Exploring perception through the eyes: from eye tracking to visual saliency and mental imagery

vorgelegt von
M. Sc.
Xi Wang

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
-Dr.-Ing.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzende:	Prof. Dr. Marianne Maertens
Gutachter:	Prof. Dr. Marc Alexa
Gutachter:	Prof. Dr. Gordon Wetzstein
Gutachter:	Prof. Dr. Kenneth Holmqvist

Tag der wissenschaftlichen Aussprache: 12. Juni 2020

Berlin 2020

Abstract

Everyday a huge amount of visual information enters the human brain through the eyes; meanwhile, a considerable amount of information is revealed by the eyes. The unique yet complex combination of the various roles of the eyes offer valuable opportunities to understand human perception, cognitive processing (e.g. visual search, recognition, and decision making), as well as mental states, and to apply this knowledge in practical applications.

This thesis studies the role of eye movements in both, the inward and outward path of the information flow. More specifically, we focus on how people look at 3D objects during perception and how people move their eyes while looking at nothing. We also investigate how eye movements could be exploited in practical scenarios. This thesis makes contributions in the intersection of fields of psychology, computer graphics and human-computer interactions, and provides the building blocks for future studies.

Part I includes developments of 3D gaze tracking system and a study focused on bias in vergence related computations. Part II presents the studies of measuring visual saliency on 3D printed objects. More precisely, it describes the data collection method which uses the system developed in Part I for 3D gaze tracking, and presents several experimental results on where people look at on 3D printed objects. Part III describes the development of computational methods to reveal information contained in eye movements while looking in front of an empty space following the well-established looking-at-nothing phenomenon in psychology. Chapter 7 describes computational methods of image retrieval based on looking-at-nothing eye movements and Chapter 8 studies the prioritization of scene elements in episodic memory by examining the recalled content during looking at nothing. The proposed computational methods allow us to explore the information revealed by eye movements during recall and provide new insights into the looking-at-nothing phenomenon.

Zusammenfassung

Einerseits strömt eine große Menge visueller Information durch die Augen in das Gehirn. Andererseits, verraten die Augen auch eine beträchtliche Menge an Informationen. Die einzigartige aber auch sehr komplexe Kombination der verschiedenen Aufgaben der Augen bietet wertvolle Möglichkeiten sowohl menschliche visuelle Wahrnehmung, kognitive Prozesse (z.B. visuelle Suche, Erkennungs- und Entscheidungsprozesse) und mentale Zustände zu verstehen, als auch dieses Wissen in praktischen Anwendungen zu nutzen.

Diese Arbeit untersucht die Rolle von Augenbewegungen während dem nach innen als auch dem nach außen gerichteten Informationsfluss. Insbesondere fokussiert die Arbeit darauf wie Menschen 3D Objekte während aktiver Wahrnehmung betrachten und wie sie ihre Augen bewegen während sich nichts betrachten. Des Weiteren untersucht diese Arbeit wie Augenbewegungen in praktischen Anwendungen genutzt werden können. Diese Arbeit liefert Beiträge an der Schnittstelle zwischen Psychologie, Computergrafik und Mensch-Maschine-Interaktion und stellt entsprechende Bausteine für zukünftige Studien zur Verfügung.

Teil I beschreibt die Entwicklung eines Systems zur Blickverfolgung in 3D und beinhaltet eine Studie zum Bias bei vergenz-bezogenen Berechnungen. Teil II präsentiert Studien bezüglich der Messung visueller Markantheit gedruckter 3D Objekte. Dieser Teil beschreibt die Methode der Datenaufnahme, welche das System aus Teil I zur 3D Blickverfolgung nutzt, und präsentiert mehrere experimentelle Ergebnisse welche Teile von gedruckten 3D Objekten Menschen am häufigsten betrachten. Teil III beschreibt die Entwicklung von Berechnungsmethoden, mit denen Informationen aufgedeckt werden, die in den Augenbewegungen enthalten sind welche ausgeführt werden wenn ein leerer Raum betrachtet wird (nach dem gut bekannten “looking-at-nothing Phänomen in der Psychologie). Kapitel 7 beschreibt Berechnungsmethoden zur Bildsuche aufbauend auf looking-at-nothing Augenbewegungen und Kapitel 8 untersucht die Priorisierung von Szenenelementen im episodischen Gedächtnis durch die Analyse des während looking-at-nothing ins Gedächtnis gerufenen Szeneninhalts. Die vorgeschlagenen Berechnungsmethoden erlauben die Informationen zu erforschen welche durch Augenbewegungen während des Erinnerns zugänglich sind und bieten neue Einsichten in das looking-at-nothing Phänomen.

Acknowledgments

I feel extremely honoured to have the opportunity to work closely together with many brilliant researchers from various disciplines. I am deeply grateful for all the supports I have got and, more importantly, for all what I have learned.

First and foremost, I would like to thank all the committee members for agreeing to review my thesis and give me valuable feedback.

I would like to thank my advisor Prof. Marc Alexa for his support and guidance through this more than five-year long journey. It's a privilege for me to have the opportunities to be advised by him and to learn a lot of things from developing research ideas to presenting the outcome. I am especially grateful for his support on all the projects, especially those that are remotely related to Computer Graphics, for the freedom to pursue "irregular" topics, and for the enthusiasm for the crazy and most often immature ideas I had.

Special thanks go to Prof. Kenneth Holmqvist for the great experience of collaborations, for being patient with us especially at the beginning the our collaboration, and for showing us different perspectives of the problems. It was Kenneth who introduced me the way into a different field and showed me how to write in the most clear manner. I had the great chance to collaborate with by Prof. Marianne Maertens at the beginning of my doctoral studies. I appreciate all her support and feedback, especially at the time when I was still trying to figuring out the ways. I also appreciate the short encounter with Prof. James Hays, for the short but fruitful and stimulating face-to-face discussions.

I had the fortune to work with many great colleagues through the years. Special thanks go to Andreas Ley for the pleasant experience of discussing problems and working together. Of course I should not forget about the after-lunch StarCraft hours. To this day, I am still amazed by his knowledge and deep understanding of various topics and the precision of how he pinpoints the problems. I will miss all kinds of discussions from the bath tab story to what is the right way to approach artificial intelligence. Many thanks go to Dr. Albert Chern for being a great mathematician friend. I still remember the first time we met, he asked whether research in our group is more like following modelling-based path or method-based way (excuse me for using the wrong terms here). At that time, even the concepts sounded new to me. I really appreciate the opportunity to collaborate on research projects as well as to talk

about all the not-work related topics. It is pity that I would not inherit your e-piano though. Too bad that it took a while until Dr. Minjing Kim and I got to know each other. I missed our girls discussions spanning from research design to cross-culture differences. It was definitely great to share those stories and thoughts. Thanks for the recommendation of Seoul Garden too. I think they even know me now, but unfortunately that probably will not bring us any discounts. I would like to thank Dr. David Lindlbauer for all his help, feedback, and efficiency through various projects. It was pity that the overlap when Ronald Richter and I were both in the group was short, but I enjoyed the discussions and chats in and out of the offices and appreciate that we continue to do so even when Ronald was no longer a cg person. I would also like to thank Dr. Christian Lessig for helping me publish my first paper in the group and for shaping it in a good form. I can barely imagine how it would be without his inputs and guidance.

I would like to thank all my friends who kept being my friends and bearing me through all these irregular years. Thanks Moritz for being a great friend who happens to play drum-sets too. I am still proud of our weekly drumming-cooking session. The unbalanced development between my drumming skill and his cooking skill sometimes makes me feel bad though. I am grateful to Martina for all the sharing and discussions. She seems to always have a big balcony in the apartment to host our long night talks. Also want to thank Mini for joining tap dance together and for listening to all my complains. I definitely need to figure out something to even catch up with the level, if that is even possible.

My heartfelt gratitude goes to my parents for their love, understanding and believing in me. As the old saying goes, parents are a child's first and most important teachers in life. Without them, I would probably still be an extremely introvert person, who might even start crying when a stranger approaches. Special thanks go to my mother, who shows me what lifelong learning means by her own example. Greatest thanks of all belong to my dear husband Ronny. Thank you for being there all the time and sharing your wisdom and experience with me. Without your love, support and guidance through all these easy or tough moments, none of the work would be possible. Thank you for all your understanding and help when I doubt about probably too many things and thank you for being on my side no matter how ridiculous I am. There is no doubt that I would be a different person without you.

Contents

Listing of figures

The eyes only see what the mind is prepared to comprehend.

Henri Bergson

1 Introduction

A tremendous amount of visual content is consumed humans every single day and all its processing replies on the input from the eyes–the sensory organ of the visual system. Visual information is gathered through eyes movements and then transformed to the brain for further processing. For example in reading information and stories are extracted from looking through lines of words. While driving on the streets, we keep looking around and focus the attention on the surroundings, so that in case of accidents, we can react in time. Sensory inputs gathered through eye movements provide information for many cognitive processes. There is a long history of studying eye movements, and extensive research has been carried out to investigate eye movements during reading[174,203,219], scene perception[51,202], visual search[96,154] and decision making[39,237].

At the same time, eye movements also act as an output device, following the center commander in the brain. Like many other human motor systems, the brain controls the movements of the eyes, makes prediction and adjustment on the fly and determines where to look next. Back to the driving example, when getting closer to a cross, our eyes are directed upwards to look for the signals of the traffic lamps. Even in situations where no traffic lamp is available, we might still look upwards. As summarized by Land and Hayhoe[162,253], the role of eye movements consists of localization, direction, guidance and checking, which are all controlled by top-down influence. In addition, studies on pupillometry[185,209] show that pupil responses are linked to high-level cognition and correspond to changes in cognitive states. For instance, increased mental effort or cognition load leads to dilated pupil. Many people

believe there are more to tell when looking into the eyes. Just as the old saying: Eyes are the window to the soul.

With the ever-increased accessibility of hardware devices, eye movement recording has become feasible in many scenarios. These advancements open new possibilities to study eye movements in natural settings and further apply in practical applications. However, the two roles simultaneously played by our eyes are tightly coupled. It presents a fundamental challenge when we attempt to separate the input utility from its output functionality in order to understand the whole information processing better.

In this thesis, we study eye movements in both inward and outward directions of the information flow. More specifically, we study how people look at 3D printed objects in space (Part II), which focuses on the role of information intake. Genuine physical stimuli are realised by 3D printing and the methods described in Part I allow us to accurately estimate gaze positions on the printed objects. Unlike most previous studies which only consider flat images as stimuli, the use of genuine physical objects brings us one step further towards understanding of natural viewing behaviour. We explore the possibility of reading information from the eyes through the looking-at-nothing (LAN) paradigm in Part III, which reveals the close link between eye movements and mental imagery. Eye movements are first used as a new modality in an image retrieval task (Chapter 7). In Chapter 8, we propose a computational method to experimentally probe what has been prioritized in episodic memory based on eye movements.

Each of these two distinct paths provides an unique view and offers the possibility to gain insights on the roles of eye movements from different perspectives. We have explored the potential of using eye movements in practical applications and these studies should provide good baselines and guidance for further research.

1.1 Eye Movements and Visual Saliency

Human viewing behavior is defined by three major systems making up the oculomotor system: the fixation-saccade system, the vestibulocular system (VOR) and the smooth pursuit system. During fixations the eyes remain relatively stationary to allow for the intake of visual information[182]. Saccades are rapid ballistic eye movements occurring between fixations[2]. The VOR stabilizes gaze during head movements. Smooth pursuit occurs when the eyes follow a smoothly moving object [226], a fact that has been exploited for interaction and to enhance eye-tracking[262], for example.

The perception of a stable visual world is achieved during fixation and a tight link between attention and eye movements has been established over the years. Concerning with human viewing behavior on static objects in a controlled environment, especially in the sense of extracting salient features, the analysis of fixations, which tells us where are the attended areas,

has been used extensively. Saccades involve no information uptake. The VOR is inactive because participants keep their heads still in most controlled experiments on visual saliency, and smooth pursuit only happens when there is a moving object.

Our visual system prioritizes visual information projected onto a small central region on the retina, the fovea. The area of the fovea corresponds to about 2° in the visual field or less than 0.03% of the whole visual field[105], yet 25% of the neurons in primary visual cortex process that foveal information. The remaining 99.9% of the visual field is used by the brain for selection of the next fixation point, and for planning body movements. The fixation-saccade system is constantly redirecting our gaze towards task-relevant and salient positions in our environment. Numerous experiments in psychology suggest that the process of selecting the peripheral elements to be looked at next is neither random nor idiosyncratic[96,224]. Humans have a common strategy which elements to fixate, and these elements must be identified in the peripheral vision. Such elements in the scene are commonly called *salient* features[20,219].

A salient visual feature is characterized by the fact that many humans direct their attention to it. Salient features have been investigated in many eye tracking experiments with images as stimuli[144,277]. While the findings are not entirely consistent, it is generally assumed that both low-level features (e.g., contrast and edges), high-level features (e. g., faces, text), and task-related features exist[33,93,114,172]. In particular, there are low-level features, which arise from the image function alone. A common setup in such an experiment includes an eye-tracker, a display which presents the stimuli as well as a chin-rest, which is required by most desktop eye trackers. The gaze position on screen is normally estimated through the built-in calibration of eye trackers. Both natural photographs and specially designed simple patterns (e.g., checkerboard) have been used as visual stimuli. Viewing time varied but is often in the order of 5 seconds and observers are mostly asked to freely explore the images. Before each trial, observers are instructed to look at a fixation cross placed at the center of a display so that the influence caused by different initial fixating points is limited.

Many saliency experiments in graphics have been conducted with 3D content being presented on screen. Besides tracking eye movements[146,165], mouse-clicking has been employed as another alternative of interacting with human observers[38,164]. Recent work has studied where people look at in virtual reality[243] or images presented on stereoscopic displays[10,268]. Both technologies have an improved 3D perception by presenting two different images to the eyes.

Questions have been arisen concerning whether models of eye movements for complex scene viewing can be generalized to behaviour in natural environments. To bridge this gap, we step towards this direction and study how people look at physical objects that are realized by 3D printing (Part II), which could potentially be used in applications of fabrication.

1.2 The Mind's Eye

In visual saliency studies, fixation is used to tell where people look, which is further taken as an indicators of what has drawn people's attention. We know that eye movements during perception are at least partially influenced by tasks and goals. But less is know about whether eye movements can be used to tell the intention or goal one has in mind. Such interpretation of the mind based on the engagement of eye movements has been formulated as part of the theory of mind[11,272].

A closely related visual experience is visual imagery, which is a conscious procedure that (re)generates visual features in a form of mental representation. Visual perception and visual imagery both rely on top-down controls. However, there are critical distinctions between them. For instance, visual imagery is less concerned about bottom-up features. Neuroimaging studies have shown that visual perception and imagery share large brain areas which are associated to high-level visual areas[55,169]. The activation of low-level visual areas during mental imagery rather depends on the level of detail and vividness of imagery content[5,47].

Visual imagery is involved in many cognitive processes, such as information retrieval, memory recall, and problem solving. Episodic memory can be considered as the storage of daily experience[259], which has autobiographical reference. To certain extends, visual imagery contains the recollection of past experience. Indeed, previous findings show that mental imagery retrieval is linked to greater activity in brain regions that are related to episodic memory[14,92].

In order to understand the role of eye movements as the mind's eye, in Part III we study the eye movements while looking in front of an empty space, following the well-established LAN paradigm. As no visual feature is presented as visual distractor during mental imagery, the influence of memory on eye movements is potentially largely amplified, offering an opportunity to read the mind's eye.

1.3 Dissertation Overview

The three complementary topics of this thesis are: (i) how to estimate gaze positions in 3D space, (ii) studying where people look at 3d printed objects and (iii) how people move their eyes while looking in front of nothing.

Part I Video-based Eye Tracking in 3D *Chapter 2* describes the approach we developed for monocular gaze estimation in 3D and *chapter 3* presents the study on 3D gaze estimation using binocular gaze data where we show that the estimated mean point of vergence is biased under projection.

Part II Comparing Eye Movements on Screen to Eye Movements in 3D *Chapter 4* first shows that saliency features do exist on 3D printed objects and the famous computational model of mesh saliency does not predict real human fixations well. Using an improved ap-

paratus, we further collect a large data set considering variations of viewing directions and printing materials. Details and analytical results are presented in *Chapter 5*.

Part III Comparing Eye Movements during Encoding to Eye Movements during Recall *Chapter 6* provides the background related to the LAN phenomena and describes the details of the data acquisition. *Chapter 7* describes the computational model we developed for eye-movement-based image retrieval system and *Chapter 8* presents an computational method which allows us to experimentally probe what has been encoded in episodic memory.

1.4 Papers Included in this Work and Contributions

This section lists the published and unpublished papers and their correspondences to the chapters in this thesis, with a note of contributions.

Chapter 2 is based on published work by: X. Wang, D. Lindlbauer, C. Lessig, and M. Alexa. *"Accuracy of Monocular Gaze Tracking on 3D Geometry"*. Burch, Michael; Chuang, Lewis; Fisher, Brian; Schmidt, Albrecht; Weiskopf, Daniel (Ed.): Eye Tracking and Visualization, Chapter 10, Springer, 2017. My contribution includes participation in the method development, implementation of the algorithm, collection of human data and numerical experiments.

Chapter 3 is based on published work by: X. Wang, K. Holmqvist and M. Alexa, *"The mean point of vergence is biased under projection"*, in Journal of Eye Movement Research, 2019. My contribution includes the collection of human data and numerical analysis.

Chapter 4 is based on published work by: X. Wang, D. Lindlbauer, C. Lessig, M. Maertens and M. Alexa, *"Measuring the Visual Salience of 3D Printed Objects"*, in IEEE Computer Graphics and Applications, 2016. My contribution includes the participation in the study design, data collection, implementation of the algorithm and numerical experiments.

Chapter 5 is based on published work by: X. Wang, S. Koch, K. Holmqvist and M. Alexa, *"Tracking the gaze on objects in 3D: how do people really look at the bunny?"*, ACM Transactions on Graphics, 2018. My contribution includes the participation in the study design, collection of human data, implementation of the algorithm and numerical analysis.

Chapter 7 is based on published work by: X. Wang, A. Ley, S. Koch, D. Lindlbauer, J. Hays, K. Holmqvist and M. Alexa, *"The Mental Image Revealed by Gaze Tracking"*, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), 2019; and unpublished work by: X. Wang, A. Ley, S. Koch, J. Hays, K. Holmqvist and M. Alexa, *"Computational discrimination between natural images based on gaze during mental imagery"*. My contribution includes the participation of the study design, collection of the human data, implementation of the algorithm and numerical analysis.

Chapter 8 is based on unpublished work by: X. Wang, K. Holmqvist and M. Alexa, *"A method to investigate which fixated content is spontaneously prioritized when recalled from*

visual episodic memory". My contribution includes the participation of the study design, collection of the human data, implementation of the algorithm and numerical analysis.

Part I

Video-based Eye Tracking in 3D

2

Accuracy of Monocular Gaze Tracking on 3D Geometry

Many applications such as data visualization or object recognition benefit from accurate knowledge of where a person is looking at. This chapter presents a system for accurately tracking gaze positions on a three dimensional object using a monocular head mounted eye tracker. It is accomplished by 1) using digital manufacturing to create stimuli whose geometry is known to high accuracy, 2) embedding fiducial markers into the manufactured objects to reliably estimate the rigid transformation of the object, and, 3) using a perspective model to relate pupil positions to 3D locations. This combination enables the efficient and accurate computation of gaze position on an object from measured pupil positions.

2.1 Introduction

Understanding the viewing behavior of humans when they look at objects plays an important role in applications such as data visualization, scene analysis, object recognition, and image generation [256]. The viewing behavior can be analyzed by measuring fixations using eye tracking. In the past, such experiments, especially for object exploration tasks, were performed with flat 2D stimuli presented on a screen [107]. However, since the human visual attention mechanism has been developed in 3D environments, depth may have an important effect on viewing behavior [151]. To understand the role of depth information, some studies [88,122,163] re-

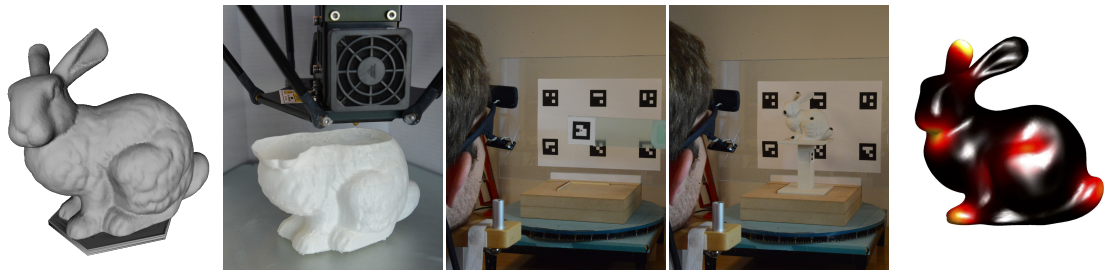


Figure 2.1: We accurately estimate 3D gaze positions by combining digital manufacturing, marker tracking and monocular eye tracking. With a simple calibration procedure we attain an angular accuracy of 0.8°.

cently combined eye tracking with stereoscopic displays. However, these displays fail to provide natural depth cues; for example they suffer from stereoscopic decoupling, the mismatch of accommodation and vergence for the displayed depth[110]. Since our research objective is to investigate the viewing behavior of humans for stimuli that are genuinely three-dimensional, we need to be able to track 3D gaze positions with high accuracy.

Standard eye tracking setups only determine human's viewing direction. The most common approach for determining viewing depth is to employ a binocular eye tracker and measure eye vergence, that is the orientation difference between the left and the right eye that ensures both are focused on the same point in space. However, as exemplified in Figure 2.2, experimentally determining depth from binocular vergence is inherently ill-conditioned. Even for an object at a modest distance the eyes and the object form a highly acute triangle so that the inevitable inaccuracies in measuring pupil positions[107] lead to large errors in the estimated depth values. Although nonlinear mappings can be employed to reduce the error[1,67,102,142,168,175,213], these require complex calibration and expensive optimization of the mapping while still leading to relatively large inaccuracies.

We base our approach on a mapping between viewing directions gathered by an eye tracker and the physical world. This is done similar to EyeSee3D[214] by tracking fiducial markers in physical space with a camera mounted on the eye tracker. We extend their approach by not only establishing which object is looked at but also determining the exact 3D gaze position on the particular object. The main components to achieve such accurate tracking are:

1. 3D stimuli are generated by digital manufacturing so that their geometry is known to high accuracy and also available in digital form without imposing restrictions on the geometry that is represented.
2. Fiducial markers are integrated into the 3D stimuli in order to reliably and accurately estimate the stimuli's 3D position relative to the head.

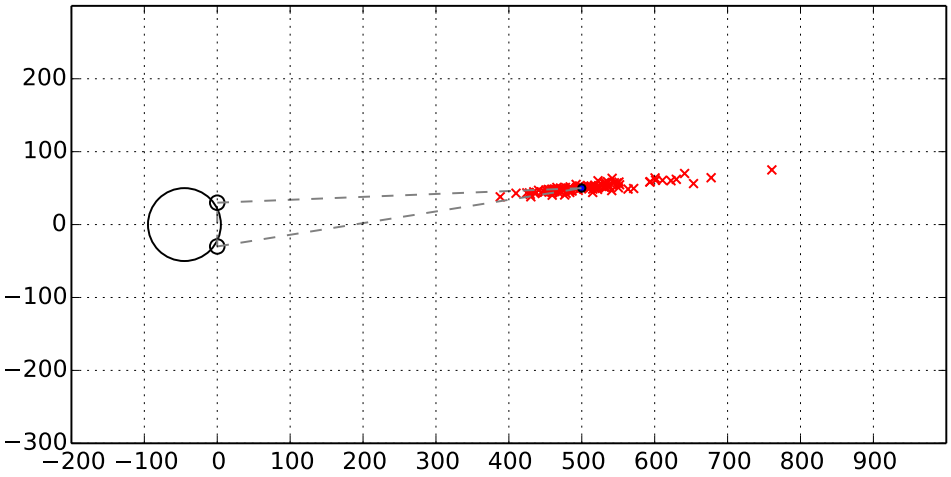


Figure 2.2: Inherent error of vergence-based depth estimation for an object at a distance of 500 mm away from the eyes. The red crosses mark estimated 3D positions for normally distributed gaze directions with mean equal to the correct angle for the object (black dot) and a variance of 0.5°. The highly acute triangle that leads to the ill-conditioning of the depth calculation is shown as dashed lines. The worst case relative error is almost 50%.

3. A simple calibration procedure that allows for an accurate computation of the perspective mapping from 3D positions to monocular pupil positions.
4. An error model for the mapping enables the computation of plausible positions on the 3D stimulus.

Our results demonstrate that for typical geometries we are able to obtain 0.8° angular resolution and reliable depth values within 1.5% of the true value, including around silhouettes where the geometry has a large slope and depth estimation is hence particularly difficult. We accomplish this with only a monocular eye tracker and an 11-point calibration procedure.

In the next section, we discuss related work on 3D gaze tracking. Subsequently, we detail our setup and explain how 3D positions can be related to pupil coordinates in Section 2.3. This is followed by a discussion of how 3D viewing positions can be obtained from pupil positions in Section 2.4. Experimental results verifying the accuracy of our approach are presented in Section 2.5. We conclude this chapter in Section 2.6 with a discussion and possible directions for future work.

2.2 Related Work

The viewing behavior of humans is typically analyzed using eye tracking by measuring a subject's fixations. However, usually only flat 2D stimuli on a screen are employed e. g. [26,132,184,217], even when one is interested in 3D objects. Only recently the first studies considering the effect of depth were performed. Lang et al. [163] collected a large eye fixation database for still images with depth information presented on a stereoscopic display. Their results show that depth can have a significant influence on a subject's fixations. Jansen et al. [122] also employed a stereoscopic display to analyze the effect of depth, demonstrating that depth information leads to an overall increase in spatial distribution of gaze positions for visual exploration tasks. Both Lang et al. [163] and Jansen et al. [122] report that visual attention shifts over time from objects closer to the viewer to those farther away. Differences in fixations between 2D and 3D stimuli were recently also investigated for stereoscopic video [88,90,112,218]. Discrepancies were mainly observed for scenes that lack on obvious (high-level) center of attention, with fixations having a larger spatial distribution when depth information is present.

Existing work investigating the role of depth information on fixation locations hence demonstrates that, at least under certain circumstances, depth has a significant effect on a subject's viewing behavior. However, so far only stereoscopic displays were employed, which do not provide all depth cues and suffer from stereoscopic decoupling [110]. Moreover, Duchowski et al. [61] showed that for stereoscopic displays the gaze depth of subjects does not fully correspond to the presented depth. Therefore, we believe that to understand viewing behavior for 3D objects, one should study stimuli that are genuinely three-dimensional. This provides the principal motivation for our work.

