The final output of the retina is the result of complex interactions between cells in all layers. Of all the retinal cell layers, the photoreceptor level is the last, most likely because photoreceptors need to be adjacent to the opaque pigment epithelium in order to access the enzymes needed for pigment regeneration [10]. To compensate for this, as light must first pass through all the other retinal cell bodies before reaching the photoreceptors, the superior layers are transparent.

Retinal structure varies spatially from the centre to the periphery. Relatively flat for the entire surface, it presents a small depression near the centre, measuring about 1500μm in diameter [11].

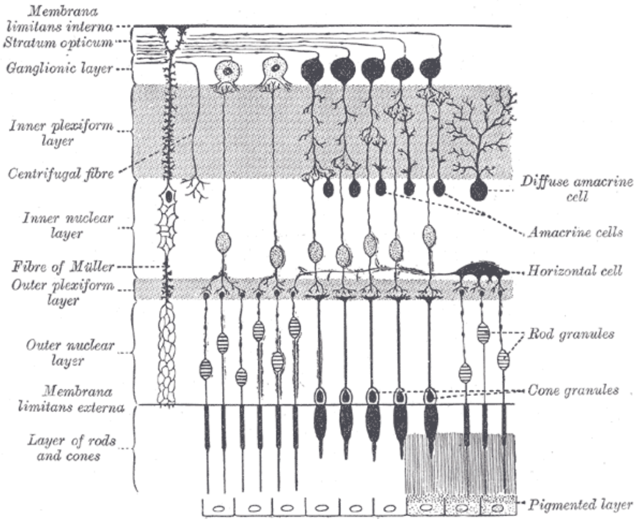


Figure 2.2: Transversal section through the retina. Light rays coming from above cross all retinal cell layers to reach the photoreceptors situated below. Information from the photoreceptor layer is passed upwards to the bipolar cell layer (the outer nuclear layer), and then to the ganglion cell layer. Horizontal and amacrine cells perform a horizontal integration of the signal from neighbouring cells, accomplishing already in the retina a basic processing of visual information. The axons of the ganglion cells provide the sole "output" as they carry the resulting neural signal through the optic nerve to other visual areas of the brain. (Illustration reproduced from *Gray's Anatomy*).

This depression is the fovea, and it corresponds to the area of maximum visual acuity. Its pit-like appearance is due to the lateral displacement of all retinal cells above the photoreceptor layer. The optic nerve leaves the eye at a small circular spot near the fovea, called the optic disc or the blind-spot. There are no photoreceptors in the blind spot, leaving it to higher visual areas to compensate for the lack of visual information from that region.

There are two types of photoreceptors, each with different functions and properties. The majority is constituted by rods, which with a count of approximately 120 million, outnumber the cones about 20 times. They are very sensitive to light, being able to detect even individual photons. Also, they saturate at high light intensities, which makes them useful only in scotopic conditions, and therefore responsible for vision under low light levels [12]. Cones operate only under photopic conditions (daylight), and although much less sensitive to light than rods, they are extremely sensitive to small intensity changes. There are three types of cones, each type responding to different light wavelengths due to containing different photopigments. This fact makes cones entirely responsible for colour vision [12].

Differences exist in the way photoreceptors are spatially distributed across the retina. The fovea contains mostly cones, making it the centre for spatial acuity and colour vision. The peripheral retina has a much higher ratio of rods to cones, and is therefore much more sensitive to light, but also essentially colour-blind [9].

Changes in photoreceptor membrane potential caused by responses to light are picked up by bipolar, but also by horizontal cells. Each bipolar cell receives input both directly from one or more photoreceptors and indirectly, via horizontal cells, from a group of neighbouring photoreceptors. The area from which the bipolar cell receives direct input constitutes the centre of its receptive field, while the area from which the cell receives indirect input constitutes its surround. The response of the cell in the centre of the receptive field is opposite to that in the surround – for example, a bipolar ON cell will be excited by light present in the centre of its receptive field, and it will be inhibited if light is present in the surround of its receptive field; the opposite will happen for an OFF bipolar cell [9]. This centre-surround receptive field organization lends the cell the ability to respond to contrast existing in the visual field, which makes edge detection part of the retinal visual processing.

The size of bipolar cell receptive fields increases towards the periphery. In the fovea there usually is a one-to-one bipolar cell-cone correspondence; in the periphery, however, each bipolar cell receives its input from a group of rods, a fact that in the end, because of several stages of spatial integration leads to higher light sensitivity, but also to reduced spatial acuity in the peripheral retina.

Further up, ganglion cells receive their input either directly from bipolar cells, or indirectly, via amacrine cells, in which case the received neural signal is subjected to some modulation. Several types of ganglion cells can be distinguished: the M (magno) cells constitute approximately 5% of the total, while the P (parvo) cells, counting about 90% of the total, constitute the majority. The remaining bipolar cells are neither of type M or of type P. When compared to P type cells, M cells have larger receptive fields, they are more sensitive to low-contrast stimuli, and they conduct action potentials much faster. Also, their responses are transient bursts as opposed to the sustained activity of P cells [9]. It is interesting to note that the output of P and M ganglion cells will be conducted and processed along separate pathways in the central visual system, a fact that will be detailed in the following section.

2.1.2 Visual Pathways in the Brain

Neural Pathways

The axons of the ganglion cells leave the eye via the optic nerve and conduct neural impulses to the optic chiasm. Here, axons carrying information from the temporal area of the retinae of the two eyes cross, so that visual information coming from the two halves of the visual field is divided between the two cerebral hemispheres [10]. After the optic chiasm, the right optic tract will only carry information from the left half of the visual field, while the left optic tract will only conduct information from the right visual hemifield (Figure 2.3). Two pathways separate on each side from the optic tract after the optic chiasm. The smaller one carries information to the brain stem, namely to the superior colliculus,

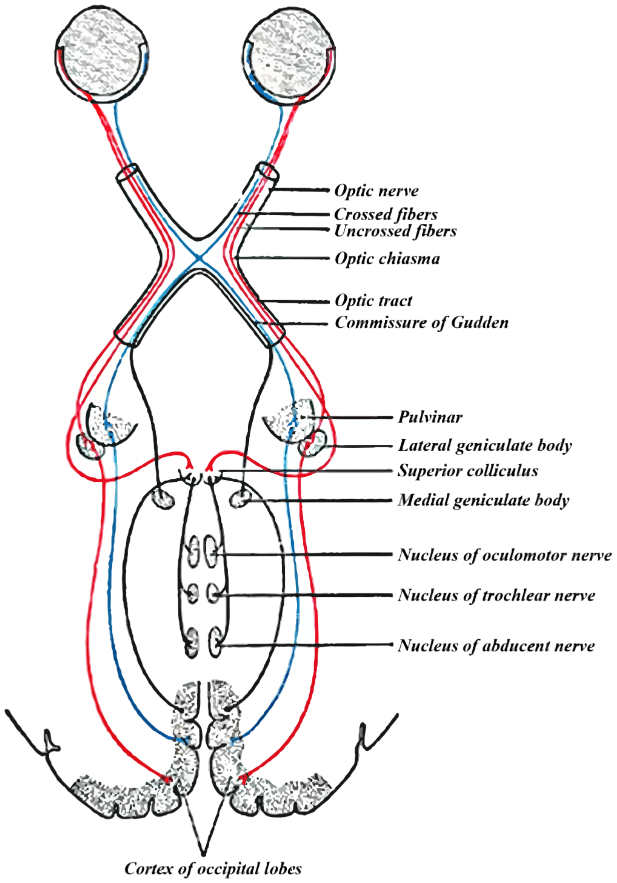


Figure 2.3: Visual pathways. The visual information from each eye is conducted by the optic nerves to the optic chiasm, where the neural fibres conducting the output of the nasal retinae cross. Because of this decussation, after the optic chiasm, the optic tract will conduct information originating from the left half of the visual field to the right hemisphere of the brain, and information from the right visual field to the left hemisphere. The greater part of the optic tract information is conducted to the lateral geniculate nucleus (LGN), and from there it reaches the visual cortex. (Illustration reproduced from *Gray's Anatomy*).

an area that appears to be involved in the control of eye movements [10]. The major pathway leads to the thalamus, to an area that will be discussed in more detail below.

Lateral Geniculate Nucleus

The main target of the optical tract axons is situated in the thalamus – a structure called the lateral geniculate nucleus (LGN). Interestingly, about 80% of the input of the LGN does not come from the retina, but from the primary visual cortex [9], suggesting that there are also feedback processes that take place at this level.

Visual information is entirely segregated at the level of the LGN, due to its laminar structure, comprising 6 distinct layers. This segregation can be seen both in terms of the source eye for the input, but also in terms of the type of the ganglion cell originating the input, as M-type cells project their axons in different layers from P-type cells. The information from the two eyes will be kept separate until the visual cortex. In fact, two separate magnocellular and parvocellular pathways can be distinguished up until the visual cortex, with the magnocellular pathway exhibiting higher conduction speed, and higher motion sensitivity [11], as well as higher spatial contrast sensitivity and gain [13, 14, 15, 16], but lower spatial frequency sensitivity than the parvocellular system [11]. This dual organization of the visual pathway suggests that one part of the visual stream will be processed in such a manner as to rapidly detect transient stimuli, while the other will be processed focusing on detail analysis and object identification – loose connection with the “vision for perception” and “vision for action” theories that we will discuss later in more detail.

Visual Cortex

Neurons from the LGN project their axons to the occipital cortex, namely to the primary visual cortex (V1).

An interesting fact about the visual pathways is that neighbouring ganglion cells in the retina transmit the neural signal to neighbouring cells in the LGN, thus preserving an approximate two-dimensional mapping of the retina onto the LGN. This retinotopic mapping is also maintained in the primary visual cortex [9], and a certain degree of retinotopy remains also in other visual cortical areas. The V1 representation of the visual field is however not symmetric [17]. The central area of the retina (corresponding to the fovea) is highly overrepresented in V1, an effect known as *cortical magnification* [18].

Similarly to the LGN, the striate cortex has a laminar structure and is anatomically divided in six principal layers [9].

In a series of now classical experiments, David Hubel and Torsten Wiesel mapped receptive fields of V1 cells first in the cat [19], and then in the monkey brain [20]. Their experiments, continued well into the 70s, helped illustrate the physiology of the primate visual cortex. In the following, some of the observations resulting from these experiments will be described.

As a function of the characteristics of their receptive field, they divided the cells they identified in V1 in three types. *Simple cells* have elongated receptive fields, with ON or OFF central areas, neighboured by an antagonistic surround either on one side (functioning akin to an edge detector) or on both (line detector). In contrast, *complex cells* do not have distinct ON and OFF regions, are highly nonlinear, are sensitive to motion, but not sensitive to the exact position of the stimulus in their receptive field. Neurons from a third cell category, named by Hubel and Wiesel *hypercomplex cells*, and known today as *end-stopped neurons* were tuned to lines up to a specific length. As a short interlude, when viewing the low and mid-level cortical processes from the perspective of the efficient coding of visual input, perspective that we will expand on in Chapter 5, the end-stopped neurons correspond to $i2D$ detectors – cells that respond to $i2D$ stimuli [21].

As it can be inferred from the shape of the receptive fields, many neurons in V1 exhibit *orientation selectivity*, in other words they respond best to a bar of light with a specific vertical, horizontal, or oblique orientation. Also, V1 cells are often *direction selective*, being tuned to a certain direction of movement.

With regard to the origin of the input, some V1 cells respond to visual stimuli presented only to the ipsilateral or the contralateral eye, while some have binocular receptive fields. Some of the binocular cells, show various degrees of dominance of the input from one of the eyes.

Neurons with similar behaviour are often grouped together. Neurons with similar preferred orientations are organized in *orientation columns*, perpendicular to the surface of the cortex. Similarly, *ocular dominance* columns have been observed.

The main target of the primary visual cortex is constituted by areas in the extrastriate cortex (V2 to V5). Although it was originally thought that each area projects sequentially to the next, it is now known that a large part of the cortical visual processing is performed in parallel, and also that each visual area projects neural fibres to several others, leading to a complex interconnection network. Moreover, these projections are often bidirectional; one can most of the time find both feedforward and feedback connections between visual areas [10].

Although the initial theory of Mishkin et al. [22] stating the existence of a dorsal "where" pathway, and a of ventral "what" pathway has proven to be controversial, it is widely accepted that the two major cortical streams exist [23]. It is now believed that the dorsal stream, passing among other areas through MT (V5) and MST appears to be involved in the analysis of motion offering the so-called "vision for action", while the ventral stream, passing through V4 and IT, has been deemed to provide "vision for recognition" [24].

The role that MT and MST have in motion processing is undeniable; neurons in these areas are motion selective neurons, and are tuned both to the direction and the velocity of the motion

[25, 26, 27]. Coming back to the computational perspective we briefly mentioned above, these motion selective cells can also be accurately modelled by the curvature-selective operators described in [21].

2.2 Eye Movements

As shown before, only a very small part of the retina samples the visual space in detail. Because of that, the eyes need to constantly move in order to supply a comprehensive sampling of the surrounding world. In the following, we will provide an overview of the main types of ocular movements, along with their roles.

The human oculomotor repertoire is fairly limited. Four main eye movement types are used, either to stabilize the fovea onto certain regions, to track targets moving at low speed, or to quickly shift gaze direction [28].

Saccades

Saccades are fast, ballistic eye movements that serve to rapidly redirect the gaze to regions of interest. Although they can be triggered voluntarily, most times they occur automatically, without being noticed by the observer. Humans can make up to 3 – 4 saccades every second [11]. During a saccade, the eye accelerates to a peak velocity, then decelerates rapidly, before returning to a stable position [29].

The amplitude of the gaze shift in a saccade is correlated with both the peak velocity and with the total duration of the saccade. In fact, a linear relationship can be observed between the duration and the amplitude of a saccade [30]. Although most saccades have small amplitudes of only a few degrees, a wide range of amplitudes exists. The largest saccades can measure more than 50 degrees. However, saccades this size are not that common as typically, for amplitudes larger than 20 – 30 degrees, head movements accompany the motion of the eye.

Another characteristic worth mentioning is the saccade latency – the time between the initiation of a stimulus and the triggering of a saccade. The saccade latency ranges between 100 and 1000 ms [29] and it is influenced by a number of factors, from the eccentricity of the stimulus, to its nature.

Despite peak velocities of up to 1000 deg/s [30], saccades do not interfere with the observer's stable perception of the surrounding world. The fact that no motion blur is consciously perceived during the ocular displacement suggests an attenuation in the visual perception mechanisms just before and during a saccade. It is not yet fully known how the visual system maintains visual stability during saccades. Some theories support the idea of *saccadic suppression* of the entire visual stream [31, 32], or only of the channels processing low-frequency high-velocity stimuli [33]. Other theories propose that saccadic motion blur is not perceived because it is masked by the preceding and subsequent fixations (this is the hypothesis of *visual masking* [34]), or that the motion blur is not perceived because saccadic retinal image motion lies outside the spatio-temporal sensitivity of the human visual system [35, 36, 37].

Fixational eye movements

Saccades alternate with *fixations*, time periods in which the eyes are relatively stationary. During these periods, the information at the fixated location is being analysed, and the analysis takes place at the same time with the selection of the next target of interest.

Even during fixations, the eye is not entirely still. Miniature eye movements such as a slow *drift*, a more rapid *tremor*, or small amplitude jumps – *microsaccades* – can be observed. It has been shown that when these small amplitude movements are eliminated by stabilizing the visual input with report to the retina, the fixated scene gradually fades from perception. This suggests that the role of these miniature eye movements is to counteract the effects of neural adaptation, by continually refreshing the image projected on the retina (see [38] for a more detailed description).

Eye movements used in target tracking

Pursuit movements allow the smooth tracking of a target moving at low speed. If the speed of the target is too high, the eye lags behind, and uses small amplitude saccades to “catch up” [39]. When the target is moving towards, or away from the observer, *vergence* movements are used to adjust the eye direction so that the target is fixated by both eyes.

2.3 Visual Attention

The highly optimized architecture of the human visual system suggests the existence of a set of complex processes that are active in the background and decide which aspects of the surrounding world need to be sampled at maximal resolution at a certain moment. It has come to be unanimously accepted that this role is carried out by visual attention.

Attention has long been a controversial topic, and when discussing it, multiple questions and theories need to be taken into consideration [40]. However, the following section will be limited to a broad overview of two of the basic attention related aspects.

First, there is the obvious question of the relation between attention and eye movements. Studies as early as that of [41] have shown that it is possible to shift attention independently of eye position. But although such a **covert** manner of orienting attention is possible, in naturalistic tasks it is more common for observers to direct their attention to the locations they fixate, in which case we talk about **overt** attention. Consensus has not yet been reached on the exact connection between overt and covert attention. However, it appears that saccade programming and visual attention are coupled [42], and there is strong evidence that attention shifts and eye movement control are correlated at a neural level [43, 44]. In the end, most theories on overt and covert attention seem to converge on the fact that there is a strong interconnection between the two, an interaction that can be briefly summarized as a covert selection of the region to subsequently receive full processing resources [45, 46].

A second question is what exactly triggers an attention shift. Is the deployment of visual attention driven solely by the content of the observed scene (**bottom-up attention**), or is it driven by the expectations and the current needs of the observer (**top-down attention**)? There is substantial evidence supporting the fact that both hypotheses are partially true, and attention is guided by a combination of top-down and bottom-up factors [47].

The voluntary component of visual attention has been highlighted by a large number of studies. Experiments such as those of [41, 48], or [49] show among other things that it is in fact possible to voluntarily shift the focus of attention. Beyond that though, the goal of the observer has a direct effect on where their attention is directed. [50] has shown that subjects viewing various images focused on areas of the scenes that were relevant for the tasks they were assigned. Also, there is extensive research on the influence of task on visual behaviour, but this aspect will be reviewed in more detail in the following chapter.

There are situations in which stimuli in the viewed scene seem to “pop out” from their surroundings. In this case we talk about an attentional capture driven by the saliency of the stimulus. Stimuli can trigger attentional capture either through visual attributes, such as colour, orientation, or motion, attributes that differentiate them from surrounding items (feature singletons), or simply through their sudden appearance in the scene (abrupt visual onsets). The two stimulus categories have slightly different attentional capture properties [51]. Although feature singletons have been shown to increase reaction times when presented as distractors in visual search tasks [52], there have been studies showing that for a full attentional capture to occur, a transient stimulus is necessary [53, 54, 55]. In addition to visual search tasks with artificial stimuli, there is a great body of research on bottom-up attention investigating oculomotor behaviour in natural scene viewing, but we will come back to this in Chapter 5.

2.4 Chapter conclusions

Everyday, we are confronted with a visual environment that is varied and rich in detail. Since it would be impossible for a brain the size of ours to analyse at their full resolution all the visual stimuli in the world around, several coping mechanisms are available.

First, the visual environment is not processed at its full resolution in all locations. Full-resolution processing is restricted to the fovea, while the resolution drops dramatically only a few degrees into the periphery. Also, rapid onsets, or motion in the periphery of our visual field reach the brain faster than information about spatial details. That is not everything: a second level of information selection exists, as the visual system uses attentional mechanisms to restrict the overwhelming visual input to a small set of stimuli needing immediate processing.

However, it can be argued that these selection mechanisms have been optimized for a world much different than today's over-urbanized environment. Indeed, as we will later show, there are circumstances when the stimulus selection is not optimal and as a consequence, important aspects of the surrounding world can be overlooked. It is in these loopholes of visual perception that an augmented vision system could prove to be of uttermost importance. Over the following chapters we will discuss designing such a system, adapted to an activity where optimal visual selection is critical.

3 Basic Methods

In our work we have used an extensive amount of methods from statistics, signal processing, and machine learning.

First, recording and analysing eye movement data constituted a crucial part of the research we conducted over the last four years. During this time, three different eye tracking systems were used for collecting the datasets that we will present in subsequent chapters. Also, because of specific properties of gaze data, we used certain statistical measures and tests that are best suited for such data.

Next, experiments regarding eye movement prediction on superimposed video clips brought the need for an adequate representation of the time-varying visual stimuli, as well as for a way to describe their geometry. Finally, machine learning techniques were used for the actual eye movement prediction part. Not last, programming tools were used to implement applications used when running the experiments, or in the data analysis phases.

The current chapter is designed to be an overview of the above mentioned theoretical notions and methods.

3.1 Eye tracking techniques

By the end of the XVIIth century, a relatively accurate representation of the anatomy of the eye was already available, and eye movements as measures of cognitive processes were beginning to attract scholarly interest. However, until the end of the XIXth century, direct observation albeit subjective and highly unreliable, was the only method to determine how an observer's eyes moved [56].

Modern eye tracking technologies can be divided in several categories.

The *electro-oculograph devices* measure potential differences between the inner and outer sides of the retina, and between the sclera and the cornea, with the help of electrodes placed on the skin near the eye. The method was introduced in the 1930s, and despite a lower accuracy, it is still widely used, especially in clinical studies [56].

Some of the oldest eye tracking devices are part of the category of *attachment eye trackers*. These involve tightly fitting a device to the eye, usually with the help of a contact lens or a suction cap that covers the cornea and part of the sclera in order to impede slippage [57]. In the earliest versions, the recording mechanism was mechanical, usually a lever-like device that would mark the movements of the eye onto a recording surface [56]. In other early devices, the eye position would be detected by using the light reflected off small mirrors attached to the cornea [58, 50]. Also among the attachment devices, the "scleral search coil" can be found. First introduced by Robinson [59], in its earliest implementation it could only detect the presence of a movement of the eyes, but after decades of improvement it is now perhaps the most accurate eye tracking method [56]. Its basic principle consists in embedding two wire coils in each contact lens, and then placing the observer within the magnetic field created by two large electromagnetic coils. Movements of the observer's eyes will cause variations in the potential difference in each coil. Despite its accuracy, the scleral search coil has all the disadvantages of attachment devices: because of the need for a very tight fit, wearing the device is highly invasive and uncomfortable, making sometimes the use of local anaesthetic necessary [60].

A third category of eye trackers, the *optical devices*, are based on detecting in an image of the eye the reflection of light rays on the cornea; based on this reflection, they can track the motion of the eye [56]. In this category, as precursors of modern trackers, the devices developed by Dodge and Cline [61] and Buswell [62] must be mentioned. Nowadays, due to the advances in computing and in image capturing techniques, optical (video-oculography based) trackers are the most widely used. Although a large variety of techniques exist, the basic idea can be summarized in a few words: a video camera films the eyes of the observer; image processing techniques are used to detect in each video frame the features used for tracking, while an algorithm converts in real time the position of these features to gaze coordinates [56]. One of the features easiest to detect is the pupil. However, using only the pupil for detection makes it impossible to separate eye movements from movements of the head. For this reason, a thorough fixation of the head, usually through a bite bar, is necessary. The most common method for compensating for head movements is to use the corneal reflex created by one, or several fixed external light sources [63]. In order to interfere as little as possible with the experiment itself, light sources emitting light in the infrared spectrum are used, together with an infrared sensitive camera. In order to account for differences in the shapes of the corneas of different observers, as well as for differences in the external conditions for each recording session, a calibration of the device is necessary before recording. Through calibration, the actual mapping function between orbital position and points in the observer's field of view is obtained [63].

All three eye tracking systems that we used in our studies fit in the category of video-oculography based trackers, and used as tracking features the dark pupil combined with one or several corneal reflexes [64, 65].

Both the experiments that we will present in Chapters 4 and 5 required high eye tracking accuracy, but without an imperative need to allow free head movements, as the experimental tasks were fully static (the free viewing of film clips and using gaze to control a computer game). Therefore, for the data recording, a *SensoMotric Instruments (SMI) Hi Speed* eye tracker was used. The device provided a monocular sampling frequency of up to 1250 Hz, and a manufacturer-specified tracking accuracy between 0.25 – 0.5 degrees; the system did not need a complete immobilization of the head of the observer, that was stabilized with the help of an adjustable height chin rest [1].

The study described in the first part of Chapter 6 was set in a PC driving simulator, that was integrated with a *SMI RED250* remote eye tracker. In comparison to the HiSpeed, the RED offered sampling rates of only up to 250 Hz, but it allowed for free movements of the head within a head box of 40x20 cm, at 70 cm distance [2].

The final study that we will describe in Chapter 6, took place in a high-fidelity driving simulator with a horizontal field of view of 225 deg. It was integrated with a *Smart Eye Pro* system that, with a flexible setup of 6 cameras and 4 infrared diodes, could track eye movements on the entire display of the simulator. The sampling frequency of the system was 60 Hz, and with the restriction of maintaining the eyes visible in at least two cameras simultaneously, the system allowed natural head motion [3]. In addition to the gaze calibration that was necessary for each of the three eye tracking systems, because of the adjustable cameras, a camera calibration was also needed in the case of the Smart Eye tracker.

3.2 Basic statistical methods

As previously said, measuring gaze data was indispensable for our work. For each study we conducted, the eye movement recording phase resulted in large quantities of data to be examined and interpreted. Just as an example, one minute of recording time on an eye tracker running at a sampling frequency of 60 Hz would result in more than 3500 gaze samples. Often, one of the most interesting questions when analysing these data was whether different subject groups or conditions could be distinguished based on the recorded eye movement behaviour. Moreover, if these differences existed, it had to be established whether they were statistically significant, or it was more likely they were due to chance. In the following, a short summary of the main statistical tools we used in our analyses is given. For a more detailed overview of these basic techniques, we refer the reader to the textbook of [66].

[2] http://www.smivision.com/fileadmin/user_upload/downloads/product_flyer/prod_smi_red250_techspecs.pdf

[3] http://www.smarteye.se/sites/smarteye/files/datasheets/smart_eye_pro.pdf

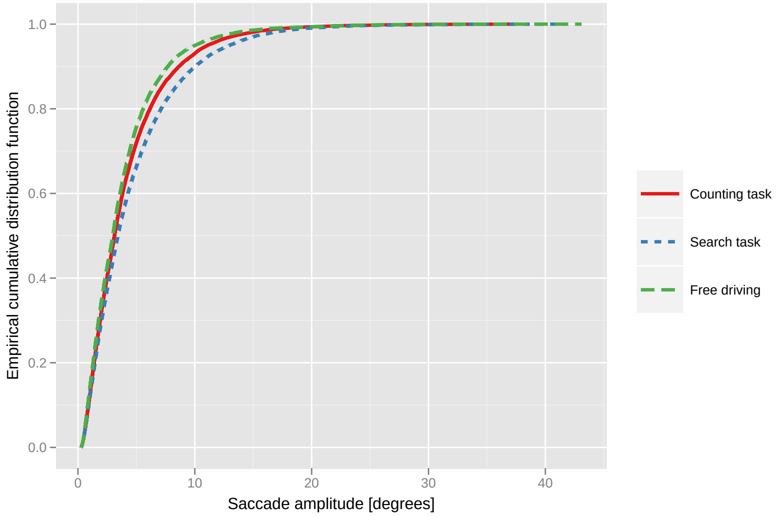


Figure 3.1: Illustration of the use of the empirical cumulative distribution function (ECDF) for the distributions of the saccade amplitudes over three different driving tasks. The plot of the ECDF highlights the differences between the three conditions. In the free driving task (uppermost curve), the subjects performed more short saccades, while in the search task (bottom curve) more large, exploratory, saccades were performed.

When using test of significance on eye movement data, several issues need to be kept in mind. First, it is not possible to assign a specific probability distribution to the gaze "generator", and therefore, non-parametric statistics need to be used.

Second, using significance testing on raw gaze data is risky, and can lead to an overestimation of the significance level. Raw gaze coordinates do not constitute independent statistical samples, since the same gaze "generator" will produce different population sizes depending on the sampling frequency. The correlations in the data are also apparent when taking into consideration the fact that, at a sampling frequency of 60 Hz, the eye position would be sampled twice or three times during an average length saccade. There are two straightforward methods to correct for the overestimation of the sample size: either use an measure that is invariant to the sampling rate of the eye tracker, such as the saccade end points, or perform an artificial subsampling to the raw gaze data. In our data analysis, we have used both methods. More details on the latter will be given in Chapter 4.

A simple approach to visualizing the trend of a sample, or the differences between several is to plot the *empirical cumulative distribution function (ECDF)*. In non-parametric approaches, the ECDF provides a simple estimate of the cumulative distribution function (CDF), in cases where only minimal assumptions can be made about the probability distribution of the data. The CDF is the integral of the probability density function; for a data point y, it denotes the proportion of data samples with values smaller than or equal to y.

For a discrete, random sample $(x_1, x_2, \ldots, x_n)$ the ECDF is defined as

$$P_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_y\{x_i\}, \tag{3.2.1}$$

where $\mathbf{1}_S$ is called an indicator function, and is defined as $\mathbf{1}_S\{x\} = \begin{cases} 1, & \text{if } x \in S \\ 0, & \text{otherwise.} \end{cases}$

To better illustrate the practical interpretation of the ECDF we give a short example from one of our studies, in which subjects drove through a simulated environment while performing several cognitive tasks. The ECDFs of the distributions of saccade amplitudes for each task reveal the differences between the conditions (Figure 3.1).

The ECDF stands at the base of a powerful non-parametric statistical significance test: the *Kolmogorov-Smirnov* test. In its two-sample form, the Kolmogorov-Smirnov test computes the distance between the ECDFs of the two given samples, $P_{1,n}$ and $P_{2,n}$, in other to decide whether they stem from the same distribution:

$$D_{n,n'} = \sup_x |P_{1,n}(x) - P_{2,n'}(x)| \tag{3.2.2}$$

The null hypothesis is rejected at the level α if $D_{n,n'}\sqrt{\frac{nn'}{n+n'}} > K_\alpha$

3.3 Multiresolution image representations

One of the first requirements for any artificial system dealing with visual stimuli is to represent the visual input in a manner that is convenient for the application it will be used in. There are many available options, from simple pixel representations, where an image is represented as a matrix of real numbers, to Fourier space, or to wavelet decompositions, each with its advantages and disadvantages [67, 68].

We have been confronted with choosing an optimal image representation in the study that we will describe in detail in Chapter 5; the practical part of the study involved recording and predicting eye movements on overlays of dynamic natural scenes. Both when creating the overlaid stimuli, and when extracting the features that were used in the prediction phase, for reasons we will later emphasize, we decomposed the time-varying scenes using image pyramids.

Image pyramids were introduced to image processing and to computer graphics (as mipmaps) around the beginning of the 80s [69, 70], and they can be considered one of the precursors of multi-scale wavelet analysis [71].

The basic idea behind pyramid analysis is the representation of an image using a set of lower resolution versions of itself. This image set is created iteratively, so that each element is generated by subsampling the previous one by a factor of two, which makes the sequence resemble a pyramid.

There is one important aspect that needs to be kept in mind when creating an image pyramid. When sampling a signal, in order to be able to perfectly reconstruct the original from the samples and to avoid the artificial introduction of additional frequency components (phenomenon known as aliasing), a certain degree of oversampling is necessary. More precisely, according to the Shannon-Nyquist sampling theorem [72], the sampling frequency must be greater than the Nyquist rate, value equal to the double of the largest frequency contained by the original signal. Therefore, before each subsampling, a low-pass filtering of the image is needed in order to eliminate frequencies higher than the half of the Nyquist rate for the new image resolution.

Since each level of an image pyramid obtained in this manner is in fact a low-pass filtered version of the original image, the structure is called a low-pass pyramid. Because often the filter used resembles a Gaussian filter, it is common for a low-pass pyramid to be called a *Gaussian pyramid*.

In a paper that is now considered a classic of the field of image processing, Burt and Adelson [69] introduce the concept of the *Laplacian pyramid* as a complete image code. Fundamentally similar to the Gaussian, the Laplacian represents the decomposition of an image into a sequence of band-pass components. In the same paper, the authors propose an algorithm for creating the Laplacian pyramid starting from the Gaussian decomposition of the image, in addition to an efficient algorithm for creating the Gaussian itself. In the following, we will go into the details of these algorithms, that are also illustrated in Figure 3.2.

Gaussian pyramid

The generation of a Gaussian pyramid always starts from the original image, I, that also constitutes g_0, the first pyramid level. Each subsequent level l is obtained by convolving the previous level $l-1$ with an averaging kernel w, and then discarding every second row and column. For a 5-by-5 weighing kernel, the value for each pixel (x, y) from level l of the pyramid can be formally written as

$$g_l(x,y) = \sum_{m=-2}^{2}\sum_{n=-2}^{2} w(m,n) g_{l-1}(2x+m, 2y+n).$$

The size of the smoothing kernel is usually decided as a compromise between filtering amount and computational cost. Additionally, several constraints are imposed on the kernel:

- first, to simplify the computations, the kernel should be separable

$$w(m,n) = \hat{w}(m)\hat{w}(n),$$

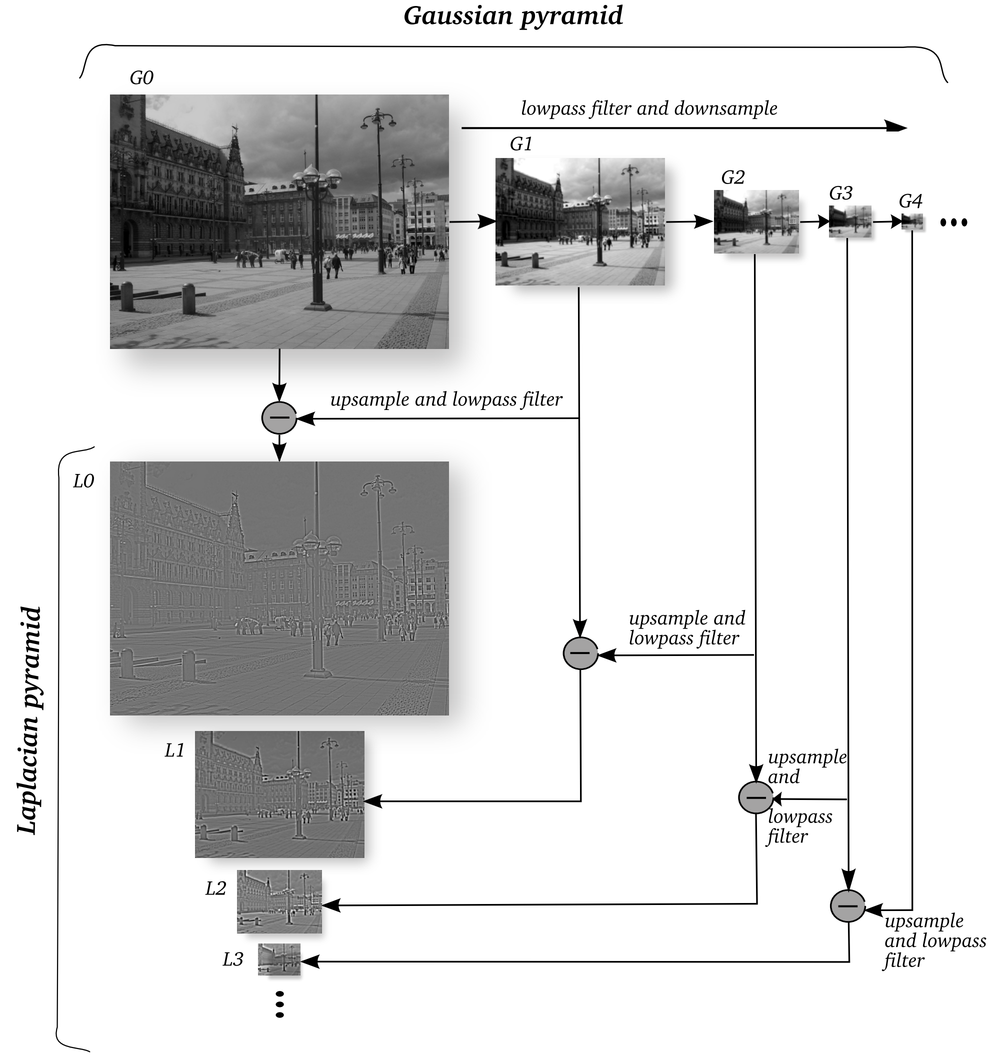

Gaussian pyramid
G0
lowpass filter and downsample
G1
G2
G3
G4
upsample and lowpass filter
upsample and lowpass filter
upsample and lowpass filter
upsample and lowpass filter
L0
L1
L2
L3
Laplacian pyramid

Figure 3.2: The process of creating the Gaussian and the Laplacian pyramid decompositions of an image. The original image constitutes the first level of the Gaussian pyramid (G_0). Each subsequent level in the Gaussian is obtained from the previous one by subsampling by a factor of two. Note that the last level of the Gaussian has a size of one, and contains the mean pixel intensity of the original image (the DC component). Before the subsampling, the image is smoothed to ensure all frequency components above the Nyquist rate for the new sampling rate are eliminated. Each level in the Laplacian pyramid is the difference between two subsequent levels of the Gaussian.

- the coefficients of the kernel should be symmetric

$$\hat{w}(i) = \hat{w}(-i), \text{ for } i = 0, 1, 2.$$

- to keep a constant energy per pixel, the coefficients of the kernel must be normalized:

$$\sum_{m=-2}^{2} \hat{w}(m) = 1$$

- also, all pixels at a given level must contribute equally to pixels from the next level

According to this algorithm, at the same time with an iterative downsampling by a factor of two, the bandwidth of the image is reduced by an octave at each pyramid level.

Laplacian pyramid

The procedure for creating a Gaussian pyramid can be reversed; in other words, a level l of the Gaussian pyramid can be expanded to obtain the level immediately above. Similarly to the generation process, the interpolation has two steps, this time in reverse order: first, each row and each column are duplicated, and then the resulting image is smoothed. The filter used is the same used in the downsampling process.

Each upsampled level $\uparrow g_l$ is however only an approximation of the original g_l level in the Gaussian pyramid, and their difference represents the level l of the corresponding Laplacian pyramid:

$$L_l = g_l - \uparrow g_{l+1}. \tag{3.3.1}$$

The Laplacian pyramid is therefore a complete image representation, as by summing all its levels, the original image is obtained:

$$g_0 = \sum L_l. \tag{3.3.2}$$

More importantly, as the difference between two low-pass images, each level of the Laplacian pyramid is in fact a band-pass image, and each level contains a different frequency range present in the original image. Analogous to the Gaussian pyramid, between two subsequent levels, the central frequency is reduced by an octave.

This brings us to the advantages offered by pyramid representations. While Gaussian pyramids are a computationally efficient method to access information situated on different scales in an image, Laplacian pyramids offer easy access to spatially-localized information from various frequency bands

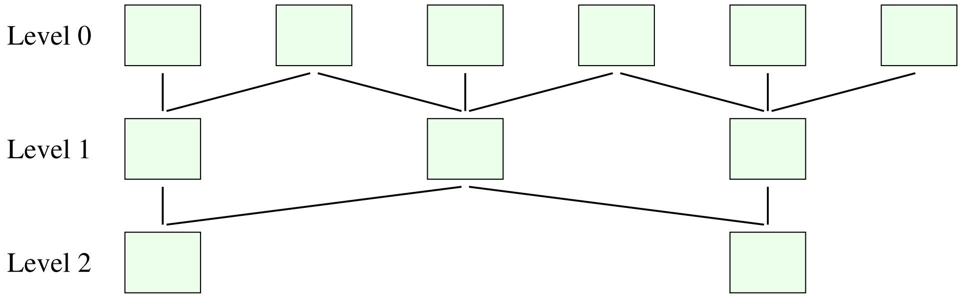


Figure 3.3: Illustration of a three level temporal pyramid. The figure also illustrates the size required for the temporal buffering window: to compute a frame from the last level of the pyramid, three history, and three lookahead frames are used in the first level of the pyramid.

in the image. This is a clear advantage when compared with the Fourier representation of an image, where all spatial information is lost during the computation of the transform.

One application of the band-pass decomposition that we have used, and that had already been suggested by Burt and Adelson [73] is the seamless merging of two images by separately equalizing the contribution of each frequency band to the final addition.

Spatio-temporal image pyramids

However, the stimuli we used in our studies were not static images, but time-varying scenes. One straightforward method to use pyramid analysis for movie clips is to decompose each frame separately, as Perry and Geisler [74] have done. However, the concept of image pyramids can easily be extended to the temporal domain, by adding to the classical spatial decomposition a temporal one [75].

A temporal domain analogue of an image pyramid is built similarly to a spatial domain Gaussian pyramid (Figure 3.3). While for a spatial pyramid at each level, every second pixel is eliminated, in the case of a temporal pyramid, it is every second frame that is removed. In this manner, on each pyramid level, the temporal resolution of the film is halved.

Of course, there are some implementation differences. Even if, due to the reduced resolution, a pyramid does not substantially increase the memory used compared to that needed by the single image (for example, storing a 5 level pyramid increases the memory load only with 0.33 compared to the original image), keeping in the memory the entire pyramid for a video clip may not always be feasible. Therefore, it is necessary to implement a buffering system, in which only the number of "history" and "lookahead" frames needed to compute all levels of the pyramid corresponding to a frame at one moment in time are analysed.

It is straightforward to combine the concepts of spatial and temporal pyramids and create a spatio-temporal Gaussian pyramid: a video can be subsampled in the temporal domain, while at the same

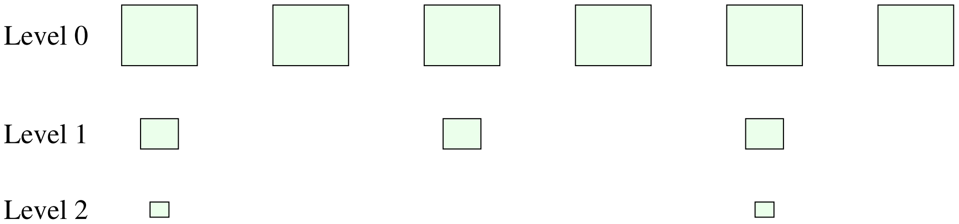


Figure 3.4: Spatio-temporal isotropic Gaussian pyramid with three levels. The concept of isotropy can be easily understood as a symmetry of the degrees of downsampling degrees in the spatial, and temporal domains: when moving "down" on the pyramid, both the temporal, and the spatial resolution decrease, so that levels with low temporal resolution contain frames with low spatial resolution, while high spatial resolution corresponds to high temporal resolution.

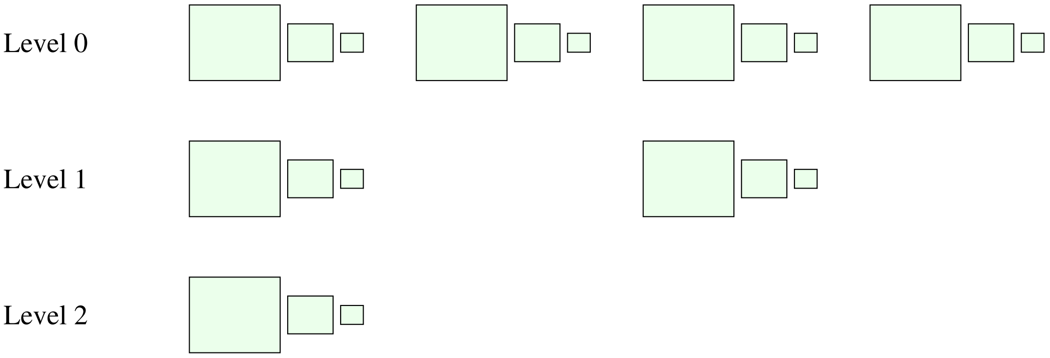


Figure 3.5: Diagram of a spatio-temporal anisotropic Gaussian pyramid with three spatial, and three temporal levels. The frames on each spatial level are further decomposed in the temporal domain. In the end, all temporal resolutions will be paired with each spatial resolution, creating a much finer partition of the frequency space.

time being subsampled in the spatial domain, on a frame by frame basis. There are two ways in which spatial and temporal filtering can be combined, and therefore, two large categories of spatio-temporal pyramids can be distinguished. If space and time are downsampled at the same time, the result will be an *isotropic pyramid* (Figure 3.4). If a spatial downsampling is performed first, and then each spatial level is further downsampled into temporal levels, the resulting pyramid is an *anisotropic pyramid* (Figure 3.5). The computational requirements for creating an anisotropic pyramid are larger than for an isotropic pyramid, but an anisotropic decomposition has the advantage of offering access to a much finer partition of the frequency spectrum.

A spatio-temporal Laplacian pyramid can be built similarly with a spatial one, but again, incurring substantial computational costs. Also, depending on how the subsampling is performed, spatio-temporal Laplacian pyramids can be *isotropic* and *anisotropic*.

3.4 Geometry of time-varying images

In the current section we will show how it is possible to describe local intensity variations in an image sequence using the concept of *intrinsic dimension*, and methods from the field of differential geometry.

Any time-varying image can be represented as a function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ that maps the three-dimensional set of spatial and temporal coordinates to the real values corresponding to the image intensity.

According to Zetzsche and Barth [76], the variation of the image function f can be characterized locally for any open region $\Omega \in \mathbb{R}^3$ in terms of its intrinsic dimension, or in other words, the number of degrees of freedom that are used for each point $(x, y, t) \in \Omega$. There are four possibilities:

- $f(x, y, t) = c$, the image is constant in all directions over Ω; this corresponds to 0 intrinsic dimension ($i0D$)

- $f(x, y, t) = g(ax+by+ct)$, the signal is constant in two directions, so the intrinsic dimension is 1 ($i1D$)

- $f(x, y, t) = g(a_1x + b_1y + c_1t, a_1x + b_1y + c_1t)$, the signal is constant in one direction, therefore it has an intrinsic dimension equal to 2 ($i2D$)

- no constant direction can be distinguished; intrinsic dimension 3 ($i3D$).

To give a only a few examples, in practice, $i0D$ signals are uniform and static image regions, $i1D$ regions correspond to stationary straight lines, or stationary straight edges, while stationary corners, or time-varying edges are $i2D$ regions. Transient corners, and motion that is not uniform constitute examples of signals that have maximal intrinsic dimension ($i3D$).

Mota and Barth [77] have shown that $i0D$ and $i1D$ image regions are redundant, and also that curved regions are unique, and sufficient for reconstructing the original image. That is of particular interest, as $i0D$ and $i1D$ regions appear more frequently in natural images [78].

Below, we will give an overview of how the intrinsic dimension of an image sequence can be estimated, as described in [79]. First, to formalize the concept of intrinsic dimension for a region Ω of a time-varying image, let us chose the maximal linear subspace $E \in \mathbb{R}^3$ for which f has a constant orientation:

$$f(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}), \forall \mathbf{x}, \mathbf{v} \text{ such that } \mathbf{x}, \mathbf{x} + \mathbf{v} \in \Omega, \text{ with } \mathbf{v} \in E. \quad (3.4.1)$$

In these conditions, the intrinsic dimension of f will be equal to $3 - dim(E)$.

Since the image sequence $f(x,y,t)$ can also be viewed as a hypersurface defined by the image intensity as a function of the spatio-temporal coordinates:

$$S(x,y,t) = x, y, t, f(x,y,t),$$

methods from differential geometry can be used to estimate its intrinsic dimension.

3.4.1 The structure tensor and the intrinsic dimension of image sequences

The equation 3.4.1 can be rewritten as

$$\frac{\partial f}{\partial \mathbf{v}} = 0, \forall \mathbf{v} \in E. \tag{3.4.2}$$

In this case, the subspace E can be estimated as the subspace spanned by the set of unity vectors that minimize the energy functional

$$\epsilon(\mathbf{v}) = \int_\Omega \| \frac{\partial f}{\partial \mathbf{v}} \| d\Omega = \mathbf{v}^T J_1 \mathbf{v}, \tag{3.4.3}$$

where J_1 is the structure tensor of f ([80, 81, 82]), and it is computed as:

$$J_1 = \int_\Omega \nabla \mathbf{f} \otimes \nabla \mathbf{f} d\Omega = \int_\Omega [f_x, f_y, f_t]^T [f_x, f_y, f_t] d\Omega, \tag{3.4.4}$$

with $\otimes$ being the tensor product, and f_x, f_y, f_t are the partial derivatives of f ($f_x = \frac{\partial f}{\partial x}$, etc.).

Equation 3.4.4 can be alternatively written as

$$J_1 = \omega * \begin{bmatrix} f_x^2 & f_{xy} & f_{xt} \\ f_{yx} & f_y^2 & f_{yt} \\ f_{tx} & f_{ty} & f_t^2 \end{bmatrix} \tag{3.4.5}$$

where ω is a low-pass spatio-temporal filtering kernel. In these conditions, E will be the subspace associated with the smallest eigenvalue of J_1, and the intrinsic dimension of f will correspond to the rank of J_1. This can either be obtained following an eigenvalue analysis of J_1, or using its symmetric invariants, H, S, and K ([83]).

$$\begin{array}{rcrcl} H &=& \frac{1}{3} trace(J_1) &=& \lambda_1 + \lambda_2 + \lambda_3 \\ S &=& M_{11} + M_{22} + M_{33} &=& \lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 \\ K &=& |J_1| &=& \lambda_1\lambda_2\lambda_3. \end{array} \tag{3.4.6}$$

M_{ii} are the minors of J_1, and are obtained from the original matrix by eliminating the row $4-i$ and the column $4 - i$.

Based on the values of the invariants of J_1, the intrinsic dimension of a region can be estimated: regions where $H > 0$, are at least $i1D$, while regions with $S > 0$ are at least $i2D$. Finally, if $K > 0$, then the region is $i3D$.

3.4.2 The generalized structure tensor and multiple orientations

The structure tensor gives an accurate description of regions that are unambiguously $i2D$, or $i3D$. However, it fails to distinguish between various additive overlays [2]. For example, the superposition of two $i1D$ signals cannot be set apart from a pure $i2D$ signal. The same happens when analysing two overlaid $i2D$ patterns, or the superposition of one $i1D$ and one $i2D$: in both cases, the structure tensor would have full rank (3), so maximal intrinsic dimension, which would make distinguishing between the two regions, or between the two regions and a simple $i3D$ signal impossible.

Nevertheless, it is possible to extend the definition of the structure tensor to *generalized structure tensors* (J_2 and J_3), that are capable to distinguish between these superpositions. In the current subsection, we will detail on how J_2 can be built and used for intrinsic dimension analysis, following the description given in [83] and [2].

J_2 is defined as:

$$J_2 = \int_{\Omega} [f_{xx}, f_{yy}, f_{xy}, f_{xt}, f_{yt}, f_{tt}]^T [f_{xx}, f_{yy}, f_{xy}, f_{xt}, f_{yt}, f_{tt}] d\Omega. \tag{3.4.7}$$

As before, computing the integral in equation 3.4.7 amounts to computing the following convolution

$$J_2 = \omega * \begin{bmatrix} f_{xx}^2 & f_{xx}f_{yy} & f_{xx}f_{xy} & f_{xx}f_{xt} & f_{xx}f_{yt} & f_{xx}f_{tt} \\ f_{yy}f_{xx} & f_{yy}^2 & f_{yy}f_{xy} & f_{yy}f_{xt} & f_{yy}f_{yt} & f_{yy}f_{tt} \\ f_{xy}f_{xx} & f_{xy}f_{yy} & f_{xy}^2 & f_{xy}f_{xt} & f_{xy}f_{yt} & f_{xy}f_{tt} \\ f_{xt}f_{xx} & f_{xt}f_{yy} & f_{xt}f_{xy} & f_{xt}^2 & f_{xt}f_{yt} & f_{xt}f_{tt} \\ f_{yt}f_{xx} & f_{yt}f_{yy} & f_{yt}f_{xy} & f_{yt}f_{xt} & f_{yt}^2 & f_{yt}f_{tt} \\ f_{tt}f_{xx} & f_{tt}f_{yy} & f_{tt}f_{xy} & f_{tt}f_{xt} & f_{tt}f_{yt} & f_{tt}^2 \end{bmatrix}, \tag{3.4.8}$$

where w is again a spatio-temporal, low-pass filtering kernel.

As before, the rank of J_2 or equivalently, its geometric invariants can be used to characterize the intrinsic dimension of a region in a time-varying imag. The invariants of J_2 are computed similarly to the invariants of J_1, the main difference being that instead of three eigenvalues, six are available. Thus, the "S" invariants of J_2 are defined as the sum all possible products of 2, 3, 4, and 5 eigenvalues. For example S_{22} will be the sum of the 15 products of eigenvalue pairs, while S_{25} will be the sum of the 6 products of 5 eigenvalues. The H and K invariants constitute the extreme cases: H_2 is

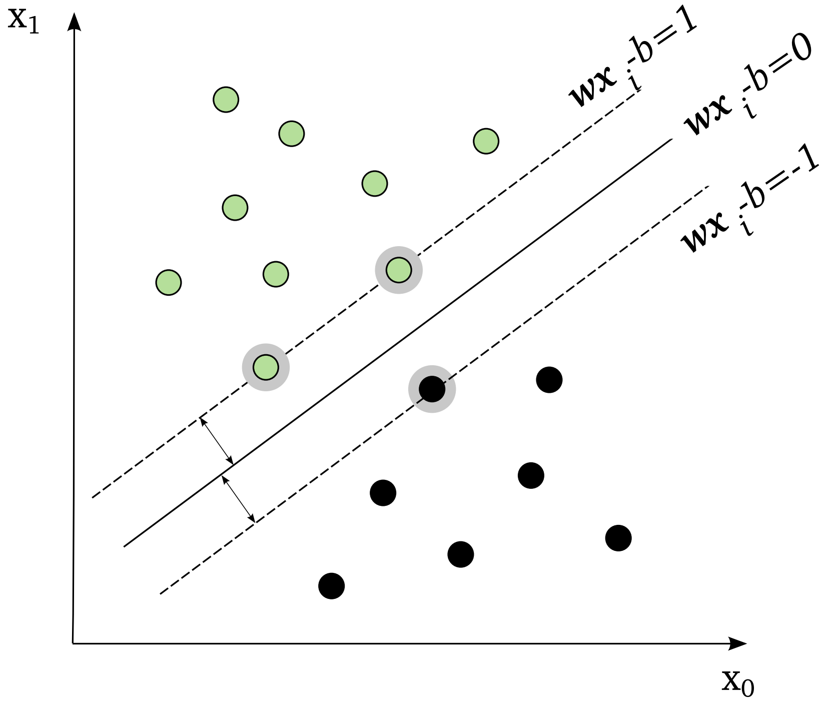


Figure 3.6: Illustration of the basic support vector machine (SVM) principle, in the two-dimensional, separable case. The goal of the classifier is to find the hyperplane that separates the two classes (represented by green, and respectively black disks) in such a manner that its distance to the closest data points is maximal. The highlighted data points represent the support vectors.

the sum of all 6 eigenvalues, where as K_2 represents their product.