With 3D stimuli, also the depth values of fixation points have to be determined. The most common approach for obtaining fixation depth is to measure the vergence using a binocular eye tracker. However, computing depth values from binocular vergence is ill-conditioned since already for modest distances minuscule measurement errors in the pupil positions lead to large depth errors, cf. Figure 2.2. To improve the accuracy, Essig et al. [67] trained a neural network that maps from eye vergence to depth values. Maggia et al. [175] proposed a somewhat simpler but also nonlinear model for the mapping from measured disparity to depth. Building on these works, current techniques [1,102,142,168,213] that employ binocular vergence to determine fixation depth obtain an error that is within 10% of the correct depth value.

Our work was inspired by existing approaches relating view directions to *known* geometry, e. g. in applications of virtual reality [44,248]. Pfeiffer and Renner used fiducial markers to align the physical world to camera space [214]. They achieved an angular accuracy of 2.25°, which correctly classifies fixation targets on the scale of whole objects. However, for investigating human viewing behavior on the surface of 3D objects, more accurate gaze tracking is required. Consequently, we create a setup with the goal of tracking visual attention on 3D objects.

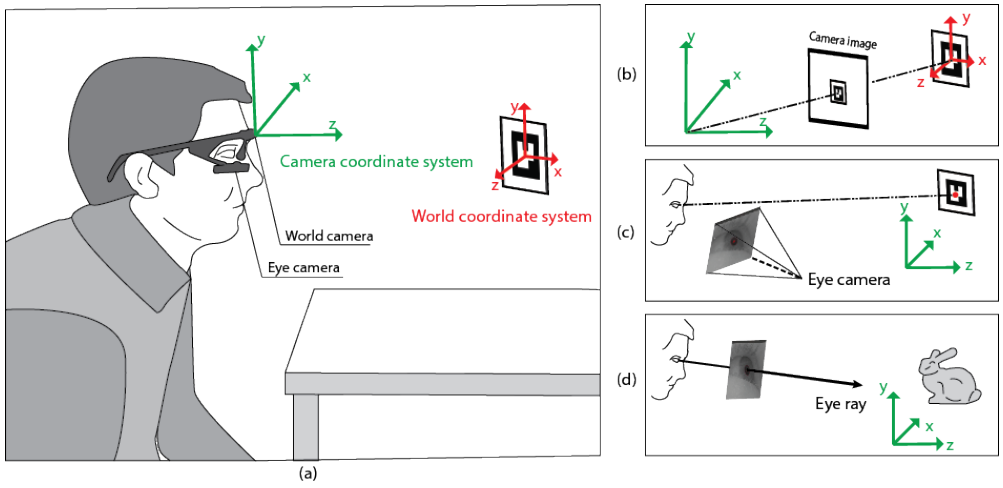


Figure 2.3: The main idea of our approach is to establish a mapping between points in 3D space (i.e. world coordinate system) and pupil coordinates in the image coordinate system of the eye camera. We consider all 3D positions relative to the coordinate system of the world camera (i.e. camera coordinate system). (b) A point in world coordinate system is first transformed into the camera coordinate system. (c) We model the mapping between pupil position in the eye camera image and a location in world camera space as projection. (d) From the estimated projective transformation, we can estimate a corresponding eye ray for each pupil position.

2.3 From 3D Positions to Pupil Coordinates

In this section we describe our setup and how it enables to accurately determine gaze positions on an object. We use a monocular head mounted eye tracking device with a front facing world camera capturing the environment and an eye facing camera capturing the pupil movement.

The world camera yields the position and orientation of fiducial markers, for example fixed to objects, relative to the subject's head relative to its reference frame. A projective mapping then relates these 3D coordinates to pupil positions relative to the camera tracking the eye. This establishes a mapping between points in 3D space and pupil coordinates (this basic idea is illustrated in Figure 2.3).

The mapping is calibrated by having a subject focus on markers at different locations, including varying depths. Once the mapping is established, 2D pupil positions can be turned into rays corresponding to gaze directions in 3D space. The gaze directions then determine the 3D positions on the object a subject is looking at, by intersecting the rays with the known 3D geometry.

In the following we will describe these steps in more detail.

2.3.1 From Local 3D Positions to World-Camera Coordinates

We employ fiducial markers to determine the 3D coordinates of locations in space in the world camera coordinate system. The mapping of a position $\mathbf{x} \in \mathbb{R}^3$, for example a point on a marker, to its projection $\mathbf{m} \in \mathbb{R}^2$ in the world camera image is given by

$$\begin{pmatrix} \mathbf{m} \\ 1 \end{pmatrix} = \mathbf{K}(\mathbf{R}\mathbf{x} + \mathbf{t}), \quad \mathbf{R}^\mathsf{T}\mathbf{R} = \mathbf{I} \tag{2.1}$$

where $\mathbf{K} : \mathbb{R}^3 \to \mathbb{R}^2$ is the intrinsic world camera matrix, modelling the perspective mapping, and $\mathbf{R}$ and $\mathbf{t}$ are the rotation and translation of the camera, forming the rigid transformation. The mapping of $\mathbf{x}$ to its representation $\mathbf{w} \in \mathbb{R}^3$ in the world camera coordinate system is hence

$$\mathbf{w} = \mathbf{R}\mathbf{x} + \mathbf{t}. \tag{2.2}$$

We determine the intrinsic world camera matrix $\mathbf{K}$, which includes both radial and tangential distortion, in a preprocessing step using the approach of Heng et al.[101]. To determine the rigid transformation given by $\mathbf{R}$ and $\mathbf{t}$ we exploit that detected marker corner points $\mathbf{m}_i \in \mathbb{R}^2$ in the camera image have known 3D locations $\mathbf{x}_i \in \mathbb{R}^3$ in the marker's local coordinate system. Given at least three such points $\mathbf{m}_i$ in the camera image, we can determine $\mathbf{R}$ and $\mathbf{t}$ by minimizing the reprojection error.

Once $\mathbf{R}$ and $\mathbf{t}$ have been estimated, we can employ Eq. (2.2) to determine the position of the center of the marker in the world camera coordinate system, as required for calibration, or to map an object with a fixed relative position to a marker into the space, as is needed to determine gaze positions.

2.3.2 From World Camera Coordinates to Pupil Positions

Given positions $\mathbf{w} \in \mathbb{R}^3$ in the world camera coordinate system, obtained as described in the last section, we have to relate these to a person's gaze direction, described by pupil positions $\mathbf{p}$ in the eye camera image. We model the mapping as a projective transformation, because the cameras and the system of the eye (i.e. the head) are in fixed relative orientation and position. In homogeneous coordinates the transformation is given by

$$s \begin{pmatrix} \mathbf{p} \\ 1 \end{pmatrix} = \mathbf{Q} \begin{pmatrix} \mathbf{w} \\ 1 \end{pmatrix} \tag{2.3}$$

where $\mathbf{Q} \in \mathbb{R}^{3\times4}$ is a projection matrix that is unique up to scale. Given a set of correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ between 3D points $\mathbf{w}_i$ in the world camera coordinate system and pupil

positions $\mathbf{p}_i$ describing the gaze direction towards $\mathbf{w}_i$, we can determine $\mathbf{Q}$ by minimizing

$$E(\mathbf{Q}) = \sum_i \left\| s_i \begin{pmatrix} \mathbf{p}_i \\ 1 \end{pmatrix} - \mathbf{Q} \begin{pmatrix} \mathbf{w}_i \\ 1 \end{pmatrix} \right\|_2^2. \tag{2.4}$$

Fixing one coefficient of $\mathbf{Q}$ to eliminate the freedom on scale (we choose $\mathbf{Q}_{3,4} = 1$), this is a standard linear least squares problem. In practice, we solve this problem using correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ obtained during calibration, as described in Section 2.5.

Since $\mathbf{Q}$ is a projective transformation we can factor it into an upper triangular intrinsic camera matrix $\mathbf{A}_Q$ and a rigid transformation matrix $\mathbf{T}_Q = (\mathbf{R}_Q, \mathbf{t}_Q)$. The factorization is given by

$$\mathbf{Q} = \mathbf{A}_Q \mathbf{T}_Q = (\mathbf{A}_Q \mathbf{R}_Q, \mathbf{A}_Q \mathbf{t}_Q) \tag{2.5}$$

and hence can be determined from the *RQ* decomposition of the left 3×3 block $\mathbf{A}_Q \mathbf{R}_Q$ of $\mathbf{Q}$. It can be computed using the *QR* decomposition as

$$\mathbf{J}(\mathbf{A}_Q \mathbf{R}_Q)^T \mathbf{J} = (\mathbf{J}\mathbf{A}_Q^T \mathbf{J})(\mathbf{J}\mathbf{R}_Q^T \mathbf{J}) \tag{2.6}$$

where $\mathbf{J}$ is the exchange matrix, which in our case is the column inversed version of the identity matrix.

2.3.3 From Pupil Positions to View Cones

So far we have related 3D locations to pupil positions. To determine a gaze point on an object we also have to relate pupil positions to a cone of positions in space. This also corresponds to the angular accuracy of our setup.

With the intrinsic eye camera matrix $\mathbf{A}_Q$, as determined in the last section, we can relate a homogeneous pupil position $\hat{\mathbf{p}} = (\mathbf{p}, 1)^T$ to an associated ray $\mathbf{r}$ in 3D world camera space:

$$\hat{\mathbf{p}} = \mathbf{A}_Q \mathbf{r}. \tag{2.7}$$

The depth along $\mathbf{r}$ is indeterminate since $\mathbf{A}_Q$ is a projection matrix. The angle between two rays $\mathbf{r}_i, \mathbf{r}_j$, represented by pupil coordinates $\mathbf{p}_i, \mathbf{p}_j$, is hence given by

$$\cos \eta_{ij} = \frac{\mathbf{r}_i^T \mathbf{r}_j}{\|\mathbf{r}_i\| \|\mathbf{r}_j\|} = \frac{\hat{\mathbf{p}}_i^\top \mathbf{A}_Q^{-\top} \mathbf{A}_Q^{-1} \hat{\mathbf{p}}_j}{\|\mathbf{A}_Q^{-1} \hat{\mathbf{p}}_i\| \|\mathbf{A}_Q^{-1} \hat{\mathbf{p}}_j\|}. \tag{2.8}$$

This suggests to interpret the matrix $\mathbf{A}_Q^{-\top} \mathbf{A}_Q^{-1}$ as an induced inner product $\mathbf{M}_Q = (\mathbf{A}_Q \mathbf{A}_Q^\top)^{-1}$

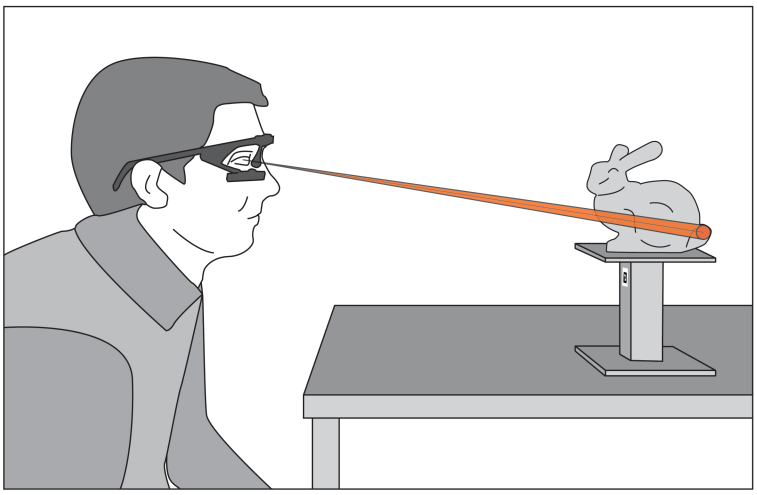


Figure 2.4: Pupil positions provide by the eye tracker correspond to cones in 3D space. The fiducial marker on the 3D printed marker allows tracking the geometry in 3D space. Intersecting the cone against the geometry yields gaze points on the object.

on homogeneous pupil coordinates. The angle η_{ij} then becomes

$$\cos \eta_{ij} = \frac{\hat{\mathbf{p}}_i^\mathsf{T} \mathbf{M}_Q \hat{\mathbf{p}}_j}{(\hat{\mathbf{p}}_i^\mathsf{T} \mathbf{M}_Q \hat{\mathbf{p}}_i)^{1/2} (\hat{\mathbf{p}}_j^\mathsf{T} \mathbf{M}_Q \hat{\mathbf{p}}_j)^{1/2}}. \tag{2.9}$$

For multiple pairs $\mathbf{p}_i$, $\mathbf{p}_j$, Eq. (2.9) can be solved efficiently when the involved matrices are precomputed.

2.4 From Pupil Coordinates to Locations on an Object

Our objective is to determine a gaze position $\bar{\mathbf{w}} \in \mathbb{R}^3$ in space from a pupil position $\bar{\mathbf{p}}$ describing a gaze direction. Central to our approach for determining $\bar{\mathbf{w}}$ is that the geometry of the observed object is known to high accuracy. This is ensured by 3D printing the object $\mathcal{M}$ from its digital representation as a triangulated surface $\mathbf{M}$. The printed object also includes a fiducial marker, which allows us to determine the rigid transformation of the object in space as described in Section 2.3.1.

As explained before, in view of inaccuracies, the pupil position $\bar{\mathbf{p}}$ describes a cone in 3D space. Consequently, we wish to identify the vertices on the object that intersect the cone and

are visible. We could then potentially identify the vertex closest to the center of the cone as the desired gaze location. The approach is illustrated in Figure 2.4.

Let

$$\hat{\mathbf{p}}_i = \mathbf{Q}(\mathbf{R}\mathbf{v}_i + \mathbf{t}) \tag{2.10}$$

be the homogeneous pupil position $\mathbf{p}_i = (p_{i_1}, p_{i_2}, p_{i_3})^\mathsf{T}$ corresponding to vertex $\mathbf{v}_i$. Then we find the set of vertices

$$\Gamma_c(\bar{\mathbf{p}}) = \left\{ \mathbf{v}_i \in \mathcal{M} \;\middle|\; \frac{\hat{\bar{\mathbf{p}}}^\mathsf{T}\mathbf{M}_Q\hat{\mathbf{p}}_i}{(\hat{\bar{\mathbf{p}}}^\mathsf{T}\mathbf{M}_Q\hat{\bar{\mathbf{p}}})^{1/2}(\hat{\mathbf{p}}_i^\mathsf{T}\mathbf{M}_Q\hat{\mathbf{p}}_i)^{1/2}} > \cos c \right\}; \tag{2.11}$$

that is, we are determining which vertices $\mathbf{v}_i$ on the object lie within the cone of angular size c centered around the eye ray corresponding to $\bar{\mathbf{p}}$. From these vertices, we consider the one closest to the eye as the intersection point. This point can be determined efficiently solely using p_{i_3}. Note that since the metric $\mathbf{M}_Q$ has a natural relation to eye ray angle, we can choose c based on the accuracy of our measurements.

2.4.1 Spatial Partitioning Tree

For finely tessellated meshes, testing all vertices based on Eq. (2.11) above results in high computational costs. Spatial partitioning can be used to speed up the computation, by avoiding to test vertices that are far away from the cone. Through experimentation we have found sphere trees to outperform other common choices of spatial data structures (such as kd-trees, which appear as a natural choice) for the necessary intersection against cones.

Each fixation on the object is the intersection of the eye ray cone with the object surface, which is represented by a triangulated surface $\mathbf{M}$. Therefore, in the first step we perform an in-cone search to find all intersected vertices. This intersection result contains both front side and back side vertices. We are, however, only interested in visible vertices that are unoccluded with respect to the eye.

A popular space-partitioning structure for organizing 3D data is K-d tree, which divides space using splitting hyperplanes, and octree, where each cell is divided into eight children of equal size. For our application, such axis-aligned space partitionings would require a cone-plane or cone-box intersection, which potentially incurs considerable computational costs. In order to avoid this, we build a space-partitioning data structure based on a sphere tree.

Sphere Tree Construction Our sphere tree is a binary tree whose construction proceeds top-down, recursively dividing the current sphere node into two child nodes. To determine the children of a node, we first apply principle component analysis and use the first principle vector, which corresponds to the largest eigenvalue of the covariance matrix, as the

splitting direction. A partitioning hyperplane orthogonal to the splitting direction is then generated so that the elements in the node are subdivided into two sets of equal cardinality. Triangle faces intersecting with the splitting hyperplane are assigned to both sets. The child nodes are finally formed as the bounding spheres of the two sets and computed as proposed in [225].

We calculate the sphere-cone intersection following the method proposed in [234]. The problem is equivalent to checking whether the sphere center is inside an extended region, which is obtained by offsetting the cone boundary by the sphere radius. Note that the extended region differs from the extend cone, and its bottom is a sector of the sphere. For each intersected leaf node, we perform the following in-cone test to find the intersected vertices.

In-cone Test A view cone is defined by an eye point $\mathbf{a}$ (i. e. the virtual eye position), a unit length view direction $\mathbf{r}$, and opening angle δ. The in-cone test allows us to determine if a given point $\mathbf{v}_i$ lies inside this cone. Given the matrix $\mathbf{M} \in \mathbb{R}^{4\times4}$

$$\mathbf{M} = \begin{pmatrix} \mathbf{S}, & -\mathbf{Sa} \\ -\mathbf{a}^T\mathbf{S}, & \mathbf{a}^T\mathbf{Sa} \end{pmatrix}, \tag{2.12}$$

where $\mathbf{S} = \mathbf{r}\mathbf{r}^T - \mathbf{d}^2\mathbf{I}$ with $\mathbf{d} = cos\delta$, the point $\mathbf{v}_i$ lies inside the cone only when $\hat{\mathbf{v}}^T\mathbf{M}\hat{\mathbf{v}} > 0$ where

$$\hat{\mathbf{v}} = \hat{\mathbf{v}}_i - \hat{\mathbf{a}} = \begin{pmatrix} \mathbf{v}_i \\ 1 \end{pmatrix} - \begin{pmatrix} \mathbf{a} \\ 1 \end{pmatrix}. \tag{2.13}$$

Visibility Test The visibility of each intersected vertex is computed by intersecting the ray from eye point to the vertex with the triangle mesh. The vertex is visible if no other intersection is closer to the eye point. We use the Möller-Trumbore ray-triangle intersection algorithm [140] for triangles in intersected bounding spheres. In our implementation, the maximum tree depth is set to 11, which allows for fast traversal and real-time performance.

2.4.2 Implementation

Our software implementation uses OpenCV [24], which was in particular employed to solve for the rigid transformations $\mathbf{R}$,$\mathbf{t}$ as described in Section 2.3.1. We determine $\mathbf{Q}$ using Eq. (2.4) with the Ceres Solver[3]. The optimization is sensitive to the initial estimate, which can result in the optimization converging to a local minimum, yielding unsatisfactory results. To overcome this problem, we use a RANSAC approach for the initial estimate, with the error being calculated following Eq. (2.14) and 1000 iterations. The result of this procedure serves as input for the later optimization using the Ceres solver.

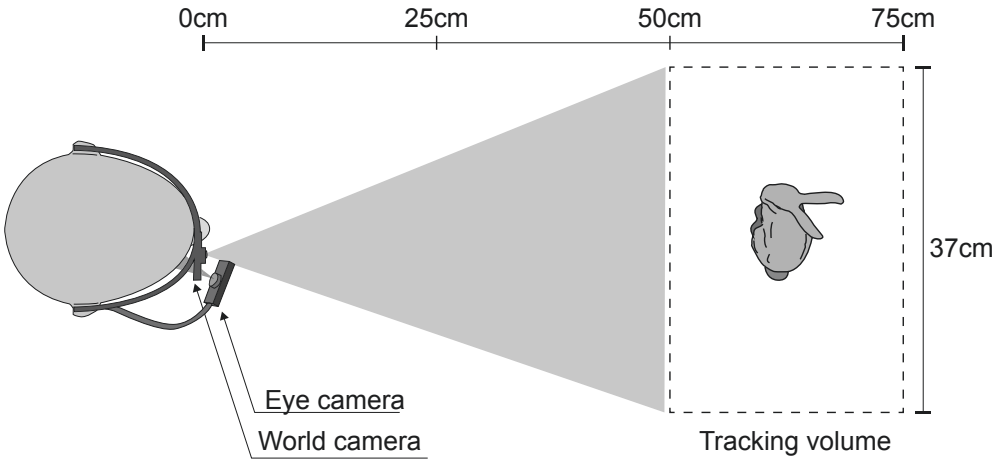


Figure 2.5: Physical setup used in our experiments.

2.5 Experiments

In the following, we will report on preliminary experimental results that validate the accuracy of our setup for tracking 3D gaze points and that demonstrate that a small number of correspondences suffices for calibration. These results were obtained using two exploratory experiments with a small number of subjects ($n = 6$).

Participants and Apparatus We recruited 6 unpaid participants (all male), all of which were students or staff from a university. Their age ranged from 26 to 39 years and all had normal or corrected-to-normal vision, based on self-reports. Four of them had previous experience with eye tracking.

The physical setup of our experiment is shown in Figure 2.5. For measuring fixations we employed the Pupil eye tracker[137] and the software pipeline described in the previous sections.

2.5.1 Accuracy of Calibration and Gaze Direction Estimation

In Section 2.3.2 we explained how the projective mapping $\mathbf{Q}$ from world camera coordinates to pupil positions can be determined by solving a linear least squares problem. As input to the problem one requires correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ between world camera coordinates $\mathbf{w}_i$ and pupil positions $\mathbf{p}_i$. The correspondences have to be determined experimentally, and hence will be noisy. The accuracy with which $\mathbf{Q}$ is determined therefore depends on the number of correspondences that is used. In our first experiment we investigated how many

correspondences are needed to obtain a robust estimate for $\mathbf{Q}$. The same data also allows us to determine the angular error of our setup.

Procedure We obtained correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ by asking a subject to focus on the center of a single fiducial marker (size 4 cm $\times$ 4 cm) while it is presented at various locations in the desired view volume (see Figure 2.1, third image). We have augmented the center of the marker with a red dot to make this task as unambiguous as possible. At each position of the marker, we estimate a single correspondence $(\mathbf{w}_i, \mathbf{p}_i)$ based on the estimation of the rigid transformation for the marker, cf. Section 2.3.1. For each participant, we recorded 100 correspondences $\{(\mathbf{w}_i, \mathbf{p}_i)\}$ for two different conditions, resulting in a total of 200 measurements per participant. In the first condition the head was fixed on a chin rest while in the second condition participants were only asked to keep facing towards the marker. For both conditions the marker was moved in a volume of 0.37 m (width) $\times$ 0.4 m (height) $\times$ 0.25 m (depth) at a distance of 0.75 m from the subject (see Figure 2.5).

Data Processing For each dataset we perform 10 trials of 2-fold cross validation and estimate the projection matrix using 7 to 49 point pairs. In each trial, the 100 correspondences are randomly divide into 2 bins of 50 point pairs each. One bin is used as training set and the other as testing set. Point pair correspondences from the training set are used to compute the projection matrix $\mathbf{Q}$ which is then employed to compute the error between the gaze direction given by the pupil position $\mathbf{p}_i$ and the true direction given by the marker center $\mathbf{w}_i$ for the points in the test data set. From Eq. (2.9) this error can be calculated as

$$\eta_i = \cos^{-1} \frac{\mathbf{p}_i^\mathsf{T} \mathbf{M}_Q \mathbf{Q} \mathbf{w}_i}{(\mathbf{p}_i^\mathsf{T} \mathbf{M}_Q \mathbf{p}_i)^{1/2} (\mathbf{w}_i^\mathsf{T} \mathbf{Q}^\mathsf{T} \mathbf{M}_Q \mathbf{Q} \mathbf{w}_i)^{1/2}}. \tag{2.14}$$

Analysis and Results In order to analyze the influence of the number of calibration points as well as the usage of the chin rest on the estimation accuracy, we performed a repeated measures ANOVA ($\alpha = .05$) on the independent variable *Chin rest* with 2 levels (with, without) and *Calibration* with 43 levels (the corresponding number of calibration points, i.e., 7 to 49). The dependent variable was the calculated angular error in degree. We used 10 rounds of cross validation for our repeated measures, with each data point being the average angular error per round. This resulted in an overall of 860 data points per participant (2 *Chin reset* $\times$ 43 *Calibration* $\times$ 10 cross validation).

Results showed a main effect for *Calibration* ($F_{42,210} = 19.296, p < .001$). The difference between 20 points ($\mathcal{M} = 0.75, SE = 0.02$) and 42, 44, 45, 46, 47 and 48 points (all $\mathcal{M} = 0.71$, $SE = 0.02$) was significantly different, as well as 22 points ($\mathcal{M} = 0.74, SE = 0.02$) compared to 45 points (all $p < .05$). No other combinations were statistically significantly different,

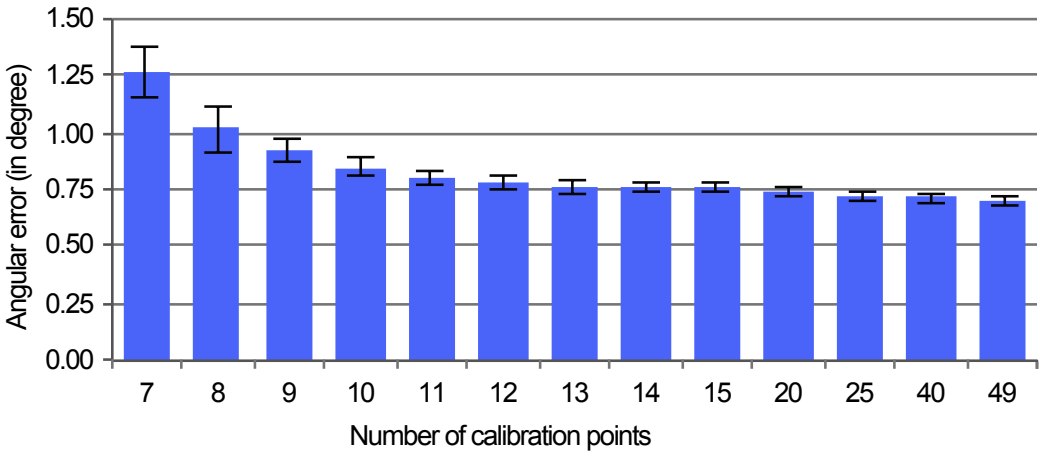


Figure 2.6: Mean values and standard errors for angular error as a function of the number of calibration points ranging from 7 to 49. No significant changes in angular error occur when using 11 or more calibration points.

arguably due to high standard deviation for lower number of calibration points. Mean values and standard errors are depicted in Figure 2.6.

When using 11 to 49 calibrations points, the angular error averages at around 0.73° ($SD = 0.02$), which is within the range of human visual accuracy and goes in line with the specifications of the pupil eye tracker for 2D gaze estimation[12,137]. The results furthermore demonstrate that even for a relatively low number of calibration points, comparable to the 9 points typically used for calibration for 2D gaze estimation[107,137], our method is sufficiently accurate.