3.5 Support Vector Machines

When predicting eye movements on superimposed video stimuli (Chapter 5) we used a *Support Vector Machine* (SVM) to distinguish between the class of attended and non-attended video locations. Over the next paragraphs, we shall briefly describe the basic idea behind maximum margin classifiers in general, and SVM in particular. For a more in-depth description, we refer to textbooks such as [84, 85], or [86].

In the simplest classification case, the two classes to be differentiated are linearly separable (Figure 3.6). If the classes are linearly separable, then a set of labelled training points

$$(y_1, \mathbf{x}_1), \ldots, (y_l, \mathbf{x}_l), y_i \in \{-1, 1\} \tag{3.5.1}$$

can be separated by a hyperplane defined as

$$\mathbf{w}_0 \cdot \mathbf{x} + b_0 = 0 \tag{3.5.2}$$

The optimal separating hyperplane has the property that it is situated at a maximal distance to the nearest training vector in any of the two classes.

Considering the requirement that the margin should be maximal, the separating hyperplane can be described by the following inequalities

$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b \geq 1 \quad & \text{if} \quad y_i = 1, \\ \mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \quad & \text{if} \quad y_i = -1 \end{aligned} \tag{3.5.3}$$

or equivalently by

$$y_i(\mathbf{w} \cdot \mathbf{x}_i) \geq 1, \quad i = 1, \ldots, l. \tag{3.5.4}$$

Keeping in mind the constraint (3.5.4), finding the optimal hyperplane translates to maximizing the expression

$$\frac{1}{\|\mathbf{w}\|} \max_i [y_i(\mathbf{x}_i - b)], \tag{3.5.5}$$

or to minimizing the functional

$$\Phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2. \tag{3.5.6}$$

In this case, the margin, defined by the *support vectors* (the points closest to the separating hyperplane) will be maximal, and it will equal $1/\|\mathbf{w}\|$.

If the two classes are not linearly separable, then no hyperplane that perfectly discriminates between them can be found without projecting the data to a higher dimensional space. A solution is simply to allow that points are misclassified during training, and to find the hyperplane that minimizes the classification error. Called *soft margin SVM*, this extension to the linearly separable case was introduced by Cortes and Vapnik [87]. The classification error associated to each data point is quantified with the help of the *slack variables*, ξ_i. The slack variables are non-negative values defined as $\xi_i = 0$ for points that have been correctly classified, and are on the right side of the margin, $0 < \xi_i \leq 1$ for points that have been correctly classified, but are inside the margin, and $\xi_i > 1$ for points that have been misclassified.

The algorithm for finding the optimal separation hyperplane can now be rewritten as minimizing the functional

$$\Phi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{l} \xi_i\right), \tag{3.5.7}$$

under the constraint

$$y_i(\mathbf{w}\mathbf{x}_i - b) \geq 1 - \xi_i. \tag{3.5.8}$$

The constant $C > 0$ appearing in (3.5.7) is a parameter controlling the trade-off between a small error penalty and a large margin.

For our analysis, we trained a soft-margin support vector machine with feature points obtained according to the method described by Vig et al. [88], using the publicly available SVM implementation contained in the LIBSVM library [89].

3.6 Software framework

As briefly mentioned in Chapter 1, for the representation of visual stimuli and the recording of eye movement data, we used an already available C++ framework. To this framework, we made several additions.

First, we programmed software that computed the generalized structure tensor and its invariants, the theoretical details of which have been described in Section 3.4. For every pixel of every frame of the input video, this required computing the second order spatio-temporal image derivatives, followed by an eigenvalue analysis of a 6-by-6 matrix. Moreover, these computations were performed on an anisotropic spatio-temporal Gaussian pyramid, increasing manifold the computational overload, and thus making indispensable the parallelization of the operations.

Second, we implemented a movie blender, that could seamlessly superimpose several input movies, on a frame by frame basis. More details on the algorithm are given in Chapter 5. The superposition was performed using anisotropic spatio-temporal Laplacian pyramids.

However, the most substantial software contribution of this thesis is constituted by the framework used to operate the driving simulator experiment described in Section 6.3. The application functioned both as a data recorder, and as a controller for the two LED arrays that were gaze-contingently toggled as a function of the location of the subject-controlled car in the virtual world. The controller received UDP streams coming from the eye tracker and the driving simulator. After analysing the packets received from both streams, the controller decided whether an activation of the LEDs was necessary. It was essential that the latency for this analysis-decision loop to be minimum, as the processes needed to be run in real-time in order to make the LED response gaze contingent.

3.7 Chapter Conclusions

This chapter, together with the first, were designed to be a basic introduction to the theoretical concepts and techniques we will use, or we will base our assumptions on in the present thesis. In addition to the eye tracking technologies we have used throughout our research, as well as the statistical measures we employed to characterize the recorded eye movement data, we briefly presented the theoretical methods behind the framework we used in Chapter 5 to predict eye movements on overlaid movies.

4

Expertise and Eye Movement Strategies

4.1 Introduction

As discussed in the previous chapter, eye movements can reveal where an observer's attention is engaged. That, combined with the strong top-down component of visual attention, results in eye movement patterns that are strongly correlated with the task being performed by the observer.

In the context of his reading research, Émile Javal was one of the first to mention the correlation between the observer's sequence of fixations and the cognitive activity that was being performed [90]. Other examples of famous early studies that found connections between the scanpath adopted by the subjects and the task assigned to them are those of Buswell [62] and of Yarbus [50]. Since then, considerable research has been conducted on the specificity of eye movement patterns associated with various tasks [91].

In the following, we will elaborate on this with a few concrete examples. One of the most straightforward cases is that of reading, where gaze-contingent display studies have allowed a thorough mapping of the perceptual span (the area in which text can be recognized in one fixation), and of how the eye moves to compensate for its narrowness [92, 93]. When reading, the observer shifts his gaze in the reading direction with the help of small saccades (the size of 8–9 letter spaces), alternating with fixations during which letters or whole words are recognized. Some words are altogether skipped, while others are fixated more than once as a small part of the saccades are regressive [94]. The connection between gaze and action remains unmistakable even in mundane everyday activities, such as making a tea [95], or preparing a sandwich [96]. These two studies highlighted how the series of fixations made by observers is linked step by step to the objects involved in executing the task; the great majority of eye movements is directed to task-relevant objects, and moreover, gaze predicts the succession of actions involved in the task model [97]. Driving is another good example: when steering, drivers use the tangent point of the inside of bends to predict the curvature of the road [98]. Fixations to locations highly informative for the performed action are also evident in sports such as cricket [99], or soccer [100, 101], and we will discuss these in the following.

In the end, what all these results have in common is the fact that when a distinct task is being executed, different observers employ a relatively consistent eye movement pattern, an eye movement pattern that, one may argue, is optimal for that action: gaze is directed towards locations that prove to be most relevant for the requirements of the task at hand [102]. This raises the question of whether eye movement strategies evolve with the learning of a task, and whether observers learn which locations are important to attend to.

Numerous studies report differences between expert and novice gaze patterns. In the previously mentioned cricket experiment, Land and McLeod find that although all players exhibit similar strategies – an anticipatory saccade close to the point where the ball would bounce, and then rapid tracking of the last part of the flight of the ball – experts show improved tracking abilities, combined with smaller latencies when responding to the appearance of the ball [99].

Also according to Savelsbergh et al., expert goalkeepers are more accurate in predicting the direction of penalty kicks, and they take their visual cues from different areas, such as the legs, or the ball area [100].
Similarly, visual strategy differences can also be observed between novice and expert drivers [103, 104, 105, 106].

However, such a learning process is not only identifiable in active tasks. Expert chess players, in addition to choosing the optimal move faster than novices, have been shown to make greater amplitude saccades, combined with fewer fixations that were often directed towards empty squares [107], suggesting that they are able to perceive more squares during one fixation [108].

The topic of evolving eye movement strategies can often be encountered also in medical imaging. To stop at only one example, Kundel and La Follette [109] compare search patterns of laymen, medical students, radiology residents, and experienced radiologists searching chest radiographs for lung nodules. They find significant differences both in the scanpaths adopted by different observer groups and in the number of fixations performed before deciding on the presence of the lesion. The example of pulmonary radiography presents itself of special interest also because the initial findings have generated relatively successful experiments aiming to find methods to improve detection performance. We must mention here a group of experiments that investigated how visual feedback to the observers' eye movements influenced later detection tasks [110, 111]. Lichtfield et al. describe an interesting study, in which under- and postgraduate radiographers searching pulmonary X-Rays for nodules were provided also with a condition in which they were presented with a preview of another radiographer's eye movements [112]. The results of the study were promising, suggesting that observers have benefited from the eye movement preview condition.

In a similar study, Dorr et al. [113] show that learning to classify fish locomotion patterns was facilitated by the visualization of the eye movements that were recorded on instructional videos, and that belonged to an expert in the field.

4.2 Expert and novice eye movement strategies in a gaze-controlled game

As listed above, there are many examples of activities containing an important visual component, for which different eye movement patterns are employed by experienced and novice actors. However, the question arises whether similar strategy contrasts would develop in an environment that is entirely gaze-controlled.

In order to investigate this issue, we observed subjects with different expertise levels playing a gaze-controlled computer game. For the experiment, we used an open source version of the popular game *Breakout* that was already adapted for gaze control for a previous study on gaze interaction [114].

4.2.1 Methods

Gaze controlled breakout

Breakout is an arcade game inspired by one of the earliest video games, *Pong* a table tennis simulation. The goal of the *Breakout* game is to eliminate several rows of bricks at the top of the screen by bouncing a ball against them (see Figure 4.1). In addition to eliminating the bricks, the player must also keep the ball from touching the bottom edge of the display by deflecting it using a mobile paddle that can be shifted horizontally.

The one dimensional input method, combined with the simplicity of the game process makes *Breakout* a perfect candidate for transforming it into a gaze-controlled game.

An instant hit when it appeared in the mid '70s, Breakout has spawned over the next 35 years a significant number of clones offering improved graphics and additional features. The clone used in the study is *LBreakout2*, an open source version of the game that was published under the GNU General Public License [115]. The GPL license allowed for modifications to be brought to the game source code, with the condition that the result is also published under the same GPL license.

The game was modified to communicate with SensoMotric Instruments eye trackers. The modifications were fairly simple. As the eye tracker would send over UDP the position of the player's gaze on the display, it was enough to implement a unit that received the gaze data, and instead of setting the paddle to the position of the mouse, it set it in absolute coordinates to the horizontal position of the player's gaze. A more detailed description of the necessary modifications, as well as of the issues encountered when adapting *LBreakout2* for gaze control can be found in the article of Dorr et al. [114].

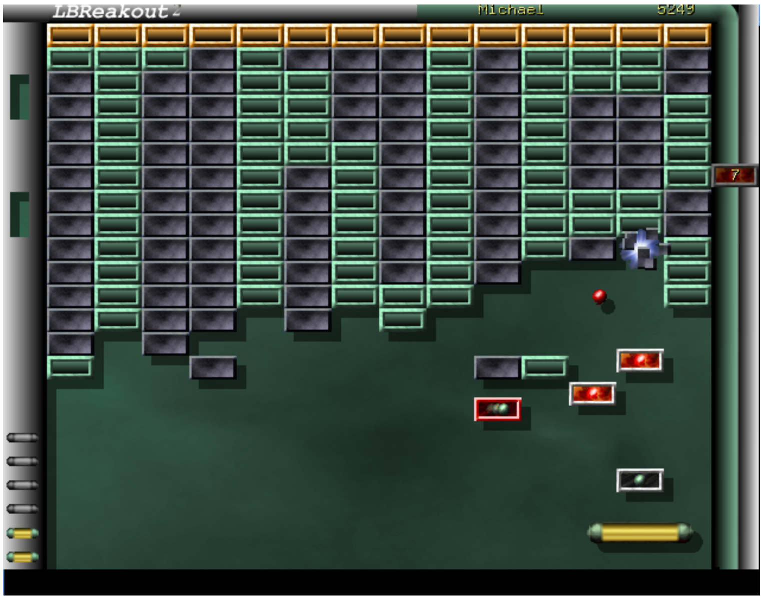


Figure 4.1: LBreakout2 screen shot. The paddle that can be seen in the bottom lower side of the image can be moved horizontally by the player. The ball, deflected by the paddle, or by the side walls of the screen causes the bricks it impacts dissolve. When hit, in addition to disappearing, some bricks release extras that can be collected with the paddle.

The experiment

We recorded data from nine volunteering observers that were playing the gaze-controlled *LBreakout2*. Five of them had considerable experience in playing the game, and thus constituted the expert group. Two of the expert subjects were actually involved in adapting the game for gaze control, but one of them was not aware of the purpose of the current experiment. The remaining four subjects had never played a gaze-controlled game before, and therefore formed the novice group.

All subjects were instructed to play until the completion of the first level, while also trying to maximize their game score – in other words to collect as many points as possible, while trying to keep all their lives. If all lives were lost, the game continued, but the score was reset to 0. It was considered that the level was complete either when all bricks were removed, or when too few bricks remained to allow for a comfortable removal. This is one of the disadvantages that playing with the gaze has: although reaction speed is improved, even small eye tracking noise, combined with the difficulty to inhibit miniature eye movements can make the fine control of the paddle more difficult. Because of this, aiming for bricks situated at extreme angles can be tedious. For all subjects, the trial took between 5 – 7 minutes to finish.

The subjects had to remove 16 rows, each containing 14 bricks (see Figure 4.1 for a screen shot). Of these, 22 randomly chosen bricks (approximately 10%) contained extra items, that could be collected with the paddle. Some extra items were “good”, for example they could bring extra score points, or elongate the paddle, or bring an extra ball. Conversely, some were “bad”, and triggered unpleasant effects such as shortening, or freezing the paddle. To avoid any bias, the types of extras were chosen randomly at the beginning of the trial.

The subjects were seated 55 cm away from a screen with a width of 40 cm and a height of 30 cm. Therefore, the game covered a visual angle of 40×30 degrees, and at the game resolution of 640×480 pixels, 1 degree of visual angle corresponded to 16 pixels on the screen.

In previous experiments involving the gaze-controlled Breakout, a remote eye tracker was used to record the eye movements of the subjects, allowing for a more natural gaming experience. However, for the current experiment, the much better accuracy of the head-fixed SMI HiSpeed eye tracker was preferred over the comfortable experience offered by the remote SMI RED-X. Although the default sampling rate of the HiSpeed was 1250 Hz, for the current experiment, such a high frequency was unnecessary. Moreover, the graphics of the display were updated at a frequency of only 120 Hz. For those reasons, the subjects' eye movements were recorded at sampling rate of 500 Hz. Before each trial, an “in game” 9 point calibration of the eye tracker was performed.

Data preprocessing

In order to analyse the eye movement strategies adopted by the subjects, we looked at which game items were mostly fixated, as well as at the distribution of the saccade landing points. We also examined the distance kept by subjects between gaze and ball.

However, computing these measures needed several preprocessing stages.

First, especially since for subjects with little eye tracking experiments experience it is relatively difficult to keep the head completely still for more than five minutes, a small degree of impulse noise was present in the gaze data. This manifested itself through implausibly large variations in the gaze position from one sample to another. To filter out the noise-affected samples, we computed the sample-to-sample velocities, and we discarded the 2% highest velocity samples. This resulted into eliminating samples with biologically implausible velocities, that often exceeded 1000 deg/s.

Secondly, when examining the gaze-to-item distance, problems arose when computing the test statistic. The gaze position was sampled at 500 Hz, a sampling rate much higher than that at which significant eye movements actually occur. For example, the eye position would be sampled at least 10 times during an average length saccade. Because of this, the distance measurements were highly correlated, a fact that violated the assumption of independent samples for statistical tests [116], and the test statistic needed to be corrected for the overestimation of sample size. For the correction, we evaluated where the autocorrelation function of the distributions dropped to 0.5, and we subsampled the distance distributions by that factor, averaged over all 9 subjects.

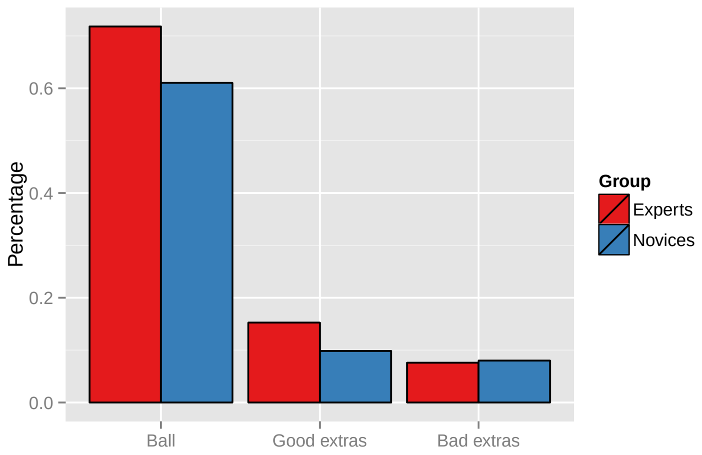


Figure 4.2: Proportion of time spent on various items in the game. The percentages are computed based on the total time in which the item was present on the screen. Experts spend more time looking at the ball and at the good extras than novices do. No differences can be seen in the case of bad extras.

4.2.2 Results

Distance between gaze and game items

Focus of fixations As a first step, we examined the time spent by subjects fixating various game items (Figure 4.2). In order to determine which game item was fixated at each moment, we computed the Euclidean distance between the gaze position on the screen and all items (balls and extras) visible at that given time. An item was classified as fixated if the distance from the item to the subject's point of regard was smaller than 5 degrees of visual angle (80 pixels).

As expected, all players spent most of the time looking at the ball. Also, expert players spent considerably more time fixating the ball than novice players did (71.8% of the total time in the case of experts, compared to 61% in the case of novices). This tendency could also be observed in the case of good extras, that were fixated by experts 15.3% of the time, and by novices 9.8%. However, for bad extras, there were no obvious differences between experts and novices (7.6% vs. 8%). It must also be noted that the percentages were computed relative to the time the item was visible on the screen. While in the case of the ball the "visibility" time was in fact equivalent to the total game time, in the case of extra items it was significantly smaller. To be more precise, for experts good extras were present 18.9%, and bad extras 10.2% of the total game time, while for novices, good extras were present 15.3%, and bad extras 9.2% of the total time.

Also, it can be noticed that for large periods of time, both for experts and for novices no obvious fixation target could be identified. Nevertheless, experts spent less time than novices looking

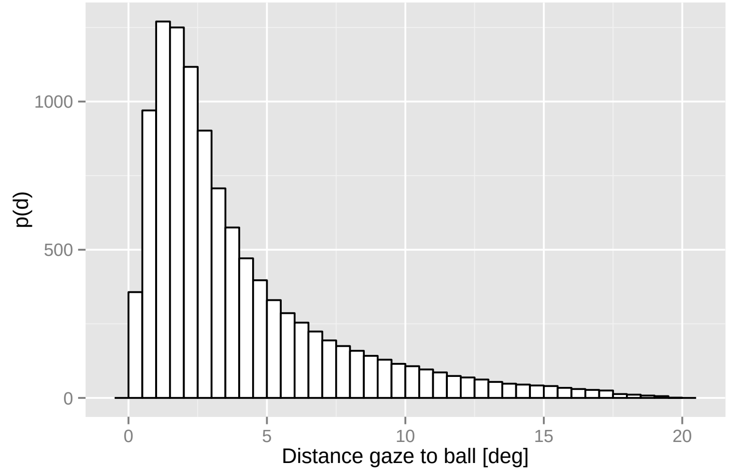


Figure 4.3: Probability distribution for the distances between gaze and ball. The computed figures contain only data for expert players, as the trend for the novice players was qualitatively similar. Note that more than 50% of the time, the subjects looked closer than 3 degrees to the ball.

nowhere in particular (24.5% vs 36.7% of the total time).

Gaze-to-ball distance We have previously shown that 71.8% of the time, expert players (61% in the case of novice players) fixate the ball. However, the question of how closely they follow the ball emerges. To answer that, we computed the Euclidean distance between the players' gaze and the position of the ball on the screen. In order to compensate for the bias caused by the non-independence of the samples, the preprocessing steps previously described were applied.

From the probability distribution plot in Figure 4.3, it can be observed that 50% of the time, experts gazed closer than 2.75 degrees from the ball (3.45 degrees in the case of novices). More than that, 75% of the time, the experts' point of regard was within 5.24 degrees to the ball (6.46 degrees for novices).

The differences in the gaze-to-ball distance between experts and novices are highly significant (Kolmogorov-Smirnov test, $p << 10^{-10}$), novices constantly keeping their gaze at a larger distance from the ball than experts (Figure 4.4).

This leaves the question of how the distance between the player's gaze and the ball varied during the game. One could expect this distance to decrease as the ball neared the lower part of the screen, to reach its minimum when the paddle should hit the ball, right at the bottom of the screen. To verify this hypothesis we plotted the horizontal distance gaze-to-ball as a function of the vertical ball position (Figure 4.5). For novice players, the expected trend can be recognized. The horizontal distance was maximal when the ball was at the top of the screen, and it decreased proportionally

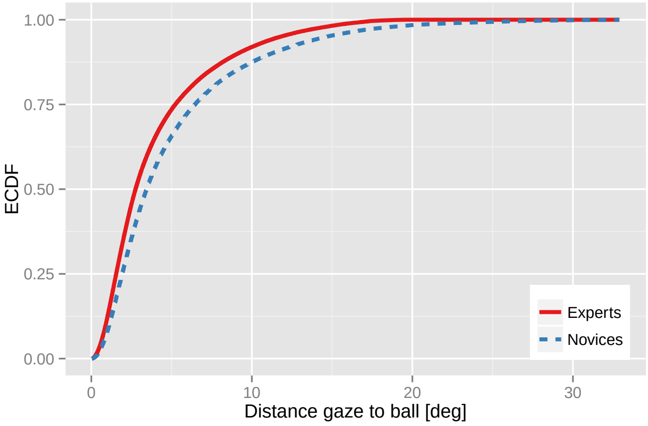


Figure 4.4: Empirical cumulative distribution of the distances between gaze and ball for expert and novice players. The differences between the two distributions are highly significant ($p << 10^{-10}$, corrected Kolmogorov-Smirnov test)

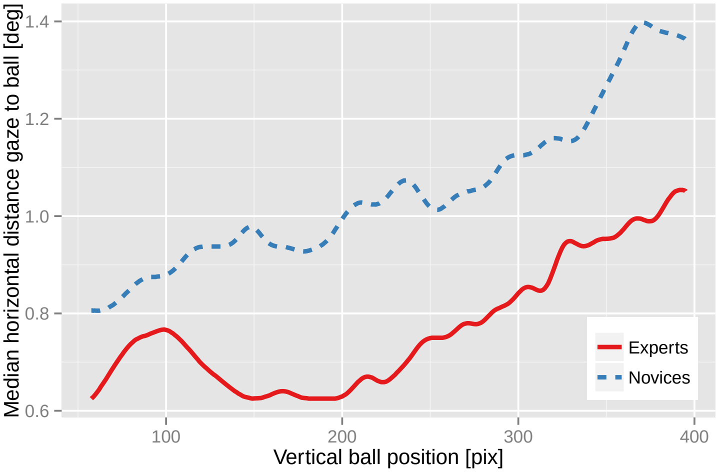


Figure 4.5: Horizontal distance between gaze and ball as a function of the vertical position of the ball on the screen. Overall, it is once more confirmed that experts gazed closer to the ball than novices did. For novices, the horizontal distance from gaze to ball decreased almost uniformly with the decrease of the vertical position of the ball. However, in the case of experts, a particular pattern could be observed, suggesting that they adopted a distinct strategy.

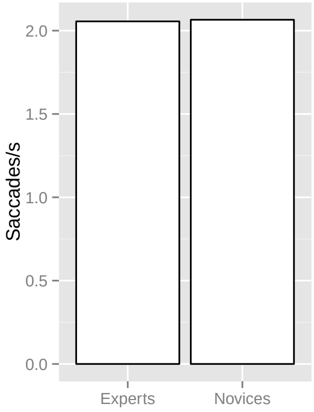


Figure 4.6: No significant difference can be observed between the saccade rate of expert subjects and that of novices.

to the vertical position of the ball, to reach its minimum when the ball was at the bottom of the screen. Several observations can be made when comparing the pattern of the novice to that of the expert horizontal gaze-to-ball distance curve. First, experts maintained throughout the game a smaller horizontal distance between their gaze and the ball. Second, both for experts and for novices, the distance was maximal when the ball was at the top of the screen. However, for experts, the distance curve reached the minimum already when the ball was close to the centre of the screen, to peak again when the ball was nearing the bottom, and finally to rapidly drop again to a minimum value when the ball was at the very bottom of the screen. It must be noted that this pattern can be individually observed for each expert subject, and it is completely absent in the case of novices. One possible explanation for the dip in the distance curve in the central area of the screen could be that experts were using the position of the ball when it was in the centre of the screen to predict where it would be when reaching the critical bottom position. This explanation makes even more sense when taking into consideration the fact that for a large part of the trial, the ball was deflected by bricks situated in the central area of the screen, this making the central part an even more informative area for the future trajectory of the ball. Lastly, the unusual distance peak just before the ball reaches the critical point could be explained as an artifice used by players in order to control which side of the paddle would hit the ball, thus controlling its future direction.

Saccades

Finally, we also extracted approximately 17,000 saccade points from the gaze data, after filtering it using the method described in the preprocessing section.

Although there were no significant differences in the number of saccades performed by novices

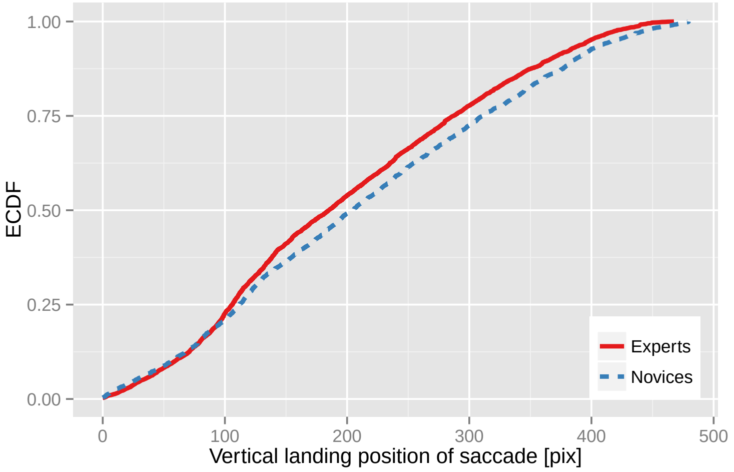


Figure 4.7: Empirical cumulative distribution function of the vertical components of the saccade landing points for both groups. Experts make significantly more saccades towards the area just above the bottom screen critical zone.

and experts (Figure 4.6), differences could be observed when analysing the vertical saccade landing points for each group. The empirical cumulative distributions of the vertical components of the saccadic landing points are visualised in Figure 4.7. Although both player groups perform a similar number of saccades directed at the bottom of the screen, experts perform more saccades that land on heights slightly above 100 pixels, so slightly above the critical bottom screen area. The differences between the behaviours of the novice and the expert group are highly significant ($p < 2.6 \cdot 10^{-5}$, Kolmogorov-Smirnov test). This result comes to support the previous observations, that experts adopt a distinct strategy just before the ball reaches the paddle.