No significant effect for *Chin rest* ($F_{1,5} = 0.408$, $p = .551$; with chin rest $\mathcal{M} = 0.73$, $SE = 0.05$; without chin rest $\mathcal{M} = 0.78$, $SE = 0.04$) was present, suggesting that the usage of a chin rest has negligible influence on the angular accuracy and our method is hence insensitive to minor head motion. This goes in line with the observation that light head motion has no effect on the relative orientation and position of eye, eye camera, and world camera. It should be noted, however, that participants, although not explicitly instructed, were mostly trying to keep their head steady, most likely due to the general setup of the experiment.

Giving participants the ability to move their head freely is an important feature for exploring objects in a natural, unconstrained manner. However, quantifying the effect of large scale motion on accuracy should be subject to further investigations.

2.5.2 Accuracy of 3D Gaze Position

In our second experiment we explored the accuracy of our approach when viewing 3D stimuli. As model we employed the Stanford bunny and marked a set of pre-defined target points on

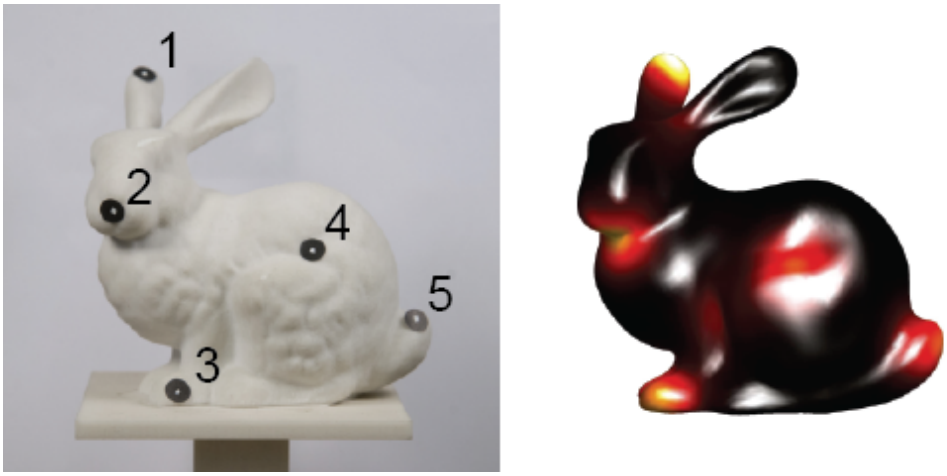


Figure 2.7: *Left*: physical bunny model with target markers (numbers indicate order); *right*: heat map of obtained gaze directions.

the 3D printed bunny as shown in Figure 2.7, left. After a calibration with 11 correspondences, as described in the last section, the test subjects were asked to focus on the individual targets (between 1 and 2 seconds). A heat map of the obtained gaze positions is shown in Figure 2.7, right. Fixations are calculated based on Eq. (2.11) where the angular size c is set to be $0.6°$. Table 2.1 shows the angular error of each target in degrees as well as the depth error in mm.

The angular error depends mostly on the tracking setup. However, since the intersection computation with eye ray cones is restricted to points on the surface (vertices in our case), we get smaller angular errors on silhouettes.

Depth accuracy, on the other hand, depends on the slope of the geometry. In particular, at grazing angles, that is when the normal of the geometry is orthogonal or almost orthogonal to the viewing direction, it could become arbitrarily large. For the situations of interest to us where we have some control over the model, the normal is orthogonal or almost orthogonal to the viewing direction mainly only around the silhouettes. Since we determine the point on the object that best corresponds to the gaze direction, we obtain accurate results also around silhouettes. This is reflected in the preliminary experimental results where we obtain an average depth error of 7.71 mm at a distance of 553.97 mm, which corresponds to a relative error of less than 2%, despite three of five targets being very close to a silhouette.

2.6 Discussion

The proposed method for estimating fixations on 3D objects is simple yet accurate. It is enabled by:

Table 2.1: Errors of individual markers on bunny.

Marker index	1	2	3	4	5
Angular error (deg.)	0.578	1.128	0.763	0.846	0.729
Depth error (mm)	7.998	8.441	10.686	3.036	8.381

- generating stimuli using digital manufacturing to obtain precisely known 3D geometry without restricting its shape;
- utilizing fiducial markers in a known relative position to the geometry to reliably determine its position relative to a subject's head;
- using a projective mapping to relate 3D positions to 2D pupil coordinates.

We experimentally verified our approach using two explorative user studies. The results demonstrate that 11 correspondences suffice to reliably calibrate the mapping from pupil coordinates to 3D gaze locations with an angular accuracy of 0.8 degree. This matches the accuracy of 2D gaze tracking. We achieve a depth accuracy of 7.7 mm at a distance of 550 mm, corresponding to a relative error of less than 1.5%.

With the popularization of 3D printing, our approach can be easily applied to a large variety of stimuli, and thus usage scenarios. At the same time, it is not restricted to 3D printed artifacts and can be employed as long as the geometry of an object is known, for example when manual measurement or 3D scanning has been performed. Our approach also generalizes to simultaneously tracking gaze with multiple objects, as long as the objects' position and orientation are unambiguously identified, e. g. by including fiducial markers. The tracking accuracy in such situations will be subject to future investigation.

3

The Mean Point of Vergence is Biased under Projection

The previous chapter introduces a system for three-dimensional gaze estimation based on monocular eye tracking data. However, the computation requires a digital representation of the scene. Theoretically, the point of interest in space can also be computed based on intersecting the two lines of sight and finding the point closest to them. This chapter studies the 3D gaze estimation based on binocular vergence. We first start by theoretical analysis with synthetic simulations. We show that the mean point of vergence is generally biased for centrally symmetric errors and that the bias depends on the horizontal vs. vertical noise distribution of the tracked eye positions. Our analysis continues with an evaluation on real experimental data. The estimated mean vergence points seem to contain different errors among individuals but they generally show the same bias towards the observer. The bias tends to be larger with an increased viewing distance. We also provide a recipe to minimize the bias, which applies to general computations of gaze estimation under projection. These findings not only have implications for choosing the calibration method in eye tracking experiments and interpreting the observed eye movements data; but also suggest that we shall consider the mathematical models of calibration as part of the experiment.

3.1 Introduction

Humans tend to direct both of their eyes at roughly the same point in 3D space. Binocular saccades and smooth pursuit between objects in a 3D scene often exhibit *vergence*, which means that two eyes move in opposite directions[31] for fixation to coincide with the intended object. In other words, vergence is the movement of both eyes towards or away from each other, depending on the relative change from the previous to the current target. It is often assumed that the fixation points of the two eyes are perfectly aligned but it has been shown that the eyes first diverge before they converge at the gaze point during fixations[41,42]. Studies on binocular coordination of eye movements during reading show that fixation points of two eyes vary during reading and disparity in both horizontal and vertical directions were observed[174,203]. There is considerable variation among participants ability to fixate the same point in depth, depending on their eye dominance and squinting, or even strabismus (when the weak eye is off-target). In addition, measurements of vergence with video-based eye-trackers are very sensitive to variations in pupil dilation[108], which leads to uncertainty over measurements of vergence.

Many approaches have been proposed to estimate the direction of gaze for each eye in physical space, based on recorded pupil positions by eye-tracking devices. Using these gaze vectors, it is possible to reconstruct the gaze point on real three-dimensional stimuli by intersecting one or both rays with the fixated object in space, assuming its geometry is known[89,188,269]. Alternatively, we can attempt to find the point where the two vectors intersect with each other in space[87,102,175,214,215], but in 3D space two gaze vectors typically do not intersect.

However, even if the observer experiences looking at a point in space with both eyes, the eye rays provided by the eye-tracker contain error, for a variety of reasons:

- Data from eye-trackers have an inaccuracy (systematic error) and introduce imprecision (variable error or noise) onto the signal[107,199,265].
- The inaccuracy is not constant, but varies with pupil size [58] and quantization of the corneal reflection (CR) in the eye camera [106].
- Ideally, human gaze direction is controlled only to bring the object into the fovea centrals[8], which has a non-negligible extent of $1.5 - 2°$.
- It is well-known that in binocular vision, many observers have a dominant eye which is more accurately directed towards the target (in about 70% of the cases, the right eye) and a weaker eye, which may be considerably off-target[41,42,66], which in extreme cases, we call strabismus.

- The resulting unknown and likely non-linear function mapping from tracked pupil and CR centers in the eye video to lines of sight is approximated using low order polynomials[34,67,107].

For these reasons, the two projected eye rays produced by eye-trackers are generally skewed and have no common intersection point in space. In order to calculate a point that approximates the expected intersection of the rays, the most natural and commonly employed solution is to compute the point that has the smallest distance to both rays in 3D. Here we call this point the vergence point. We derive the necessary equations for this computation next and then use it for simulating the reconstruction of vergence points in the presence of systematic (accuracy) and variable (precision) errors. We then develop the mathematical description of inaccuracy and imprecision of gaze vectors, that are used to simulate the effect on estimated intersection point of inaccuracy (offsets) and imprecision (noise). Hooge et al.[108] present human data of vergence error as an effect of inaccuracy from changes in pupil dilation. In this chapter, we focus on the effect of imprecision and present recorded human vergence data that validate the simulation on noise. Finally, we present a method to better estimate the intersection of eye rays and reconstruct the position of the fixated object in 3D, given noisy vergence data. Note that the whole analysis applies to eye tracking not only in space but also on flat surface, as long as the underling projective mapping is used in the model.

3.2 Part 1: Mathematical Model and Simulation

When recording binocular gaze in 3D, the two gaze vectors can be thought of as originating in the centers of the two eyes of the observer. Two vectors in three- dimensional space are generally *skew*, i.e. they have no common (intersection) point. For the reasons mentioned in the introduction, the two gaze vectors are commonly far from intersecting. In order to assign a *point of interest* given two gaze vectors that have no intersection point, the dominant strategy is to compute the point in 3D space that has the smallest sum of squared distances to the two gaze lines. Here we show how to compute distances of a point to a line by using the formulation of projector, and then how to find the point of interest.

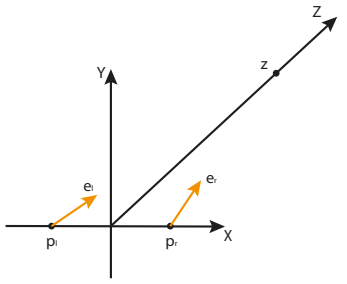


Figure 3.1: Geometric setup. $\mathbf{p}_l$ and $\mathbf{p}_r$ are the centers of the left and right eyes, and $\mathbf{e}_l$ and $\mathbf{e}_r$ are gaze vectors of unit length, i.e. the eye rays. We want to calculate the point $\mathbf{z}$ with the minimal distance to both rays.

We choose the coordinate system such that centers of the eyes are displaced symmetrically from the origin along the first coordinate direction. The up direction defines the second coordinate direction and the target is placed on the third direction pointing away from

the observer. This means the centers of the left and right eyes are

$$\begin{pmatrix} -a \\ 0 \\ 0 \end{pmatrix} = \mathbf{p}_l, \qquad \begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix} = \mathbf{p}_r. \tag{3.1}$$

Here, and in the following, boldface lower-case characters denote *column* vectors in Euclidean space. Define an up direction as $\mathbf{u} = (0, 1, 0)^\mathsf{T}$. We refer to the first coordinate axis as *horizontal* and the second one (i. e. along the up direction) as *vertical*. Objects of interest located at $\mathbf{z} \in \mathbb{R}^3$ displaced from the eyes mostly along the third coordinate axis, i. e. $\mathbf{z} = (\approx 0, \approx 0, z)$. Then the normalized (unit-length) vectors from eye to interest point are

$$\mathbf{e}_l = \frac{\mathbf{z} - \mathbf{p}_l}{\|\mathbf{z} - \mathbf{p}_l\|}, \qquad \mathbf{e}_r = \frac{\mathbf{z} - \mathbf{p}_r}{\|\mathbf{z} - \mathbf{p}_r\|}, \tag{3.2}$$

where the subscripts l and r refer to the left and right eye. Normalization ensures that the vectors have unit length:

$$1 = \|\mathbf{e}_{l,r}\| = \|\mathbf{e}_{l,r}\|^2 = \langle \mathbf{e}_{l,r}, \mathbf{e}_{l,r} \rangle = \mathbf{e}_{l,r}^\mathsf{T} \mathbf{e}_{l,r}. \tag{3.3}$$

Note that latter notation for the inner product follows from $\mathbf{e}_{l,r}$ being a column vector and the usual conventions for matrix multiplication; we use this notation in what follows.

Given only the positions of the eyes and unit gaze vectors, we want to compute the point closest to both eye rays. One way to do so is measuring the squared distances to the rays and finding the point that minimizes them. Let's first consider a ray through the origin. We define it by specifying a unit direction vector $\mathbf{v} \in \mathbb{R}^3$, $\|\mathbf{v}\| = \mathbf{v}^\mathsf{T}\mathbf{v} = 1$. Then the points on the ray in the direction $\mathbf{v}$ are given by $\lambda\mathbf{v}$, where $\lambda \in \mathbb{R}$ is a scalar parameterizing the ray.

Consider the (symmetric) matrix $\mathbf{V} = \mathbf{I} - \mathbf{v}\mathbf{v}^\mathsf{T}$ Here $\mathbf{I}$ denotes the 3×3 identity matrix (we generally use uppercase bold-face letter for matrices) and $\mathbf{v}\mathbf{v}^\mathsf{T}$ is an *outer* product, following directly from the common rules for matrix multiplication:

$$\mathbf{V} = \mathbf{I} - \begin{pmatrix} v_0 \\ v_1 \\ v_2 \end{pmatrix} (v_0, v_1, v_2) = \begin{pmatrix} 1 - v_0^2 & -v_0 v_1 & -v_0 v_2 \\ -v_0 v_1 & 1 - v_1^2 & -v_1 v_2 \\ -v_0 v_2 & -v_1 v_2 & 1 - v_2^2 \end{pmatrix}. \tag{3.4}$$

Multiplication of this matrix with any point $\lambda\mathbf{v}$ on the ray yields

$$\mathbf{V}\lambda\mathbf{v} = \left(\mathbf{I} - \mathbf{v}\mathbf{v}^\mathsf{T}\right)\lambda\mathbf{v} = \lambda\left(\mathbf{v} - \mathbf{v}\,\mathbf{v}^\mathsf{T}\mathbf{v}\right) = \lambda\left(\mathbf{v} - \mathbf{v}\right) = 0, \tag{3.5}$$

while multiplying with a vector $\mathbf{w} \in \mathbb{R}^3$ of arbitrary length but orthogonal to $\mathbf{v}$ ($\mathbf{v}^\mathsf{T}\mathbf{w} = 0$),

yields

$$\mathbf{V}\mathbf{w} = (\mathbf{I} - \mathbf{v}\mathbf{v}^\mathsf{T})\mathbf{w} = \mathbf{w} - \mathbf{v}\,\mathbf{v}^\mathsf{T}\mathbf{w} = \mathbf{w}. \tag{3.6}$$

So the matrix $\mathbf{V}$ annihilates components in direction $\mathbf{v}$ and leaves directions orthogonal to $\mathbf{v}$ unchanged. It is commonly called a *projector* for the direction $\mathbf{v}$. Similarly define the projectors for the gaze vectors $\mathbf{E}_l$, $\mathbf{E}_r$.

Multiplying a point $\mathbf{x} \in \mathbb{R}^3$ from the left and the right, i. e. $\mathbf{x}^\mathsf{T}\mathbf{V}\mathbf{x}$, results in taking the inner product of the component orthogonal to the vector or, in other words, measuring the squared distance of $\mathbf{x}$ to the line along $\mathbf{v}$ through the origin. If the line is not through the origin, all we need to know is a point $\mathbf{p}$ on the line. Then we translate everything so that $\mathbf{p}$ is in the origin, meaning we get the squared distance of $\mathbf{x}$ to the line along $\mathbf{v}$ through $\mathbf{p}$ as

$$(\mathbf{x} - \mathbf{p})^\mathsf{T}\mathbf{V}(\mathbf{x} - \mathbf{p}). \tag{3.7}$$

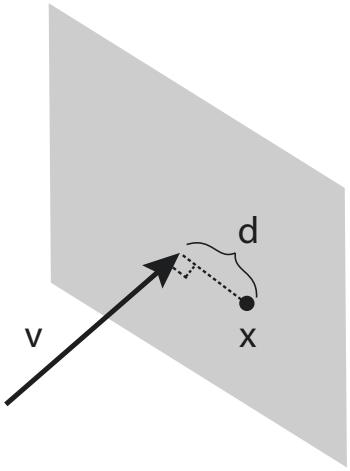


Figure 3.2: Distance vector d from point x to vector v lies in the plane that is perpendicular to v.

With this way of measuring the distance to a ray, the sum of the squared distances to the eye rays for any point $\mathbf{x}$ in space can be written as:

$$d^2(\mathbf{x}) = (\mathbf{x} - \mathbf{p}_l)^\mathsf{T}\mathbf{E}_l(\mathbf{x} - \mathbf{p}_l) + (\mathbf{x} - \mathbf{p}_r)^\mathsf{T}\mathbf{E}_r(\mathbf{x} - \mathbf{p}_r) \tag{3.8}$$

To find the point in space that minimizes this sum of squared distances compute the gradient of this function (with respect to $\mathbf{x}$)

$$\nabla d^2 = 2\mathbf{E}_l(\mathbf{x} - \mathbf{p}_l) + 2\mathbf{E}_r(\mathbf{x} - \mathbf{p}_r) \tag{3.9}$$

and set it to zero:

$$(\mathbf{E}_l + \mathbf{E}_r)\mathbf{x} = \mathbf{E}_l\mathbf{p}_l + \mathbf{E}_r\mathbf{p}_r = (\mathbf{E}_r - \mathbf{E}_l)\,\mathbf{p}_r. \tag{3.10}$$

In this way the point of interest $\mathbf{x}$ is defined as the solution of a 3×3 linear system. The system has a unique solution as long as the sum $\mathbf{E}_l + \mathbf{E}_r$ is non-singular. Each of the two matrices $\mathbf{E}_{l,r}$ has a one-dimensional kernel: the ray direction $\mathbf{e}_{l,r}$ is an eigenvector with zero eigenvalue. If the two gaze vectors are parallel, then the projectors are identical and $\mathbf{E}_l + \mathbf{E}_r = 2\mathbf{E}_l = 2\mathbf{E}_r$ is singular. This is quite intuitive, as there is no unique point with smallest distance to two parallel lines.

If the eye rays are not parallel, however, the sum $\mathbf{E}_l + \mathbf{E}_r$ is non-singular. This is also geometrically intuitive as there is a unique point minimizing the squared distances to the two lines; and this fact can be proven rigorously[255] Corollary 2.5. Thus, the point of interest is defined as

$$\mathbf{e}_l \neq \lambda \mathbf{e}_r, \lambda \in \mathbb{R} \quad \Longrightarrow \quad \mathbf{x} = (\mathbf{E}_l + \mathbf{E}_r)^{-1}(\mathbf{E}_r - \mathbf{E}_l)\mathbf{p}_r. \tag{3.11}$$

3.2.1 Ray Errors

Firstly, we introduce *variable error (imprecision, noise)* $_l$, $_r$ into the eye rays, with separate horizontal (η) and vertical (ν) noise as well as separate noise for left and right eyes:

$$\varepsilon_l = \begin{pmatrix} \eta_l \\ \nu_l \\ 0 \end{pmatrix}, \quad \varepsilon_r = \begin{pmatrix} \eta_r \\ \nu_r \\ 0 \end{pmatrix}. \tag{3.12}$$

While errors are usually represented in terms of angular deviation (i. e. radians), for small enough values the linear approximation $\sin\varphi \approx \varphi$ is very good and adding the error vectors to the eye ray vectors has the same effect as rotating the eye rays. Including renormalization this yields:

$$\mathbf{e}_l' = \frac{\mathbf{e}_l + \varepsilon_l}{\|\mathbf{e}_l + \varepsilon_l\|}, \quad \mathbf{e}_r' = \frac{\mathbf{e}_r + \varepsilon_r}{\|\mathbf{e}_r + \varepsilon_r\|}. \tag{3.13}$$

With this model we simulate how noise affects the computation of the vergence point. Following the results in[200] (which show that noise in eye trackers is mostly Gaussian distributed), we use a zero-mean Gaussian distribution, i. e.

$$p(\varepsilon_{l,r}) = |2\pi\Sigma_{l,r}|_+^{-1/2} \exp\left(-\frac{1}{2}\varepsilon_{l,r}^\top \Sigma_{l,r}^{-1}\varepsilon_{l,r}\right). \tag{3.14}$$

with the horizontal and vertical deviations being uncorrelated (meaning noise distribution in each direction varies independently)

$$\Sigma_{l,r} = \begin{pmatrix} \sigma^2_{\eta_{l,r}} & 0 & 0 \\ 0 & \sigma^2_{\nu_{l,r}} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.15}$$

In this setup, error vectors can be generated by simply drawing the components $\eta_{l,r}, \nu_{l,r}$ independently from univariate normal (Gaussian) distributions with zero mean and standard deviation $\sigma_{\eta_{l,r}}, \sigma_{\nu_{l,r}}$.

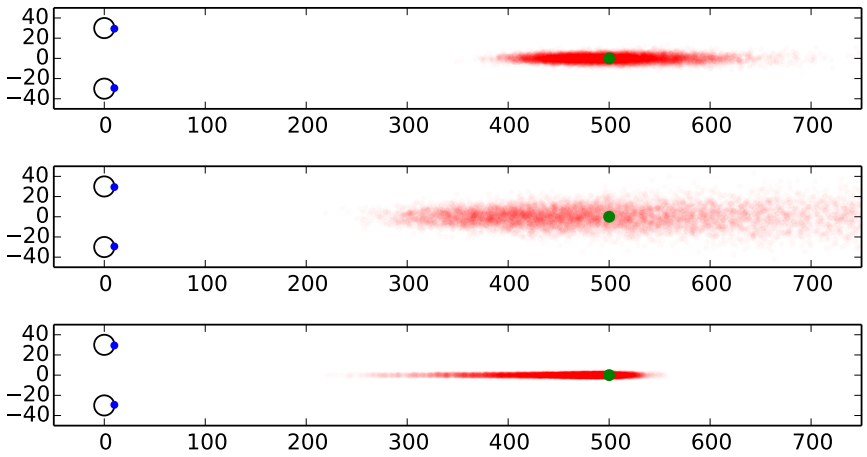


Figure 3.3: Vergence points from eye rays distorted with zero-mean normally distributed errors (i.e. noise). Distance between two eyes is 6 cm and the object is at the distance of 50cm. Top row: the noise distribution has standard deviation of 1.5 degrees in both horizontal and vertical direction. Despite the relatively small error, the variation in depth is quite large. Middle row: standard deviation in vertical direction is only 0.16, but in horizontal direction it is 1.5. The noise in horizontal direction gets larger and the mean estimated point of vergence is shifted towards larger depths. Bottom row: standard deviation is 1.5 degrees in vertical direction and 0.16 in horizontal direction. Horizontal noise is small, but the mean estimated point of vergence shifts significantly towards the viewer. Same number of samples is drawn in each simulation. 10,000 points are current shown in the figure.

A point of interest $\mathbf{z}$ defines the unbiased eye rays $\mathbf{e}_{l,r}$, which align with the lines of sight. To generate points of vergence in the presence of precision error, we draw error vectors $\varepsilon_{l,r}$, modify the eye rays accordingly and reconstruct the vergence point using the linear system above. It outputs the mean vergence point and offers graphical output, such as the one shown in Figure 3.3. We chose two standard deviations of 0.16° and 1.5° for the Gaussian distributions (which are motivated by experimental data described later). 0.16° corresponds to the average precision of fixations and 1.5° corresponds to difficult eye tracking situations for example when observers wear glasses. The simulation shows a large error in direction of depth – this is quite intuitive given the short base line relative to the distance to the object.

More severely, *the mean vergence point appears to be biased* systematically towards or away from the observer. This is clearly visible when the noise distribution has different variance in horizontal and vertical directions as shown in Figure 3: if variance is larger in the horizontal direction, the mean vergence points shifts away from the viewer; if variance is larger in the vertical direction, the mean vergence point shifts towards the viewer. Table 3.1 shows detailed simulation results when the distance between observer and the fixation target is considered as another variable. We perform the simulation when the fixation target is placed at three different distances, namely 50 cm, 70 cm and 110 cm. The further away the target is, the larger

distance	σ_x	σ_y	$\bar{p}_z$	$\bar{error}$	$\bar{std}$	$\bar{error}_x$	$\bar{error}_y$	$\bar{error}_z$	$\bar{std}_x$	$\bar{std}_y$	$\bar{std}_z$
50	1.5	1.5	38.6	14.0	8.1	0.6	0.6	13.9	0.5	0.5	8.2
50	1.5	0.16	41.3	12.8	8.6	0.6	0.6	12.8	0.5	0.6	8.7
50	0.16	1.5	36.8	13.2	2.5	0.06	0.5	13.2	0.04	0.4	2.6
70	1.5	1.5	49.5	25.4	17.1	0.7	0.7	25.4	07	0.7	17.1
70	1.5	0.16	56.1	23.9	28.6	0.8	0.09	23.9	0.9	0.1	28.7
70	0.16	1.5	45.8	24.2	4.7	0.1	0.7	24.2	0.05	0.5	4.7
110	1.5	1.5	63.5	54.7	44.8	1.0	1.0	54.6	1.1	1.1	44.8
110	1.5	0.16	82.9	58.4	101.4	1.3	0.1	58.4	2.1	0.2	101.4
110	0.16	1.5	59.0	51.0	9.1	0.09	0.9	51.0	0.07	0.7	9.1

Table 3.1: Simulation errors with target placed at three different distance 50 cm, 70 cm and 110 cm. σ_x and σ_y represents the standard deviation in horizontal and vertical directions measured in radians. $\bar{p}_z$ is the mean position of estimated vergence points in depth. Averaged distance errors and standard deviations are reported in cm with x represents the horizontal direction, y the vertical direction, and z in depth. Notice that the small baseline of distance between the eyes leads to large errors when the target point is further away. On average, the errors of the estimations are 27% (50 cm), 35% (70 cm), and 49% (110 cm) in percentage of the corresponding distances.

the bias is. Note that this bias is non-linear due to the underlining projective relation. Within each condition, the bias is mainly in depth but large variation in horizontal direction leads to large bias while the same amount of variation in vertical direction corresponds to smaller bias.

Secondly, although it is not the focus of this chapter, our noise formulation could also be used to investigate the effect of *systematic errors (inaccuracy, offset)* by introducing constant offset for the eye positions in (3.12). As shown in the simulation results in Figure 3.4, unsurprisingly, the resulting errors, i.e. the distances between the estimated mean vergence point and the true target point, are larger when inaccuracy is introduced compared to when only noise is added to data. Furthermore, systematic offsets in the horizontal direction shift the mean of the estimation no matter whether the offsets in the two eyes converge or diverge. Meanwhile, vertical systematic offsets in the same direction lead to larger estimation errors without shifting the distribution mean much. Again, the estimated mean is biased towards the observer when systematic offsets in opposite vertical directions are introduced, similar to the bias we observed when only noise was introduced where the estimation of the mean vergence point is always closer to the observer than the true target point.