4.2.3 Conclusions

Observers engaged in playing a gaze-controlled version of the game *Breakout* display distinct eye movement strategies depending on their experience in playing the game. This result confirms that in gaze-controlled environments, just as in the case of other activities with an important visual component such as drawing, driving, or reading, eye movement strategies change with expertise.

4.3 Chapter Conclusions

In this chapter we emphasized the fact that in many activities eye movements follow stable patterns that have been optimized for accomplishing that particular activity. However, these eye movement strategies are not available in their definite form from the first contact of the observer with the activity;

they evolve concurrently with the process of learning the task. This sets the premises for using gaze guidance as an aid to learning an optimal eye movement strategy, but also as a safety net for the case where the strategy employed by the observer proves inadequate to cope with the context at hand.

5

Prediction of Eye Movements

At the end of Chapter 4, we proposed the hypothesis that guiding gaze could be used as an aid in learning and performing tasks with complex visual components.

However, attempting to elaborate on this idea already unveils a first issue: an augmented vision system capable of gaze guidance needs to be able to capture gaze efficiently, and also in a relatively unobtrusive manner. The best way to accomplish that is by understanding what attracts eye movements to certain locations under naturalistic conditions. In order for the gaze guidance system to achieve unobtrusiveness, these natural mechanisms must later be emulated as closely as possible. But even before that, the first step in guiding eye movements is being able to predict them.

As described in the previous chapter, top-down mechanisms have a strong influence on where an observer fixates. Therefore, in the context of a well-defined task that is known to elicit a specific eye movement behaviour, it is easy to narrow down the "interesting" locations that will be fixated. Unfortunately, top-down mechanisms are most of the time difficult to study – it is not an easy task to establish the observer's internal agenda. However, above-chance eye movement prediction levels can be obtained using only salient characteristics of the observed scene, and there is strong evidence suggesting that in real life activities there is a strong interdependence between top-down and bottom-up mechanisms, where the latter act as modulators for the former.

In the following, we will focus on bottom-up saliency, and its use for predicting eye movements. After a very short overview of existing research on eye movements prediction on natural scenes, we will present our results in predicting eye movements on superimposed video clips. More details on our results can be found in [2, 5].

5.1 Introduction

As briefly discussed in Chapter 2, the characteristics of the viewed scene can influence eye movements, and areas in the environment that are salient can attract the observer's gaze. There is a wideranging body of research in the field of bottom-up gaze allocation. The following overview is designed only to give an idea of the main research directions that exist in saliency modelling, and it is not meant to be in any way exhaustive.

Many of today's successful computational frameworks of bottom-up gaze allocation have a starting point in the feature integration theory, proposed by Treisman and Gelade in [117]. This theory hypothesizes that features such as colour, orientation, shape, or movement are analysed separately, within parallel channels over the entire visual field. The feature analysis would occur automatically in the early stages of visual processing, while object identification and more complex analyses would take place at a later stage, in the presence of focused attention. Koch and Ullman introduce the notion of the saliency map, that encodes the global saliency in a scene, and that is obtained by integrating over the separate feature maps [118]. In their model, the location to be attended next is selected from the saliency map using "winner-take-all" mechanisms. Subsequently, this model has been used as a basis for many implementations and extensions. One of the first is that of Itti et al., who implement a neural network that can select attended locations from the saliency map built by combining multiscale image features [119]. Parkhurst et al. use the saliency model introduced by Koch and Ullman to investigate the correlation between computed stimulus salience and eye movement data recorded on static natural and artificial scenes, under normal viewing conditions [120]. Their results confirmed that the correlation between scene salience and eye movements was higher than chance. Itti [121] extended the saliency map to the temporal domain, and using recorded gaze data concluded that eye movements were more accurately predicted by areas with temporal changes than by colour or intensity.

Other studies infer basic features by studying the characteristics of the patches fixated by observers freely viewing natural scenes. For example, Reinagel and Zador study image statistics at the centre of gaze, and conclude that fixated patches show higher contrast and lower spatial correlations [122]. In a similar manner, Tatler et al. show that fixated locations tend to exhibit higher spatial frequencies, suggesting that contrast and edge information are preferentially fixated [123].

A slightly different approach stems from a set of hypotheses introduced in the context of the development of information theory research in the fifties/early sixties by Attneave [124] and Barlow [125]. These hypotheses view sensory systems as information-handling devices that therefore must be built in a way that reduces redundancies that exist in their inputs. In terms of human vision, this translates as the efficient coding hypothesis: the mechanism that drives eye movements selects its targets in a scene in such a manner as to efficiently encode the visual input, by maximizing the information gained following fixation.

One example in this category is the model developed by Bruce and Tsotsos [126], a model built using computational constraints that have the purpose to maximize the sampled information.

Another example is the model for early visual coding proposed by Barth and Watson [21], and developed in [127, 79, 2]. This model is based on the intrinsic dimension of the visual input; areas with low intrinsic dimensions, denoting uniform regions, or straight lines are deemed to be redundant, and therefore suppressed by the visual system. This is the model we used in our eye movement prediction research.

5.2 Eye movements on overlaid film clips

We stated in Chapter 1, that to a significant degree, our work continues and extends research conducted within the GazeCom project. That is specifically the case of our eye movement prediction research.

Attempting to create a simple saliency model, that makes as few a priori assumptions as possible, Vig et al. have used the intrinsic dimension of the visual input to predict eye movements on movie clips [128]. Their framework was based on the fact that a large number of video regions are uniform, or vary in only one direction, and thus are highly redundant (see the theoretical overview in Chapter 3), while regions with local variations of the signal are informative, and attract eye movements. They tested their framework on a large dataset of eye movements recorded on naturalistic movie clips, and found that all three geometrical invariants of the structure tensor give good prediction results. Invariant *K*, that segmented regions of the input signal where the intrinsic dimension was maximal (*i3D*) was shown to result in significantly better prediction both when compared to the other invariants, and when compared to state-of-the-art saliency models.

As we have mentioned in Section 3.4, the structure tensor can be extended to distinguish between higher order signal variations caused by the overlays of transparent motions. In the experiment that we will describe in the current chapter, we investigate how the transparent overlays influence eye movements, and also whether the results obtained by Vig et al. [128] with the structure tensor can be scaled on multiple motions with the help of the invariants of the generalized structure tensor.

5.2.1 Motivation

Although previously used in the lab as probing techniques for the mechanisms that handle motion processing in the brain (see [129] for a comprehensive review), transparent overlays, occlusions, and reflections are ubiquitous in nature. Leaves moving in the wind, layered clouds moving across the sky, or reflections on a transparent window are simple and very common examples. Because of this, eye movements on superimposed patterns constitute an interesting questions. In addition to increased complexity, the semantic content of the scene formed by overlays is reduced compared to that of its components. What we already know of the way multiple motions are perceived is that, as expected, the observers' performance in separating and recognizing overlaid patterns is lower than for single motions. Mulligan [130, 131] reports that subjects are able to distinguish up to two superimposed moving patterns, while Andrews and Schluppeck show that with sufficiently large difference between the movement directions, three superimposed patterns can be separated [132]. This is confirmed by the results of Dorr et al. [133], who show that although difficult, it is still possible to distinguish between 3 and 4 motions. Indeed, later studies show that the cost of perceiving multiple transparent motions depends on the difference between the directions of the motion patterns [134].

5.2.2 Data recording

We recorded eye movements from subjects that viewed a set of superimposed video clips (see Figure 5.1 for an example). The 19 video clips were obtained by blending movie pairs randomly chosen from a set of 14 high-resolution videos of outdoor scenes. The original clips are described in more detail in Dorr et al. [135]. Each movie had a resolution of 1280 by 720 pixels, and a duration of 19 seconds. They were recordings of natural dynamic scenes from various environments. Their content was fairly diverse, and it ranged from scenes with large uniform spatial regions, combined with low temporal variation, to scenes with high temporal and high spatial variation (such as pedestrians walking on a busy street).

To compensate for the event in which two films with very different frequency content would be paired, we equalized the contribution of each original clip to the final overlay, by performing a separate weighted addition on each frequency band. A comparison of the superposition results with, and without equalization is illustrated in Figure 5.2. Both component movies of a pair were decomposed using a spatio-temporal anisotropic Laplacian pyramid with 5 temporal and 5 spatial layers. The overlay was realized in two steps. First, the standard deviation for each frequency band of each of the paired movies was computed. Then, the blending weights for each frequency band were set separately, inversely proportional to the standard deviations computed in the first step. Only after this, every layer of the two pyramids was added using the computed weights, and the final stimulus movie was synthesised from the pyramid.

The duration of each of the final movies was 17 s, 2 seconds shorter than that of the originals because of their creation on a spatio-temporal pyramid.

The videos were displayed on a 22” Iiyama Vison Master Pro 514 CRT display, with an actual viewable diagonal of 20”. The subjects were seated 50 cm away from the display, and viewed the stimuli under a 43×23 degrees angle. As the aspect ratio of the display did not match the one of the videos, they were displayed in a “letterboxed” format, framed by black strips at the top and at the bottom of the screen.

The 10 subjects that took part in the experiment were instructed to freely view the set of 19 movies. The observers were all volunteers, and had normal or corrected to normal vision. Their eye movements were recorded using a SMI HiSpeed eye tracker running at a sampling frequency of

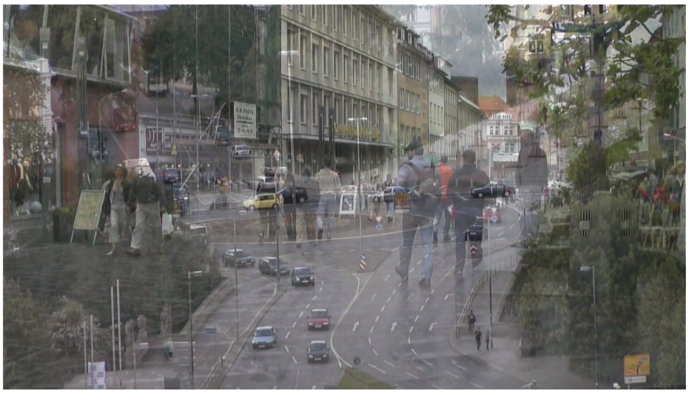
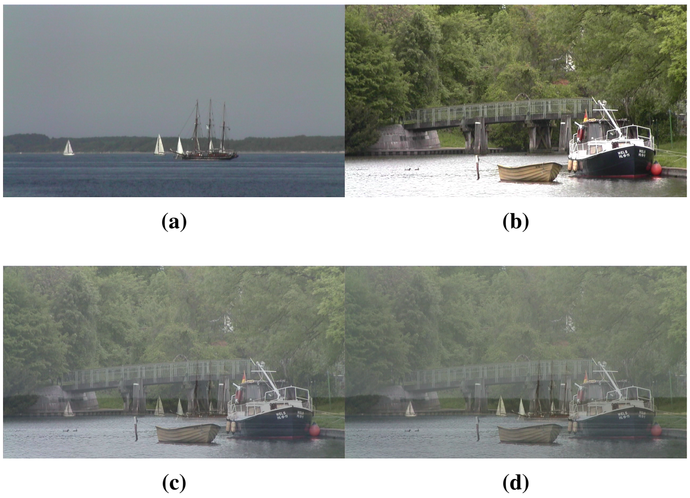


Figure 5.1: Still shot from one of the stimulus video clips. The stimuli were obtained by blending frame by frame randomly selected pairs from a set of 14 high-resolution film clips of outdoor scenes.



(a) (b)

(c) (d)

Figure 5.2: Blending process: (a) and (b) show stills from the original component films. (c) illustrates the result of a simple addition of the two component frames. The (b) clip visibly dominates the result of the addition. (d) An anisotropic spatio-temporal Laplacian pyramid is used; each pyramid level is blended separately, and the blending weights for each frequency band are inversely proportional to its standard deviation. The contribution of the two clips in the superposition is now equal. The effect of the energy equalization is apparent when observing the high-frequency foliage details corresponding to clip (b).

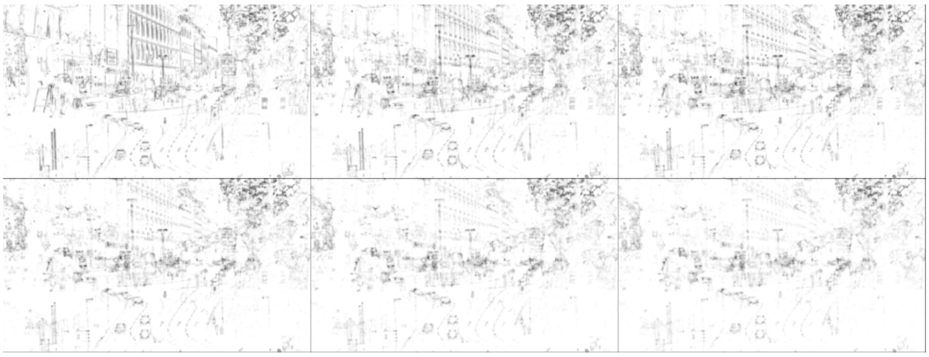


Figure 5.3: The six geometric invariants of J_2, computed for the still shown in Figure 5.1. From top left to bottom right, their order is $H_2, S_{22}, S_{23}, S_{24}, S_{25}, K_2$. The first spatio-temporal level on the spatio-temporal pyramid is shown. White areas denote regions where the value of the invariant is 0.

1250 Hz. Before each trial, a full 9 point calibration of the eye tracker was performed. Also, after each film clip a drift correction was run. After half the movies were displayed, the recording was interrupted, and the subjects were allowed to take a break. Before resuming the experiment, a new full calibration of the eye tracker was run. The movies were displayed for each subject in random order.

From the recorded eye movement data, we extracted saccades using the two step velocity-based algorithm described in Böhme et al. [136]. The resulting saccades were further filtered to eliminate samples recorded during blinks. After the filtering, a dataset of over 10,000 remained. The extracted saccade landing points were then used as test and training data in the prediction framework we will describe in the following section.

5.2.3 Prediction framework

For the prediction, we used the methods developed by Vig et al. [88], in which we replaced the invariants of the structure tensor (J_1) with those of the generalized structure tensor (J_2). Therefore, with the exception of the steps required for computing the invariants, the algorithm and the parameters we use are the same with those described in the previously mentioned paper.

Features

Invariants

To obtain representations on multiple spatio-temporal scales, we computed the invariants of J_2 on each level of an anisotropic spatio-temporal Gaussian pyramid with 5 spatial and 5 temporal levels. The computation of J_2 can be separated in three distinct stages.

- First, we created the second order derivatives. This was achieved iteratively: initially, we computed the first order derivatives by using one dimensional difference kernels of the type $[-1, 0, 1]$, on the input sequence that was previously smoothed with a 5-tap spatio-temporal binomial kernel. Then, we repeated the procedure above on the formerly computed first derivative to obtain the second differentiation.

- Second, we computed the 21 partial derivative product terms that are used in Equation 3.4.8 to estimate J_2.
- The last step was constituted by the convolution of the nonlinear product terms with a Gaussian smoothing kernel ω.

From J_2 it was then possible to obtain the invariants. We found during the implementation phase that it was no longer computationally effective to compute the minors of the tensor. Instead, we used a publicly available scientific library for C++ (gsl[1]) to perform the eigenvalue analysis on the tensor. Afterwards, we computed the invariants of J_2 as sums of products of eigenvalues, as described in Section 3.4. It must be noted though, in order to equalize the differences between the six invariants, instead of simply using products of eigenvalues, we used their geometric means. In Figure 5.3 we illustrate the invariants of J_2 for the same frame that is depicted in Figure 5.1.

In order to compare the prediction results obtained using the invariants of J_2 with those obtained using the invariants of J_1, we computed also $H_1, S_1,$ and K_1 using the steps described in [88]. As opposed to J_2, in the case of J_1 the eigenvalue analysis was not performed; instead, the invariants were computed using the minors of the structure tensor.

Signal energy

From this stage, we could have fed directly the invariant data to the classification framework. However, there are several reasons why it is more reasonable to use different features that, instead of using raw pixel data, offer a measure of the selected window around a saccade landing points. First, it is possible to have minor shifts in the eye movement data, caused both by eye tracker noise, but also by the imprecision of saccade planning. Second, using all the pixels in a movie patch, in addition to being computationally expensive, would lead to a huge increase of dimensionality of the resulting feature space. For these reasons, instead of using raw pixel data, we computed the signal energy in a spatio-temporal window around each saccade landing point (x, y):

$$e_{s,t} = \sqrt{\frac{1}{w_s^2} \sum_{i,j=-w_s/2}^{w_s/2} I_{s,t}^2(x_s - i, y_s - j)}. \tag{5.2.1}$$

In the previous equation, $I_{s,t}$ is the frame that corresponds to the spatial level s and the temporal

[1] `http://www.gnu.org/software/gsl/`

level t of the pyramid decomposition of an invariant I. As with every spatial level of the pyramid, the spatial resolution decreases by a factor of two, we have $(x_s, y_s) = (x/2^s, y/2^s)$, and also $w_s = w/2^s$. w_0 was set to 64 pixels. Also, the time window was approximately one second.

Following the considerations described above, each feature vector corresponding to a saccade landing point contained 25 components. Each of these components was the signal energy of the chosen invariant at a particular spatio-temporal scale, in a window around the considered saccade landing coordinate.

Data labelling and classification

From the computed feature data, we created a set of attended and a set of non-attended locations. The set of attended movie patches was built around the landing points of saccades previously extracted, while for the set of non-attended patches we shuffled movies and scanpaths. In this manner, the non-attended regions that corresponded to a movie were selected using attended locations that corresponded to another. Although this method did not ensure that the two classes were 100% non-overlapping, it guaranteed that the negative examples were also drawn from a distribution of specific human scanpaths. In addition to this, it made sure that artefacts caused by the centre-fixation bias ([137]) were removed.

From the positive and negative datasets we separated a training and a test subset. The training set contained data from two thirds of the ten subjects, recorded on all the movies, while the test set contained the remaining third. We made sure that data from any subject is only found in one of the two subsets. This way, in the test phase we could predict behaviour for "new" subjects.

For learning how to separate the attended and non-attended classes we used a soft-margin Support Vector Machine, with a Gaussian kernel. We performed the above analyses on all invariants of J_1 and J_2, for 20 separate subdivisions of the data in test and training subsets (20 training and test realizations).

5.2.4 Prediction results

In order to characterize the performance of our classifiers, we used the receiver operating characteristic (ROC curve). The ROC curve is a statistic that estimates the range of ratios between the true and false positive rates of a classifier. An ROC score of 0.5 means that the used classifier resembles a random classifier, while perfect discrimination is equivalent to an ROC score of 1.

Figure 5.4 illustrates the box plots of the ROC scores for the invariants of J_1 and of J_2. We used Wilcoxon's signed rank test to verify the significance of the differences between the ROC scores of various invariants. It can be noticed that K_1 has a higher ROC score than both H_1 and S_1. These differences are highly significant ($p(H_1, S_1) = 0.0019$, and $p(S_1, K_1) = 0.0089$). At the same time, the prediction rates are further improved in the case of the invariants of J_2. These improvements are highly significant starting from S_{24} ($p(K_1, S_{24}) = 0.0014$). All the invariants give very good prediction rates (over 72%), and also, all the higher order invariants, including also K_1 give median prediction rates over 78%.

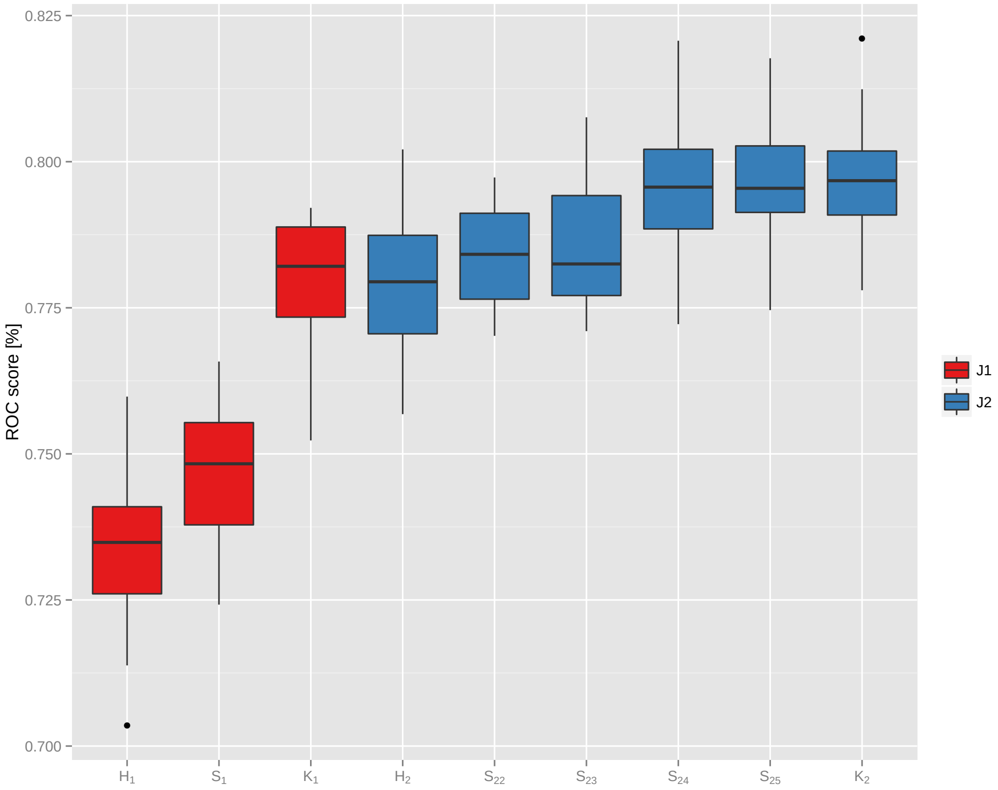


Figure 5.4: Box plot that compares the ROC scores of the eye movement prediction obtained using the invariants of J_1 and J_2 on overlaid movies over 20 training/test set realizations. The median, the lower and upper quartiles, as well as the minimum and the maximum value for a result set are shown in the plot. Outliers are represented by filled circles. Note a general tendency for improvement of the ROC score when moving towards the right on the horizontal scale, which also corresponds to an increase of the invariant order.

5.2.5 Discussion

Overall, the results confirm that the predictability of eye movements increases when features with a higher intrinsic dimension are used. This is apparent when observing the prediction rates obtained with the invariants of J_1, but the same effect is visible for the invariants of J_2. Moreover, the predictions based on the invariants of J_2, are significantly better than those based on J_1, confirming the hypothesis that redundancies are suppressed even in the complex case of overlaid motions. Further details on this can be found in [2].

5.3 Chapter conclusions

In the current chapter we have shown that eye movements on natural overlaid stimuli can be accurately predicted using the invariants of the generalized structure tensors and that higher order invariants are significantly more predictive.

6 Gaze Guidance in Driving

This chapter summarizes the main results of the current dissertation.

As discussed in detail in Chapter 4, there is a strong connection between the task performed at a certain moment by an observer, and their eye movements. As a short summary: activities with a strong visual component can very often be associated with very specific eye movement patterns that are consistent for different observers, and that evolve towards an optimal pattern during the process of learning that activity.

One of the best examples of a highly complex visually guided activity is driving. Every moment, the driver must have an overview of the often busy and highly dynamic visual scene around the car. From this, he must recognize and interpret the actions of other traffic participants, while at the same time keeping track of traffic signs and road characteristics. Often, in addition to the driving itself, the motorist is confronted to external distractions, such as conversations with a passenger in the car, or the operation of various electronic devices.

With all this in mind, it is clear that the limited attentional resources of the driver have to be optimally allocated. Even in “normal” driving conditions, decisions regarding various manoeuvres have to be taken at a rapid rate, and in a constant manner. At the same time, the driver must be able to immediately disrupt their normal course of driving, and to react in an adequate manner in the case of any unexpected event. Such critical events are unfortunately not uncommon; a pedestrian suddenly running in front of the vehicle, another car suddenly braking or changing direction are just a few examples.

Here, we show that gaze guidance can be highly effective in helping drivers prevent accidents in such critical conditions. The results, highlighted in Figure 6.5 are the first to show a clear beneficial effect of a gaze-guiding system.

6.1 Introduction

Over the last century, driving has become the main transportation modality in many countries, with the number of motor vehicles per capita still increasing. However, through extensive safety measures, it has been possible to achieve a significant reduction in traffic fatalities. A NHTSA research note reports the year 2010 as having the lowest number of deaths resulting from vehicle crashes in the US since 1949, a 2.9% decrease since the previous year. Nevertheless, this is the equivalent of 30,246 fatalities in the US only [138]. Also, according to the same note, more than 2,2 million have been injured in traffic crashes over the year 2010; of these, approximately 130,000 were not occupants of a motor vehicle.

Of the 33,883 fatalities registered in motor vehicle crashes in 2009, 16% occurred in accidents for which driver distraction was cited as the main cause [139].

Among the distractions most often reported are the use of mobile phones, or that of in-car technology. According to Klauer et al. [140], it is enough for the driver not to look at the road for a little over 2 seconds to significantly increase the risk of a collision. This verifies the fact that in-car electronic devices that interfere, even briefly, with the driver visually monitoring the environment have a strong impairing effect.

However, as Harbluk et al. [141] conclude following an on-road study, even when devices that do not interfere with the visual behaviour of the driver (such as hands-free mobile phones) are used, significant changes could be observed both in the visual scanning behaviour and in the driving behaviour of the drivers. This confirms that the additional cognitive load created by the use of in-car electronic devices is in itself sufficient to affect driving safety. Substantial research exists on increased cognitive load in driving. Some studies focus on investigating its effects on driving safety, while others search for novel methods to assess the amount of cognitive load a driver is subjected to. In the following, we will enumerate just a few examples. Recarte and Nunes [142] observed the influence of an increased cognitive load created by having subjects perform several mental tasks while driving in naturalistic conditions. Following a study that combined instrumented vehicle recordings with driving simulator experiments, Engström et al. [143] showed that increasing the cognitive load led to less visual scanning, more precisely to the concentration of the subjects' gaze towards the centre of the road. In a more recent experiment, Brookhuis and de Waard [144] used physiological measures such as heart rate and EEG data to estimate the drivers' mental workload.

Driving errors can sometimes be caused by visual perception failures. Often after an accident, phrases as "failed to look" and "looked but failed to see" are mentioned [145, 146]. It is common for such accidents to occur at intersections. "Looked but failed to see" (LBFTS) errors, where the driver fails to detect another traffic participant despite having scanned the area where that participant was, have generated a vast amount of research over the last 30 years [147, 148]. One explanation for such perception errors can be found in the phenomenon of change blindness, phenomenon that consists in a failure to recognize changes that occurred in a scene if they were accompanied by visual disruptions, such as blinking, saccades, or short interruptions in the scene [149, 150]. McCarley et al. [151] and Galpin et al. [152] investigate LBFTS driver errors from the perspective of the change blindness paradigm. In a similar experiment in a driving simulator, Zheng and McConkie [153] showed that local changes of the scene are frequently not noticed when accompanied by brief blankings (brief intervals in which a grey frame is shown).

Another direction of research on safer driving deals with modalities to redirect a distracted driver's attention to the road. Ho and Spence [154], for example, have concluded that spatially predictive auditory cues could be effective in capturing attention and signaling the fast approach of another car. In a similar manner, Wang et al. [155] conducted a driving simulator study of the effects and the interaction between a visual side collision-avoidance signal and an auditory cue. The visual cue consisted of an arrow pointing either to the left or the right of the display, and coincided either to the direction of the threat or with the direction of escape, while the auditory signal was triggered before the appearance of a threatening vehicle. They showed that subjects' responses to a visual signal consistent with the direction of the threat were significantly quicker than for trials where the location of the warning signal was incompatible with that of the threat. They also suggested that the cueing signal may in fact only direct the attention of the drivers towards the location of the threat, instead of triggering an immediate steering reaction.

With the latest advances in technology, higher-fidelity driving simulators started to become widely available. The number of driving studies performed in a simulator has increased significantly over the past decade [156]. Although driving simulators do have important disadvantages compared to naturalistic studies [157], the advantages of a simulator experiment are undeniable. Studies on distraction during driving can sometimes be problematic to perform. A study in naturalistic settings might be impractical for collecting sufficient data in a reasonable amount of time, and also cannot offer the same level of control on experimental parameters as an investigation performed in a laboratory. Also, ethics and safety aspects cannot be overlooked. Experiments involving subjects with visual impairments, such as the one performed by Bowers et al. [158] to determine how hemianopia affects the detection of pedestrians in hazardous driving conditions, but also benign driving under increased cognitive load experiments can potentially place both the subject and other traffic participants at serious risk when performed in real traffic conditions.

Nevertheless, the research on safer driving has not remained only on a theoretical level. Over recent years, more and more Advanced Driver Assistance Systems (ADAS) have been developed, and have been included as features in real vehicles.

Some ADAS focus on monitoring the driver and they issue warnings when he or she appears to be distracted or drowsy. Some systems use driving parameters (for example the steering behaviour) to monitor whether the driver's reactions are inconsistent with a safe driving pattern. Implementations of such driver assistance systems are already commercially available[1,2]. Others can detect directly when the driver's head has been turned away from the driving direction for a certain period of time, or when physiological measures, such as eyelid opening or blinking rate suggest that he or she is sleepy. For example, a system developed by SAAB detects the driver's eye and head direction, their blinking rate, and their eyelid closure, and issues an audible warning when the driver appears tired or distracted[3]. The Lexus driving monitoring system functions in a similar manner, and often comes integrated with a pre-crash application, which can detect if there is an obstacle approaching the car. If at the same time with the presence of an obstacle, the monitoring system concludes that the driver's head has been turned away for too long, warnings are triggered[4].

Other driver assistance applications focus solely on monitoring the street, and warn the driver when an immediate danger has been detected. An ADAS released by Volvo detects pedestrians ahead of the vehicle and issues acoustic and visual warnings if pedestrians are about to walk in front of the car; if no action is taken by the driver, the brakes are automatically applied[5]. The Subaru "EyeSight" system operates similarly[6].

Another direction in ADAS development focuses on creating enhanced-vision systems. Night view systems constitute a good example of this category; they render, either on a display on the dashboard, or directly on a portion of the windshield as a head-up display, an infrared view of the street ahead, sometimes enhanced with pedestrian detection. Night view systems have already been implemented using different technologies by car manufacturers such as Toyota, Mercedes or BMW[7,8]. Finally, head-up displays (HUD) are becoming a common feature offered by the automotive industry. Typically, they are used to display vehicle information such as speed or driving directions in a small portion of the windshield[9,10], but there are also attempts to extend HUDs to the use of the entire windshield. In 2010, General Motors disclosed that research is currently conducted on such a system; as envisioned use, they mention highlighting important aspects of the scene ahead of the car,

[1] http://www.daimler.com/technology-and-innovation/safety-technologies/driver

[2] http://corporate.ford.com/innovation/innovation-features/innovation-detail/ford-new-lane-technology

[3] http://www.saabnet.com/tsn/press/071102.html

[4] http://www.lexus.eu/range/ls/key-features/safety/safety-driver-monitoring-system.aspx

[5] http://www.volvocars.com/en-ca/top/about/news-events/pages/default.aspx?itemid=17

[6] http://subaru.com.au/about/eyesight

[7] http://www.wired.com/science/discoveries/news/2006/02/70182?currentPage=1

[8] http://www.toyota-global.com/innovation/safety_technology_quality/safety_technology/technology_file/active/night_view.html

[9] http://www.lexus.eu/range/rx/key-features/interior/interior-head-up-display.aspx

[10] http://www.bmw.com/com/en/insights/technology/connecteddrive/2010/safety/vision_assistance/head_up_display_information.html#more

such as the edges of the road, or traffic signs [11].

Nevertheless, there are downsides to the existing advanced driver assistance systems. Audible warnings are often triggered without taking into consideration the driver's intentions or the current traffic context, and because of that they are perceived as annoying and are turned off. Visual warnings add to an already significant existing visual demand, and even when they do not require the driver to take their eyes off the road, by increasing the cognitive load the driver is subjected to, they risk becoming a source of distraction themselves.

What we propose is to build enhanced-vision systems that can unobtrusively direct the drivers' eye movements towards critical events. Previous work has already shown that by using gaze-contingent interactive displays to render visual information with increased salience in selected regions, a gaze guidance effect can be obtained [159, 160].

An early proposal to use unobtrusive gaze guidance for better driving was made in [161]. Meanwhile, a simple unobtrusive gaze guidance system has been demonstrated and implemented in a prototype car [12] at the Volkswagen AG [162].

In the following sections, we will describe experiments that took place in special, gaze-contingent driving simulators, with the purpose to investigate how gaze guidance can be used to help drivers to more efficiently distribute their attentional resources and drive more safely.

6.2 Gaze guidance in a desktop driving simulator

In order to investigate whether it is possible to guide gaze during a driving task, and more importantly, whether beneficial effects could be therewith obtained, we conducted a series of experiments in a PC-based driving simulator. For more details, we refer the reader to [1, 3, 6, 7, 163].

In each of these experiments we used scenarios of normal driving in urban environments, in which we introduced potentially critical events, namely pedestrians unexpectedly crossing the street. The first study introduced drivers to gaze-contingent cues that highlighted directly the high-risk pedestrians, while the following two attempted to extend the results to more general cue types.

6.2.1 Experimental setup

The desktop driving simulator in which the experiments took place was integrated with a high speed remote eye tracker and allowed the gaze-contingent placing of gaze-capturing events (Figure 6.2).

[11] `http://media.gm.com/content/media/us/en/news/news_detail.brand_gm.html/content/Pages/news/us/en/2010/Mar/0317_hud`

[12] `http://www.spiegel.de/auto/aktuell/neue-warnsysteme-lichtorgel-statt-piepshow-a-562943.html`

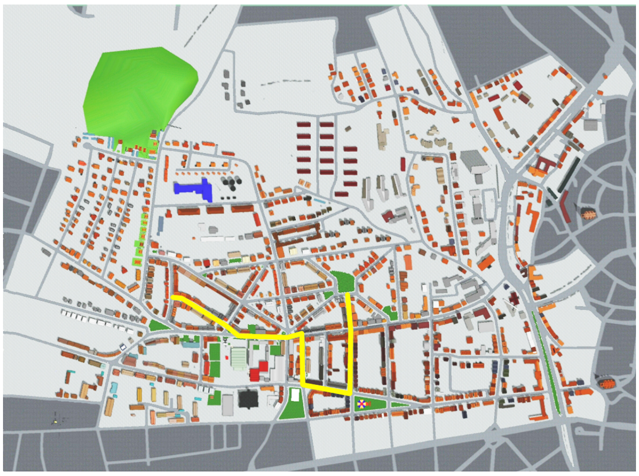


Figure 6.1: The map of the virtual city. One of the courses used in the first experiment is traced in yellow. The map was used when creating traffic scenarios to read the coordinates that defined all trajectories and event triggering points.

Virtual environment

The simulated environment modelled an existing urban area (the city of Osnabrück) with its roads and buildings. The virtual world was rendered under an angle of 73 degrees. The subjects viewed the virtual city from a car driver's perspective, and they controlled the car (the egocar) using a pedal/steering wheel system.

The graphical simulation provided a basic static content layer that consisted of streets, buildings, and green areas. To the static layer it was possible to add dynamic content, that comprised pedestrians and cars, road signs, and traffic lights. In addition to the traffic participants and the traffic regulation items, it was also possible to add external overlays to the simulation. These overlays could be added to the virtual environment, either relative to the simulator display, or "attached" to a traffic participant. Such overlays were used during trials both as gaze-contingent cues, and to guide the drivers along desired routes through the simulated city.

Pedestrians could be chosen from a set of eight distinct characters. The vehicle set comprised ten items, but since both variations in type and colour were possible, a large available set could be generated.

The traffic was guided and regulated with the help of a coherent traffic sign network. The available traffic sign set contained fifteen regulatory signs that controlled the right of way and the allowed traffic direction. Although it was also possible to create a traffic light network, because of limitations of the simulator engine, scenarios in which the subject was forced to stop at a red light could not be created. Therefore, only a very limited number of light signals were added, and they were always designed to turn green as soon as the subject approached.

Both for pedestrians and for cars, their attributes as well as the trajectory they followed had to be preprogrammed. Only the moment where a traffic participant would begin to move could be controlled, as an event triggered either by a certain position of the egocar in the virtual world, or by a certain position of the driver's gaze. For that reason, it was not possible to alter the behaviour of other traffic participants in response to specific actions of the egocar. Therefore their actions were triggered as events, in response to the 2D position of the egocar on the map of the virtual city. Because of this, all events were triggered in a similar fashion for all subjects, making the trials fully comparable.

The trajectories for each traffic participant were specified in terms of start and end coordinates. The coordinates were manually read from the 2D map of the virtual city (see Figure 6.1). The number of traffic participants was not explicitly limited, but several factors played a role in restricting the complexity of the planned scenarios. To mention just a few such limitations: the process of specifying the trajectories and event triggering points was in itself laborious and time consuming, and also, with a large quantity of dynamic content delays appeared in the triggering of the events.

Another characteristic of the driving simulator that needs to be mentioned is that both the acceleration and the brake pedal were very sensitive, so it was extremely difficult to adjust the speed of the egocar to an intermediate value. Although sometimes problematic for the drivers, due to this fact, the speed of the egocar was relatively uniform between subjects, making it possible to synchronize the event triggering from one driver to another. Otherwise, although all events were triggered at the same distance from a fixed location, because of speed variations, they would not have encountered the event-triggered pedestrian in the same position.

There were eight available speed levels, between which the subject or the experimenter could switch manually, by pressing a key on the keyboard. Since it proved quite difficult for subjects unacquainted to the driving simulator to control the egocar at higher speeds, the first experiment was run entirely in the first "gear", at a maximum speed of under 30 km/h. For subsequent experiments, in which higher speed levels were used, an extensive training stage preceded the actual trials.

Physical setup

The setup consisted of two computer workstations: one ran and displayed the simulation, while the other acted as a server controlling the events and the eye tracking device, a SensoMotoric Instruments RED250 remote eye tracker (Figure 6.2). The server and the simulator communicated through a direct ethernet interface, and the eye tracker was connected to the server via USB.

The participants were seated 70 cm away from the 22" display. The display had a spatial resolution of 1680x1050 pixels, and the viewing angle was approximately 38x24 degrees. The eye tracking was running at a sampling frequency of 250 Hz, and it was calibrated before each trial, using a 9 point

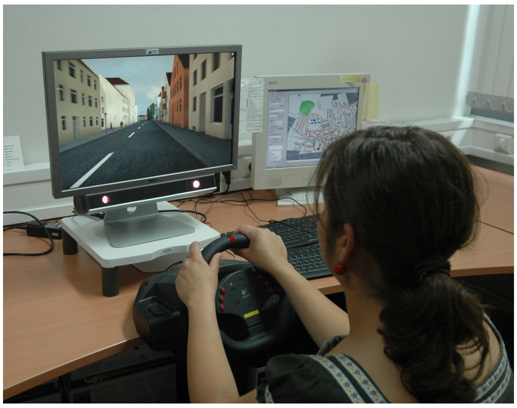


Figure 6.2: Still shot from a simulator recording. The remote eye tracker that was integrated to the driving simulator is visible below the screen. The subjects controlled the simulation using a pedal/steering wheel system. The DS server, displaying the map of the simulated environment, is visible to the right of the image.

calibration.

6.2.2 First experiment: pedestrian-centred cues.

In the experiment that we will describe below, the first of a series of two, we investigated whether distracted drivers are aided by gaze-contingent cues (GCCs) overlaid on high-risk pedestrians.

Methods

Subjects drove along pre-established courses inside the simulated city while performing additional cognitive tasks. Three distinct routes were selected, each of them stretching on average over a distance of 900 m. The drivers were guided along these routes by transparent directional arrows overlaid on the road at intersections.

Amidst benign traffic scenarios, each route had four or five potentially critical sections consisting of pedestrians unexpectedly crossing or coming close to the street. In total for the three routes, seven of the fourteen potentially critical sections would result in a collision between the egocar and a pedestrian in the absence of a prompt reaction from the driver.

The additional cognitive tasks were designed to act as a distractor, thus contributing to a more realistic driving experience. In the first task, the subjects were instructed to count the floors on all buildings along the route, and to remember the approximate location of the tallest one. In the second task, they had to search for an item (e.g. a copy shop) on the route, and to report how many occurrences of it they observed, and where they were located. In the last trials they were told to drive freely, but they were verbally distracted.

An experiment consisted of nine trials resulting from the combination of each route with each

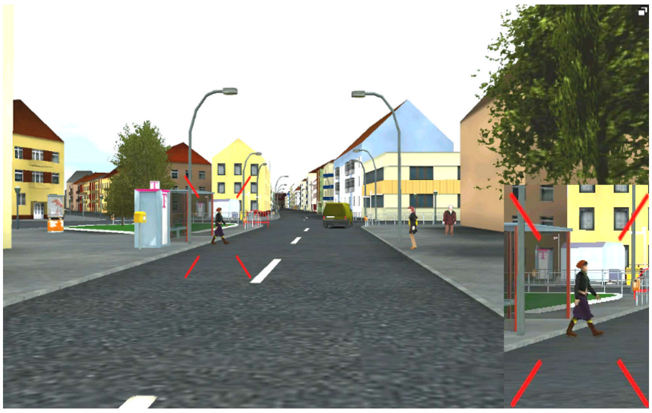


Figure 6.3: Simulator scene. Because the driver is looking away (red 3D marker), the pedestrian beginning to cross the street is highlighted with a gaze-contingent cue. Bottom right: gaze-contingent cue enlarged for better visibility (not shown during the experiment.)

task. In other words, each subject drove each route three times, every time with a different cognitive load. As the events on each course would repeat themselves in a very similar fashion, we tried to minimize the habituation effect for each subject by maximizing the time interval between two repetitions of the same route. Also, we made sure that the maximum level of distraction would be attained during the first repetitions. To that end, the tasks were repeated in decreasing order of their difficulty. In the first stage, all three routes were repeated with the counting task, in the second stage the subjects drove again all three routes, this time performing the visual search task, while in the last stage, they were allowed to drive freely, only with conversation acting as a distractor. The task difficulty was assessed empirically during preliminary trials. The route sequence was always presented in the same order.

Before the experiment, subjects were instructed to drive through the city following the directional arrows, while acting as if they were driving a real car through an inhabited city. They were told that it was of utmost importance to follow traffic regulations and to drive as safely as possible. Nonetheless, they were not explicitly warned about the possibility of pedestrians attempting to unexpectedly cross the street. All the experiments began with a short training route, in which the drivers were allowed to drive freely in a remote part of the city. Only when the subject was able to drive safely on the simulation road, the actual trials would begin. Including instructions and training, an experiment lasted on average thirty minutes.

For one subject group, the potentially critical events were highlighted with gaze-contingent cues (GCC) attached to the risk pedestrian and overlaid on the simulator scene. Several cue shapes, colours and transparencies were tested in pilot experiments in order to select a cue as unobtrusive as possible that would still be salient enough to capture the subjects' gaze. The chosen cue material was a simple

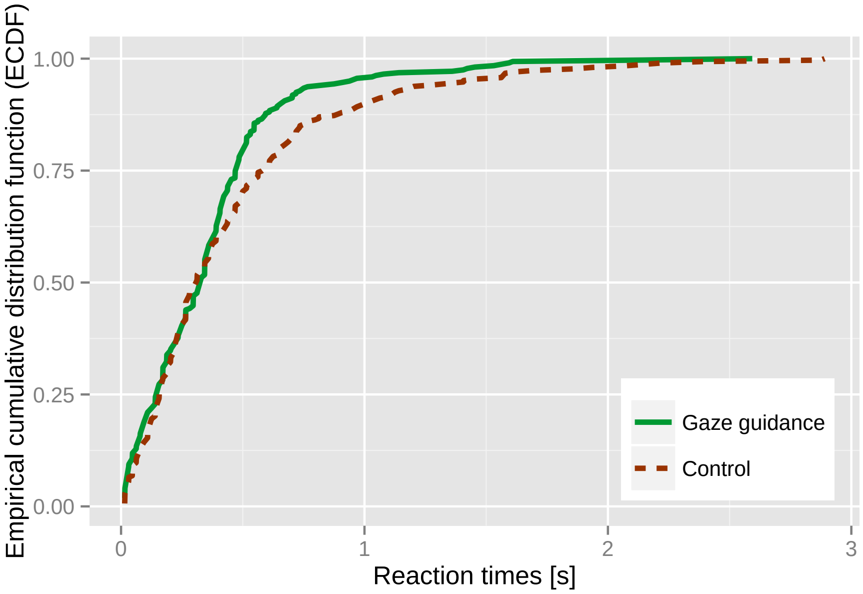


Figure 6.4: Empirical cumulative distribution function (ECDF) of the time needed by the subjects to first fixate a critical scenario pedestrian. Gaze guidance subjects show decreased reaction times, meaning that they fixate on the pedestrian sooner (mean/s.d. 0.375/0.37 s for GCC subjects, 0.487/0.72 s controls). The results are significant ($p = 0.011$, Kolmogorov-Smirnov test).

red overlay shaped like four rays converging on the pedestrian (see Figure 6.3). The cue would be triggered only when the subject was looking away from the danger element, and would be triggered off as an immediate result of the subject looking at it. The control group was not exposed to any GCCs.

We recorded data from thirty volunteering subjects with normal or corrected to normal vision (ten female and twenty male, with ages between 20 - 55 years). All had a driving licence with at least one year driving experience and variable computer gaming experience. Fifteen subjects were part of the gaze guidance aided group, while the remaining fifteen were controls.

From the over 400 minutes of gaze data, more than 75,000 saccades were extracted using the velocity-based algorithm described in [136]. The simulator also recorded driving parameters such as speed, pedal position and steering wheel inclination at a frequency of 60 Hz.

Results

In the analysis of the influence of GCCs on the driving performance of the subjects, no distinction was made between the different cognitive tasks. The data were pooled for each subject group over all three conditions.

Reaction times

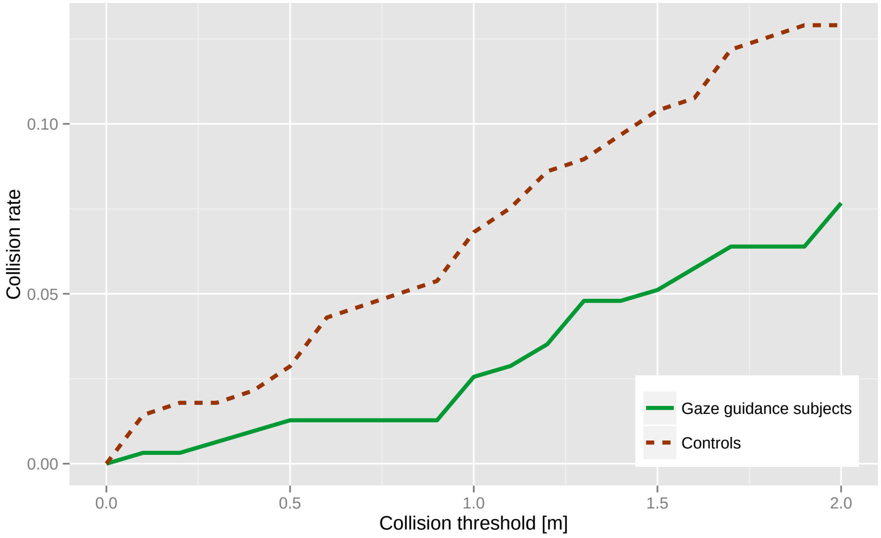


Figure 6.5: Collision rate as a function of distance threshold. The collision rate was computed as the fraction of critical event sequences where the minimum distance between the centre of the egocar and the critical pedestrian was smaller than the collision threshold. These data show that the collision rate does not critically depend on the distance threshold; for further analyses it was set to 1 m.

We examined the reaction time measured between the triggering of a critical event (a pedestrian crossing) and the first gaze hit on the pedestrian of interest, comparing the GCC and the control group.

A tendency for shorter reaction times can be recognized for subjects aided by gaze guidance. When examining the ECDF curves of the reaction times for the two subject groups, presented in Figure 6.4, 80% (0.8 on the y-axis) of the control subjects (blue dashed curve) had a reaction time of 650 ms or less. For the gaze-guidance group (red solid curve), the 80% mark was reached earlier, already at 500 ms, i.e. gaze-guided subjects reacted faster. According to the Kolmogorov-Smirnov test, the distance between the two curves is statistically significant, at $D = 0.1291, p = 0.011.$

Provoked accidents

As mentioned earlier, of the total number of events for which a GCC would be triggered if the driver looked away, seven had the potential to lead to a collision of the subject-driven car with the critical event pedestrian. To evaluate the subjects' driving performance, we looked at the number of accidents caused in the experiment.

Since the simulator did not provide any collision feedback, we used a distance-based algorithm to detect pedestrian-egocar collisions. We computed the distances between the centre of the egocar and the critical scenario pedestrian. Based on the dimensions of an average city car, we set the distance threshold for a collision to 1 m from the car centre. Nevertheless, the following result holds

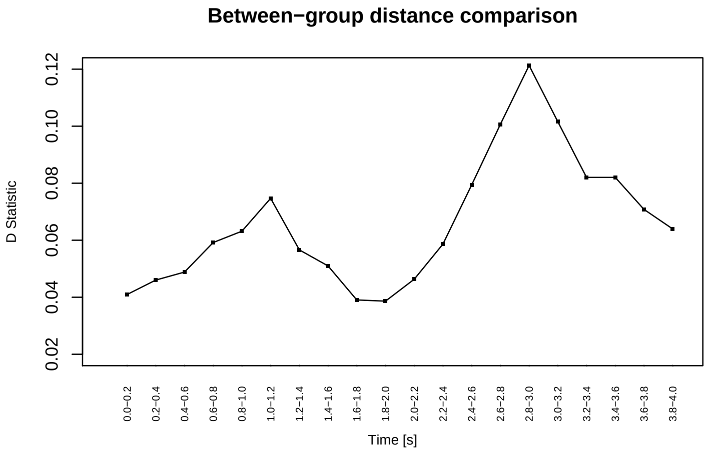


Figure 6.6: Maximum vertical distance between the ECDF curves of the egocar-pedestrian distance distribution (D statistic). The distance distribution in each time window contains data pooled from all seven potentially critical events.

qualitatively also for other distance thresholds (Figure 6.5).

We computed the accident rate as the total number of accidents for the group, divided by the total number of events which could have led to an accident for that group. We found that the accident rate is strongly reduced for the gaze guidance subjects (0.026), being less than half of that of the controls (0.068). This reduction is highly significant (99.8% confidence interval). In order to check whether the significance of the accident rate reduction holds also for thresholds other than that of 1 m, we computed the 95% confidence intervals for all distances larger than 0.5 m, up to 2 m, in increments of 10 cm. The results confirmed that for any of these thresholds, the collision rates for the two subject groups were significantly different.

Because of the reduced sample size, the differences between accident rates for each task did not reach statistical significance. However, it is interesting to note that for the control group, the largest collision rate was registered during the counting task (0.10), followed by the free driving task (0.06). The smallest collision rate was recorded for the search task (0.04). For the gaze guidance group, the order of the accident rates over tasks was the same (counting task, 0.05; free driving, 0.02; search 0.01).