3.2.2 Qualitative Analytical Analysis of Bias

In the section above (and examples shown in Figure 3.3), we use a simple numerical simulation to visualize the distribution of computed points of interest given an assumption on the probability distribution of the noise. This simulation suggested that the mean vergence point

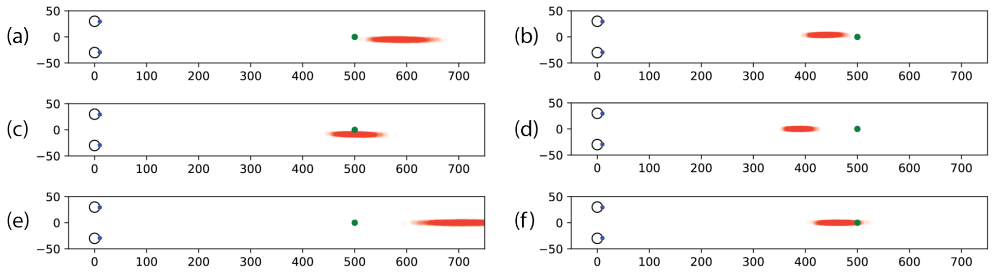


Figure 3.4: Effects of systematic offsets (inaccuracy) on the distribution of the vergence estimations. The distance between the two eyes is 6 cm and the object is at a distance of 50 cm. Each eye ray is distorted with zero-mean Gaussian noise with standard deviation of 0.16 degree. In the plots, we see the effects when a systematic offset (inaccuracy) of 1 degree is applied, not an uncommon effect size from a small change in pupil dilation. (a) Left eye is horizontally rotated by one degree outwards. (b) Left eye is horizontally rotated towards the right eye direction by one degree. Symmetrical distributions are obtained when we apply a systematic offset to the right eye only. (c) Both left and right eyes are rotated in the same direction by one degree. Figures (d) and (e) show the distribution of vergence points when both eyes are systematically rotated inwards and outwards by one degree respectively. Systematic offsets in vertical direction lead to larger estimation errors without obvious shifts of the distribution mean, except when the offsets in opposite vertical directions are applied as shown in (f). Similar to the previous simulation of variable error (noise), vertical offsets lead to a bias towards the observer, i.e. the estimated mean vergence point is closer to the observer than the target.

is biased, depending on the standard deviation of the noise distribution in horizontal vs. vertical directions. We wish to analyze this behavior analytically. For this we attempt to compute the mean (or *expected*) vergence point for a given probability distribution. As we will see, the general problem is difficult to approach. Yet, by assuming the geometric situation exhibits symmetry (see App. A for details) we are able to show that the trend we have observed in the numerical simulation holds qualitatively for wide classes of practically relevant scenarios.

The expected value for a discrete error distribution would be the sum of values multiplied with the respective probabilities for the input parameters. In the continuous case, this sum turns into an integral over the product of the computed value and the probability density distribution for the input parameters.

The projectors $\mathbf{E}'_{l,r}(\varepsilon_{l,r})$ are generated from the noisy eye rays $\mathbf{e}'_{l,r}$. Assuming a probability distribution $p_{l,r}(\varepsilon_{l,r})$ the expected intersection point is:

$$\mathbf{x} = \iint (\mathbf{E}'_l(\varepsilon_l) + \mathbf{E}'_r(\varepsilon_r))^{-1}(\mathbf{E}'_l(\varepsilon_l)\mathbf{p}_l + \mathbf{E}'_r(\varepsilon_r)\mathbf{p}_r) \\ p_l(\varepsilon_l)p_r(\varepsilon_r)\, d\varepsilon_l\, d\varepsilon_r \tag{3.16}$$

This integral, in general, cannot be treated analytically. However, note that the resulting mean is a linear function of the individual points of vergence. This means we can generate

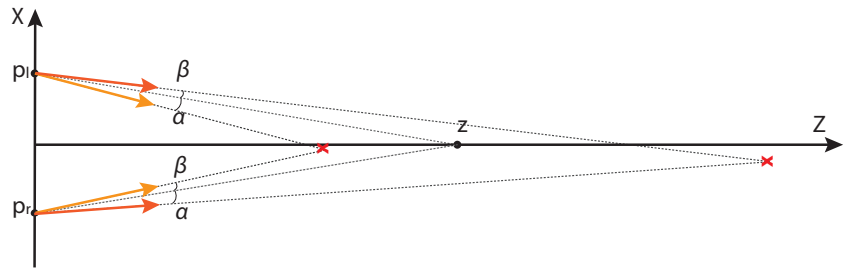


Figure 3.5: Illustration of two intersection points with inverted pair of errors α, β. Inverted pair of errors leads to an intersection point in the opposite directions; however, the mean vergence point is biased to be further away from the target point $\mathbf{z}$. In this illustration, $\alpha = 5^\circ$ and $\beta = -3^\circ$, where positive angle corresponds to the clockwise direction.

the mean of several instances of the distribution and then integrate over these local means. We will make use of this property in the following.

From the perspective of applications, we are mostly interested in understanding the bias with respect to different distances in depth. This allows us to concentrate on geometric configurations that exhibit symmetry. First, we assume the fixated object is on the symmetry line and at unit distance, i. e. $\mathbf{z} = (0, 0, 1)^\mathsf{T}$. Note that by setting the interocular distance $\mathbf{p}_r - \mathbf{p}_l = 2a$ still considers arbitrary distances because geometrically it makes no difference if we change the distance of the target or the distance between the eyes. Second, we assume the probability distributions for left and right eyes are identical. We denote this by dropping the subscript: $p = p_l = p_r$. Third, we may additionally assume the probability distributions are symmetric around the origin. Point symmetry of the distribution means the probabilities of the variable errors $+\varepsilon$ and $-\varepsilon$ are the same:$p(+\varepsilon) = p(-\varepsilon)$. This would allow us to consider pairs of points based on the error vectors $\varepsilon_l, \varepsilon_r$ and $-\varepsilon_r, -\varepsilon_l$, which are symmetric around the line $(0, 0, 1)^\mathsf{T}$ (i. e. one error vector is the reflection of the other by the line). Since their probabilities are the same, their mean is on this line. Pairing all instances of variable errors in this way shows that the mean would be on the symmetry line, *for all probability distributions with point symmetry*.

Consider the pair of error vectors $\varepsilon_l, \varepsilon_r$ *and* the reversed pair $\varepsilon_r, \varepsilon_l$. By our assumptions, these instances have the same probability. The two pairs give rise to two points of interest. Figure 3.5 illustrates the case of horizontal error only. In this case, one of the two intersection points is closer to the observer then the target, while the other one further away. Their mean will never be closer then the real target, indicating that the mean over all instances with horizontal noise only will be biased to be further away than the target – as observed in the numerical simulations. We now prove that this intuitive reasoning is correct.

To compute the expected depth value, we start by analytically solving the least squares estimation (LSE) in equation (3.10) for the last component. This can be done by elementary

computations. In this way we can express the depth for a certain pair of error vectors ε_l, ε_r as well as for the reversed pair ε_r, ε_l. Then we take the mean of the two depth values. All computations can be carried out by a computer algebra system such as Maple or Mathematica. The resulting expression is lengthy and unsuitable for direct inspection. However, considering the case of horizontal noise only leads to the following simple expression:

$$\frac{z'}{z} = \frac{4a^2}{4a^2 - (\eta_l - \eta_r)^2} \tag{3.17}$$

As defined in equation (3.1) and equation (3.12), $2a$ equals to the interocular distance and $\eta_{l/r}$ represents the horizontal component of variable error $epsilon_{l/r}$. It makes sense to assume that the distance between the displacements $\eta_{l/r}$ is smaller than the interocular distance $2a$ (otherwise the intersection of the eye rays would be behind the observer). In this case the denominator is positive but never larger than the numerator, so the depth is larger. Because this is true for any pair of instances (see more in App. A), this is also true for the mean resulting from an arbitrary probability distribution (as long as they are they same for both eyes). This confirms our intuitive geometric conclusion from Figure 3.5.

Next we consider only vertical error $\nu_{l/r}$. In this case it is convenient to consider the pair of error vectors ε_l, ε_r and $-\varepsilon_r$, $-\varepsilon_l$. Note that compared to the former case we now also reverse the signs, which is admissible if the probability distribution is symmetric. Using computer algebra as before, we find that the relative mean depth for the two intersection points is

$$\frac{z'}{z} = \frac{4a^2 + a^2(\nu_l + \nu_r)^2}{(\nu_l - \nu_r)^2 + 4a^2 + a^2(\nu_l + \nu_r)^2} \tag{3.18}$$

Comparing numerator and denominator reveals that $z' \leq z$ with equality only for $\nu_l = \nu_r$ regardless of the probability distribution and without any restrictions on the vertical errors.

Together, these two results confirm our observations that errors in horizontal direction bias the mean depth towards larger values, while errors in vertical direction bias it towards the observer. This result holds for all probability distributions, as long as they are identical for both eyes (and exhibit symmetry in the vertical direction).

In any realistic scenario, however, noise errors have horizontal and vertical components. The resulting bias will depend on the relative magnitude of these errors. If the magnitude of horizontal and vertical noise error is equal, then the bias is toward smaller depth and the magnitude of this effect is dominated by the squared difference $(\nu_l - \nu_r)^2$. In general, however, the variance along the horizontal versus vertical axes in the probability distribution determines the bias in depth. If the variance is higher in horizontal direction, the depth will be biased towards greater values; if the variance is higher in vertical direction, depth will be biased closer to the viewer. As before, this result holds for probability distributions as long as they are identical for both eyes and the point closest to any two eye rays in the distributions is in front

of the viewer (not behind).

3.3 Part 2: Human Data

The numerical simulations above show that variable and systematic errors in eye rays have a significant influence on the estimated mean point of vergence under projection. In this section, we collect human vergence data to study the real noise distributions when using a video- based eye-tracker. Eye-trackers primarily output mapped gaze positions, which in 2D are points on a screen, described in pixel coordinates. This mapping is established using a calibration routine, where participants are asked to look at several targets on the screen while features of their eyes are tracked in the camera image. These pre-calibrated features - the pupil centre and the centre of the reflection in the cornea - are mapped to gaze positions on the monitor during calibration. With an established mapping, any eye positions can be mapped into the target space, which corresponds to the gaze point.

Several sources of errors are known to interfere in this procedure, in particular when both eyes are being calibrated. Despite existing research on vergence eye movements[41,42,66], there is no established consensus on how to calibrate binocularly. Nuthmann & Kliegl[203] have brought up the question whether binocular calibration (i. e. calibrating both eyes at the same time) or monocular calibration (i. e. calibrate one eye at one time while the view of the other eye is occluded) is better suited for binocular eye movements study. Besides the intrinsic properties of eye movements in binocular viewing, the estimated mapping function in calibration also introduces errors in the estimated point of interest. The estimation of parameters in the mapping function is a minimization procedure. In practice, low order polynomials are often used to map pupil and corneal reflection positions onto screen coordinates, and the mapping parameters are approximated through an optimization procedure (e. g. least squares), which inevitably contains modeling errors.

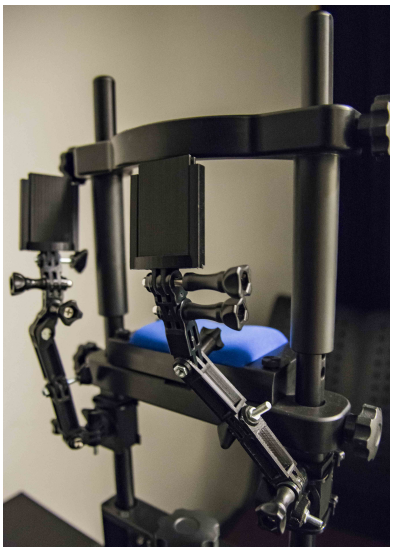
Therefore, we designed a data collection including both binocular and monocular viewing conditions and analyzed the detected raw pupil and corneal reflection positions without mapping them to the target space. Two depth variations were included and we used symmetric presentation of stimuli to counterbalance any potential behavioral differences due to spatial dependency.

3.3.1 Participants

25 participants from TU Berlin (students and staff) joined our experiment (mean age = 27, SD = 7, 4 females). They all had normal or corrected to normal vision and provided informed consent. Five of them had previous experience with eye tracking experiments. Their time was compensated. Additionally we performed a monocular eye examination for each participant

and the averaged acuity is 0.93 (SD=0.26) measured in the decimal system. We also measured the eye dominance of observers following the standard sighting eye dominance test, as our experiment does not involve any interocular conflict. Observers were asked to look at a distant point though a small hole placed at arm length, and then to close their eye one after another. Only the dominant eye supposes to see the point while viewing monocularly. Among all 25 observers, only 7 of them have a left dominant eye while 18 of them reported to have a right dominant eye. We discuss on the eye dominance test in the discussion section.

3.3.2 Apparatus and Recording Setup



The data collection was conducted in a quiet and dark room. We used an EyeLink 1000 desktop mount system in the experiment and binocular eye movements were tracked at a sampling rate of 1000 Hz. A chin- forehead rest was used to stabilize observers' head positions. A 24-inch display (0.52 m $\times$0.32 m, 1920 $\times$ 1200 pixels) was placed at two distances of 0.7 m and 1.1 m from the eyes. Stimulus presentation was controlled using PyLink provided by SR Research. Note that any eye trackers can be used for the recordings as our analysis is based on detected eye positions in the camera frames.

Two custom-built eye covers fabricated by 3D printing were mounted on the chin-forehead rest (see Figure above). Each of them can be rotated by 180° to open or block the view of one eye.

3.3.3 Design and Procedure

In order to familiarize observers with fixational eye movements, and to verify the setup that observers' eyes were clearly tracked, we ran a default 9-point calibration routine before the data collection started. A calibration is followed by a validation to verify its accuracy. The experiment was continued only when the achieved validation accuracy is on average smaller than 1.0° of visual angle. Otherwise we repeated the calibration and validation routine either immediately or after a short break. However, the resulting gaze data are not used. All further analyses are purely based on the pre-calibrated pupil minus CR (corneal reflection) data, to minimize influence from the calibration mathematics of the EyeLink.

The data collection consisted of two distances and each distance had five repetitions. In each repetition, we presented three viewing conditions, namely monocular viewing with left

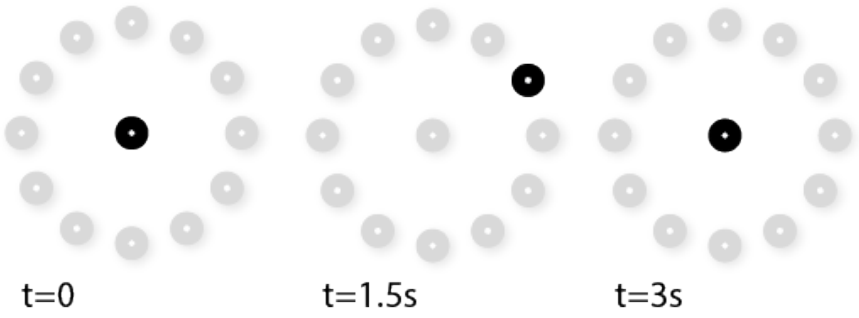


Figure 3.6: An exemplary trial. The center was always presented at the beginning of each trial. A randomly selected marker on the ring was then presented for 1.5s followed by the presentation of the center marker for another 1.5s. Then another random selected marker on the ring was presented. The presentation continues in such way until all markers on the ring were presented once. Gray markers are drawn here for illustration purpose and were invisible during experiment.

eye, monocular viewing with right eye, and binocular viewing. Eye covers were rotated by the instructor to occlude/reveal the eyes in different viewing conditions. Meanwhile, the EyeLink was continuously tracking in binocular mode throughout.

For each distance, repetition and condition, the participant was asked to fixate each one out of 12 targets in the form of a ring with an addition central point (Figure 3.6). Participants were instructed to follow the marker and fixate the white dot at the center as accurately as possible. Black circles with their center marked by a white dot were used as the target. The radius of the ring corresponded to 7° of visual angle. Each marker was presented for 1.5 s and a beep sound was used to signal the start. Between each fixation trial, participants refixated the center point. In total, there were 12 targets fixations on the periphery targets and 12 at the center target.

Each block of five repetitions for a distance took about 20 minutes and participants were asked to take a break before the second block. In total it lasted about one hour including a simple trial run at the beginning.

3.3.4 Analysis Methods and Results

Our analysis was based on raw samples of detected eye positions, i. e., the pupil minus CR positions, represented as pixel coordinates in the eye camera. The only use of calibrated gaze was our use of the built-in velocity- based algorithm of the EyeLink software to detect fixations. In order to find fixations on targets, we filtered out all short fixations that are less than 100ms in a preprocessing step as short fixations very often are the result of undershoots (Holmqvist & Andersson [105], p.222-223) and are quickly followed by a correction saccade onto targets. Conservatively, among fixations on the same target, we considered those that are two

standard deviations away from the cluster center as outliers. The remaining fixations were from then on only processed as uncalibrated pupil minus CR.

For each dataset of one observer, we aggregated all fixations (pupil-CR) in the five repetitions for each target in the same viewing condition. On average there are 56 fixations in each such block (SD = 20), and each target has 2.3 fixations as we have 24 fixation targets in each repetition. Here fixations from individual eye are considered independently. The averaged duration of fixations is 821ms (SD = 494ms). There is no significant difference among different viewing conditions at two distances.

3.3.5 Comparing Sets of Covariance Matrices

In the previous simulation (see Section "Ray Errors"), fixation positions have independent zero-mean Gaussian distributions for variable errors in horizontal and vertical directions, which can be represented by a 2×2 diagonal covariance matrix Σ. Diagonal elements in Σ describe the variances in horizontal and vertical directions separately and off-diagonal elements show the correlation between them. Therefore, Σ is a positive semi-definite symmetric matrix, and it can be visualized as an error ellipse with its axes pointing into the directions of its eigenvectors. The lengths of the semi-axes are proportional to the square roots of the corresponding eigenvalues λ_i. We used $\sqrt{5.991}\sqrt{\lambda_i}$ as the semi-axis length (derived from a Chi-Square distribution), which gave us the ellipse that covers a 95% confidence interval. We continued using covariance matrices to analyze the error distributions and visualized them as ellipses in the following. Essentially a 2D covariance matrix measures the precision in two dimensions, as we are interested in its spatial distribution.

In our data collection, we had a set of markers on screen and a covariance matrix represented fixation distribution at each marker position. To model the variability among different covariance matrices and to compare the differences among sets of covariance matrices, we computed distances between all pairs of covariance matrices in a set and then compared the resulting distributions. Despite the raising amount of applications of analyzing the variance among covariance matrices, there is still no consensus on how to analyze the covariance structures[221]. We settled on a logarithm-based distance estimator[72], a Riemannian metric, defined as

$$d(\Sigma_1, \Sigma_2) = \left\| \log\left(\Sigma_1^{-\frac{1}{2}} \Sigma_2 \Sigma_1^{-\frac{1}{2}}\right) \right\|, \tag{3.19}$$

where the logarithm is given by $\log \Sigma = \mathbf{U} \log(\mathbf{S}) \mathbf{V}$ and $\mathbf{U}, \mathbf{S}, \mathbf{V}$ can be factorized from singular-value decomposition (SVD) as $\Sigma = \mathbf{USV}$. $\mathbf{S}$ is a diagonal matrix of singular values of Σ. $\|\mathbf{X}\|$ is the Euclidean norm (also called Frobenius norm) and it can be computed by the matrix trace $\|\mathbf{X}\| = \sqrt{\mathrm{Tr}(\mathbf{X}^T\mathbf{X})}$. In case of covariance matrix, the trace measures the total variation in each dimension without considering the correlations among variables.

As shown in [72], this distance measure is symmetric and non-negative, and it is invariant under affine transformation and inversion. Intuitively speaking, by multiplying with $\Sigma_1^{-\frac{1}{2}}$ in bilinear form, we transform Σ_2 into a new basis and $\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}$ is a perfect circle. The distance measures the relative ratio of eigenvalues in the new basis and the largest eigenvalue of $\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}$ corresponds to the ratio of maximum variance between two groups [221]. To aggregate a set of covariance matrices, we used the arithmetic average as the mean covariance matrix. Dryden et al. [59] compared many covariance distance measures especially in the context of shape interpolation; however, it is not clear which one suits best in our case. Here we only want to be able to compare sets of covariance matrices.

In the next step, the distance distributions of two sets of covariance matrices were compared using the Kolmogorov-Smirnov test (KS-test) [228] and a p-value was computed to determine whether the two distributions differ significantly. The KS-test computes the vertical distance between two cumulative fraction functions that are used to represent two distributions and takes the largest distance as the statistic. Therefore, it is robust with respect to variance types of distributions.

3.3.6 Results

Below, we examined whether the bias was present also in the human data. We first compared the distributions of variable errors among all individual observers. Then we aggregated all datasets to examine whether the distributions has any spatial dependency. In the last step, we used the averaged variable error distribution to sample eye positions following the procedure in the previous simulation (see Section Ray Error) to investigate whether there is a bias in the direction of the observer in human data as there was in simulated data.

Eye Dominance and Acuity do not Seem to Matter

Each dataset (of one observer) had five repetitions of each viewing condition. Each eye had two viewing conditions, namely monocular viewing and binocular viewing. Targets were presented at two different distances. We first aligned the five repetitions in the same condition and computed the covariance matrix of fixations at each marker position. Following equation (3.19), each covariance matrix was compared to the other eleven covariance matrices. Note that we discarded all fixations at the center marker. One exemplar dataset is shown in Figure 3.7. Covariance matrices are visualized as ellipses and distributions of all pair-wise distances are plotted as histograms.

Using the KS-test, we compared histograms of the distance distributions of each individual dataset to the others and results are shown in Figure 3.8. Red line marks the significance level of 0.05. Note that any statistic value *larger* than the level (i. e. above the line) corresponds

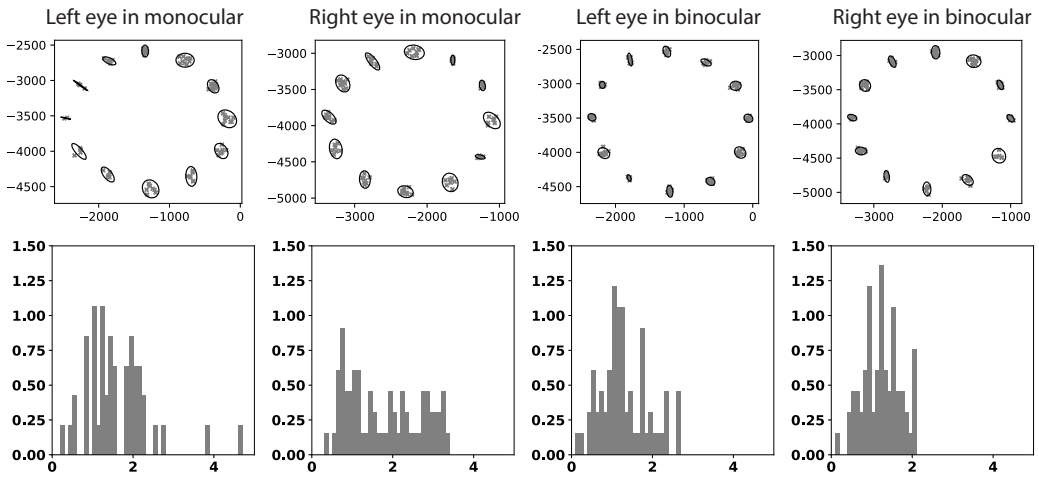


Figure 3.7: Example of one-dataset samples at the distance of 70 cm. In the first row, fixations at each marker position collected over five repeats are scattered and corresponding covariance matrices are visualized as ellipses. Second row shows the histograms of distances between all pairs of ellipses in each condition (summed over five repeats). Observer of this dataset has right dominant eye and eye acuity is 0.9 for left eye and 1.0 for right eye.

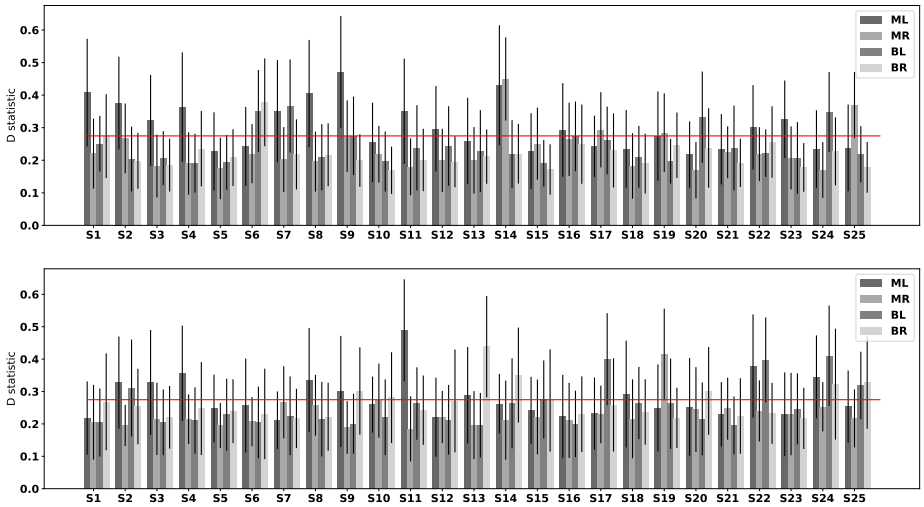


Figure 3.8: Statistics of KS-test on individual differences. Upper plot shows the statistics of data with a display placed at a distance of 70 cm and the lower one shows the result when the display was at a distance of 110 cm. Red lines mark the 0.05 significance level. ML stands for the condition of monocular viewing of left eye, MR monocular viewing of right eye, BL binocular viewing of left eye and BR binocular viewing of right eye.

to a significantly different distribution. Considering each eye separately, we had four viewing conditions in one dataset, namely monocular viewing of left eye (ML), monocular viewing of right eye (MR), binocular viewing of left eye (BL) and binocular viewing of right eye (BR). At the distance of 70 cm, 13 out of 25 datasets have significantly different fixation distributions in ML, 4 in MR, 5 in BL and 1 in BR. According to the categorization of eye dominance, we have 8 significantly different distributions of dominant eye in monocular viewing and 12 significantly different distributions of non-dominant eye. In binocular viewing condition, 2 distributions of dominant eye significantly differ from the others and 6 of non-dominant eye. At the distance of 110 cm, the numbers of datasets, which are significantly different from the others, are 10 (ML), 1 (MR), 6 (BL), and 7 (BR). In monocular viewing, 3 and 9 datasets are significantly different for distributions of dominant eye and non-dominant eye respectively. In binocular viewing condition, the number of different datasets is 9 for dominant eye and 6 for non- dominant eye. Note that this counting is based on the mean statistic value. And we do observe large variances in each dataset as shown in Figure 3.8.

In conclusion, the differences between dominant and non-dominant eye are so small and varied that we cannot conclude that eye dominance matters. Neither did we observe any correlation between the distribution differences and eye acuity, which indicates that eye acuity does not contribute to the differences in distributions. It is likely that other factors, such as eye colour, may explain part of the noise.

Noise does not Vary Depending on Where Participants Look

It is known that noise varies across the measurement space (Holmqvist & Andersson [105], p.182). To test whether there is any such spatial dependency of noise distributions of fixations, we computed the distance distribution at each marker position by comparing each pair of covariance matrices from 25 datasets. Similarly, we applied the KS-test to compare the distance distributions. Only one distribution from monocular viewing of right eye when screen was at a distance of 70 cm is significantly different. To visualize the variance at each marker position, we computed the mean of 25 covariance matrices for each marker in one viewing condition and plotted corresponding ellipses in Figure 3.9. Note that comparing the two dataset from two different distances, covariance ellipses are similar in both sizes and orientations. But ellipses in binocular viewing condition have smaller sizes but still similar orientations. In conclusion, the amount of noise does not seem to differ between the conditions, nor between positions. There is however a tendency that vertical noise is larger than horizontal noise.

But There Is a Bias in the Human Data

After we have established that noise does not vary across eye dominance, acuity and fixation position, we can now aggregate the covariance matrices at all marker positions, which gives

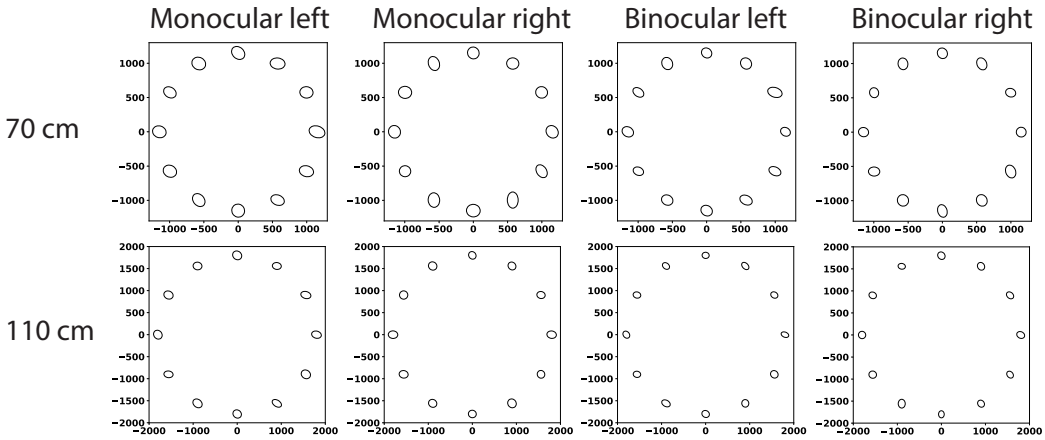


Figure 3.9: Visualization of mean covariance matrices at each marker position. First row shows error distributions of data collected at a distance of 70 cm and second row shows data collected at a distance of 110 cm. In each viewing condition at one distance, data from all observers were aggregated together and covariances at each marker position are visualized as ellipses. For instance, a left-tilted ellipse means that vertical noise is larger than horizontal noise. The size of the ellipse thus corresponds to the total variation while orientation indicates the correlation between horizontal and vertical directions.

us an overall representation of noise distribution of fixations in each condition. Using the distributions – a covariance matrix for each of the 12 fixation point in each condition - rather than the data themselves introduces no bias, but is computationally easier, since it allows us to calculate the mean easily without being biased by unevenly taken data (i. e. uneven contribution of individual observers because of fixations of differing lengths).

We used the arithmetic mean of the 12 covariance matrices (of 12 targets) to represent the fixation distribution in each viewing condition and obtained two covariance matrices, one for each eye in either monocular or binocular viewing condition at one distance. As shown in Figure 3.10, an ellipse is used to represent a covariance matrix. Errors seem to be larger when targets move from 70 cm to 110 cm with increased sizes of ellipses. Noise in binocular viewing condition is smaller comparing to that in monocular viewing condition, evidenced by smaller ellipses in most positions in all conditions. The radius of the ring corresponds to 7° of visual angle and opposite sample positions span a visual angle of 14°. Following this ratio, we converted sample units (measured in pixel) into degrees of visual angle and applied the analysis framework used in the simulation above. Variances of error distributions shown in Figure 3.10 approximate to 1° of visual angle, which is commonly used as a calibration threshold in 2D eye tracking experiment. Additionally we also experiment with 2° of visual angle, which is equivalent to the start-of-art eye tracking accuracy in 3D space[87,214,269]. Simulated results based on sampling from real error distributions are plotted in Figure 3.10. On the

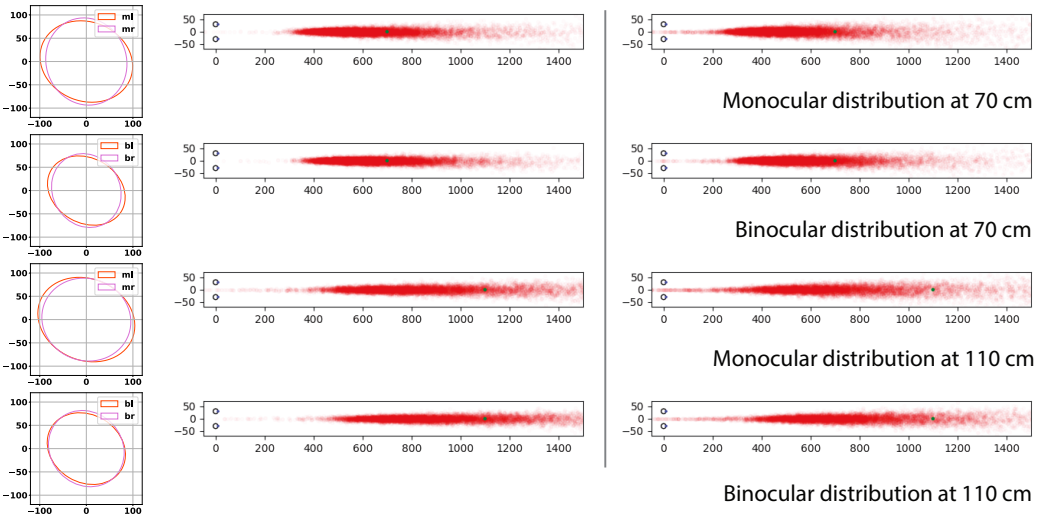


Figure 3.10: Estimation of mean vergence point based on distributions formed from real human data. Left column shows the mean covariance matrices of left eye (l) and right eye (r) in monocular viewing (m) and binocular viewing (b). First two rows show results when target is placed at a distance of 70 cm and last two rows show results at a distance of 110 cm. Right column shows the distribution of estimated vergence points in space where variance (visualized in the left column) was converted to 1° and 2° of visual angle respectively. See Table 3.2 for more detailed statistics.

left side we have spatial distributions of vergence points when variance was converted to 1° of visual angle and on the right side we see the results after converting to 2° of visual angle. Detailed statistic results are given in Table 3.2. The error in estimated mean vergence point is acceptable when everything is perfect within 1° of visual angle, however, achieving such accuracy in 3D is rather challenging. Even with an acceptable 2° accuracy, the bias towards viewer in the estimated mean point of vergence is already obvious.

3.4 Minimizing the Uncertainty in Vergence Point Estimation

When researching vergence using a video-based pupil and corneal reflection eye-tracker such as the Eyelink, what can you do to minimize biases and errors? We assume that the human participant has a negligible difference in gaze directions between left and right eye, that luminance conditions are fixed and no other effects on pupil dilation are present, and that the only remaining issue is to minimize the bias from the noise in the signal.

As this bias is an effect of the projective mapping, the non-linear mapping that is commonly used to estimate the vergence point in space, there is not much you can do if you use the calibration routine shipped with the eye-tracker. The single fixation per calibration

distance	σ	condition	$\bar{p}_z$	$\bar{error}$	std	$\bar{error}_x$	$\bar{error}_y$	$\bar{error}_z$	std_x	std_y	std_z
70	1.0	monocular	69.1	22.4	32.0	0.88	0.86	22.3	0.88	0.90	32.0
		binocular	70.4	18.9	27.8	0.76	0.74	18.8	0.79	0.77	27.8
	2.0	monocular	60.0	35.1	82.6	1.2	1.2	34.9	2.8	2.1	82.6
		binocular	64.3	28.6	53.7	1.1	1.0	28.5	1.3	1.3	53.8
110	1.0	monocular	98.5	49.9	102.3	1.2	1.1	49.8	1.5	1.7	102.3
		binocular	104.4	41.6	88.8	9.9	9.5	41.5	1.4	1.2	88.8
	2.0	monocular	73.3	70.1	141.2	1.5	1.3	70.0	2.5	2.8	141.2
		binocular	86.7	60.2	171.7	1.3	1.3	60.1	2.3	2.7	171.7

Table 3.2: Vergence errors sampled from real noise distributions. σ represents converted standard deviation in degree of visual angle. $\bar{p}_z$ represents the mean point of vergence in depth measured in cm. Averaged errors and standard deviations are reported in cm with x represents the horizontal direction, y the vertical direction, and z in depth. On average, the errors in percentage of the corresponding distance are 30% ($\sigma = 1$) and 46% ($\sigma = 2$) when target distance is 70 cm, and 41% ($\sigma = 1$) and 50% ($\sigma = 2$)when target distance is 110 cm.

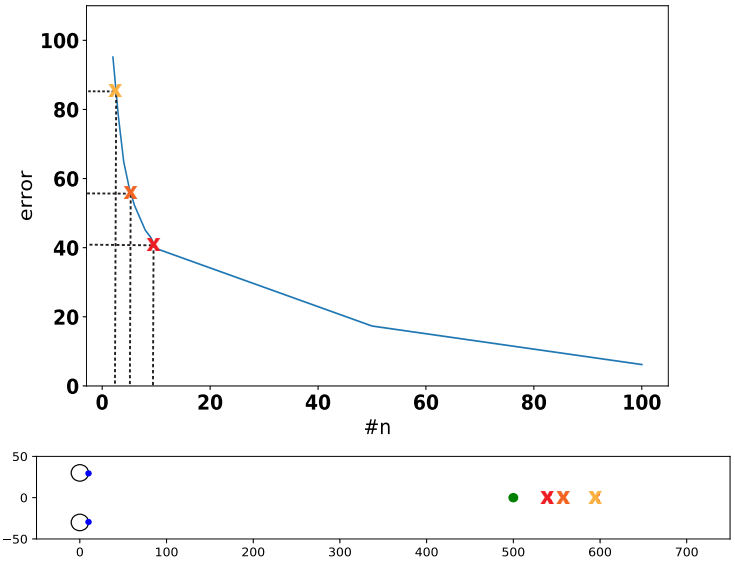


Figure 3.11: More fixations per target leads to better estimation of mean vergence point. At the top of the figure, offset error is plotted as a function of number of fixations per calibration point. Three color-coded crosses mark the offset errors when number of fixations equal to 2, 5, and 10. These three offsets are visualized in a top view plot at the bottom. We used standard deviation of 1.5 in horizontal direction and 0.16 in vertical direction.

point will have an inaccuracy and noise that increases the bias. However, it is possible to instead record *multiple* fixations on each point in your own set of calibration targets. Then take the average of the data samples in the several fixations per fixation target and use that average to calibrate the eye-tracker. The key is to use many fixations per calibration point. The same principle applies to real experiment after calibration: collect many fixations if possible. For each point of interest, taking the average of fixation samples in the camera space leads to better estimation in space. In practice, this could mean a repetition of the same experimental condition, for example, where observers are asked to refixate on the same targets in a vergence study.

The reason this method works is that for geometrical reasons, it is better to average in the calibration data than in the depth data of the intersection points. Assume that noise in the eye samples are Gaussian distributed, taking the average in the calibration data leads to better approximation to the noise-free eye samples. However, due to the non-linearity of the mapping, averaging in the depth data only leads to a bias as we see before. Figure 3.11 shows for simulated data how the offset in depth decreases with an increasing number of fixations per calibration point.

This solution works on the DPI, the EyeLink and the SMI eye-trackers, which all provide access to the center point of the corneal reflection and the pupil (for EyeLink and SMI) or the 4th Purkinje (for the DPI) in the data. Note that the recipe works with raw positions of the eyes and does not depend on specific calibration model given by any eye trackers.

3.5 Discussion

The major finding in this chapter is that vergence data from eye-trackers exhibit a bias, depending on the noise level in horizontal vs. vertical directions. Using a mathematical simulation, we show that noise in gaze vectors leads to a bias in the estimated vergence point under projection. In particular noise in the vertical direction bias the estimated mean point of vergence towards the participant. We further collected human data using a high-end video-based eye-tracker and studied the noise distributions in these real data. We showed that the estimated mean point of vergence is biased also in the human data, and that the bias increases with an increased viewing distance. This applies to the estimation of point of vergence in three-dimensional space, as well as to the vergence estimation on planar surface in two-dimensional space. As long as projective model is used in the mapping, the estimated mean vergence point will be biased.

3.5.1 Reading Research and Calibration Algorithms

The bias we have found is in line with Nuthmann & Kliegl [203], who reported that the fixations during reading were almost always crossed, peaking 2.6 cm in front of the plane of text, which is very much in line with the error reported in Table 3.2. However, instead of resulting in a discussion about potential biases from the measurement technology itself, the subsequent papers instead investigated whether monocular vs. binocular calibration could cause the crossing of fixations.

Liversedge et al. [174] had found that "When the points of fixation were disparate, the lines of gaze were generally diverged (uncrossed) relative to the text (93% of fixations),..., but occasionally converged (crossed) (7% of fixations)." using a monocular calibration. Later study by Kirkby et al. [149] replicated the finding that - after monocular calibration - the majority of fixations are uncrossed, both with the EyeLink and the DPI.

Our results using the collected human data show a significant bias towards observers for both monocular and binocular calibration, corresponding to crossed fixations in reading, and we could show that the bias is a result of the geometry of the recording situation and the noise in the data. This bias must have existed also for the studies by Liversedge et al. [174] and Kirkby et al. [149], so it is surprising that they found crossing result in the opposite direction of the bias.

Švede et al. [250] argued that monocular calibration is the only physiologically correct form in the sense that it preserves the difference in gaze direction between the two eyes. It could be the case that the average gaze differences between left and right eye after monocular calibration are so large that they consume the whole bias and nevertheless can remain uncrossed.

Despite the fact that the bias in the estimated gaze positions in binocular viewing is smaller than the bias in monocular viewing condition, the ratio in depth between target position and the estimated mean vergence point is around 1.2.

3.5.2 Eye Dominance, Interpupillary Distance and Fixation Disparity

The collected human data show a large variance among individual observers, especially the fixation distributions of left eye in monocular viewing condition tend to differ from each other. However, neither eye dominance nor acuity leads to significantly different noise distribution and where participants look also does not seem to matter. Nevertheless we should be aware that eye dominance information is based on observer's self-report. Moreover, the eye dominance is determined by the so-called sighting eye dominance test. A recent study[56] brought up the question whether only one type of eye dominance exists and their results indicate disagreements among different eye dominance test methods, especially the difference between sighting eye dominance test and binocular rivalry based test. Even though our experiment does not involve any interocular conflict, the actual dominant eye during the experiment might still be different from the measured one, which might further explain the

observed no-impact findings.

It is not clear so far how fixation disparity may contribute to our observations. Liversedge et al.[174] reported that fixation disparity decreased over time during reading, and it is tightly linked to vergence eye movements[124,229,249] as well as binocular vision[123]. Additionally, vision training may improve stereo perception and eliminate fixation disparity[49]. We neither performed any fixation disparity test nor measured the binocular version in our study, although how to accurately measure fixation disparity seems to be an on-going effort[52]. Future studies should include measurements such as the near point of convergence, positive and negative fusional range, dissociated phoria at near and far, stereopsis and amplitude of accommodation. It would also be interesting to test whether there exists a correlation between fixation disparity and the bias in individual's dataset.

3.5.3 Limitations of Our Study

Independent Gaussian distributions were used for each dimension. It would be interesting to study the dependency between noise in horizontal and vertical directions and to see how a multivariate distribution influences the bias.

In our experiment targets were displayed on a flat plane without the need of focus change. But noise distribution might be different due to the dynamic changes of focus in 3D. For example we might need to take into account the changes of pupil size. How to collect enough fixation data in 3D while maintaining the same precision is another practical but challenging problem.

To minimize the bias from the signal noise, we suggest to collect many fixations per calibration target, and then use the mean fixation point to calculate the mapping. Future work should validate this proposal with real human data. Note that this proposed recipe does not account for fixation disparity, i.e. the alignment difference between the dominant eye and the weak eye.

3.5.4 Suggestion for Future Experiments

There seems to be no good reason to prefer one viewing condition to the other in calibration. In monocular calibration, each individual eye is forced to fixate on targets without the interference of binocular fusion and eye dominance. However, visual acuity is also limited in monocular viewing and possibly decreases over distances. Precision of eye movements in binocular viewing condition seems to be higher.

We believe it is important to be aware that the propagation of noise may lead to a bias in the estimated mean vergence point in space. Such bias can easily influence the observed data. It is also very important for future eye movement studies, especially for vergence studies, to provide the validation error of calibration not just as a single scalar, but also in the form of a

spatial distribution (i. e. an error ellipse). It provides a confidence level of the observed data and puts their interpretation into perspective. There is very little information on this in the literature. It is not even clear if the error distribution would be roughly similar for all types of tracking devices and experiments, or if the distribution changed with the type of experimental task. If so, resulting mean vergence depth would vary with experimental setup.

Moreover, our results suggest that, for researchers using eye-tracking devices, it is good to think about the procedure, instead of being only concerned about the data after calibration. The mathematical models behind calibration might provide additional information as being part of the experiment.

Part II

Comparing Eye Movements on Screen to Eye Movements in 3D

I shut my eyes and all the world drops dead; I lift my eyes and all is born again

Sylvia Plath

4

Measuring Visual Saliency of 3D Printed Objects

This chapter introduces our first study towards measuring visual saliency of 3D printed objects, as an effort to validate the assumption that the saliency found in flat stimuli can be related to a 3D scene. We set up an experiment that examines if visually salient features exist for genuine 3D stimuli. Using the method described in Chapter 2, we gather fixations on the presented stimuli. This data is used to validate assumptions regarding visual saliency that so far have experimentally only been analyzed for flat stimuli. We provide a way to compare fixation sequences from different subjects as well as a model for generating test sequences of fixations unrelated to the stimuli. This enables us to provide statistics suggesting that human observers agree in their fixations for the same object under similar viewing conditions – as expected based on similar results for flat stimuli. We then validate computational models for the visual saliency of 3D objects and our results indicate that popular models of mesh saliency based on center surround patterns fail to predict fixations.

4.1 Introduction

Visual saliency describes the idea that certain features in a visual stimulus stand out more than others and are more likely to attract an observer's attention. For flat stimuli such as 2D images, numerous experiments have shown that human observers are more likely to shift their gaze

to such visually salient features[23]. A common assumption in the literature is that the saliency found in flat stimuli can be related to the underlying 3D scene. Although this assumption might seem intuitive, it has not been validated experimentally.

To evaluate this assumption, we set up an experiment that examines if visually salient features exist for *genuine 3D stimuli*. In this chapter, we describe that experiment and its analysis. Specifically, we asked whether different human observers consistently fixate on similar points on the surface of a given physical object under constant surface reflectance and fixed illumination. We used 3D printed objects as stimuli and tracked the observers' gazes while they were inspecting the objects' surfaces.

From the pupil position data, we then extracted the fixations during the first few seconds of the visual inspection. Because we know the object's geometry, we can relate the observers' fixations to gaze positions on the objects. Next, we analyzed the fixation data to address two questions. First, we tested whether human observers show consistency in their fixation patterns for the same object. Such consistency would be expected if visual salient features exist for physical objects and if these features guide fixation behavior. Second, we tested whether an algorithmic model of visual saliency, known as *mesh saliency*, can accurately predict human fixations.

To test the consistency between different human observers, we propose an analysis in which the observed fixation patterns on an object are tested against sequences of fixations that are unrelated to the geometry, yet are psychophysically plausible. These test sequences were generated from fixation sequences recorded for the same subject but *using a different object*. The resulting fixation sequences are realistic because they share the underlying oculomotor characteristics, but they are unrelated to the geometry of the object for which they serve as test sequences. To compare the generated and real fixation sequences, we quantified their similarity by computing the difference in eye-ray space.

Our results show a higher amount of agreement for fixations on the same object between observers compared with generated test sequences. This indicates the existence of visually salient features on 3D objects. The result also suggests that gaze directions systematically and meaningfully vary with the external stimulus, which is a necessary precondition for further analyses of human viewing behavior for 3D stimuli.

Because the data appears to be meaningful, we used it to investigate the predictive power of *mesh saliency*[167,244,276], which is the only model with a psychophysical experiment supporting it. Mesh saliency extracts a measure of visual saliency from the local geometry of an object's mesh representation, not taking surface scattering and view-dependent phenomena into account. The method has been validated against human fixations on flat stimuli of synthetically generated images of objects[146] and against points manually selected by human subjects[62]. As pointed out in earlier work[38], the latter definition of feature points should not be mistaken for visually salient points in the sense of features that would trigger low-level human vision.

Figure 4.1: Experimental setup. Participant viewing one test object (*left*). The calibration rig (*right*) was used to establish a mapping between gaze position and the object space.

Lastly, to validate mesh saliency using fixation data for genuinely 3D stimuli, we compared the algorithmic predictions against permutations of the values across the mesh's vertices. If mesh saliency was indeed a good predictor of the fixation positions, then it would perform better than the permutations of itself. Our results show that this is not the case.

We believe that our experiment is an important first test of the assumption that theoretical concepts of human perception derived from experiments with 2D images also hold for the perception of 3D objects.

4.2 Experimental Method

For our experiment, we recruited 30 unpaid participants (eight female and 22 male), all of whom were students. Their ages ranged from 19 to 31 years (median = 25), and all had normal or corrected-to-normal vision (based on self-reports). Four participants had previously participated in experiments involving eye tracking.

Stimuli and Apparatus The experiment was conducted in a quiet room. The stimuli were presented inside a 1 m × 1 m × 1 m box that was placed on a table and covered in white fabric (see Figure 4.1). The box was evenly illuminated and illumination was kept constant across all objects and participants.

The participants were seated in front of the box, and their heads were placed on a chin rest 70 cm away from the stimuli. To measure gaze position, participants were equipped with the monocular eye tracker from Pupil Labs[137]. The eye tracker and software were calibrated to establish the subject-dependent correspondence between the raw gaze data and 3D object space using a custom procedure, similar to the one described in Chapter 2.

The 3D objects used in the experiments were 3D printed to ensure we had highly accurate realizations of their geometries (approximately 0.1 mm) and the input mesh used for printing was readily available digitally. We used a Contex DESIGNmate Cx SLS 3D printer, which produces homogeneous and highly diffused surfaces. Our experiment included 15 different objects to cover a range of different stimuli, from abstract to humanoid. Each object was presented in one of three predefined viewing directions. Orientations were randomly distributed among participants and the different objects and were ensured through a custom-built base with a dent for each orientation.

Null Hypothesis and Design The goal of our experiment was to study human observers' fixation patterns while they were visually exploring 3D objects. In particular, we designed the experiment to test the following hypothesis, stated as null hypothesis:

Fixations on a 3D object are distributed randomly across the object's surface.

Experimentally, the hypothesis claims that observers produce their own idiosyncratic fixation patterns. A similar null hypothesis for visual saliency[117,150] has been repeatedly refuted when *images* were used as stimuli [23]. Not being able to refute this hypothesis for objects would be worrisome and would seriously call into question the validity of images as experimental stimuli[152].

Task The participants were instructed to look at and visually explore each object. The objects were grouped into blocks consisting of the successive presentation of three different objects. After each block, participants were asked one question about one of the objects. An example mock question is "Was the dragon's mouth open?" We introduced this task to reduce the variability in the viewing patterns between observers, who might otherwise invent their own tasks[107] Ch. 3.2.3.

On average, the participants answered 79 percent of the questions correctly, with a standard deviation (SD) of 18 percent. Participants responded with "I am not sure" in nine out of 150 questions. We did not observe any systematic inability to answer questions for individual participants.

Procedure At the beginning of the experiment, each observer completed a demographic questionnaire. Thereafter, participants were introduced to the experimental setup and equipped with the eye tracker. The experimenter explained the calibration routine and the sequence of events during the experiment, which was as follows.

The objects were presented one at a time for a duration of approximately five seconds in blocks of three objects, with short breaks between subsequent blocks. Each participant completed five blocks, resulting in the presentation of 15 objects per participant.

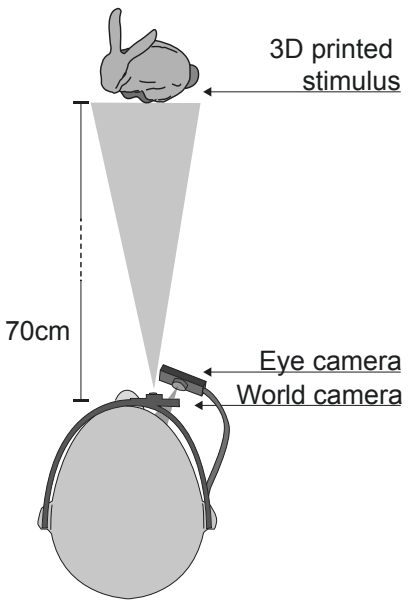
Prior to each block, we calibrated the experimental setup. The calibration procedure consisted of two subsequent sequences of looking at the calibration targets (see Figure 4.1, *right*) with the first one being the actual calibration and the second one a verification. If the error exceeded a threshold of 1.5 degrees, the calibration was deemed too inaccurate and was repeated. To prevent participants from seeing the objects prior to data collection, we blocked their view with a movable screen after the calibration and in between the presentations.

The objects were presented in random order. Each observer was presented with each object only once to avoid potentially confounding effects of habituation or boredom for repeated presentations of the same object. We familiarized participants with the procedure by having them complete a brief training session with one sample object (a teapot) prior to the first block.

Gaze position was recorded for the first few seconds after the onset of the stimulus. The exact time is a parameter in the data-processing pipeline.

Data Collection and Representation We collected both video data from the eye tracker's scene camera (30 fps) and gaze data (120fps). The raw gaze data is in the form of estimated pupil center positions in the pupil camera's coordinate system at discrete time values.