In the following, we attempted to establish whether the differences suggested by the collision rates were consistent for the entire subject group, or whether they only apply for isolated cases which came near the critical accident distance. We also sought for evidence of the effect of gaze guidance in differences between the eye movement distributions of the two groups. To this end, we analysed the data recorded during a four second interval from the triggering of the critical event, i.e. the

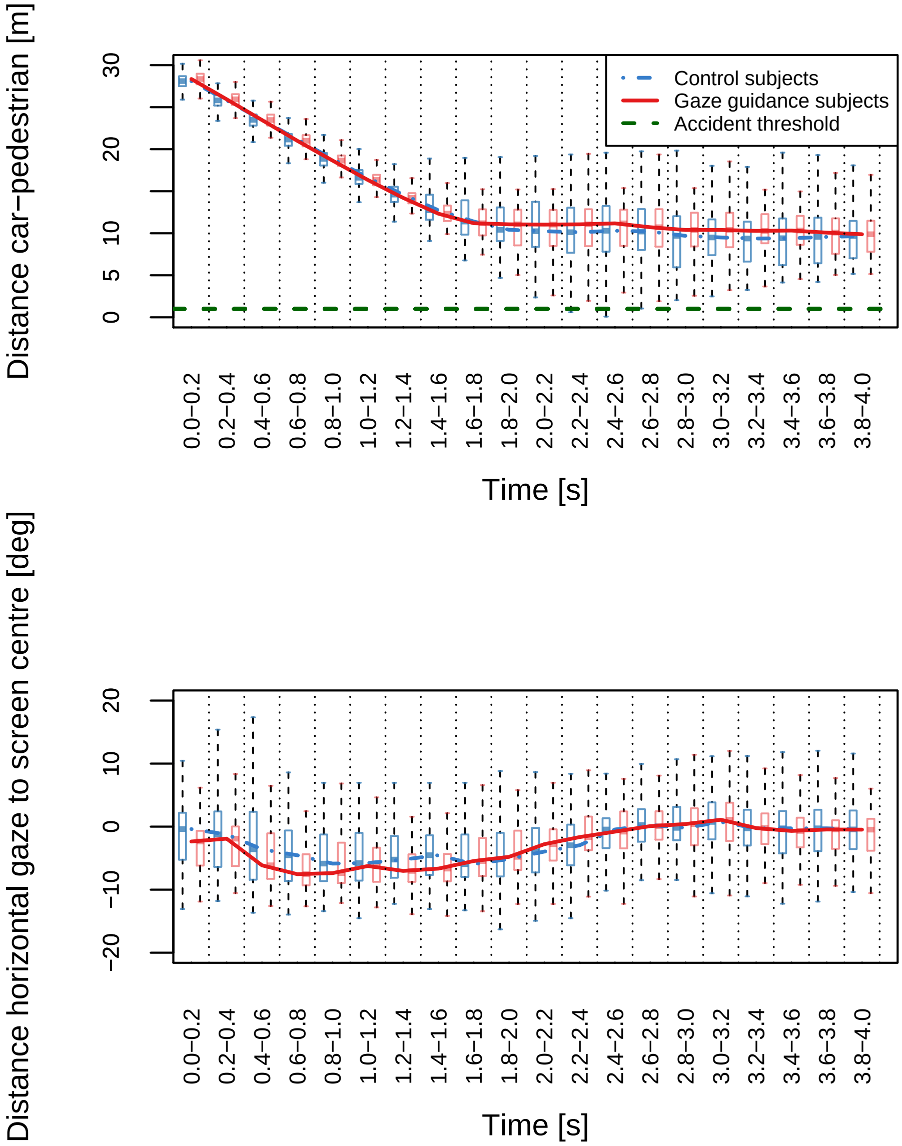


Figure 6.7: Example of car-pedestrian distance and horizontal gaze position during a high-risk event (see text for more details).

interval corresponding to the entire duration of the event (after which the pedestrian had crossed the street). We divided this interval into 200 ms time windows, and for each event, we plotted all egocar-pedestrian distances in the corresponding time window. To evaluate the eye movements performed by the subjects during critical events, we also analysed the horizontal gaze component for the same time intervals. A plot for a single event is shown in Figure 6.7. For better visualization, outliers have not been plotted.

In order to check whether the differences between the distance distributions for control and for gaze guidance subjects were significant, we pooled in each time window the data from all seven potentially critical events and then for each time window we computed and plotted the D statistic (Figure 6.6). The variation of the maximum distance between the two distributions confirms the tendencies illustrated in Figure 6.8.

Because of the relatively short duration of a critical event (a maximum of 4 seconds), not enough saccade samples were available to establish the statistical significance of the differences between the eye movement distribution for each time window. That is why, concerning gaze distributions, we will only describe tendencies.

There are slight variations from event to event, depending on various factors such as the direction and the distance from which the pedestrian appears, or the characteristics of the scene at event onset. Nevertheless, some observations remain valid for all events. For the first part of the event, the egocar-pedestrian distance decreases in an approximately linear fashion for all subjects, and the variance of the distance distributions is small. After approximately 1 s, typical reaction time, the egocar-pedestrian distance stabilizes near a constant value which, depending on the nature of the event can be close to the collision threshold. In this latter section, the distance tends to be larger for gaze guidance subjects.

Certain tendencies can also be distinguished in the horizontal gaze distribution. The variability of the gaze distributions for the first part of the event tends to be smaller for GG subjects. Also a shift in gaze position between the two subject groups can be noticed for that interval.

Next, we analysed the cumulated data over all events. For each time window of each event, we computed the difference between the medians of the distances, and also between the statistical dispersion (see below) of the gaze positions (Figure 6.8).

We chose the median as a measure of the central tendency of the distance distributions because of its resilience to outliers. The trend of the curve remains the same when using the mean instead of the median. GG subjects maintain larger distances to the pedestrian; the effect is particularly strong in time windows where accidents occur.

To quantify the statistical dispersion of the eye movement coordinates we also used a robust measure with regard to outliers, specifically the median absolute deviation (MAD). The MAD is computed as the median of all the absolute deviations from the sample's median. In the time windows

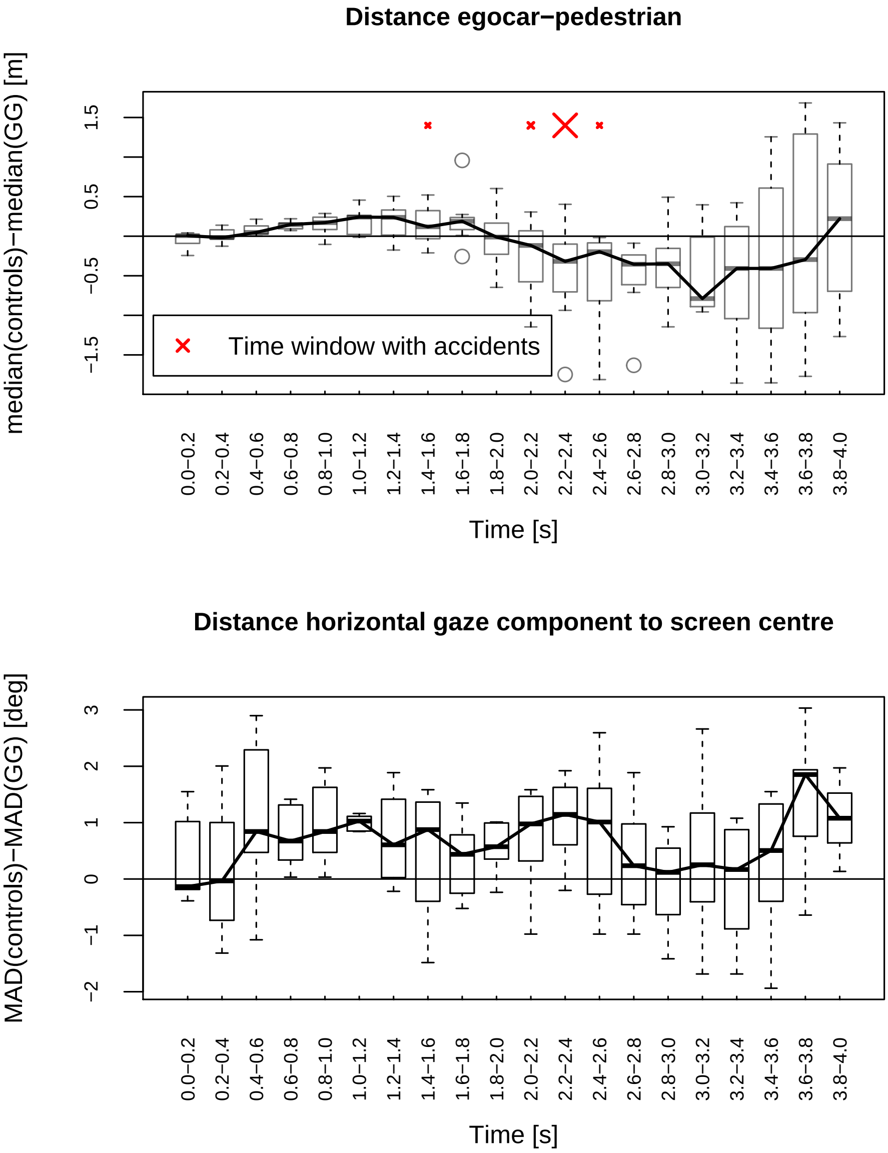


Figure 6.8: Analysis as in Figure 6.7 for aggregated data on all events. The top plot shows the difference between the medians of the distance samples for each time window during a critical event. Crosses indicate time windows where accidents took place. Note that cross size is proportional to the number of accidents in the corresponding time window. The bottom plot illustrates the differences between the variabilities (median absolute deviations - MAD) of the eye movement horizontal positions for the corresponding time windows.

between 600 ms and 12000 ms, the variability of the gaze positions of the GG subjects is smaller than that of controls. This time interval also corresponds to a local maximum in the distances between the medians of the horizontal eye movement distributions of the two groups (plot not shown here).

Discussion

Subjects in the gaze guidance group showed shorter reaction times and looked at the critical events sooner than control subjects. Still, as shown in Figure 6.4, approximately 60% of all subjects immediately fixate the pedestrian. However, the gaze guidance did help reduce the upper bound of the reaction time distribution, suggesting that for the 40% inattentive drivers, gaze guidance would have made a significant difference in a real scenario. We would argue that compared to a real traffic scene, the graphical environment of the simulation offers a relatively low number of cars and pedestrians, no real street life, etc. and therefore, the subjects were more likely to look at the pedestrians anyway. Also, the high frequency of "accident-generating scenarios" occurring during the trials increased significantly the gaze-capturing potential of the pedestrians that walked into the street. In real driving scenarios it is not likely that seven critical situations with pedestrians occur in less than 30 minutes of driving. Should that nevertheless happen, the driver would allocate significantly more cognitive resources to anticipating similar events. Therefore, we could expect an even stronger effect of the gaze-capturing cues in real life situations.

A further gaze guidance effect that we found was that before a potential accident eye movements were less variable in case of the GG group compared to the controls. Since the GG and the control conditions differ only in the presence of the GCC, the reduced variability of the gaze positions must be attributed to the GCC. We could further assume that the GG subjects were better focusing on the pedestrians, but we cannot verify this assumption because the position of the pedestrian on the screen is not precisely known.

The major finding, however, is that gaze guidance led to safer driving as GG subjects braked earlier and thus maintained a larger distance to the pedestrian. It must be noted that this increased safety zone cannot be observed during normal driving but just before the potential collisions. Overall this change in driving behaviour due to gaze guidance led to a major reduction in the number of accidents.

To conclude, we found that when safety-critical events were highlighted with briefly flashed, gaze-contingent cues, drivers attended to these events more quickly. More importantly, such gaze guidance led to a safer driving behaviour and a significantly reduced number of accidents, although subjects reported that they were not distracted by the cues, part of which went unnoticed.

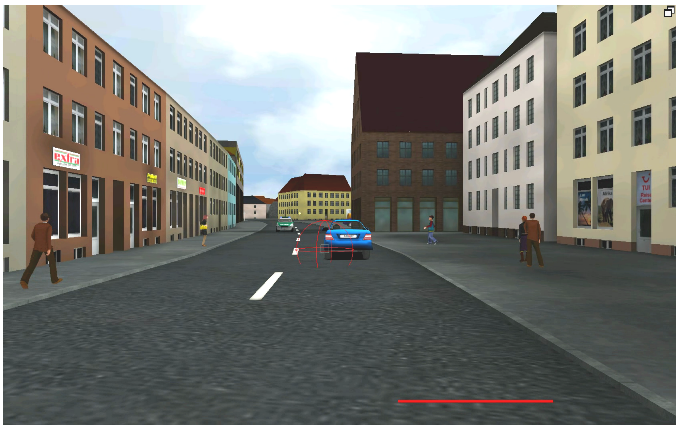


Figure 6.9: Stillshot from the second experiment. The GCC highlights the direction from which the high-risk pedestrian is walking. As before, the cues are only activated if the driver is not looking in the direction of the critical event, and they disappear when the driver fixates the pedestrian in question.

6.2.3 Second experiment: directional cues

The results of the previously presented experiment (Section 6.2.2) showed a significant reduction in the number of accidents for drivers aided by pedestrian-centred gaze-contingent cues. However, following that study, a series of questions needed to be addressed.

First, it had been argued that the possibility to implement gaze-guiding cues in an actual car is currently limited; both sufficiently accurate pedestrian detection, as well as head-up displays advanced enough for precise highlighting of traffic participants are not yet widely available.

Second, several issues regarding the methodology of the experiment, such as the division into distinct control and gaze guidance subject groups, and the degree to which the tasks influenced subject driving behaviour, needed to be investigated.

In order to address these points, we devised a second study that focused on evaluating the impact of simpler cues, that only indicated the horizontal direction of the critical event. Also, the cognitive tasks were eliminated. Without additional cognitive load though, the driving scenarios combined with the simulator driving conditions were not demanding enough for the subjects to justify the need for any driving aid. The only available method to increase the complexity of the driving task was to allow higher speeds of the egocar – as previously mentioned, in the first experiment, the maximum speed of the egocar did not exceed 30 km/h. However, it was not clear whether, in combination with the geometry of the simulated environment, controlling the car at higher speeds would not prove to be too challenging for subjects.

Therefore, before the actual experiment, we conducted a pilot study to ensure that changing the speed level would not substantially alter the driving behaviour of the subjects. The preliminary study also tested what was the maximum speed for which the gaze-contingent cues were still effective, and explored issues that would arise in timing the triggering of the events with more speed variation allowed.

In the following sections, we will describe this study, as well as the actual experiment, that was planned and conducted using the results of the preliminary investigation.

Preliminary stage: driving with varying speed levels

The pilot experiment largely followed the structure of the main study. We selected two courses through the simulated city. To give a realistic feel to the simulation, these courses were populated with cars and pedestrians involved in everyday-like traffic scenarios. In total for the two routes, 21 of all the traffic scenarios were safety-critical, and, as before, consisted of pedestrians unexpectedly crossing the street.

Each route was divided in three sections, each corresponding to a different driving gear. The maximal egocar velocities for each gear were approximately 25 km/h, 38 km/h, and 50 km/h. The gear was changed manually by the person conducting the experiment; the gear change points were pre-established, and they coincided for all subjects. The routes measured on average 1.3 km, and the subjects needed in total approximately 3.3 minutes to travel them. This trial duration included also the stops that needed to be made along the course.

In addition to the two test routes, each recording began with a training session to help subjects understand the mechanics of the simulator. The training route was much longer than the test routes, and it did not contain any other traffic participants. In order to get accustomed with controlling the car at higher speeds, the subjects drove sections of the training route in each of the three gears, and also practised full-braking.

Again, a control and a gaze-guidance data set were recorded. The same ray-shaped markers as before were used as gaze-contingent cues. However, in contrast to the first experiment, each of the 32 subjects contributed both to the control, and to the gaze guidance data set. More details on this aspect will be given in the following section. At the end of the trial, each subject filled out a questionnaire that, among other queries, asked for an evaluation of the perceived realism of the simulator, as well as of the difficulties encountered in controlling the car.

Results of the pilot study

As expected, higher speeds significantly increased the difficulty of the driving task. The majority of the subjects found controlling the car in the third gear to be considerably challenging. Although a 50 km/h speed would not pose any problems in controlling a real vehicle, the narrowness of the internal simulation field-of-view could explain why the speed was felt to be much higher than in reality. The subjects performed well when driving in the first two gears. Also, a precise synchronization for events met while driving in the third gear was extremely laborious.

Concerning the effects of the GCCs, fewer accidents were caused with pedestrians that were highlighted. This tendency was visible for each of the three gears. Surprisingly, more accidents happened while the subjects were driving in the second gear (12, compared to 3 in the first gear, and 4 in the third).

Methods

The main idea of the second experiment was similar to that of studies ran before: subjects were instructed to drive normally along predetermined routes inside the simulated environment, while taking care to respect traffic regulations and to avoid causing any accident. As before, along the routes, among normal traffic scenarios, potentially critical situations consisting in pedestrians coming close to, or crossing the street, were created.

Minor differences existed in the planning of the routes and of the traffic scenarios. In order to eliminate any bias that may be caused by cognitive tasks, or by learning effects due to the repetition of the routes, each subject drove along the three courses only once, and without any additional tasks. The courses were significantly longer (the average route length was 2850 m, more than three times the average length of the courses in the first experiment), and driving each route took on average 3.6 minutes. Also, to compensate for the absence of the additional tasks, all the courses were driven in the second gear, creating a more realistic, and a more demanding driving experience. To minimize any learning bias, the order in which the subjects drove the routes was randomly chosen.

We collected eye movements and driving data from 18 volunteering subjects. One trial consisted of driving along all three distinct routes, each of them containing 16 critical scenarios. Another difference to the first experiment was constituted by the way the control and the gaze guidance data sets were created. For each route, a number of 8 randomly picked critical events were highlighted using a gaze-contingent marker, while the remaining 8 were not cued, and served as control scenarios. The subjects were paired, so that every second drove a route version where the cueing of the events was mirrored with respect to the subject before. For example, if for the first driver, the first and the fourth events were cued, then for the second driver, they were control events. This strategy was used to eliminate potential differences in driving behaviour between subjects. The marker was a temporally transient red horizontal line overlaid at the bottom of the screen, in the half of the display corresponding to the direction from which the pedestrian was approaching (Figure 6.9).

After completing the experiment, each subject filled out a questionnaire, designed to explore whether the markers were immediately identified with the critical events and whether they were found to be distracting, as well as how realistic the simulation and the driving scenarios were perceived. The results of this were for the most part not relevant. However, they do hint to the fact that the subjects were not disturbed by the markers, but also to the fact that there is a lack of realism in the

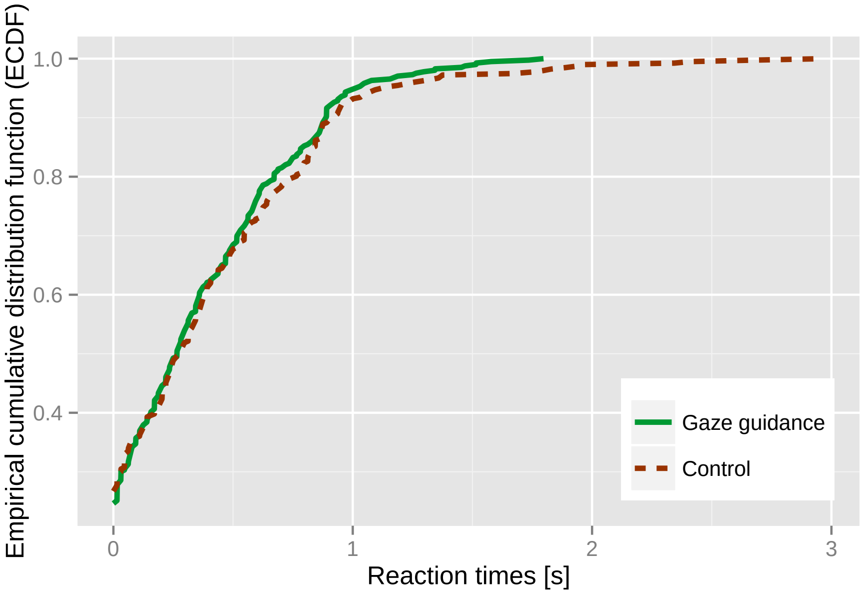


Figure 6.10: The analysis of the time needed by each driver to fixate the ”risk” pedestrian shows a tendency for shorter reaction times in the case of cued events (mean reaction time 0.583s for control, and 0.445 for GG events).

simulator.

Results

In total, more than three hours of driving data were recorded. Of this, a data set containing 408 gaze guidance, and 408 control events was built, and subjected to the same analyses as in the case of the first experiment (Section 6.2.2).

Reaction times

The ECDFs of the distributions of the time needed by the subjects to fixate the high-risk pedestrian are shown in Figure 6.10. Again, when analysing the reaction times, a tendency for shorter latencies can be observed in the case of GG events (mean/standard deviation: 0.583 s/1.21 s for control events, and 0.445 s/0.69 s). However, this time the significance threshold is not reached ($p = 0.5$, Kolmogorov-Smirnov test).

Provoked accidents

The collisions of the egocar with the high-risk pedestrians were computed using the same distance-based algorithm as before. The accident rates, expressed as the number of collisions in an event set, over the total number of events in that set, are illustrated in Figure 6.11. A significant drop in the accident rate for gaze-guidance events can be noticed: 0.0522 in the case of control events, vs. 0.0287

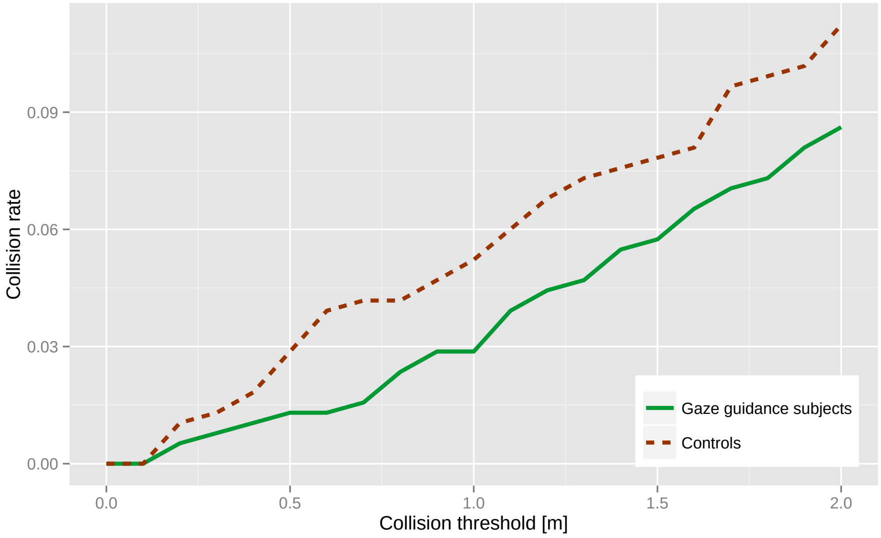


Figure 6.11: Collision rate as a function of the distance between the car and the pedestrian. With gaze guidance, the probability of an accident is reduced by approximately 45% (the accident rate for controls is 0.0522, as opposed to 0.0287 for gaze-guidance events). As before, the collision threshold was set to 1 m, but as the figure illustrates, the differences between accident rates remain qualitatively the same for different distance thresholds.

in the case of GG events (confidence interval: [0.0313, 0.0757]).

We also analysed the driving and the eye movement behaviour immediately after the triggering of a critical event. Although there are variations from event to event, several observations remain generally valid (one such example is illustrated in Figure 6.12). For the first part of the event, the egocar-pedestrian distance decreases in an approximately linear fashion, and the variance of both distance distributions is small. After approximately one second, as the egocar approaches the accident threshold, the distance egocar-pedestrian stabilizes near a constant value that tends to be larger in the case of cued events.

When examining the eye movement behaviour, a shift in the direction of the cues can be noticed, together with a tendency for a smaller variability for cued scenarios.

Discussion

Although the cues were not plainly connected to a traffic participant, drivers still had a tendency to fixate the critical pedestrian sooner after a direction cue was visible. However, when comparing the reaction times distributions between the group of cued and the group of control events, the differences fail to reach the significance level. As noted in the previous section, this could in part be due to the size of the dataset (approximately 600 samples for the first study, as opposed to approximately 400 for

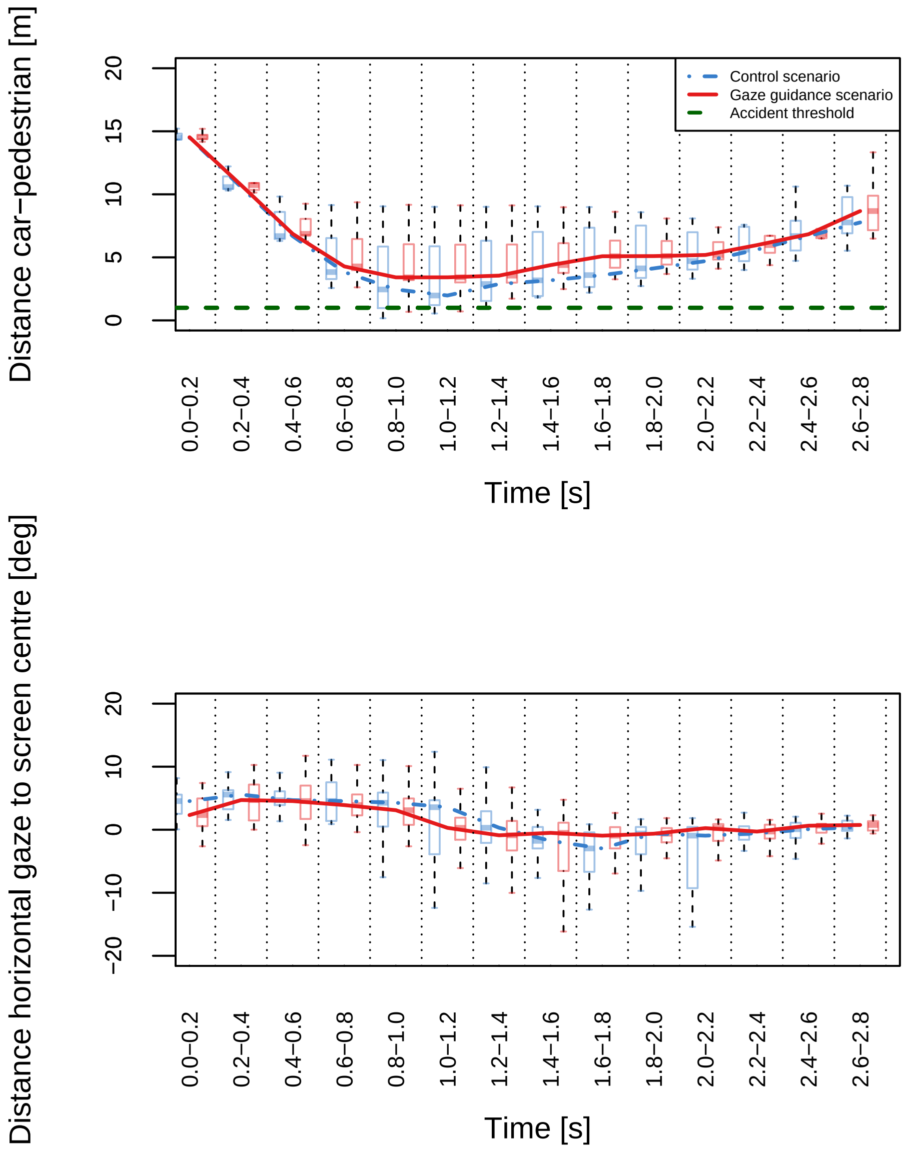


Figure 6.12: Example of eye movements and car-pedestrian distance during a critical event triggered from the left side of the road. The eye-movement data end the car-pedestrian distances were analysed during a 2.8 second time interval after the beginning of the event.

the current experiment). Also, as some of the subjects did not immediately recognize the connection between the risk pedestrian and the cues, this may also contribute to the small differences between the distributions of the reaction times. Nevertheless, when examining the number of accidents caused with cued and non-cued pedestrians, there is a significant reduction in the collision rate when the event was highlighted with a gaze contingent marker. Therefore it can be concluded that guiding the gaze of the driver in the critical direction was already sufficient to induce a safer driving behaviour.

6.2.4 Gaze guidance in a desktop simulator – Conclusions

In the series of experiments presented above, we used a desktop driving simulator with integrated eye tracking that offered the possibility to add gaze-contingent markers to its virtual environment. Using the simulation engine, we designed driving scenarios in which normal driving situations were intercalated with critical events consisting of pedestrians that unexpectedly crossed the street and thus forced the driver to take immediate action in order to avoid an accident. We have shown that gaze-contingent markers that were either overlaid on the high-risk traffic participants, or that merely indicated the direction from the dangerous event would emerge are successful in significantly reducing the number of accidents with those high-risk pedestrians.