In this chapter, we refer to data according to the following scheme. Superscripts starting with a lowercase i indicate the object, and subscripts starting with n indicate the subject. This conforms to a top view of the experimental setup, with the observer being on the bottom of the image and the object on the top (see Figure above). Hence, pupil data for object i collected from subject n at time sample t_k is $\mathbf{g}_n^i(t_k) : \mathcal{N} \mapsto \mathbb{R}^2$.



4.3 From Pupil Positions to 3D Gaze Locations

Our procedure that relates measured pupil positions to 3D gaze locations on an object relies on an eye tracker with a world and eye camera, as already described, and a fiducial marker in a fixed relative position to the object (see Figure 4.1). We analyzed measured pupil positions in the eye camera image and extracted fixations as positions that remain within a fixed radius ϱ for a sufficiently long period of time τ (the minimum fixation duration). The fiducial marker's image in the eye tracker's world camera is used to determine the mapping from the object space to the

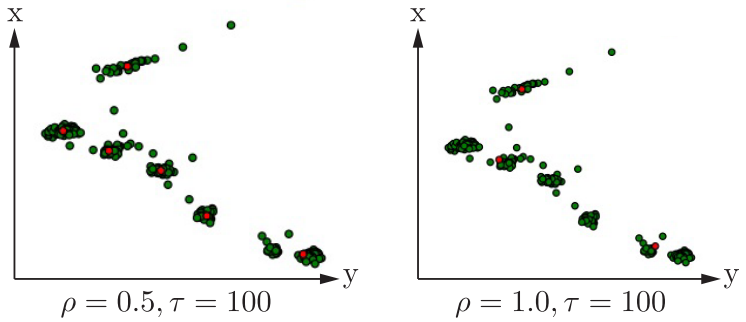


Figure 4.2: Fixation classification for two values of ϱ. Red dots indicate classified fixations, and green dots show raw data points. On the left a good clustering of fixations is obtained for $\varrho = 0.5^\circ$ degrees, whereas $\varrho = 1.0^\circ$ results in merged groups that visually do not belong together shown on the right. The data here is a subset of fixations on one object from one participant.

world camera coordinate system. The mapping parameters from the world to the eye camera coordinates are determined in the calibration phase. Together with the known geometry, this setup lets us determine highly accurate gaze positions on the object.

Here, we provide an overview of the mapping from measured pupil positions to gaze locations on an object. For a detailed discussion, see Chapter 2.

From Pupil Positions to Fixations Using this parameter setting resulted in an average fixation duration of 315 ms (SD = 41 ms) for $t = 100$ *ms* and 346 ms (SD = 41 ms) for $t = 150$ *ms*. Additionally, we chose t = 1.5 seconds. On average, these parameter settings result in 3.75 fixations (SD = 0.31) per object and participant.

To determine a fixation position $\mathbf{f}_n^i \in \mathbb{R}^2$ for object i and subject n in the eye camera image, we consider a sequence of measured pupil positions $\mathbf{g}_n^i(t_k) : \mathbb{N} \mapsto \mathbb{R}^2$ at consecutive sampling times. If $\mathbf{g}_n^i(t_k)$ remain in a small region of radius ϱ for a duration of at least τ milliseconds, then these are considered a fixation $\mathbf{f}_n^i$ at the mean location of the $\mathbf{g}_n^i(t_k)$. We consider only the first t seconds after stimuli onset for fixations because we are interested in spontaneous visual reactions. Together, the parameters ϱ, τ and t hence control the mapping from measured pupil positions $\mathbf{g}_n^i(t)$ to fixations $\mathbf{f}_n^i$. For our analysis, we chose $\varrho = 0.5^\circ$, which gave us a good balance between fixation length and data stability (illustrated in Figure 4.2). Using this parameter setting resulted in an average fixation duration of 315 ms (SD $=$ 41 ms) for $\tau = 100$ ms and 346 ms (SD $=$ 41 ms) for $\tau = 150$ ms. Additionally, we chose $t = 1.5$ seconds. On average, these parameter settings result in 3.75 fixations (SD $=$ 0.31) per object and participant.

From Fixations to 3D Gaze Positions We employ the fiducial marker's image in the world camera and its known relative position to the object to estimate the position $\mathbf{t}_n^i$ and orientation $\mathbf{R}_n^i$ of object i relative to subject n. This provides a mapping from points $\mathbf{x}_i \in \mathbb{R}^3$ in object space to points $\mathbf{w}_n^i \in \mathbb{R}^3$ in the world camera co- ordinate system of subject n:

$$\mathbf{w}_n^i = \mathbf{R}_n^i \mathbf{x}_i + \mathbf{t}_n^i. \tag{4.1}$$

To relate a fixation $\mathbf{f}_n^i$ to the corresponding gaze location on object i, we use the inverse of the mapping from 3D points $\mathbf{w}_n \in \mathbb{R}^3$ in the world camera coordinates to 2D pupil positions $\mathbf{p}_n$ in the eye camera image. This mapping is independent of the object because the world camera coordinate system serves as a reference, with the object dependence being described by equation (4.1). The mapping is projective and in homogeneous coordinates and hence is given by

$$s \begin{pmatrix} \mathbf{p}_n \\ 1 \end{pmatrix} = \mathbf{Q}_n \begin{pmatrix} \mathbf{w}_n \\ 1 \end{pmatrix} \quad , \quad \mathbf{Q}_n \in \mathbb{R}^{3\times 4}, \tag{4.2}$$

where s is a scaling factor. We determine $\mathbf{Q}_n$ from a set of correspondences $\{\mathbf{p}_i, \mathbf{w}_i\}$ obtained during calibration, where subjects are asked to fixate on fiducial markers on a custom build 3D rig (see Figure 4.1). Because $\mathbf{Q}_n$ is a projective transformation, it can be factored into an intrinsic camera matrix $\mathbf{A}_n^{\mathcal{Q}}$ and a rigid transformation $\mathbf{T}_n^{\mathcal{Q}} = (\mathbf{R}_n^{\mathcal{Q}}, \mathbf{t}_n^{\mathcal{Q}}) \in \mathbb{R}^{3\times 4}$ consisting of a rotation $\mathbf{R}_n^{\mathcal{Q}}$ and a translation $\mathbf{t}_n^{\mathcal{Q}}$. This allows us to associate with each fixation $\mathbf{f}_n^i$ a ray $\mathbf{r}_n^i$ in world camera space whose direction is defined by

$$\mathbf{f}_n^i = \mathbf{A}_n^{\mathcal{Q}} \mathbf{R}_n^{\mathcal{Q}} \mathbf{r}_n^i. \tag{4.3}$$

The ray origin $\mathbf{o}_n$ is $\mathbf{t}_n^{\mathcal{Q}}$, and the depth along it is indeterminate because $\mathbf{A}_n^{\mathcal{Q}}$ is a projection. Pupil positions have limited accuracy for determining viewing direction[107] and eye tracking introduces additional measurement errors, so each fixation is in fact associated with a cone of possible gaze positions. The opening angle can be determined experimentally[269], and we will denote it by c.

To find a subject's gaze location $\mathbf{v}_n^i$ on object i for a fixation $\mathbf{f}_n^i$, we determine the vertex $\mathbf{v}_a$ of the object's mesh representation $\mathcal{M}$ whose mapping

$$\hat{\mathbf{p}}_a = \mathbf{Q}_n\big(\mathbf{R}_n^i \mathbf{v}_a + \mathbf{t}_n^i\big) \tag{4.4}$$

to pupil coordinates is closest to the fixation $\mathbf{f}_n^i$. For this, we first consider the set

$$\Gamma_c(\mathbf{f}_n^i) = \left\{ \mathbf{v}_a \in \mathcal{M} \left| \frac{(\hat{\mathbf{f}}_n^i)^T \mathbf{M}_n^Q \hat{\mathbf{p}}_a}{((\hat{\mathbf{f}}_n^i)^T \mathbf{M}_n^Q \hat{\mathbf{f}}_n^i)^{1/2} (\hat{\mathbf{p}}_a^T \mathbf{M}_n^Q \hat{\mathbf{p}}_a)^{1/2}} > \cos c \right. \right\}$$

of all vertices $\mathbf{v}_a$ in the cone with the opening angle c, where $\hat{\mathbf{f}}_n^i = (\mathbf{f}_n^i, 1)$ is the fixation in homogeneous coordinates. The sought-after gaze location $\mathbf{v}_n^i$ corresponding to fixation $\mathbf{f}_n^i$ is then the vertex in $\Gamma_c(\mathbf{f}_n^i)$ closest to the eye, which is given by $\mathbf{M}_n^Q = (\mathbf{A}_n^Q \mathbf{A}_n^Q)^{-1}$ (see Chapter 2 for a derivation).

4.4 Analysis

We tested the processed data against the null hypothesis (fixations are randomly distributed) in three steps: First, we defined a measure of agreement between two fixation sequences on the same object. Second, we generated test fixation sequences that are unrelated to a particular object in the experiment. And third, we test whether there is more agreement between real fixations than between real and test fixations.

Comparing Fixation Sequences on the Same Object Our analysis is based on fixation sequences. A fixation sequence v_n^i is given as a set of vertex indices of the mesh representing object i. With v_n^i, v_m^i being two such sets, we want to measure the amount of agreement between the two sequences. We denote this similarity as $s(v_n^i, v_m^i)$.

It is generally difficult to measure the distance of geometric sets. We therefore accept that the measure is asymmetric–that is, in general $s(v_n^i, v_m^i) \neq s(v_m^i, v_n^i)$. Now, recall that the vertices that are computed from measured data are known to be inaccurate. This means that the computed vertices are unlikely to be exactly identical for identical fixations, and small distances between two fixations in either image space or on the object's surface are unlikely to be meaningful. Consequently, it is most meaningful to consider two fixations identical when they are closer than some threshold.

This threshold is well defined in terms of angular deviation between eye rays, which we have experimentally determined to be 0.8° on average. Therefore, we will relate two vertices to the angle between the rays from the eye to each of the vertices.

The rays from eye to vertex depend on the eye's position relative to the object coordinate system, and different subjects have different eye positions. Fortunately, given the geometry of our setup, a slight change in eye position has only a negligible effect on the angle between two eye rays. This is because the distance between eye and object is much larger than the distance between the two vertices on the object (at least for vertices potentially considered as the same fixation). We exploit this observation and measure the angle between eye rays from one of

the two eye positions connected to the two data sets. With $\mathbf{o}_n$ as the center of rays for subject n, we can generate the eye ray for any point $\mathbf{x}$ in world coordinates based on the eye center as

$$\mathbf{r}_n(\mathbf{x}) = \frac{\mathbf{o}_n - \mathbf{x}}{\|\mathbf{o}_n - \mathbf{x}\|}. \tag{4.5}$$

To compare two vertices in the sequences v_n^i, v_m^i, we consider the rays from the eye of subject n for a vertex $\mathbf{v} \in v_n^i$ in the fixation sequence for object i as well as a vertex $\mathbf{w} \in v_m^i$ in the fixation sequence of subject m, whose eye center is different.

The cosine of the angle between the view rays is

$$\mathbf{r}_j(\mathbf{v})^\mathsf{T} \mathbf{r}_j(\mathbf{w}). \tag{4.6}$$

For vertices with associated eye rays that differ by an angle smaller than the defined threshold δ, we consider the corresponding fixations identical.

Because we used a free viewing paradigm, we do not assume fixations to occur in an orderly sequence. We therefore compare each fixation in sequence v_n^i to all fixations in sequence v_m^i within the same time window τ regardless of their actual temporal position in their own sequence.

Our similarity measure is therefore

$$s(v_n^i, v_m^i) = \sum \left| \left\{ \mathbf{v} \in v_n^i, \mathbf{w} \in v_m^i \;\middle|\; \mathbf{r}_j(\mathbf{v})^\mathsf{T} \mathbf{r}_i(\mathbf{w}) > \cos(\delta) \right\} \right|. \tag{4.7}$$

As with other parameters, we varied δ to verify the stability of our result with respect to the particular choice made. Unless otherwise mentioned, we use $\delta = 1^\circ$. We generally suggest to choose δ to be slightly larger than the measured angular deviation ($\delta = 0.8^\circ$ in our case). The number of fixations might vary with changes in δ, but for the similarity measure, the test and real sequences must have the same number of fixations. Each sequence pair is compared twice in both directions due to the measurement asymmetry.

Generation of Test Sequences To analyze viewing behaviour, we compare the fixation sequence $\mathbf{f}_n^i$ from subject n for object i against a mock fixation sequence. The mock fixation sequences are generated by projecting the observed sequence $\mathbf{f}_n^i$ onto another object j. We denote $v_n^{i \to j}$ the *test sequence for object j* generated by intersecting the eye rays computed from the fixation sequences f_n^i with object j.

When projecting onto the true object, we ignore fixations that have no intersection with the object. This provides us with a set of fixated vertices that is unrelated to the visual stimulus created by j (assuming that objects i and j are sufficiently different from each other). The mock sequences exhibit all properties of fixation sequences that arise from normal oculomo-

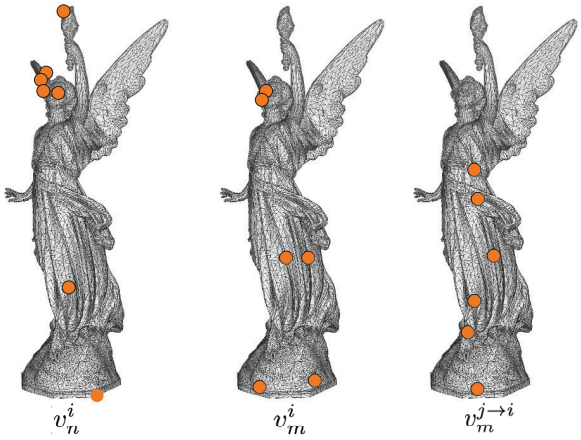


Figure 4.3: Sample fixations for participants $n = 8$ and $m = 13$. The right image shows a sequence generated by taking a fixation sequence gathered while a different object j was presented to subject m and projecting this data on the current object i.

tor functioning, such as characteristic dwell times and velocities, and hence are more realistic than randomly generated sequences. In particular, because the physiological processes underlying the generation of eye movements are not well understood, it would be difficult to generate fair random sequences.

Agreement of Fixations for the Same Visual Stimulus To test whether different human observers tend to generate similar fixations for the same visual stimulus, we compare the fixation sequences of two observers n and m for a fixed object i. Specifically, we ask whether the real sequence v_n^i is more similar to the real sequence v_m^i or to a test sequence $v_m^{j\to i}$ generated for observer m from an object $j \neq i$. This is illustrated in Figure 4.3, where the real sequence of subject n is compared to the real and test sequences of subject m.

The comparison was insensitive to the number of fixations because we used only sequences $v_m^{j\to i}$ that had the same number of fixations as the real sequence v_m^i. Summarizing, a single trial evaluates

$$\operatorname{sgn}(s(v_n^i, v_m^i) - s(v_n^i, v_m^{j\to i})), \quad \text{s.t. } |v_m^i| = |v_m^{j\to i}|. \tag{4.8}$$

Note that this expression may evaluate to zero because ties are possible. According to the null hypothesis, there should be no difference between the test and real data, implying that there is no preference for the sign value in the trial.

In the experiment, each participant viewed each of the 15 objects only once. This results in $90 = \mathbf{P}_2^{10}$ trials per object (all pairwise permutations because the comparisons are asymmet-

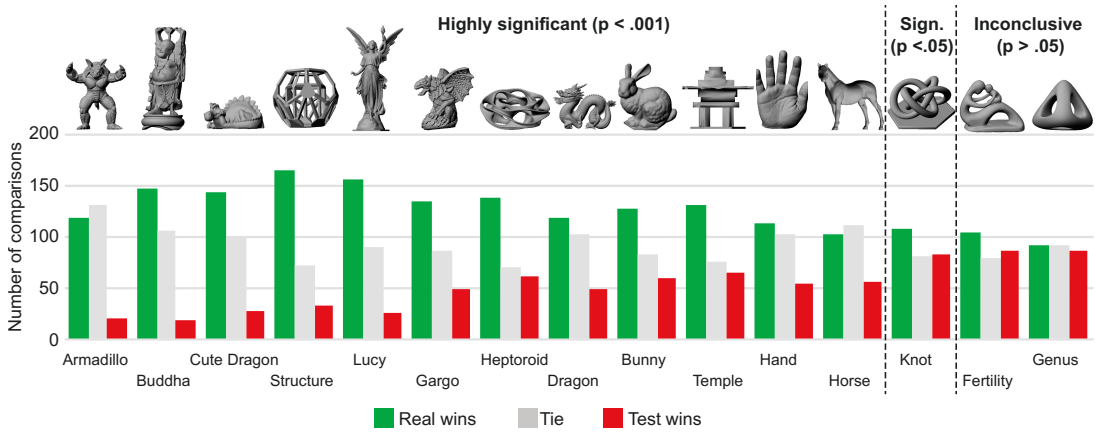


Figure 4.4: Results of the trials between the measured data from one subject and measured or generated data for subject m (over all participants). The green bars represent the trials in which measured data agreed more with measured data, the red bars show higher agreement between measured data and test data, and gray represents ties. For all but three objects, we find very significant agreement across subjects for nearly identical stimuli. For nearly featureless objects, the results are less significant or inconclusive.

ric). For each trial, we randomly selected a test sequence that satisfied the condition that the number of fixations is equal. Discounting the ties, we compared all remaining trial results to a cumulated binomial distribution to estimate the chance probability of that result.

We deliberately designed our similarity measure so that repeated fixations would result in a higher score. One reason for this is that repeated fixations of the same position might indicate a higher visual saliency of that object part. It is also possible that, because of our accuracy limits, two close features might not be properly resolved. In other words, spatially close features can result in the same measurement. We therefore repeated the experiment for all objects varying the parameters ϱ, τ, t, and δ. In general, the results are stable with respect to these parameters. Figure 4.4 illustrates the resulting p values.

For most of the objects, fixation patterns agree between observers. This is in particular the case for the more complex shapes with distinctive features. Only a few cases were less conclusive. We suspect that smoother or simpler shapes have fewer features that stand out so that fixations might be attracted by the occluding contours of the objects, and thus the agreement between observers drops to chance levels.

4.5 Validating Computational Saliency Models

Our results indicate that the fixations we gathered in our experiment agree across subjects for similar stimuli. This means that the data is suitable for testing the validity of computational models of saliency.

The main tool we suggest for such an analysis is permutation of the values generated by the computational model. If a computational model has predictive power for fixations on the object it will provide significantly higher saliency values for our data than its own permutations. Permutation tests are generally considered as strong tests for significance[170], and they directly yield p values as the normalized rank of the original data relative to the permutations.

Unbiased Permutations of Saliency Values Computational saliency models provide, in one way or another, a scalar value for each point on an object's surface that describes the point's visual saliency. Our goal is to provide another scalar function over the surface such that the expected value for a sample drawn *under the same conditions*, as in the experiment, is unchanged. There seem to be (at least) two possible assumptions on how mesh saliency values are sampled:

- Samples are drawn uniformly from the surface. This assumption reflects that saliency models are object-based, and permutations could be interpreted as alternative view-independent, global, object-based saliency models (such as mesh saliency[167]).
- Samples are drawn uniformly from the visible part of the surface. This assumption reflects that fixations are based on a single view (that is, the probability of fixating on vertices on the backside of the object is zero).

We derive unbiased permutations for these two assumptions.

We assume here that the object is represented by a triangle mesh $\mathcal{M} = (\mathbf{V}, \mathcal{T}) \subset \mathbb{R}^3$. Quantities defined on a mesh are commonly given for each vertex or for each triangle. Although often not explicitly specified, these values can be extended to every point on the surface using basis functions $b_l(\mathbf{z}), \mathbf{z} \in \mathcal{M}$, where l is the index of the mesh element–that is, a vertex or a triangle. To make this concrete, consider values given in vertices, which are linearly interpolated over triangles (more complex basis functions are possible and can be treated similarly).

The basis functions, together with the saliency values s_l per vertex l, define the saliency over the piecewise linear surface as

$$s(\mathbf{z}) = \sum_l s_l b_l(\mathbf{z}), \quad \mathbf{z} \in \mathcal{M}. \tag{4.9}$$

The expected value for a sample drawn uniformly from the surface $\mathcal{M}$ is then

$$E_{\mathcal{M}} = \int_{\mathcal{M}} \sum_l s_l b_l(\mathbf{z})\, d\mathbf{z} = \sum_l s_l \int_{\mathcal{M}} b_l(\mathbf{z})\, d\mathbf{z}. \tag{4.10}$$

To compute the expectation in the projection, we need the surface normals $\mathbf{n}(\mathbf{z})$, the eye ray $\mathbf{r}(\mathbf{z})$ (see (4.5)), and the binary information $v(\mathbf{z}) \in \{0, 1\}$ on whether the surface point $\mathbf{z}$ is visible. Then we can adjust (4.10) for projection and visibility and get

$$E_P(\mathbf{z}) = \sum_l s_l \int_{\mathcal{M}} \mathbf{n}(\mathbf{z})^\top \mathbf{r}(\mathbf{z})\, v(\mathbf{z}) b_l(\mathbf{z})\, d\mathbf{z}. \tag{4.11}$$

To make this concrete for linear interpolation, consider the *area vector* to vertex l:

$$\mathbf{a}_l = \frac{1}{2} \sum_{(l,l',l'') \in \mathcal{T}} (\mathbf{v}_l - \mathbf{v}_{l'}) \times (\mathbf{v}_l - \mathbf{v}_{l''}) \tag{4.12}$$

encoding both vertex normal and associated area[6]. The area vector lets us succinctly write the integrals in these expectations for the case of piecewise linear basis functions. We have

$$B_l = \int_{\mathcal{M}} b_l(\mathbf{z})\, d\mathbf{z} = \frac{1}{3} \|\mathbf{a}_l\| \tag{4.13}$$

and we approximate the case for projections under the assumption that directional variation of the ray from the origin to the surface is small for a single triangle (and considering only visible vertices) as

$$B_l = \int_{\mathcal{M}} \mathbf{n}(\mathbf{z})^\top \mathbf{r}(\mathbf{z})\, b_l(\mathbf{z})\, d\mathbf{z} = \frac{1}{3} \mathbf{a}_l^\top \mathbf{r}(\mathbf{v}_l). \tag{4.14}$$

Note how equation (4.13) and equation (4.14) differ only in the projection of the area vector.

We can easily keep the expected saliency value constant by taking permutations of the expected values $\{s_l B_l\}$ for the individual elements. Let $\{\pi[l]\}$ be a permutation of mesh elements. Then we assign permuted saliency values as

$$s'_l = s_{\Pi[l]} \frac{B_{\Pi[l]}}{B_l}. \tag{4.15}$$

Testing Mesh Saliency We compute mesh saliency values following the original approach of Chang Ha Lee and his colleagues[167] and obtain saliency values over the entire mesh using piecewise linear basis functions. Given a set of fixations v_{ij}, we need to compute a saliency score. Although it is clear that we want to sum over the contributions $\mathbf{v} \in v_n^i$ of individual fixations, there is no single correct way for considering each fixation[146].

We opt for simply taking the saliency values $s(\mathbf{v})$ and summing them up over the set of

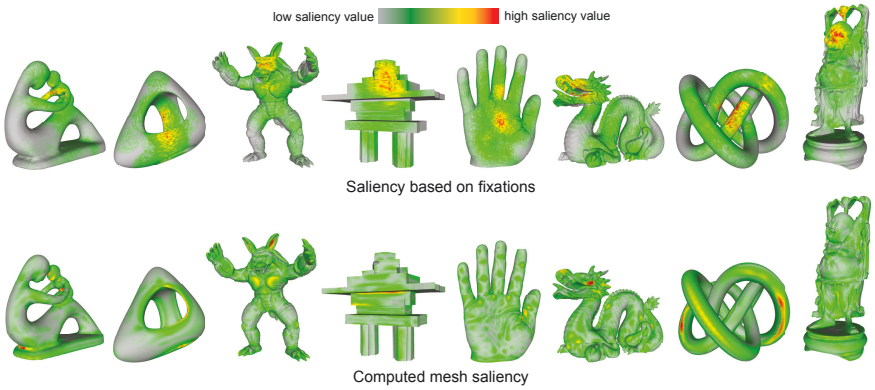


Figure 4.5: Mesh saliency test. We compared the heat maps generated from the collected fixation data (*top*) to the computed mesh saliency values (*bottom*) on a representative set of objects used in the experiment.

fixations v_n^i. The reason is that the permutations we compute should yield the same expected value for *point-wise samples*, yet not necessarily for other means of collecting values. In particular, integrating over a small area on the surface (such as within a cone related to the measurement error) has different characteristics for mesh saliency compared to its permutations: mesh saliency provides smoothly varying values (see Figure 4.5, *bottom*), so all saliency values are rather similar, while the permutations generate high-frequency noise. Area integrals for the permutations tend to be the same everywhere, whereas area integrals for the true mesh saliency do depend on the fixation. This would introduce bias.

Based on summing up the saliency values of the closest vertices for all fixation sequences on an object, we calculate the score for mesh saliency values and 100,000 of its permutations. The rank of the mesh saliency score among all of its permutations yields the p value. We perform this experiment for permutations adjusted to the object's surface as well as its projections.

As a sanity check, we created heat maps from the fixation data (see Figure 4.5, *top*) and verified that they perform significantly better than their permutations. The results of this test for mesh saliency show that it generally fails to outperform its own permutations in a statistically significant way. As an example, Figure 4.6 shows the result for our preferred value for dispersion $\varrho = 0.5°$ and permutations adjusted for area. We considered objects in different viewing directions individually. Permutations adjusted for the visible projected area lead to comparable results, but they contained higher variations because of the large influence of the small projected area of fixations close to occluding contours. Results from using permutations of mesh saliency on the whole object yielded similar results.

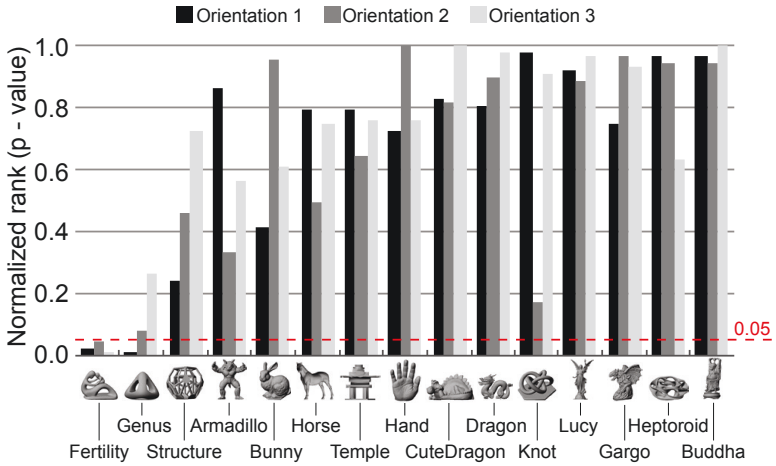


Figure 4.6: Mesh saliency matching our recorded gaze data. This figure shows data for a dispersion value of 0.5°. Permutations were computed with respect to the surface area.