Although the pedestrian-centred cues had a clear gaze-capturing effect and were easily correlated with the nature of the critical event, their implementation in a real car is challenging, even with the technology advances available today. However, simpler cues also proved effective in reducing the number of pedestrian accidents. It can also be argued that until more subtle gaze guidance techniques are developed, such cues would be less distracting in a complex scene, as well as easier to implement.

6.3 Gaze guidance in a high-fidelity driving simulator

The results of the studies presented above were highly encouraging; they showed that gaze guidance has indeed the potential of serving as an effective driving aid. However, because of limitations of the experimental setup, these results cannot be easily generalized to real world environments. Some of these limitations, such as the lack of complexity of the simulated environment, and a certain lack of realism of the driving conditions were already described in more detail in the previous section.

Therefore, the main goal of the current study is to further investigate the usefulness of gaze-contingent cues in the context of a realistic driving task performed in a wide field-of-view driving simulator. The results of the study will be made public in [4].

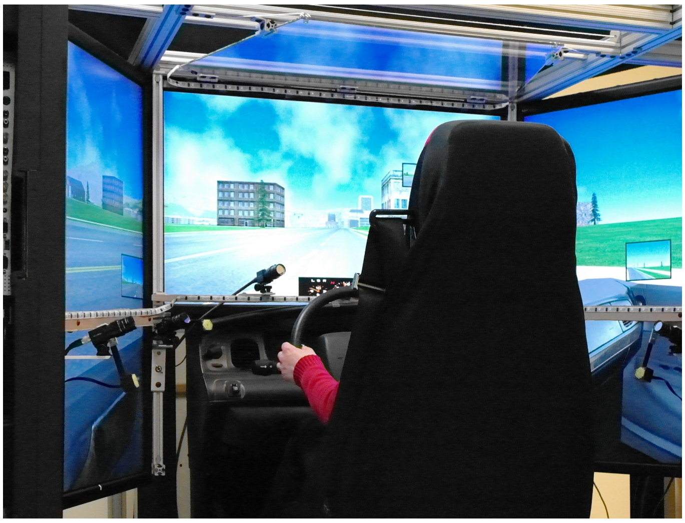


Figure 6.13: View of the state-of-the-art FAAC driving simulator. The system emulates the interior of a car, in which the windshield, covering a 225 degrees horizontal field-of-view, is created using five LCD panels. The simulator is integrated with a six-camera Smart Eye PRO eye tracking system.

6.3.1 Setup and data collection

The experiments took place in a state-of-the-art, wide field-of-view driving simulator. Unlike in the case of the previously used desktop simulator, no modifications could be made directly to the virtual environment, so an external solution for implementing gaze-contingent cues needed to be found.

In the current section we will first give an overview of the characteristics of the driving simulator itself, and then we will describe the technical details of the warning implementation. In the end, aspects connected to the planning of the experimental trials will be discussed.

Driving simulator

The study took place in a high-fidelity DE-1500 driving simulator (FAAC Inc, Ann Arbor, Mi). FAAC is a privately held US company specialized in distributed interactive simulations and also in motor vehicle simulations both for civil and military use[13].

The DE-1500 simulator is composed of an “open air” driving station, illustrated in Figure 6.13. The windshield was simulated by a multidisplay system composed of five wide-screen LCD displays, and it covered a visual angle of approximately 225 by 38 degrees. Each LCD display had a resolution of 1360 × 768 pixels. The controls of the simulator reproduced those of an automatic transmission vehicle. A real car – Ford Crown Victoria – was used to model the dashboard and the driving seat.

In addition to the visual component, the simulator also delivered auditory and haptic feedback.

[13] http://www.faac.com

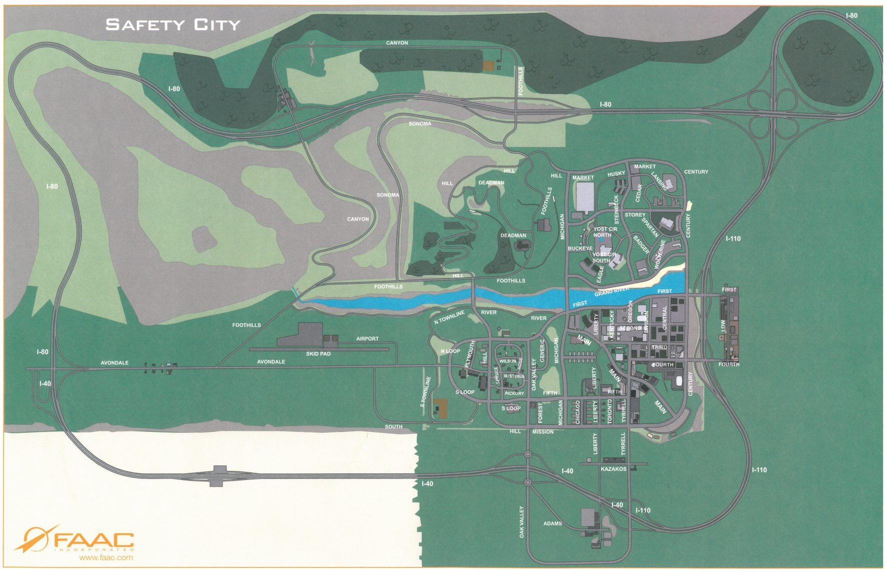


Figure 6.14: The map of the virtual environment. The map combines urban, suburban, and industrial areas with hilly and desert terrain. A substantial highway stretch is also available at the exterior of the simulated "city".

A set of speakers reproduced sounds such as engine and road noise, tire skidding, or braking noises. The force-feedback steering wheel offered variable resistance in function of the characteristics of the simulated road. Tactile feedback was also provided on curb strikes. The driving seat attempted to recreate the acceleration/deceleration feedback available in a real motor vehicle. Also, on collisions with objects or persons in the virtual environment both auditory and haptic feedback was given. An adjustable air flow originating from the dashboard lent a further degree of authenticity to the driving experience.

The driving environment could be chosen from several available maps. We used the "Safety City" (Figure 6.14), a general setting that included a variety of terrains and settlement types. Just to give an example, the virtual world contained urban, suburban, industrial and rural areas. The "inhabited" regions were encircled by a highway network. The terrain was at times flat, at times hilly, with a set of mountain roads available in the north-east of the map. A section of the highway crossed a desert area.

The basic layer of the virtual world contained roads, landscape elements, buildings, together with some predefined cars and pedestrians, all static. To this fundamental layer, one could add actors from several available categories such as vehicles, or walkers using a graphical interface – the "Scenario Development Toolbox". Each category contained a broad range of available actors. Besides adding active traffic participants, within the toolbox it was also possible to add supplementary static content such as extra buildings, trees, animals, vehicles, or pedestrians. These would remain unchanged throughout the trials, and serve as scenery, increasing the degree of complexity and realism of the environment.

The simulator provided two approaches to adding traffic to the virtual environment. The first, very similar to that available in the desktop simulator we first used, meant manually adding and setting the behaviour of every traffic actor. Again, the actions of each participants needed to be fully scripted. The trajectories and also the triggering events could be defined within the Scenario Development Toolbox. The second approach was to use the preexistent autonomous traffic mode. When the autonomous traffic option was activated, the simulator engine tended to populating the area around the egocar with other moving vehicles. In autonomous traffic mode, all that needed to be set were the density and the aggressiveness level of the autonomous vehicles. These could be varied on a scale from 0 to 9. With some precautions, autonomous and scripted vehicles could be combined.

It was possible to externally manipulate the weather conditions in the virtual world. In addition to the fundamental "clear day" setting, the other four available weather conditions were night, fog, rain, snow, and dust. They could be added to the simulation using an intensity scale that varied from 0 to 8. Any number of these conditions could be combined, in order to obtain the desired result.

The eye movements of the subjects were recorded using a six-camera SmartEye Pro remote eye tracking system, that covered the entire windshield. The eye tracker delivered the gaze position on the multidisplay area at a sampling frequency of 60 Hz, and with a manufacturer-specified tracking accuracy of 0.5 degrees.

The virtual world projected on each of the displays was created by five image-generator computers, that were in their turn controlled by a master computer that acted as the "brain" of the system, and ran the simulation. The master unit also received and logged gaze data from the eye tracking PC, which it integrated to its own stream of driving data. The driving data was recorded at a frequency of approximately 30 Hz, and it contained statistics such as coordinates in the virtual world, speed, or acceleration.

Cue implementation

As previously mentioned, the virtual environment could not be externally modified, and also no mechanism for adding textures or cues inside the simulation was available. However this suited our intentions to explore realistic cues that could be implemented with minimal effort into an actual vehicle. The Volkswagen prototype mentioned in the introductory section of the current chapter implements warnings using a row of lights underneath the windshield. The LEDs can be selectively turned on. According to press reports, they appear to already lead to a guiding effect, and after some

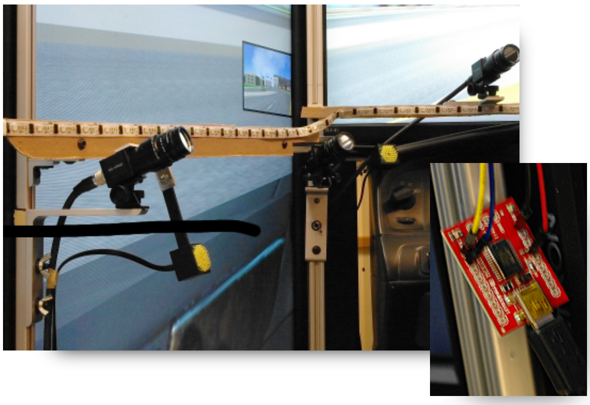


Figure 6.15: LED strip mounted on aluminium profiles across the bottom edge of the simulated windshield. Thumb image: the breakout board used to connect the LED array to the controller unit.

time, as the driver learns to consciously ignore them, they become mostly unobtrusive.

We decided to take a similar approach, and use light-emitting diode (LED) arrays to implement horizontal directional cues similar to the ones used in the second desktop simulator study. Unfortunately, the simulation engine resembled a black box as it only allowed access to the data contained in the output driving stream. This contained no information about the scene visible in the virtual environment, and therefore, it was impossible to connect the position of the egocar with what was being projected on the simulator display. Because of that, it was difficult to correlate the warnings with specific events involving other traffic participants.

For this reason, unlike in the first series of experiments, the cues were not correlated to any traffic event in the simulation. Instead, we chose to cue specific locations in the environment, namely intersections. As earlier discussed, intersections are recognized to be sites with higher risk exposure for non-motorised traffic participants such as cyclists or pedestrians, but also for the motorists themselves. Many accidents taking place at crossroads are caused by the driver's failure to properly scan all directions from which oncoming traffic might arrive, and thus fail to identify potential threats. We planned the warnings in such a manner so that in each intersection they would cue the driver towards a predetermined direction, left or right, but only if the driver is not already looking in cued direction.

In order to implement the gaze-contingent warnings, the simulator was outfitted with two arrays of light-emitting diodes, one at the bottom of the windshield, covering the central 150 degrees of the visual field, and one at its top, covering only the 67 degrees corresponding to the centre screen. The arrays were mounted on a set of 80/20 T-slotted aluminium profiles fitted transversally across the two vertically-set side displays (see Figure 6.15).

To create the LED arrays, we used 3 addressable RGB LED strips of one meter length, each with

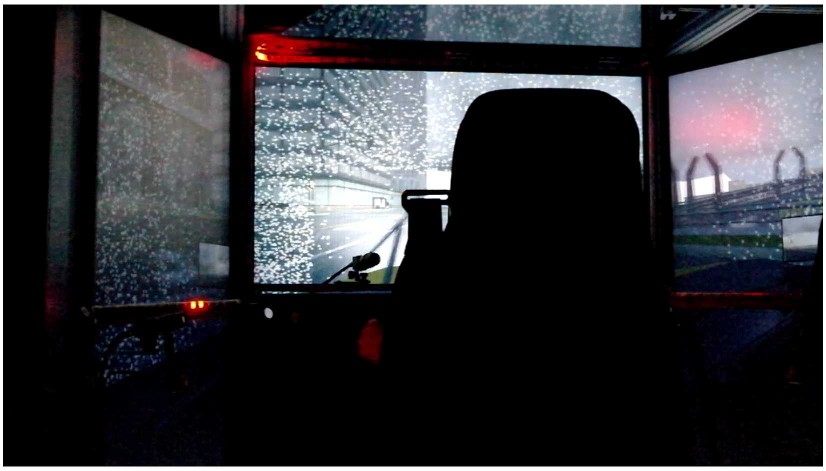


Figure 6.16: A visible cue to the left during a data recording session. The subject was driving in an urban environment, in combined night and rain conditions. We did not vary the colour of the LEDs during the experiments, the only used colour was red, as it offered a good compromise between visibility of the cue, and little impact on the vision of the driver.

32 LEDs. For the top array we used a single strip, while for the bottom one we connected two of them together. The strips could be connected to a computer via a serial interface, thus allowing each of the LEDs to be individually controlled. The connection was achieved with the help of a USB to serial printed circuit board (PCB).

The LEDs were controlled by a separate PC unit to which they were directly connected. As they had to be triggered as a function of position in the simulated world, but also as a function of horizontal gaze position on the multidisplay system, the PC unit had to receive data streams both from the simulator master computer and from the eye tracking computer. The master script that triggered the LEDs also synchronized the two data streams it received via UDP. (Figure 6.16).

In the following, we will very briefly describe the algorithm behind the cue triggering. Each packet from the driving simulator data stream received by the master script contained all the driving statistics, including the position in the virtual city, and the speed of the egocar. In parallel, the master had access to (and ran through) a list with all the cued locations. If the distance between the egocar and the current critical location was smaller than a predefined constant, the controller could take the decision to trigger a warning. In order to compensate for large speed variations between different drivers, the distance measure was combined with an additional restriction regarding the estimated time before reaching the critical location.

Once it was established that the egocar had reached a cued location, the first eye movement sample was checked, in order to determine which LEDs would be activated during the warning. On the top array, it was always the last two LEDs in the predefined cueing direction that were briefly flashed. However, on the bottom array, the cue was dynamic, and it followed a “chasing lights” pattern: each of the LEDs between the start position of the cue and the end of the array would light on and off in a rapid succession. The start LED was computed as the LED that corresponded to the horizontal component of the last read gaze sample, to which a predefined offset in the direction of the cue, equal to half the resolution of the centre screen, was added. If the subject’s gaze was already in the cued direction, than no LED was lit.

Data recording

We selected four courses inside the Safety City world, each set in diverse environments: two were situated mostly in urban and suburban areas, while the other two focused mostly on highway sections. The route length ranged between 3 and 8 km. In addition to varying the setting, we also varied the weather conditions. Two courses were set in daylight conditions, one in a clear day, the other in a day with level 6 fog, while the remaining two courses were both set in identical, combined, night and rain conditions.

Following the method presented in Section 6.2.3, for each subject, only approximately half of the intersections were highlighted by a cue, while the remaining ones serving as control locations. Each route was paired with a mirrored version, in which the control and the GCC sets were swapped. Each complete trial consisting of the four pre-programmed routes contained approximately 40 control and 40 GCC locations. We set the cueing directions randomly, and each cued location was paired with a fixed direction.

The order in which the routes were driven was fixed: daylight – highway and urban environment, rainy night – suburban and urban environment, foggy day – urban environment and highway, and again rainy night – urban and suburban environments. The driving order, together with the route environment were deliberately set. This was done taking into consideration that it had not been uncommon for subjects from previous studies that took place in the same settings to experience simulator sickness.

Simulator sickness is a relatively common occurrence, and has been accordingly documented since the appearance of the first virtual environments (see [164, 165, 166] for an overview). Its symptoms are for the most part similar to those of motion sickness, and are commonly evaluated using a simulator sickness questionnaire [167, 168]. There are several hypothesis on the causes of simulator sickness, but no certain explanation is known. One of the earliest and most widespread theories is that of the cue conflict: it is suggested that the sickness is caused by a discrepancy between sensory cues, for example visual and motion cues. Another plausible explanation is that of postural instability [169].

Although the exact cause for the simulator sickness often encountered in the FAAC simulator could not be determined, we speculate that the pairing between some minor geometry faults of the simulated world and some inaccurate motion feedback through the driving seat are the principal triggers. In any case, a short accommodation period in which the subject drove on a straight and flat route, combined with a gradual introduction to steering manoeuvres and to altitude differences have been observed to reduce the phenomenon. Also, having a somewhat reduced visibility for sections with many turns and ramps appeared to help.

Despite these precautions, of the 13 volunteering subjects we recorded, one had to end the experiment without completing a single course, while two other had to stop after the first course.

Regarding the additional traffic on the routes, we chose to use the autonomous traffic mode at high density, and moderate aggressivity levels. Although using scripted vehicles would have ensured to a higher degree that the trials are comparable between different subjects, a number of problems arose. First, although the speed of the scripted vehicles could to a limited degree be adapted to that of the egocar, large variations of the egocar velocity and of the driving behaviour would have made the similar timing of events from subject to subject impossible. Second, no interaction between scripted vehicles or scripted vehicles and egocar was possible, so very simple driving scenarios needed extensive fine tuning that could still fail at an unpredicted manoeuvre of the subject. Such aspects were not an issue with autonomous traffic, and as the AI of the simulator created the autonomous traffic following a similar algorithm every time, a certain degree of similarity was still present between different trials.

6.3.2 Results

The aspect of most interest when analysing the recorded data was whether the cues succeeded in guiding the gaze of the drivers in the desired direction. From the more than 5 hours of recorded driving data, we extracted approximately 45,000 saccades.

In a first step, we analysed the saccade behaviour over the first two seconds after reaching each critical location. In total, over all the two-second post-cue time intervals, more than 3500 saccades were available. For each saccade landing point, we computed the distance between its horizontal coordinate and the centre of the multidisplay system. The differences between the gaze distributions built as such for control and cued locations are highly significant, and suggest a shift in gaze position in the direction of the warning for cued locations.

For visualisation purposes, we divided the 4 s interval of interest in 200 ms time bins. For each event we computed and plotted the mean gaze position corresponding to each time bin (Figure 6.17), as the average of the horizontal component of all the saccade landing points from the respective time interval. A significant shift of the gaze horizontal position in the direction of the cue can be distinguished starting after approximately 400 ms from the cue triggering (Mann-Whitney U test:

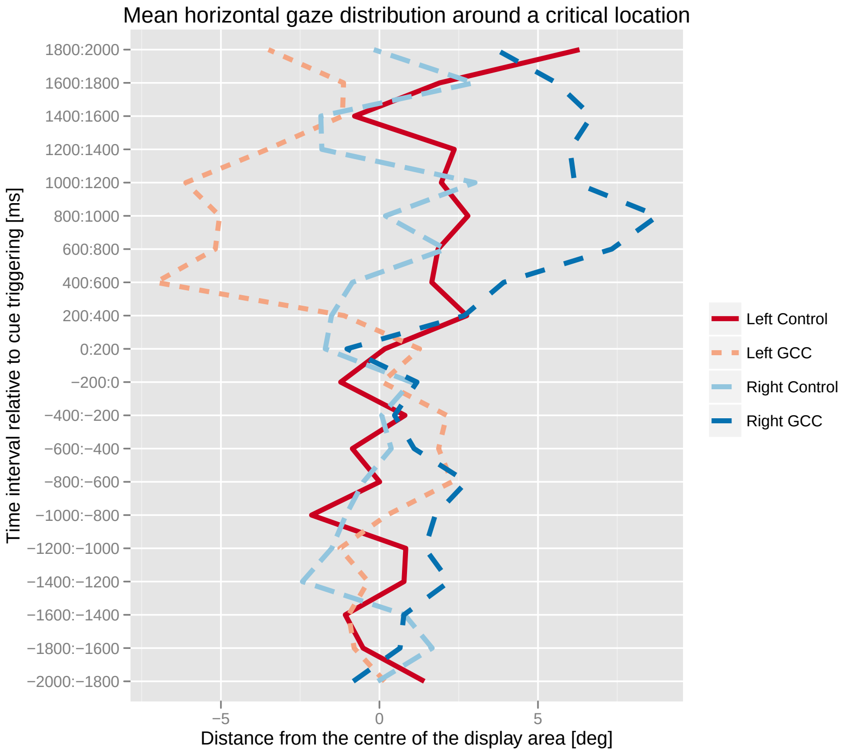


Figure 6.17: Mean horizontal saccade landing point distribution two seconds before, and two seconds after event triggering. It can be noticed that there is a significant shift of the gaze in the direction of the cue.

$p - value < 10^{-7}$ for cues to the right, and $p - value < 10^{-13}$ for cues to the left). This shift is not present before the event triggering or in the case of control locations.

In order to ensure that the cues did not directly attract the gaze, but guided the driver to look in their general direction, we repeated the analysis above, but this time with the vertical coordinates of the landing points belonging to the saccades immediately before and after a cue. No significant shifts have been found towards the bottom arrays ($p - value > 0.4$, Mann-Whitney U test), suggesting the subjects did not tend to look at the moving cue itself. However, a weak shift towards the top of the displays was visible ($p - value$ 0.095, Mann-Whitney U test).

6.3.3 Discussion

Gaze-contingent cues that horizontally highlight certain regions of the driving space are successful in influencing eye movements in a state of the art driving simulator, without disrupting the driving activity.

6.4 Chapter conclusions

In the current chapter, we have shown that it is possible to create an augmented vision system that uses gaze guidance in order to help drivers better allocate their attention resources.

First, using a simple desktop driving simulator, we showed that gaze-contingent cues that highlight a high-risk pedestrian, either directly or just by indicating his direction of walking, are effective in reducing the number of crashes with that pedestrian. We extend these results to a more realistic driving environment, and we show that simple directional gaze contingent cues can influence eye movements without disrupting the driving activity in a state-of-the-art driving simulator.

Building such a system would involve already available components such as pedestrian detection, pre-crash sensing and attention monitoring, and maybe also novel developments such as a wide-angle gaze-contingent head-up display. Naturally, in order for such a system to be useful it should be 100% reliable, give no false warnings and signal all potential dangers, but that holds true not only for gaze guidance, but for any driver assistance system.

The actual benefits of unobtrusive gaze guidance can be fully comprehended when accepting that visual perception is, as we have tried to show over the previous chapters, widely determined by our expectations, i.e. by a model. This model is only selectively updated because of limited attentional resources. Warnings such as beeping attempt to avoid critical situations by forcing the driver to change their current model, whereas subliminal guidance would change the way the model is updated. The latter process is faster and less distracting, and therefore more efficient.

7 Conclusions

At the beginning of this dissertation we set out to explore aspects related to the efficacy and to the practical implementation of an augmented vision system based on gaze guidance techniques and adapted to driving.

In a series of experiments conducted in a desktop driving simulator, we showed that fewer accidents are registered with pedestrians involved in safety-critical scenarios when these pedestrians are highlighted using gaze-contingent cues. These results were first obtained using fairly complex, pedestrian-centred markers, and were later confirmed in an experiment that used simpler cues, that only highlighted the direction of the high-risk pedestrian. In the final part of our work, we investigated the effects of similar directional gaze-contingent cues in a more realistic driving environment. The cues, implemented using LED arrays adjoined to a wide-field-of-view driving simulator, led to a clear gaze guidance effect when triggered randomly at intersections in the virtual world.

Moreover, following an experiment that compared the eye movement strategies adopted by novice and expert players engaged in playing a gaze-operated game, we could confirm that both task and experience have a strong influence on eye movements. Next, as being able to guide eye movements implies being able to predict what actually attracts them in a scene, we looked at bottom-up factors that influence gaze allocation. We used low-level features of the visual input to accurately predict eye movements on complex stimuli consisting of time-varying superimposed natural scenes. We showed that eye movements prefer informative areas of the visual input, namely areas with a higher intrinsic dimension.

To summarize our main results, we have shown that gaze-contingent cues do have a gaze-capturing effect, and can influence where drivers look without being disruptive towards the driving activity. Moreover, the gaze guidance effect is present even if the cues are not directly associated with a traffic participant, but instead highlight a general direction of danger. This significantly simplifies the implementation of such cues into a real vehicle. Our most important contribution though, is showing that in actual safety-critical situations occurring in a driving simulator, drivers aided by gaze-contingent cues commit significantly fewer accidents.

Nevertheless, there are numerous questions that we did not address, leaving them open to future research. To mention just a few, the optimal timing of the cues needs to be investigated, as well as what exactly needs to be cued in a real-world situation.

Of course, one may argue that with the current advances in technology, the future belongs to autonomous driving, and human-driven cars will soon be completely eliminated. Although this is indeed a significant direction in automotive research, building a self-driving car poses a number of critical problems, and it is unlikely that all will be addressed in the near future. As this did not directly concern our research, we only mention here one obvious, but challenging requirement. In addition to accurately detecting without any false positives all the potentially critical events near the car, the automated system must also have the ability to identify what the risk elements are, and to decide on an adequate set of actions to respond to them.

Seen from this perspective, a driving assistance system that integrates gaze guidance with a highly-accurate detection component, provides the optimal compromise between two imperfect components: an artificial brain that does not get distracted, and can process concurrently all the available stimuli, and a human component that can effortlessly assess whether an event poses an immediate risk.

Bibliography

[1] L. Pomarjanschi, M. Dorr, and E. Barth, "Gaze guidance reduces the number of collisions with pedestrians in a driving simulator," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 1, no. 2, p. 8, 2012.

[2] E. Barth, M. Dorr, E. Vig, L. Pomarjanschi, and C. Mota, "Efficient coding and multiple motions," *Vision research*, vol. 50, no. 22, pp. 2190–2199, 2010.

[3] L. Pomarjanschi, M. Dorr, C. Rasche, and E. Barth, "Safer driving with gaze guidance," in *BIONETICS 2010*, ser. LNICST, J. Suzuki and T. Nakano, Eds., vol. 87. Springer, 2012, pp. 581–586.

[4] L. Pomarjanschi, M. Dorr, P. J. Bex, and E. Barth, "Simple gaze-contingent cues guide eye movements in a realistic driving simulator," in *Proceedings of SPIE, Human Vision and Electronic Imaging XVIII*, In Press.

[5] L. Pomarjanschi, M. Dorr, and E. Barth, "Eye movements on blended natural videos," vol. 38, 2009, p. 45.

[6] L. Pomarjanschi, C. Rasche, M. Dorr, E. Vig, and E. Barth, "Safer driving with gaze guidance," vol. 39, 2010, p. 83.

[7] L. Pomarjanschi, M. Dorr, and E. Barth, "Gaze guidance effective in reducing collisions in a driving simulator," vol. 40, 2011, p. 177.

[8] M. Dorr, L. Pomarjanschi, and E. Barth, "Gaze beats mouse: A case study on a gaze-controlled breakout," *PsychNology Journal*, vol. 7, no. 2, pp. 197–211, 2009.