4.6 Discussion

The process of gathering fixations on 3D objects is more complex than for flat stimuli. We believe that the (expected) result of agreement between subjects for the fixations for most objects indicates that our experimental setup is meaningful and avoids excessive noise.

The data is already useful in its current form, but we believe that varying experimental conditions further is important to learning about the invariant properties of visual saliency. In particular, our future work will look into varying the illumination conditions and the material properties of the stimuli. Whether fixations would still agree across such changes is still an open question.

From this perspective, one might expect that similar parameters are varied in the generation of flat stimuli based on rendering. Compared with 3D printing or setting up lights physically, it appears more convenient to change material properties or lights in rendering. Yet, to our knowledge, this is not usually done. We suspect that such parameters might also affect the results of on-screen eye tracking experiments. Such differences in the choice of the experimental setup might explain why our results differ from those of previous research with respect to the predictive power of mesh saliency[146].

Quite generally, we believe that predicting visual saliency independent of viewing conditions will be ill-posed; we probably need at least the orientation of the object toward the observer. These considerations are the main reason why we resisted the temptation to simply fit a saliency model to our data.

5

Tracking the Gaze on Objects in 3D: How do People Really Look at the Bunny?

The previous chapter shows results that suggest the existence of salient features on genuine 3D objects and provides evidence that traditional geometry-based computational methods fail to capture where people look at on 3D objects. But the study is limited to a single viewing direction. In this chapter we aim to study the influence of viewing directions on eye movements. We collect the first large dataset of human fixations on physical 3D objects presented in varying viewing conditions and made of different materials. Our experimental setup is carefully designed to allow for accurate calibration and measurement. We estimate a mapping from the pair of pupil positions to 3D coordinates in space and register the presented shape with the eye tracking setup. By modeling the fixated positions on 3D shapes as a probability distribution, we analyse the similarities among different conditions. The resulting data indicates that salient features depend on the viewing direction. Stable features across different viewing directions seem to be connected to semantically meaningful parts. We also show that it is possible to estimate the gaze density maps from view dependent data. The dataset provides the necessary ground truth data for computational models of human perception in 3D.

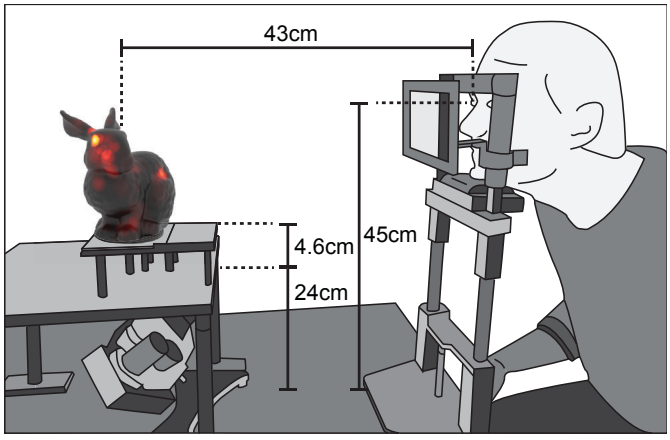


Figure 5.1: Schematic of the experimental setup. Shapes are placed approximated 100 mm below the eyes.

5.1 Introduction

A large part of geometry processing in computer graphics is based on *perceptually-based metrics*[166] and *visually salient shape features*[167,244]. Salient features are usually defined as objects or regions that draw attention of human observers. Interestingly, most approaches are based entirely on geometric or information theoretic measures. Those that are based on experiments almost exclusively use renderings of the shapes presented on a screen for evaluation[27,63,70,146]. We find that data derived from human observers inspecting physical manifestations of 3D shapes would provide a firmer ground for computational models of human perception. In this chapter we present an experimental setup for this task and gather data from over 70 participants on 16 shapes presented in 14 conditions.

The original notion of salient visual features derives from eye tracking experiments using images presented on a screen as visual stimuli. The main idea is that humans tend to attend to the most important parts of a scene first. Computational models of saliency[117] were developed only after some consensus had been reached on the local image characteristics that seemed to evoke attention.

Presenting the stimulus on a screen leads to a simple experimental setup. It has been argued[154] that only the visual percept on the retina matters, so restricting stimuli to images might suffice to learn about saliency of features. This point of view is questioned more and more[96,115,253]. If 3D shapes are restricted to virtual environments, such as being only presented on screens, then screen-based experiments naturally provide the necessary insight. While it may be true that computer graphics researchers rather deal with teapots and bunnies on-screen, 3D computer graphics and, more specifically, geometry processing derive their importance from the fact 3D shapes describe the "real world". The recent trend of direct digi-

tal manufacturing (aka. 3D printing) should remind us that a purely virtual existence of 3D shapes is the exception rather than the rule. It also provides an ample number of reasons for basing visual saliency on experiments with real 3D data.

Collecting points on real 3D shapes from human viewing behavior is significantly more involved than experiments using a screen for presentation. The experiments we are aware of[270] are limited in the variation of viewing conditions. We believe an important question is if low-level geometric saliency exists at all. This would mean that a region on a shape is attended to across different human observers, different surface reflection properties and different viewing directions. For this reason we have put effort into varying viewing directions (7 directions 15° degrees apart) and material (diffuse powder and comparatively glossy plastic) for a number of different shapes (see Section 5.2 for details). Illumination is restricted to one diffuse light source at a fixed location. The data will be generally useful to evaluate existing computational models for geometric saliency[167,241,244,251] and, if possible, directly generate such models from the data similar to recent approaches for images[125,156,157].

Eye tracking on 3D requires establishing a mapping between the pupil positions and positions on the shape. We do this using a setup (see Figure 5.1) that allows estimating a mapping from pairs of pupil positions to points in 3D and then intersecting registered 3D shapes in this environment. The mapping allows us to create gaze density maps, a probability representation of eye tracking data over the surfaces of the shapes for further analysis.

Data collected in this setup from over 70 human observers seems to suggest that salient features depend on the viewing direction, but not on the two different materials we used. Visual inspection of regions that are fixated in all viewing directions appear to be connected to semantically meaningful parts. These observations indicate that visual saliency is difficult to predict from geometric features alone. Based on these observations we build a small convolutional network that is able to predict the gaze density maps generated from our experiments for a given shape. Consistent with our experimental findings, it fails to generalize across shapes, yet is still better in predicting saliency than geometric approaches such as mesh saliency[167].

In summary, we make the following contributions:

- We develop a setup for eye tracking experiments on real 3D shapes, including an accurate registration, calibration procedure and automatic mapping from binocular eye tracking data to the surface of 3D shapes.
- We provide the first large data set with fixations on 3D shapes. The data set will be useful for assessing perceptual metrics and saliency measures.
- We develop a novel method to analyze distributions of fixations on 3D shapes.
- We show that stability of features depends on distance in viewing angle.

- We develop a machine learning approach that allows predicting human visual saliency on objects based on view-dependent geometry information.

5.2 Design and Setup of the Experiment

Our experiment follows the established protocol of eye tracking experiments for detect salient regions in image stimuli [21,131]: in a first step, calibration targets with known positions are presented to the observer, allowing to establish a mapping from pupil positions to the coordinate space of the calibration targets. Then, stimuli are presented for a short amount of time in the same coordinate frame and observer's eye movements are recorded. Fixations are detected from eye movement sequences and can be mapped to the stimulus for further analysis. The fixations shortly after the onset of the stimulus are indicative of salient regions in visual scenes.

The main idea of our experiment is to present physical 3D shapes as stimuli. Besides carefully adapting the standard setup, this comes with a few challenges, such as accurately aligning the coordinate spaces of the calibration targets and shapes as well as presenting the shapes at once. Moreover, the experiment should reflect our main question, namely if geometric features of the shape may be salient, i. e., attract attention regardless of viewing conditions.

Setup For eye tracking, we use an EyeLink 1000 table top device, which is routinely used in a variety of eye tracking experiments. The eye tracker consists of a camera and an integrated IR illumination (as shown in Figure 5.2). The camera and the light source need to have free view on the eyes, with the angle to the line of sight being limited. As eye tracking has limited angular accuracy, the spatial accuracy decreases with the distance to the observer. This motivates us to bring the shape in the experiment close to the observer such that the error in relating the gaze to the shape is small, while still keeping the eye tracker in its working range with distinct corneal reflections.

We accomplish the requirements of the eye tracking device and our goal to place the shape close to the observer by placing the shapes onto a fixture that allows placing the eye tracker under it (see Figure 5.2). The fixture is placed with its front edge at a distance of 320 mm to the observer, allowing the presentation of objects at an average distance of about 430 mm (see Figure 5.1 for a schematic illustration of related distances). The fixture is made of aluminum. The base is a block with dimensions 300 $mm \times 300\ mm \times 12\ mm$. It is mounted onto four cylindrical legs with a diameter of 20 mm, which fit into sockets permanently mounted to the table. There are two copies of the fixture. One base plate contains a raster of 9×9 screw mounts with grid constant 40 mm. The screw mounts serve to hold the legs as well as 20 tubes with calibration targets as shown in Figure 5.2. The other base plate only has the four corner mounts for the legs and 4 holes to hold a connector for the base of the shapes as shown

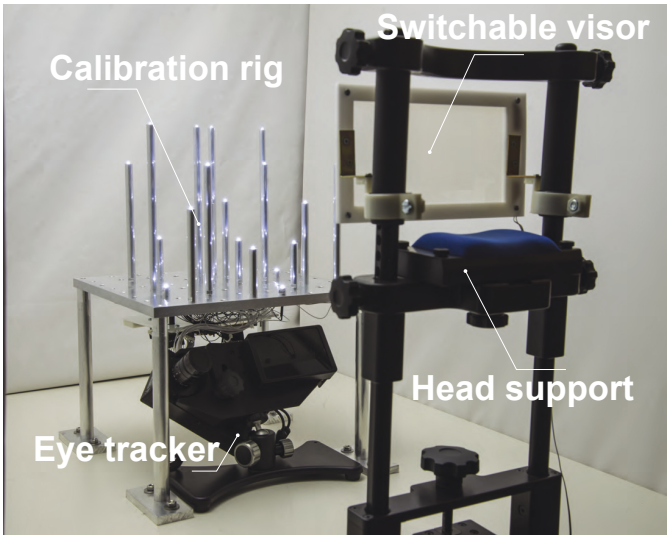


Figure 5.2: Calibration setup. EyeLink 1000 is used to track the eye movements and a chin-forehead rest is used for stabilization. Coordinates space of the calibration targets and shapes are aligned with sockets permanently mounted on the table. A switchable visor is used to control the on-site of stimuli. A custom-built calibration rig is used with 20 LEDs mounted as calibration targets.

in Figure 5.4. Machining precision for these parts is reportedly on the order of 2/10 mm. This allows presenting the shapes in a coordinate frame that is very well aligned with the coordinate frame of the calibration targets.

The calibration targets are LEDs. Each LED is mounted onto the top of an aluminum tube, wired through the tube and the screw hole. The tubes have different lengths and the LEDs cover a volume of 150 mm^3 (consistent with the size of the shapes, see below). LEDs are arranged in space as evenly as possible while not being occluded. They are controlled by an Arduino board, so that the active time of each of the LEDs can be recorded and aligned with the data from the eye tracker. While the accuracy of the positions of the tubes on the base plate is high, the exact heights of the LEDs relative to the top of the tubes vary slightly, and the angular deviation of the screw mounts potentially translates into significant displacement at the top of the tube. To compensate for this, we measure the positions of the LEDs using a recent structure from motion tool [238]. We took 10 photographs with constant camera parameters of the calibration rig while all LEDs are illuminated. In each image, we identify the four corners of the base plate by fitting lines to the edges of the base and intersecting them. The front left corner serves as the origin of the coordinate system. We fit a quadratic function to the smoothly varying brightness of the LEDs in the photographs. This yields the LED centers with sub-pixel accuracy. The resulting reconstruction has a reported average accuracy of 0.6 mm in the positions of the LEDs. The reconstructed positions are consistent with the

Figure 5.3: The whole stimuli set of 16 shapes printed in sandstone.

design of the fixture.

The whole setup is enclosed by a box with diffuse white walls to avoid presenting visually interesting features apart from the stimulus. The front side of the box is open, leaving space for a head and chin rest.

Selection of Stimuli It is well known that both low-level features, such as contrast and edges, and high-level features, such as faces, consistently attract visual attention[81,98,253]. In order to best investigate how low-level features generated by the geometry of a region and high-level features embedded in the shapes contribute to the visual saliency, we try to select a set of models that represent a broad generalization. We include shapes with both smooth surface and sharp corners. Symmetrical shapes, including those with repetitive geometric features, are also selected, although we suspect that repetitive features could make it difficult to find a consistency among observers. Even if such features draw attention, the number of fixations on each of them could still be small. Inspired by[164], we also include man-made artifacts (e. g., teapot and spanner), which might have task-related affordances (e. g., grabbing) that attract attention. Shapes with discernible semantic features like the BUNNY-object are also included in the set to have a generalized representation. Based on these principles, we selected 16 shapes (shown in Figure 5.3). The number 16 is a compromise between providing enough variation and the duration required for each experiment session.

Using direct digital manufacturing for creating the physical stimulus has several important advantages (cf. Wang et al. [269,270]):

1. Because we start from the digital version and manufacturing devices are reported to have high geometric accuracy, the geometry of the physical artifact is known.

2. Digital modeling allows us to add geometry to the bottom of the shape, enabling a connection to the experimental setup in a controllable way.

3. The material is homogeneous.

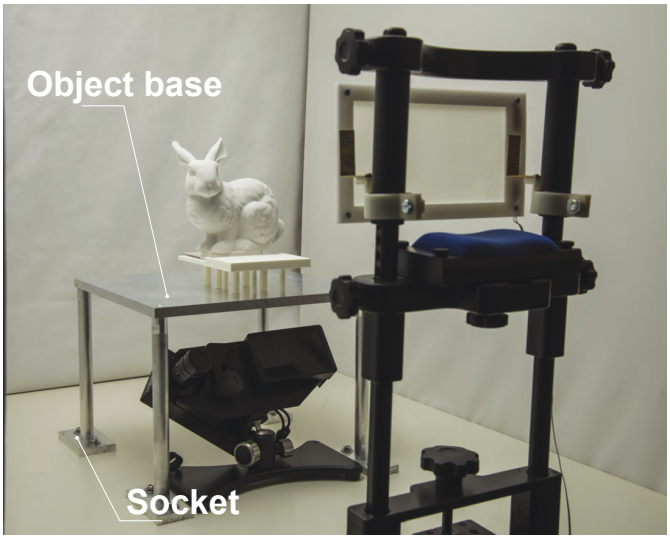


Figure 5.4: During viewing, 3D printed stimuli are placed in front on a fixture, which is mounted into the sockets on the table. The coordinate frame of shapes is aligned with the calibrated space by mounting the two identical fixtures in the sockets that are permanently mounted onto the table.

The only potential problems result from some manufacturing techniques being limited in terms of the minimal thickness of parts in the shape as well as the largest dimensions because of limited build volume. The size of the shapes results from covering a large visual angle without being uncomfortable for humans to inspect the object while not moving their head. Other experiments suggest that an acceptable visual angle is 20°, resulting in an average size of 150 mm along the largest dimension. This size is still compatible with mass-market 3D printing.

For evaluating constancy of features against change in material, we choose to manufacture each shape in two materials, using two different manufacturing devices. One set is generated using the Stratasys Uprint SE Plus fused deposition modeling device available in our lab with ABS* as filament, resulting in a slightly shiny and smooth appearance. Another set is manufactured commercially using 3D ink-based printing with a diffuse material†. Figure 5.5 shows a visual comparison of the BUNNY-object printed in two materials.

To test the variation in viewing behavior, we present each shape in several orientations. For each shape we decide on an up-direction. The different orientations result from rotating around the up-axis. Rotation by very large angles would lead to occlusion or disocclusion of features. We feel a total range of 90° is sufficient. One may expect that for very small angles

*ABSplus P430XL

†We printed at Shapeways using SANDSTONE.

Figure 5.5: Experimental conditions of one stimulus. Each shape is printed in two materials and presented in 7 viewing directions. Here we see an example of the Stanford BUNNY printed in ABS shown in the first row. The second row shows the shape printed in sandstone. From left to right we see all seven viewing directions presented in the experiment.

of rotation, the resulting visual stimulus in the experiment hardly changes, so this would add little information. We split the 90° into steps of 15° (see Figure 5.5 for example). To facilitate an accurate presentation at different angles, we add a flat 24-gon to the base of the shape. Adding this 24-gon to the shape before manufacturing has the advantage that the angle of the vertices of the polygon relative to the geometry is well-defined.

The set of 7 orientations together with the two different materials leads to 14 different experimental conditions for each of the 16 shapes.

Presentation We believe a lighting situation that is common for humans leads to the most meaningful results. Consequently, a single light source is placed above and slightly to the left (see illustration in Figure 5.4) of the shapes. This leads to different surface scattering properties of shapes printed in ABS comparing to shapes printed in sandstone. We use a luminance meter to measure the amount of light reflected from the surface and for shapes printed in ABS (acrylonitrile butadiene styrene) it is 74 cd/m^2 and for shapes printed in sandstone it is 42 cd/m^2. In future work, it would be interesting to include more lighting conditions by varying the number of light sources, directions and intensities. Determining a good set of conditions to study variations for human perception is an interesting question.

It is important that each visual stimulus is presented *at once*. The underlying idea of analyzing saliency by eye tracking is that an unknown stimulus is explored, and the first milliseconds after the stimulus became present are indicative for the most important features. This can only be achieved by blocking the observers view while setting up the shape on the fixture. We wish to avoid any moving objects in front of the observer, as moving objects tend to draw attention. We would also like to avoid any evasive motion of the observer's head, which would invalidate the calibration. For this reasons we mount a sheet of polymer-dispersed liquid crys-

Figure 5.6: View of an observer during stimuli presentation. An occluded view when the switchable diffuser is opaque is shown in the bottom corner.

tal (PDLC) switchable diffuser on the chin-forehead rest and the diffuser is controlled by an Arduino circuit. In its transparent condition, PDLC switchable diffuser is reported to have 90% transmission. In opaque state, the material exhibits approximately 80% haze (i. e. scatters incoming visible light), making it virtually impossible for participants to see through[173]. Arduino control allows us to record the time of the onset of the stimulus and to synchronize with the recorded eye positions. Figure 5.6 shows the view of an observer when the BUNNY-object is presented and the occluded view is shown in the right corner. No significant change of pupil size is observed when the diffuser is switched between its two conditions.

5.3 Data Collection

Observers We recruited $n = 78$ participants (mean age = 24, SD = 4.5, 32 females) for the experiment. They had normal or corrected-to-normal visual acuity and no (known) color deficiencies. 8 observers failed to calibrate the eye-tracker with the required accuracy, which left us with a dataset of 70 observers viewing 16 shapes. Importantly, all participants were naive with respect to the purpose of the experiment. Consent was given before the experiment and participants were compensated for their time.

Eye Movement Recording The experiment was conducted in a quiet room and shapes were presented on the fixture 430 mm in front of the observer. The largest visual span is 20°, resulting from 150 mm being the largest dimension of all shapes. Binocular eye movements were tracked with an EyeLink 1000 in remote mode and calibration was performed with our custom-built calibration fixture.

In calibration 20 LEDs were lit up one after another in random order with the first one being repeated once at the end, resulting in 21 targets in total. Recorded eye movements for the first LED is discarded and we only use the more reliable data from the second repeat.

Task Observers read the written task beforehand and were instructed to look at and inspect the shapes. The exact task is written as "Look at each object. See if anything is unusual or odd about the object. At the end of the experiment we will ask you to point out any observations you made. We will show the objects again, so you do not have to memorize them.". We do so to encourage observers actively viewing each shape without introducing an additional task. As an experimental task in eye tracking based perception studies is often designed as a trade-off between motivating observers to actively perceive the stimuli without introducing systematic bias and reducing the influence of noise and fatigue, we introduced such visual search task in the experiment. Observers might interpret the task differently but we do not observe any bias in the collected data, which coincides with the visual search literature as well[80,193]. Most observers reported that nothing is unusual except there are several objects which they were unable to identify. All of them could describe details of the viewed shapes and report their perceived aspects.

Procedure After reading the task, observers were introduced to the experimental setup and the detailed experimental routine. 16 objects were divided in two blocks with each being viewed for 5 seconds. Each observer only views one object in one condition and viewing order is randomized for each observer. Calibration and validation were conducted before each viewing block while validation is essentially a repeated procedure of calibration. We verify the calibration accuracy in validation and it took approximately 6 minutes for each block on average. As each configuration of one object is viewed for 5 seconds, we can easily take any subset for analysis. Although viewing order is only randomized without guaranteeing that the space of all possible viewing orders are sampled evenly, such simple randomization is more than sufficient to investigate whether viewing behavior changes over time.

One practice block was conducted at the beginning, which consists of calibration, validation and one shape (a horse) for viewing. Through the practice block, observers are familiarized with the experimental procedure as well as the tasks they need to perform.

We use the velocity-based fixation detection algorithm provided by EyeLink and on average there are 15 fixations in each trail of viewing one shape. Material, viewing direction and shapes all have no significant influence on the amount of fixations.

5.4 Mapping

An appealing feature of the 2D to 2D mapping approach is that it can be developed from minimal assumptions: identical pupil positions identify identical positions on the stimulus; and small displacements of stimuli induce small displacements of pupil positions. Mathematically, this means the mapping can be approximated by a smooth function, and practice shows that low order polynomials are sufficient. In particular, while some models are derived from additional assumptions on the geometry or physiology of the problem, their success is independent of the validity of the assumptions. This is important, because in many cases such assumptions are difficult to test experimentally.

Our goal is to relate *pairs* of pupil positions to the attended points in space. We believe this is possible because of vergence. We wish to also base our approach on minimal assumptions. In particular we want to avoid identifying individual pupil positions with eye rays and then intersecting these rays, because in this approach calibration is usually not directly optimized for the resulting positions in 3D but rather for the directions of the rays. In the following we develop a model that allows directly optimizing for the positions of the calibration targets.

Based on the established mapping between pairs of pupil positions and calibration targets, we analyze the error and model it as a Gaussian distribution. We can then estimate the probability distribution of a fixation on the provided three-dimensional object, simply as the restriction of the Gaussian distribution of the fixation in space to the object's surface.

5.4.1 Mapping Function

We consider a pair of pupil positions

$$\mathbf{p} = \begin{pmatrix} \mathbf{p}_l \\ \mathbf{p}_r \end{pmatrix} \in \mathbb{R}^4, \tag{5.1}$$

where the subscripts l and r refer to the left eye and the right eye, respectively. Our goal is to establish a mapping $\mathbf{f} : \mathbb{R}^4 \mapsto \mathbb{R}^3$ that identifies pairs of pupil positions with fixated points in space directly. The parameters governing $\mathbf{f}$ should be estimated directly from the known positions of the calibration coordinates $\mathbf{x}_i \in \mathbb{R}^3$ and the corresponding pairs of pupil positions $\mathbf{p}_i$ measured in the calibration phase.

We develop a parametric model for $\mathbf{f}$ based on geometric reasoning. As mentioned before, as long as the mapping provides sufficient accuracy, it is irrelevant whether our geometric assumptions are valid. Still, it makes sense to provide at least the precision of an idealized situation.

First, we assume that lines of sight have a common center for each eye and denote them by $\mathbf{e}_l, \mathbf{e}_r \in \mathbb{R}^3$. The pupil positions are mapped to affine planes in $\mathbb{R}^3$ using homogeneous

pupil positions $(\mathbf{p}_l, 1)^\mathsf{T}, (\mathbf{p}_r, 1)^\mathsf{T}$ and transformations $\mathbf{T}_l, \mathbf{T}_r \in \mathbb{R}^{3\times 3}$. Then the two half-lines emanating from the centers are defined as the lines passing through the eye centers and the pupil position mapped to the affine plane:

$$\begin{aligned} \mathbf{h}_l(\lambda_l) &= \mathbf{e}_l + \lambda_l \mathbf{T}_l \mathbf{p}_l, \quad \lambda_l > 0 \\ \mathbf{h}_r(\lambda_r) &= \mathbf{e}_r + \lambda_r \mathbf{T}_r \mathbf{p}_r, \quad \lambda_l > 0. \end{aligned} \tag{5.2}$$

We may ask that the two affine planes for mapping the pupil positions coincide, and that the recovered geometry for the eye centers and the affine planes are consistent with the desired world coordinate system. Because the planes coincide, for any point $\mathbf{x}$ in space we find

$$\mathbf{x} = \mathbf{h}_l(\lambda_l) = \mathbf{h}_r(\lambda_r) \Longrightarrow \lambda_l = \lambda_r = \lambda, \tag{5.3}$$

and the parameter λ is a linear function of the distance of the point $\mathbf{x}$ to the eyes. When solving for λ we have

$$\mathbf{e}_r - \mathbf{e}_l + \lambda \left(\mathbf{T}_r \mathbf{p}_r - \mathbf{T}_l \mathbf{p}_l \right), \tag{5.4}$$

and this is a rational function with constant nominator and a denominator that is linear in the pair of pupil positions. Plugging this expression back into the equations for the half lines to find the point in space $\mathbf{x}$, which is a function that is linear in λ, leads to rational linear function in the pair of pupil positions. This means we can write the mapping as

$$\mathbf{f} : \mathbb{R}^4 \mapsto \mathbb{R}^3, \quad \mathbf{f}(\mathbf{p}) = \frac{\mathbf{A} \begin{pmatrix} \mathbf{p} \\ 1 \end{pmatrix}}{\mathbf{b} \begin{pmatrix} \mathbf{p} \\ 1 \end{pmatrix}}, \quad \mathbf{A} \in \mathbb{R}^{3\times 5}, \mathbf{b} \in \mathbb{R}^5. \tag{5.5}$$

There are 20 parameters in $\mathbf{A}$ and $\mathbf{b}$, however, they share a common scale factor, leaving us with 19 degrees of freedom. Since each point in space provides 3 constraints, this means we need at least 7 calibration targets to estimate the mapping – usually we use more. To estimate the parameters with more constraints than unknowns we consider the residuals

$$\mathbf{r}_i = \mathbf{x}_i - \frac{\mathbf{A} \begin{pmatrix} \mathbf{p}_i \\ 1 \end{pmatrix}}{\mathbf{b} \begin{pmatrix} \mathbf{p}_i \\ 1 \end{pmatrix}}. \tag{5.6}$$

A common optimization goal is to minimize the sum of the squared lengths of the residuals. Based on our geometric motivation, however, we really want the residuals to have non-

uniform lengths: the error in pupil positions is measured on a plane; it is proportional to the error in space, however, by a factor that depends on the distance to the center of projection. In other words, we want the error to be proportional to the distance to the observer.