[9] M. F. Bear, B. W. Connors, and M. A. Paradiso, *Neuroscience: Exploring the brain.* Lippincott Williams & Wilkins, 2007.

[10] S. E. Palmer, *Vision science: Photons to phenomenology.* MIT Press, 1999, vol. 1.

[11] J. M. Findlay and I. D. Gilchrist, *Active vision. The psychology of looking and seeing*. Oxford University Press, 2003.

[12] M. E. Burns and T. D. Lamb, "Visual transduction by rod and cone photoreceptors," in *The Visual Neurosciences*, L. M. Chalupa and J. S. Werner, Eds. MIT Press, 2004, pp. 215–233.

[13] R. Shapley, E. Kaplan, and R. Soodak, "Spatial summation and contrast sensitivity of x and y cells in the lateral geniculate nucleus of the macaque," *Nature*, vol. 292, no. 5823, pp. 543–545, 1981.

[14] E. Kaplan and R. M. Shapley, “X and Y cells in the lateral geniculate nucleus of macaque monkeys.” *The Journal of Physiology*, vol. 330, no. 1, pp. 125–143, 1982.

[15] M. J. Hawken and A. J. Parker, “Contrast sensitivity and orientation selectivity in lamina IV of the striate cortex of old world monkeys,” *Experimental Brain Research*, vol. 54, no. 2, pp. 367–372, 1984.

[16] E. Kaplan and R. M. Shapley, “The primate retina contains two types of ganglion cells, with high and low contrast sensitivity,” *Proceedings of the National Academy of Sciences*, vol. 83, no. 8, pp. 2755–2757, 1986.

[17] D. C. V. Essen, W. T. Newsome, and J. H.R.Maunsell, “The visual field representation in striate cortex of the macaque monkey: asymmetries, anisotropies, and individual variability,” *Vision research*, vol. 24, no. 5, pp. 429–448, 1984.

[18] P. Daniel and D. Whitteridge, “The representation of the visual field on the cerebral cortex in monkeys,” *The Journal of Physiology*, vol. 159, no. 2, pp. 203–221, 1961.

[19] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[20] ——, “Receptive fields and functional architecture of monkey striate cortex,” *The Journal of physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[21] E. Barth and A. B. Watson, “A geometric framework for nonlinear visual coding,” *Optics Express*, vol. 7, no. 4, pp. 155–165, 2000.

[22] M. Mishkin, L. G. Ungerleider, and K. A. Macko, “Object vision and spatial vision: Two cortical pathways,” *Trends in neurosciences*, vol. 6, pp. 414–417, 1983.

[23] M. A. Goodale and D. A. Westwood, “An evolving view of duplex vision: separate but interacting cortical pathways for perception and action,” *Current opinion in neurobiology*, vol. 14, no. 2, pp. 203–211, 2004.

[24] M. A. Goodale and A. D. Milner, “Separate visual pathways for perception and action,” *Trends in neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[25] T. D. Albright, “Direction and orientation selectivity of neurons in visual area MT of the macaque,” *Journal of Neurophysiology*, vol. 52, no. 6, pp. 1106–1130, 1984.

[26] A. Mikami, W. T. Newsome, and R. H. Wurtz, “Motion selectivity in macaque visual cortex. mechanisms of direction and speed selectivity in extrastriate area MT,” *Journal of Neurophysiology*, vol. 55, no. 6, pp. 1308–1327, 1986.

[27] H. R. Rodman and T. D. Albright, “Coding of visual stimulus velocity in area MT of the macaque,” *Vision research*, vol. 27, no. 12, pp. 2035–2048, 1987.

[28] M. F. Land, “Oculomotor behaviour in vertebrates and invertebrates,” in *The Oxford Handbook of Eye Movements*, S. P. Liversedge, I. D. Gilchrist, and S. Everling, Eds. Oxford University Press, 2011, pp. 215–233.

[29] I. D. Gilchrist, “Saccades,” in *The Oxford Handbook of Eye Movements*, S. P. Liversedge, I. D. Gilchrist, and S. Everling, Eds. Oxford University Press, 2011, pp. 215–233.

[30] W. Becker, “Saccades,” *Eye movements*, vol. 8, 1991.

[31] E. Holt, “Eye-movement and central anaesthesia,” *Psychological Monographs*, 1903.

[32] B. Bridgeman, D. Hendry, and L. Stark, “Failure to detect displacement of the visual world during saccadic eye movements,” *Vision research*, vol. 15, no. 6, pp. 719–722, 1975.

[33] D. C. Burr, M. C. Morrone, and J. Ross, “Selective suppression of the magnocellular visual pathway during saccadic eye movements,” *Nature*, vol. 371, no. 6497, pp. 511–513, 1994.

[34] E. Matin, “Saccadic suppression: a review and an analysis.” *Psychological bulletin*, vol. 81, no. 12, p. 899, 1974.

[35] E. Castet and G. S. Masson, “Motion perception during saccadic eye movements,” *Nature Neuroscience*, vol. 3, pp. 177–183, 2000.

[36] M. A. Garcıa-Pérez and E. Peli, “Intrasaccadic perception,” *The Journal of Neuroscience*, vol. 21, no. 18, pp. 7313–7322, 2001.

[37] M. Dorr and P. J. Bex, “Peri-saccadic visual sensitivity while freely-viewing natural movies,” *Journal of Vision*, vol. 11, no. 11, pp. 517–517, 2011.

[38] S. Martinez-Conde, S. L. Macknik, and D. H. Hubel, “Microsaccadic eye movements and firing of single cells in the striate cortex of macaque monkeys,” *Nature Neuroscience*, vol. 3, no. 3, pp. 251–258, 2000.

[39] G. R. Barnes, “Ocular pursuit,” *Oxford Handbook of Eye Movements*, p. 115, 2011.

[40] J. K. Tsotsos, L. Itti, and G. Rees, “A brief and selective history of attention,” in *Neurobiology of attention*. Elsevier, San Diego, CA, 2005.

[41] H. von Helmholtz, “Concerning the perceptions in general,” *Treatise on physiological optics*, vol. 3, pp. 1–37, 1866.

[42] H. Deubel and W. X. Schneider, “Saccade target selection and object recognition: Evidence for a common attentional mechanism,” *Vision research*, vol. 36, no. 12, pp. 1827–1837, 1996.

[43] A. A. Kustov and D. L. Robinson, “Shared neural control of attentional shifts and eye movements,” *Nature*, 1996.

[44] M. Corbetta, E. Akbudak, T. E. Conturo, A. Z. Snyder, J. M. Ollinger, H. A. Drury, M. R. Linenweber, S. E. Petersen, M. E. Raichle, and D. C. V. Essen, “A common network of functional areas for attention and eye movements,” *Neuron*, vol. 21, no. 4, pp. 761–773, 1998.

[45] J. M. Henderson, A. Pollatsek, and K. Keith Rayner, “Covert visual attention and extrafoveal information use during object identification,” *Attention, Perception, & Psychophysics*, vol. 45, no. 3, pp. 196–208, 1989.

[46] G. Rizzolatti, L. Riggio, I. Dascola, and C. Umiltá, “Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention,” *Neuropsychologia*, vol. 25, no. 1, pp. 31–40, 1987.

[47] C. E. Connor, H. E. Egeth, and S. Yantis, “Visual attention: bottom-up versus top-down,” *Current Biology*, vol. 14, no. 19, pp. R850–R852, 2004.

[48] J. Mertens, “Influence of knowledge of target location upon the probability of observation of peripherally observable test flashes,” *JOSA*, vol. 46, no. 12, pp. 1069–1070, 1956.

[49] M. I. Posner, “Orienting of attention,” *Quarterly journal of experimental psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[50] A. Yarbus, *Eye movements and vision*. Plenum press, 1967.

[51] H. E. Egeth and S. Yantis, “Visual attention: Control, representation, and time course,” *Annual review of psychology*, vol. 48, no. 1, pp. 269–297, 1997.

[52] J. Theeuwes, “Perceptual selectivity for color and form,” *Attention, Perception, & Psychophysics*, vol. 51, no. 6, pp. 599–606, 1992.

[53] S. Yantis and J. Jonides, “Abrupt visual onsets and selective attention: evidence from visual search.” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, no. 5, p. 601, 1984.

[54] J. Theeuwes, A. F. Kramer, S. Hahn, and D. E. Irwin, “Our eyes do not always go where we want them to go: Capture of the eyes by new objects,” *Psychological Science*, vol. 9, no. 5, pp. 379–385, 1998.

[55] A. Hollingworth, D. J. Simons, and S. L. Franconeri, “New objects do not capture attention without a sensory transient,” *Attention, Perception, & Psychophysics*, vol. 72, no. 5, pp. 1298–1310, 2010.

[56] N. J. Wade and B. W. Tatler, *The moving tablet of the eye: The origins of modern eye movement research*. Oxford University Press, USA, 2005.

[57] L. R. Young and D. Sheena, “Survey of eye movement recording methods,” *Behavior Research Methods*, vol. 7, no. 5, pp. 397–429, 1975.

[58] J. Orschansky, “Eine Methode die Augenbewegungen direkt zu untersuchen,” *Centralblatt für Physiologie*, vol. 12, pp. 785–790, 1899.

[59] D. A. Robinson, “A method of measuring eye movement using a scleral search coil in a magnetic field,” *Bio-medical Electronics, IEEE Transactions on*, vol. 10, no. 4, pp. 137–145, 1963.

[60] E. L. Irving, J. E. Zacher, R. S. Allison, and M. G. Callender, “Effects of scleral search coil wear on visual function,” *Investigative ophthalmology & visual science*, vol. 44, no. 5, pp. 1933–1938, 2003.

[61] R. Dodge and T. S. Cline, “The angle velocity of eye movements.” *Psychological Review*, vol. 8, no. 2, p. 145, 1901.

[62] G. T. Buswell, “How people look at pictures,” 1935.

[63] A. T. Duchowski, *Eye tracking methodology: Theory and practice*. Springer, 2007.

[64] *iViewX System Manual*, 2011.

[65] *SmartEye Pro 5.9. User Manual.*

[66] J. E. Gentle, *Elements of Computational Statistics*, 2002.

[67] R. C. Gonzales and R. E. Woods, *Digital image processing*. Prentice-Hall, 2002.

[68] B. Jähne, *Digital image processing*. IOP Publishing, 2002, vol. 13, no. 9.

[69] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *Communications, IEEE Transactions on*, vol. 31, no. 4, pp. 532–540, 1983.

[70] L. Williams, "Pyramidal parametrics," in *ACM Siggraph Computer Graphics*, vol. 17, no. 3. ACM, 1983, pp. 1–11.

[71] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[72] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[73] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," *ACM Transactions on Graphics (TOG)*, vol. 2, no. 4, pp. 217–236, 1983.

[74] J. S. Perry and W. S. Geisler, "Gaze-contingent real-time simulation of arbitrary visual fields," in *Human Vision and Electronic Imaging: Proceedings of SPIE, San Jose, CA*, vol. 4662. Citeseer, 2002, pp. 57–69.

[75] M. Böhme, M. Dorr, T. Martinetz, and E. Barth, "A temporal multiresolution pyramid for gaze-contingent: Manipulation of natural video," in *Passive Eye Monitoring*, R. I. Hammoud, Ed. Springer, 2008, pp. 225–243.

[76] C. Zetzsche and E. Barth, "Fundamental limits of linear filters in the visual processing of two-dimensional signals," *Vision Research*, vol. 30, no. 7, pp. 1111–1117, 1990.

[77] C. Mota and E. Barth, "On the uniqueness of curvature features," *Proceedings in Artificial Intelligence (Dynamische Perzeption). Köln: Infix Verlag*, vol. 9, pp. 175–178, 2000.

[78] C. Zetzsche, E. Barth, and B. Wegmann, "The importance of intrinsically two-dimensional image features in biological vision and picture coding," in *Digital images and human vision*. MIT Press, 1993, pp. 109–138.

[79] C. Mota, I. Stuke, and E. Barth, "The intrinsic dimension of multispectral images," in *MICCAI workshop on biophotonics imaging for diagnostics and treatment*, 2006, pp. 93–100.

[80] J. Bigün and G. H. Granlund, "Optimal orientation detection of linear symmetry," in *First International Conference on Computer Vision, ICCV (London)*, 1987, pp. 433–438.

[81] H. Knutsson, “Representing local structure using tensors,” in *Proceedings of 6th Scandinavian Conference on Image Analysis, Oulu: Oulu University*, 1989, pp. 244–251.

[82] H. Haußecker and B. Jähne, “A tensor approach for local structure analysis in multidimensional images,” *Proceedings 3D Image Analysis and Synthesis' 96, Universität Erlangen-Nürnberg*, 1996.

[83] C. Mota, I. Stuke, and E. Barth, “Analytic solutions for multiple motions,” in *Proceedings of the International Conference on Image Processing*, vol. 2. IEEE, 2001, pp. 917–920.

[84] V. N. Vapnik, *The nature of statistical learning theory*. Springer, 2000.

[85] E. Alpaydin, *Introduction to machine learning*. The MIT Press, 2004.

[86] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[87] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[88] E. Vig, M. Dorr, and E. Barth, “Efficient visual coding and the predictability of eye movements on natural movies,” *Spatial vision*, vol. 22, no. 5, pp. 397–408, 2009.

[89] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[90] É. Javal, *Physiologie de la lecture et de l'écriture*. F. Alcan, 1905, vol. 105.

[91] M. F. Land and B. W. Tatler, *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, USA, 2009.

[92] S. M. Reder, “On-line monitoring of eye-position signals in contingent and noncontingent paradigms,” *Behavior Research Methods*, vol. 5, no. 2, pp. 218–228, 1973.

[93] G. W. McConkie and K. Rayner, “The span of the effective stimulus during a fixation in reading,” *Attention, Perception, & Psychophysics*, vol. 17, no. 6, pp. 578–586, 1975.

[94] K. Rayner, “Eye movements in reading and information processing: 20 years of research.” *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.

[95] M. F. Land, N. Mennie, and J. Rusted, “The roles of vision and eye movements in the control of activities of daily living,” *PERCEPTION-LONDON-*, vol. 28, no. 11, pp. 1311–1328, 1999.

[96] M. Hayhoe, “Vision using routines: A functional account of vision,” *Visual Cognition*, vol. 7, no. 1-3, pp. 43–64, 2000.

[97] M. F. Land and M. Hayhoe, “In what ways do eye movements contribute to everyday activities?” *Vision research*, vol. 41, no. 25-26, pp. 3559–3565, 2001.

[98] M. F. Land and D. Lee, “Where we look when we steer,” *Nature*, vol. 369, no. 6483, pp. 742–744, 1994.

[99] M. F. Land and P. McLeod, “From eye movements to actions: how batsmen hit the ball,” *Nature neuroscience*, vol. 3, no. 12, pp. 1340–1345, 2000.

[100] G. J. Savelsbergh, A. M. Williams, J. V. D. Kamp, and P. Ward, “Visual search, anticipation and expertise in soccer goalkeepers,” *Journal of sports sciences*, vol. 20, no. 3, pp. 279–287, 2002.

[101] G. J. Savelsbergh, J. V. der Kamp, A. M. Williams, and P. Ward, “Anticipation and visual search behaviour in expert soccer goalkeepers,” *Ergonomics*, vol. 48, no. 11-14, pp. 1686–1697, 2005.

[102] M. Hayhoe and D. Ballard, “Eye movements in natural behavior,” *Trends in cognitive sciences*, vol. 9, no. 4, pp. 188–194, 2005.

[103] R. R. Mourant and T. H. Rockwell, “Strategies of visual search by novice and experienced drivers,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 14, no. 4, pp. 325–335, 1972.

[104] P. Chapman, G. Underwood *et al.*, “Visual search of driving situations: Danger and experience,” *Perception-London*, vol. 27, no. 8, pp. 951–964, 1998.

[105] G. Underwood, P. Chapman, N. Brocklehurst, J. Underwood, and D. Crundall, “Visual attention while driving: sequences of eye fixations made by experienced and novice drivers,” *Ergonomics*, vol. 46, no. 6, pp. 629–646, 2003.

[106] G. Underwood, “Visual attention and the transition from novice to advanced driver,” *Ergonomics*, vol. 50, no. 8, pp. 1235–1249, 2007.

[107] N. Charness, E. M. Reingold, M. Pomplun, and D. M. Stampe, “The perceptual aspect of skilled performance in chess: Evidence from eye movements,” *Memory & Cognition*, vol. 29, no. 8, pp. 1146–1152, 2001.

[108] P. Blignaut, T. Beelders, and C. So, “The visual span of chess players,” in *Proceedings of the 2008 symposium on Eye tracking research & applications*. ACM, 2008, pp. 165–171.

[109] H. L. Kundel and P. S. L. Follette, “Visual search patterns and experience with radiological images,” *Radiology*, vol. 103, no. 3, pp. 523–528, 1972.

[110] H. L. Kundel, C. F. Nodine, and E. A. Krupinski, “Computer-displayed eye position as a visual aid to pulmonary nodule interpretation,” *Investigative radiology*, vol. 25, no. 8, p. 890, 1990.

[111] T. Donovan, D. J. Manning, and T. J. Crawford, “Performance changes in lung nodule detection following perceptual feedback of eye movements.” *Medical Imaging 2008: Image Perception, Observer Performance, and Technology Assessment (Proceedings Volume)*, vol. 6917, pp. 691 703–1, 2008.

[112] D. Litchfield, L. J. Ball, T. Donovan, D. J. Manning, and T. Crawford, “Learning from others: Effects of viewing another person’s eye movements while searching for chest nodules,” in *Proceedings of SPIE medical imaging*, 2008.

[113] M. Dorr, H. Jarodzka, and E. Barth, “Space-variant spatio-temporal filtering of video for gaze visualization and perceptual learning.” ACM, 2010, pp. 307–314.

[114] M. Dorr, M. Böhme, T. Martinetz, and E. Barth, “Gaze beats mouse: a case study,” in *Proceedings of the 2nd Conference on Communication by Gaze Interaction*, 2007, pp. 16–19.

[115] M. Speck, “LBreakout2,” http://lgames.sourceforge.net/index.php?project=LBreakout2.

[116] M. S. Weiss, “Modification of the Kolmogorov-Smirnov statistic for use with correlated data,” *Journal of the American Statistical Association*, pp. 872–875, 1978.

[117] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[118] C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” *Hum Neurobiol*, vol. 4, no. 4, pp. 219–27, 1985.

[119] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[120] D. Parkhurst, K. Law, and E. Niebur, “Modeling the role of salience in the allocation of overt visual attention,” *Vision research*, vol. 42, no. 1, pp. 107–124, 2002.

[121] L. Itti, “Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes,” *Visual Cognition*, vol. 12, no. 6, pp. 1093–1123, 2005.

[122] P. Reinagel and A. M. Zador, “Natural scene statistics at the centre of gaze,” *Network: Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.

[123] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, “Visual correlates of fixation selection: Effects of scale and time,” *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[124] F. Attneave, “Some informational aspects of visual perception.” *Psychological review*, vol. 61, no. 3, p. 183, 1954.

[125] H. Barlow, “Possible principles underlying the transformation of sensory messages,” *Sensory communication*, pp. 217–234, 1961.

[126] N. D. Bruce and J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” *Journal of Vision*, vol. 9, no. 3, 2009.

[127] E. Barth, “A geometric view on early and middle level visual coding,” *Spatial vision*, vol. 13, no. 2-3, pp. 2–3, 2000.

[128] E. Vig, M. Dorr, T. Martinetz, and E. Barth, “Intrinsic dimensionality predicts the saliency of natural dynamic scenes,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 99, pp. 1–1, 2011.

[129] O. Braddick and N. Qian, “The organization of global motion and transparency,” *Motion vision–computational, neural, and ecological constraints*, pp. 85–112, 2001.

[130] J. B. Mulligan, “Motion transparency is restricted to two planes,” *Investigative Ophthalmology and Visual Science*, vol. 33, no. 4, p. 1049, 1992.

[131] ——, “Nonlinear combination rules and the perception of visual motion transparency,” *Vision research*, vol. 33, no. 14, pp. 2021–2030, 1993.

[132] T. J. Andrews and D. Schluppeck, “Ambiguity in the perception of moving stimuli is resolved in favour of the cardinal axes,” *Vision research*, vol. 40, no. 25, pp. 3485–3493, 2000.

[133] M. Dorr, I. Stuke, C. Mota, and E. Barth, “Mathematical and perceptual analysis of multiple motions,” *TWK 2001 Beiträge zur 4. Tübinger Wahrnehmungskonferenz*, p. 173, 2001.

[134] N. Suzuki and O. Watanabe, “Perceptual costs for motion transparency evaluated by two performance measures,” *Vision research*, vol. 49, no. 17, pp. 2217–2224, 2009.

[135] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, “Variability of eye movements when viewing dynamic natural scenes,” *Journal of vision*, vol. 10, no. 10, 2010.

[136] M. Böhme, M. Dorr, C. Krause, T. Martinetz, and E. Barth, “Eye movement predictions on natural videos,” *Neurocomputing*, vol. 69, no. 16–18, pp. 1996–2004, 2006.

[137] B. W. Tatler, “The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions,” *Journal of Vision*, vol. 7, no. 14, 2007.

[138] National Highway Traffic Safety Administration, “2010 motor vehicle crashes: Overview,” 2011.

[139] ——, “Distracted driving 2009,” 2010.

[140] S. Klauer, T. Dingus, V. Neale, J. Sudweeks, and D. Ramsey, “The impact of driver inattention on near crash/crash risk: an analysis using the 100-car naturalistic driving study data,” National Highway Traffic Safety Administration, US Department of Transportation, Tech. Rep., 2006.

[141] J. L. Harbluk, Y. I. Noy, and M. Eizenman, “The impact of cognitive distraction on driver visual behaviour and vehicle control,” 2002.

[142] M. A. Recarte and L. M. Nunes, “Mental workload while driving: Effects on visual search, discrimination, and decision making,” *Journal of Experimental Psychology*, vol. 9, no. 2, pp. 119–137, 2003.

[143] J. Engström, E. Johansson, and J. Östlund, “Effects of visual and cognitive load in real and simulated motorway driving,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 2, pp. 97–120, 2005.

[144] K. A. Brookhuis and D. de Waard, “Monitoring drivers’ mental workload in driving simulators using physiological measures,” *Accident Analysis and Prevention*, no. 42, pp. 898–903, 2010.

[145] J. Mosedale, A. Purdy, and E. Clarkson, “Contributory factors to road accidents,” Department for Transport, UK, Tech. Rep., 2006.

[146] National Highway Traffic Safety Administration, “An examination of driver distractions as recorded in NHTSA database,” 2009.

[147] M.-B. Herslund and N. O. Jørgensen, “Looked-but-failed-to-see-errors in traffic,” *Accident Analysis & Prevention*, vol. 35, no. 6, pp. 885–891, 2003.

[148] A. Koustanaï, E. Boloix, P. V. Elslande, and C. Bastien, “Statistical analysis of “looked-but-failed-to-see” accidents: Highlighting the involvement of two distinct mechanisms,” *Accident Analysis & Prevention*, vol. 40, no. 2, pp. 461–469, 2008.

[149] R. A. Rensink, J. K. O’Regan, and J. J. Clark, “To see or not to see: The need for attention to perceive changes in scenes,” *Psychological Science*, vol. 8, no. 5, pp. 368–373, 1997.

[150] D. J. Simons and D. T. Levin, “Change blindness,” *Trends in cognitive sciences*, vol. 1, no. 7, pp. 261–267, 1997.

[151] J. S. McCarley, M. J. Vais, H. Pringle, A. F. Kramer, D. E. Irwin, and D. L. Strayer, “Conversation disrupts change detection in complex traffic scenes,” *Human Factors*, vol. 46, no. 3, pp. 424–36, 2004.

[152] A. Galpin, G. Underwood, and D. Crundall, “Change blindness in driving scenes,” *Transportation Research Part F*, no. 12, pp. 179–185, 2009.

[153] X. S. Zheng and G. W. McConkie, “Two visual systems in monitoring of dynamic traffic: Effects of visual disruption,” *Accident Analysis and Prevention*, vol. 42, pp. 921–928, 2010.

[154] C. Ho and C. Spence, “Assessing the effectiveness of various auditory cues in capturing a driver’s visual attention,” *Journal of Experimental Psychology: Applied*, vol. 11, no. 3, pp. 157–174, 2005.

[155] D.-Y. D. Wang, D. F. Pick, R. W. Proctor, and Y. Ye, “Effect of a side collision-avoidance signal on simulated driving with a navigation system,” in *Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2007.

[156] L. N. Boyle and J. D. Lee, “Using driving simulators to assess driving safety,” *Accident Analysis and Prevention*, no. 42, pp. 785–787, 2010.

[157] A. Kemeny and F. Panerai, “Evaluating perception in driving simulator experiments,” *Trends in Cognitive Sciences*, vol. 7, no. 1, 2003.

[158] A. R. Bowers, A. J. Mandel, R. B. Goldstein, and E. Peli, “Driving with hemianopia, I: Detection performance in a driving simulator,” *Investigative Ophtalmology and Visual Science*, vol. 50, no. 11, 2009.

[159] E. Barth, M. Dorr, M. Böhme, K. R. Gegenfurtner, and T. Martinetz, “Guiding eye movements for better communication and augmented vision,” in *Perception and Interactive Technologies*, ser. Lecture Notes in Artificial Intelligence, vol. 4021. Springer, 2006, pp. 1–8.

[160] P. de Graef, G. Hamon, M. Dorr, E. Barth, and K. Verfaillie, "Virtual navigation training and gaze guidance," 2009, the Eye and the Auto, Detroit, USA.

[161] E. Barth, "Information technology for active perception: Itap," in *First GRP-Symposium, Sehen und Aufmerksamkeit im Alter, Benediktbeuren, December 2001*, 2001.

[162] M. Kiss, "Driver assistance in the peripheral field of view," 2009, science Beyond Fiction: European Future Technologies Conference, Prague. Session on Modelling and guiding attention in an increasingly complex world.

[163] L. Pomarjanschi, M. Dorr, and E. Barth, "Gaze guidance reduces the number of vehicle-pedestrian collisions in a driving simulator," in *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*. ACM, 2011, pp. 119–119.

[164] J. G. Casali, "Vehicular simulation-induced sickness: Volume I: An overview," *Orlando, FL: Naval Training Systems Cez. ter. NTIS No. AD A173-904*, 1986.

[165] E. M. Kolasinski, "Simulator sickness in virtual environments." DTIC Document, Tech. Rep., 1995.

[166] D. M. Johnson, "Simulator sickness research summary," DTIC Document, Tech. Rep., 2007.

[167] R. S. Kennedy, N. E. Lane, M. G. Lilienthal, K. S. Berbaum, and L. J. Hettinger, "Profile analysis of simulator sickness symptoms- application to virtual environment systems," *Presence: Teleoperators and Virtual Environments*, vol. 1, no. 3, pp. 295–301, 1992.

[168] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The international journal of aviation psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[169] G. E. Riccio and T. A. Stoffregen, "An ecological theory of motion sickness and postural instability," *Ecological psychology*, vol. 3, no. 3, pp. 195–240, 1991.