One way of solving this problem is to weigh the residuals with the inverse of the known distance of the calibration targets $\mathbf{x}_i$ to the observer and then solve the resulting non-linear least squares problem using an appropriate solver (e. g., Ceres Solver[3]). Another solution arises from the observation that the parameter λ is proportional to the distance from the observer. Recall that λ is a constant function divided by $\mathbf{b}(\mathbf{p}, 1)^\mathsf{T}$. This means we introduce a weighted residual by multiplying with $\mathbf{b}(\mathbf{p}, 1)^\mathsf{T}$ to get

$$\mathbf{r}'_i = \mathbf{b}\begin{pmatrix}\mathbf{p}_i\\1\end{pmatrix}\mathbf{r}_i = \mathbf{b}\begin{pmatrix}\mathbf{p}_i\\1\end{pmatrix}\mathbf{x}_i - \mathbf{A}\begin{pmatrix}\mathbf{p}_i\\1\end{pmatrix}. \tag{5.7}$$

These residuals are a pure linear function in the unknown coefficients of $\mathbf{A}$ and $\mathbf{b}$, so minimizing the squares leads to a *homogeneous* linear system. We compute the parameters using the singular-value decomposition (SVD) of the resulting system by taking the singular vector corresponding to the smallest singular value.

Based on validation we have found that the best results are achieved by optimizing the non-linear function, however, using the values computed with the SVD for initialization.

5.4.2 Selecting the Fixations from Calibration

During calibration, observers are asked to direct their gaze at the illuminated calibration markers. This usually leads to more than one fixation per calibration target. A common strategy among manufacturers of eye tracking devices is to select the fixations that lead to smallest residual in the estimated mapping function. We believe this approach is questionable, as it is based on the unfounded assumption that the mathematical mapping is an accurate model of the real world behavior.

We base our selection on the idea that in repeated presentation of the same calibration target, accurate fixation should likely reappear, while fixations that are slightly off-target should be independently distributed and are unlikely to be repeated. Our protocol consists of repeating the calibration procedure, with the main idea of having data to validate the estimated mapping. We use the validation cycle to compute distance between fixations for corresponding calibration targets and select the pair with the smallest euclidean distance. Formally, let $\mathbf{p}_i^j, j \in \{0, 1, \ldots\}$ be the pupil positions for calibration target with index i in the calibration phase, and $\mathbf{q}_i^k, k \in \{0, 1, \ldots\}$ the data from the validation phase. Then we select the pair

$$\underset{j,k}{\operatorname{argmin}} \|\mathbf{p}_i^j - \mathbf{q}_i^k\| \tag{5.8}$$

The reported precision of EyeLink 1000 is 0.1° root mean square (RMS)[‡] but there is no measured precision in the camera coordinates. In our implementation, we use four times the standard deviation of raw eye samples within a fixation as the threshold.

5.4.3 Error

We estimate a mapping from the pupil positions in calibration, selected as explained above, and the corresponding locations of the calibration targets. We then estimate an *error* for this mapping by taking the fixation data for the validation session. Again, this is based on the above selection. The mapped pupil position and the known calibration target yield a sequence of error vectors $\mathbf{v}_i$. We use this set of vectors to generate a first order model of the error for this mapping.

Our assumption is that the error should really grow linearly with the distance to the observer. Based on this idea we suggest to consider the error per unit distance (from the observer). For this we divide the error vectors by the distance of the corresponding target:

$$\mathbf{v}'_i = \frac{1}{\mathbf{z}_i}\left(\mathbf{x}_i - \frac{\mathbf{A}\begin{pmatrix}\mathbf{p}_i\\1\end{pmatrix}}{\mathbf{b}\begin{pmatrix}\mathbf{p}_i\\1\end{pmatrix}}\right). \tag{5.9}$$

Here, z_i is the depth value of x_i. Let m be the number of scaled error vectors (this number is 20 in most cases). Then compute the mean $= m^{-1}\sum_i \mathbf{v}'_i$ and covariance matrix

$$\mathbf{C} = \frac{1}{m}\sum_i (\mathbf{v}'_i -)(\mathbf{v}'_i -)^\mathsf{T} \tag{5.10}$$

for the mapping. The eigendecomposition of this matrix allows us to define an *error ellipsoid*:

$$\mathbf{C} = \mathbf{Q}\Lambda\mathbf{Q}^\mathsf{T} = \mathbf{Q}\Lambda^{1/2}\,\Lambda^{1/2}\mathbf{Q}^\mathsf{T} = \mathbf{M}\mathbf{M}^\mathsf{T} \tag{5.11}$$

where the matrix $\mathbf{M}$ contains the semi-axes of the error ellipsoid.

Both, the mean and the error ellipsoid need to be understood as functions of the distance to the observer, since we have defined them based on first dividing by depth. Putting everything together, the mean and error ellipsoid are defined as

$$z\mu, \quad \sigma z\mathbf{M}. \tag{5.12}$$

[‡]EyeLink 1000 User manual http://sr-research.jp/support/EyeLink%201000%20User%20Manual%201.5.0.pdf

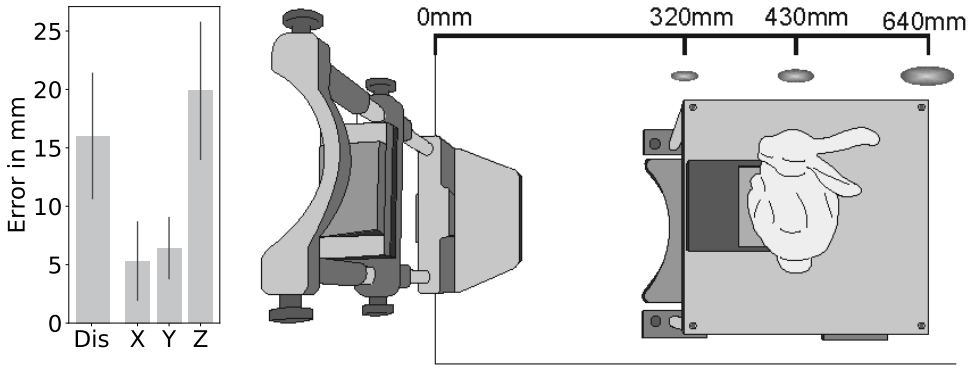


Figure 5.7: Mapping accuracy. Averaged errors measured in *mm* together with the mean absolute errors in x, y, z direction are plotted on the left. X is the horizontal direction, Y the vertical direction and Z points to the depth direction. Error ellipsoids are visualized on the right in a top view of the experimental setup when bunny is used as the stimulus.

The depth z can be taken either from the calibration targets when we want to evaluate the quality of the estimated mapping, or from the estimated viewing point by applying the mapping to the pupil positions. With σ we can adjust the size of the ellipsoid to account for a desired confidence that the ellipse contains the observed points in the validation. It is common to assume a chi-squared distribution, so we can compute the confidence interval using the cumulative chi-squared distribution for three dimensions applied to σ^2. We choose $\sigma = 2$, corresponding to an approximately 75% confidence interval.

5.4.4 Accuracy of the Mapping

We use the smallest singular vector of the system of linear equations described in equation (5.7) as the initialization and further optimize the solution with Ceres solver. Applying our data to the mapping procedure reveals results that are on a par with or better than other results reported in the literature. The averaged distance between estimated positions and target points is 16.02 mm ($SD = 5.42$), with the largest inaccuracy in depth. The mean absolute residuals in horizontal, vertical, and depth direction are 5.31 mm, 19.88 mm, 6.42 mm respectively (corresponding *SDs* are 3.41, 5.92, 2.67). The mean absolute residual per mm distance over all participants is 0.050 ($SD = 0.016$). This translates to a mean absolute error of 15.03 mm at 300 mm distance and 25.05 mm at 500 mm distance (see Figure 5.7 for a comparison with bunny). The error ellipsoid for the 75% confidence interval has a mean semi-axes length of 0.106, 0.027, 0.037 per mm distance (corresponding *SDs* are 0.076, 0.021, 0.031).

Accuracy in the planes orthogonal to the dominant view direction is comparable to accuracy reported for eye tracking experiments on displays, only that the mapping we compute

for 3D needs to accommodate the potential variation of this mapping along the depth axis. Our numbers are consistent with video-based eye tracking experiments – where we would stress that numbers provided by manufacturers of eye tracking devices are usually based on the residuals from the fitting procedure and not from independently collected data. This way of reporting the data is highly dependent on the degrees of freedom in the model and fails to account for the inaccuracy of repeat fixations for the same target.

The error in depth is significantly larger. This is to be expected because of the small interocular base line relative to the distance of the stimulus. It is difficult to find meaningful points of comparison, because the majority of 3D eye tracking experiments are done either using some type of 3D display (e. g. red-green glasses[67] or stereoscopic displays[267]) or they operate on a single plane[178]. This may lead to slightly different results for relating vergence to positions in 3D because vergence is controlled not just by binocular disparity but also other depth cues[264,274]. Gutierrez Mlot et al.[87] appear to fit a series of mappings for stimuli presented at varying depth and then report the error in depth for each of them. This would mean, their mappings are conditioned on estimating depth around a fixed value, while the mapping we generate applies generally to all depths at once. Nonetheless our numbers are comparable.

We believe using the error ellipsoid is the correct approach from a statistical point of view (see below) for counting the number of valid fixations. One may argue that very small and very large errors lead to unintuitive results: for an observer with a calibration that turned out to be highly accurate on the validation, the error ellipses are small. This means that the fixations of highly accurate observers are counted as being on the surface only when they are very close to the surface, which is implausible given the sources of error influencing the absolute positional accuracy of our setup. Conversely, observers with a large deviation between calibration and validation get assigned to very large ellipses, which tend to intersect the surface almost regardless of their position in space. Out of that perspective, one might want to also check how ellipses of constant size intersect the surface. For this, we adjust the longest semi-axis of the unit-distance ellipsoid to a fixed value in the interval $[.01, .15]$. These values translate to the longest semi-axes of 4 mm - 60 mm at the target distance of 400 mm. Keep in mind that the longest axis is usually along the depth direction and that errors on the order of 10 mm - 50 mm have to be accepted based on the accuracy of the eye tracker.

As we provide all the data to the public, we are certain the inevitable minor problems that still remain will soon be discovered and the data adjusted accordingly.

5.5 Analysis

We base the analysis on *gaze density maps*, generated from fixations on the surface of the object. For this we interpret the Gaussian error distribution of an individual fixation as a density and restrict it to the surface, and then sum over the fixations. We consider different sets of fix-

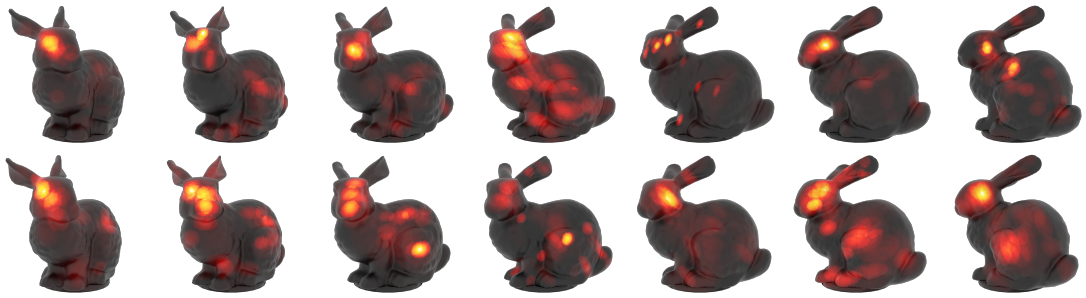


Figure 5.8: Gaze density maps for the individual conditions resulting by assigning Gaussian probability density functions over the volume to each fixation and then combining them using the relative durations as probabilities. The volumetric functions are sampled on the surface and then used to assigned color values. Columns correspond to the 7 viewing directions, upper row shows the results for ABS (slightly glossy), lower row for sandstone (diffuse).

ations to account for different assumptions. The resulting density maps are compared using Bhattacharyya distance. We perform several analyses on a per-object basis: first, we compare pairs of observers to find out if the variability of per-observer gaze densities is smaller within conditions than across conditions. Then, we analyze the dependence of gaze density maps on the conditions (viewing direction and material), i. e., does gaze behavior change for different viewing directions or materials? Lastly, we provide a visualization of regions that are attended across conditions.

5.5.1 Generating Gaze Density Maps on Objects

It is common to aggregate fixations into gaze density maps. For this, each fixation is associated with a density function, and the density functions are summed up over the relevant fixations, weighted by the duration of the fixations [21,131].

Based on the error analysis in the preceding section, we model the distribution of an individual fixation as a Gaussian in space: given the unit distance mean μ and error ellipsoid $\mathbf{M}$ for an observer and fixation position $\mathbf{x}$ with duration t computed from the eye tracking sequence, we define the distribution as

$$\frac{t}{|\mathbf{M}|} \exp\left(-\sigma^2(\mathbf{x} - x_2)^\mathsf{T}\mathbf{M}^\mathsf{T}\mathbf{M}(\mathbf{x} - x_2)\right) \tag{5.13}$$

The normalization factor $t/|\mathbf{M}|$ accounts for the fixation duration and volume of the ellipsoid, such that the resulting distribution integrates to a fixed constant proportional to t. Note that the volume of the ellipsoid is proportional to the determinant of $\mathbf{M}$ and that it exhibits the error of the mapping. Larger error, i. e., larger volume, should not result in more weight being given to a fixation.

To map this distribution over $\mathbb{R}^3$ onto the surface we take the *restriction*: we consider the values of the distribution in space only in the positions of the surface. Since the surface is given as a mesh in our case, we sample the values in the vertices. Vertices are only considered if they are within the 75% confidence interval. This interval defines an ellipsoid in space. To effectively collect the vertices in this ellipsoid we use an axis-aligned-bounding-box-tree[82] and filter vertices in the axis aligned bounding box around the ellipsoid[234]. The density map resulting from the fixations is stored as a vector $\mathbf{f} \in \mathbb{R}^{\mathcal{V}}$, where $\mathcal{V}$ is the number of vertices in the mesh representing the stimulus object.

Several fixations are combined into one gaze density map on the surface simply by adding the values in the vertices, i. e., the gaze density representation results from fixations $\mathbf{f}_i$ as $\sum_i \mathbf{f}_i$. The density function on the surface is modeled as piecewise constant. This means, we need a measure of area that is associated to each vertex. We take the barycentric area measure[191], and denote the diagonal matrix of vertex areas as $\mathbf{A}$. The aggregated gaze density representation is normalized, so that the density integrates to one over the surface. Based on our model assumption, the integrated gazed density is the result of multiplying with the area matrix $\mathbf{A}$ and then summing up the vertex values. So the normalized gaze density map resulting from a set of fixations $\mathbf{f}_i$ is

$$\mathbf{g} = \frac{\mathbf{A}\left(\sum_i \mathbf{f}_i\right)}{\left\|\mathbf{A}\left(\sum_i \mathbf{f}_i\right)\right\|_1}, \tag{5.14}$$

where the 1-norm $\|\cdot\|_1$ implements the summation over vertices.

Naturally we combine fixation data of the same condition, i. e., the same view on the same stimulus made out of the same material. Figure 5.8 provides a color coded visualization of the gaze density maps for the 14 conditions of the Bunny-object used as stimulus. For color coding we use a perceptually uniform heat map from the color maps provided by Kovesi[153].

5.5.2 Measuring and Visualizing the Distance of Gaze Distributions

In order to analyze the *dependence* on the conditions we need a way to compare different gaze distribution functions. We suggest to use the Bhattacharyya distance[4]. Let g, g' be two continuous densities, then distance is defined as $-\log \int \sqrt{gg'}$. This means the densities are multiplied in each point in the domain, then the square root is taken in each point, end the resulting function is integrated over the domain. For the discrete model we define the *similarity* vector of two (normalized) gaze density maps $\mathbf{g}, \mathbf{g}'$ as

$$\mathbf{s}(\mathbf{g}, \mathbf{g}') = \left(\sqrt{g_0 g'_0}, \sqrt{g_1 g'_1}, \ldots\right)^{\mathsf{T}} \in \mathbb{R}^{\mathcal{V}}, \tag{5.15}$$

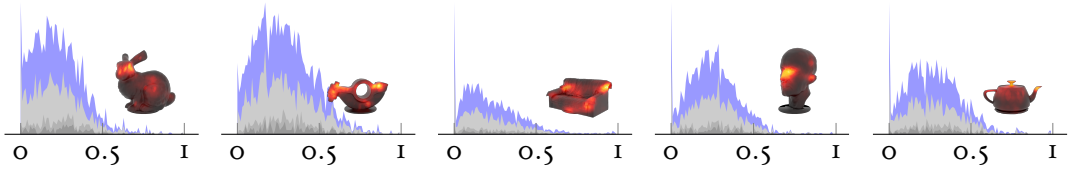


Figure 5.9: Distributions of the distance between pairs of gaze density maps (computed as Bhattacharyya distance) per stimulus object. The blue distribution contains all possible pairs. The gray distributions are the subsets of pairs that belong to the same condition, where we distinguish between same material, same direction, and same material and direction. The distributions appear to be rather similar, suggesting that the inter-observer variation of gaze density maps is generally high.

encoding the point-wise similarity in each vertex. This representation allows us to write the distance as

$$d(\mathbf{g}, \mathbf{g}') = -\log \|\mathbf{s}(\mathbf{g}, \mathbf{g}')\|_1, \tag{5.16}$$

where the 1-norm $\|\cdot\|_1$ is a discrete version of integrating the piecewise constant function defined in the vertices over the surface.

We prefer the Bhattacharyya distance as a measure over other possible ways for comparing gaze distribution functions because it results in large distance if fixations are disjoint from each other. What is particularly nice is that $\mathbf{s}(\mathbf{g}, \mathbf{g}')$ in itself nicely visualizes *why* two functions are similar, if they are. Only regions where *both* gaze densities are likely to contain fixations will have non-zero values.

The concept of similarity can be extended to more than two gaze densities: Matusita[187] introduced a measure of affinity that is based on the geometric mean of the densities (see also Toussaint[257]). In our context this means we extend the similarity representation to a set of m gaze density maps $\mathbf{g}^0, \ldots, \mathbf{g}^{m-1}$ as

$$\mathbf{s}(\mathbf{g}^0, \ldots, \mathbf{g}^{m-1}) = \begin{pmatrix} (g_0^0 \cdot \ldots \cdot \mathbf{g}_0^{m-1})^{1/m} \\ (g_1^0 \cdot \ldots \cdot \mathbf{g}_1^{m-1})^{1/m} \\ \vdots \end{pmatrix}. \tag{5.17}$$

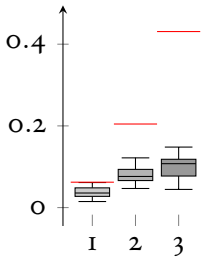
In analogy to $\mathbf{s}(\mathbf{g}, \mathbf{g}')$, this extension can be used to visualize the regions that have been attended to in all gaze patterns, i. e., it highlights stable surface features. The sum of the values in the vector representation provides a measure of similarity among the gaze distributions.

5.5.3 Inter-observer Variation

Wang et al.[270] have provided evidence indicating that the variation across observers is smaller for the same object as stimulus than for different objects. Here we refine this question to

the variation for the same object as stimulus, but different viewing conditions. Specifically we ask: Is the difference among different observers looking at the same object in the same condition smaller than looking at the same object in different conditions?

To do this we compute all pairwise differences of two observers on the same stimulus. There are 70 observers, resulting in $\binom{70}{2} = 2415$ pairs for each object. For each object, we distinguish the 7 viewing directions and 2 materials. We consider three classes: 1) the 14 different conditions resulting from directions and materials, 2) the 7 conditions differentiating the viewing direction, but ignoring the difference in material, and 3) the 2 material conditions, ignoring the viewing direction. Figure 5.9 shows the resulting distributions for a subset of the stimulus objects. The distribution in blue shows all pairs, independent of condition. The three distributions in gray are pairs that are limited so that both observers are within the same class, corresponding to the classes mentioned above. Visual inspection suggests that the distributions are similar, meaning the distance between gaze density maps of two observers is *not* smaller for the same condition.



To test this claim statistically we apply the Kolmogorov-Smirnov test on the pairs within one of the conditions defined by the three classes vs. the distribution of all pairs. The inset to the right shows the resulting KS test statistic for the same material (1), same direction (2), and same material and direction (3). The red lines illustrate the threshold for significance at the $p = 0.05$-level. None of the within class distributions differ significantly from the distribution of all pairs.

5.5.4 Dependence on view direction

As the inter-observer variation is high, we analyze the dependence on direction by considering all fixations for one condition, both with and without considering the difference in material. This means we are generating three different sets of gaze density maps $\mathbf{g}(\varphi), \mathbf{g}_a(\varphi), \mathbf{g}_s(\varphi)$, where the subscripts a and s identify the materials ABS and SANDSTONE, and the parameter φ takes on discrete values for the seven viewing directions.

The following analysis applies identically to the three sets–we describe it only for the set $\mathbf{g}(\varphi)$. We compute all $\binom{7}{2} = 21$ differences between pairs $\mathbf{g}(\varphi), \mathbf{g}(\psi), \varphi \neq \psi$. The resulting values are illustrated in Figure 5.10, in the form of a triangular matrix. We are asking: is the difference of the gaze density maps dependent on the pair or, more specifically, is the distance smaller for small differences $|\varphi - \psi|$ in viewing direction and growing for larger such differences? In order to answer this question we perform linear regression on the data points $(\varphi, \psi, d(\mathbf{g}(\varphi), \mathbf{g}(\psi))$. The null hypothesis is that linear regressor is flat, i.e., that the fitted plane has normal $(0, 0, 1)$. The plane normal is found by generating the co-variance matrix of

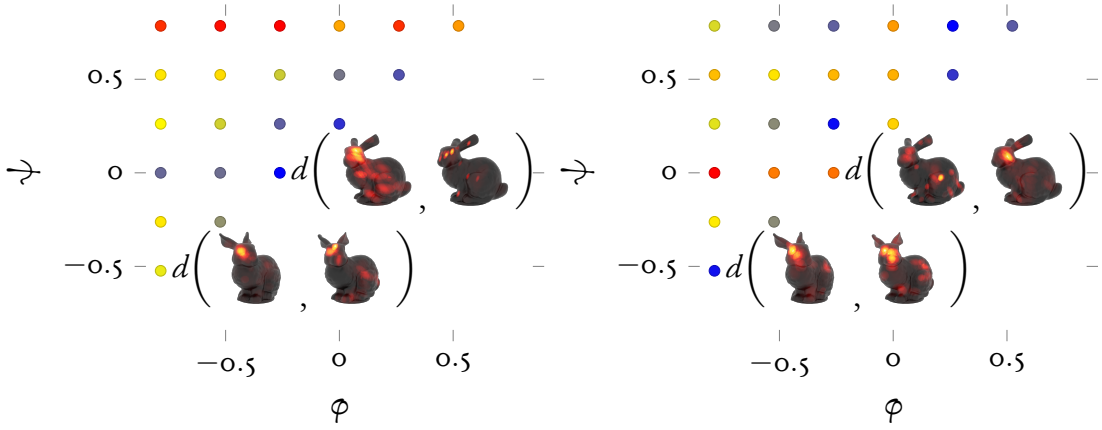


Figure 5.10: The distances between two gaze density maps for different viewing directions form a symmetric matrix. We consider the upper half of the matrix and fit a linear model to the distance. We then ask if the linear model has a significant tilt away from the diagonal, meaning that larger angular distances result in large distances between the gaze density maps. The two materials are considered separately (upper and lower illustration), as well as combined (not shown here). The distance of the gaze densities is color coded, ranging from blue for small distance to red for large distance. Note the similar trend in the data, but different variance.

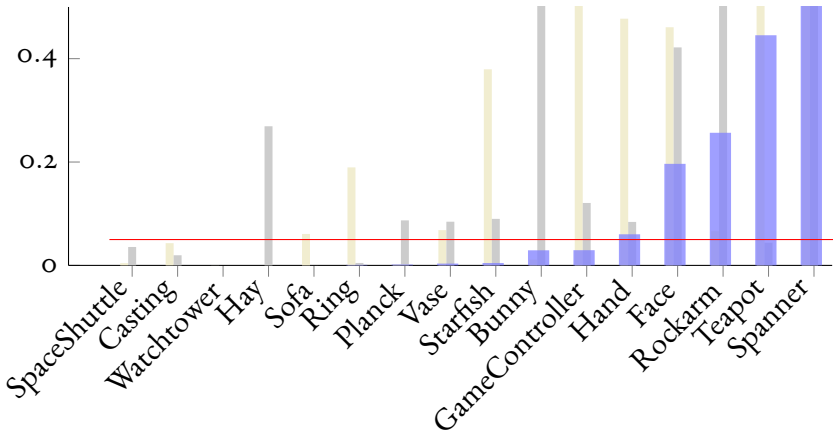


Figure 5.11: The red line shows the p-value of the linear regressor exhibiting a gradient in the direction of increasing angular difference. Blue bars indicate the result for combining the fixations from the two material conditions, the lighter bars depict restrictions to one material.

the 21 points and then taking the eigenvector corresponding to the smallest eigenvalue. The eigenvalue provides the variance σ, and the standard error is then $\sigma/\sqrt{n}$, where $n = 21$.

We wish to understand if the resulting normal (with standard error) is significantly different from $(0, 0, 1)$. For this we need to compute how likely it is to observe a tilted normal by chance – we need a probability distribution for the plane normals. This probability distribution is likely not available analytically, so we sample it: we take the same set of fixations from the 70 observers (35 in case we restrict to one of the two materials), and split it randomly into 7 groups of 10 (5) observers each. We combine the fixations and consider them as set of 7 'directions' (only now they are independent of the directions used in the experiment). We perform the linear regression on the distances of the 21 pairs of 'directions'. This process is done to generate 10,000 samples of random distance matrices similar to the ones illustrated in Figure 5.10, emanating from the same distribution underlying the distances among the 7 directions. We find that, as expected, the mean normal of this distribution is numerically close to $(0, 0, 1)$.

Based on the sampled distribution of normals and standard errors, we can then provide a significance level for the data generated from $\mathbf{g}(\varphi)$. For this we consider the direction and magnitude of the component of the normal orthogonal to $(0, 0, 1)^\mathsf{T}$ [64]. We derive a one-dimensional probability distribution from the random sample for the magnitude of these vectors. In this way we can test if the tilt of the plane is significant, meaning it is unlikely the result of a chance event. In addition we check that the direction is consistent with our assumption that larger difference in viewing angle results in larger distances. This test could be interpreted as first performing a two-tailed test and then restricting to one of the tails, so it is more conservative.

Figure 5.11 shows the results. We find that for 11 out of 16 objects there is a statistically significant dependence of distance on difference in angle at the $p = 0.05$-level. When we consider the material, fewer objects reach the significance level. Inspecting the linear regression results reveals that the planes are quite similar for the three different cases, only the smaller amount of data leads to larger standard errors for the individual materials (see, for example, the illustration for the Bunny in Figure 5.10 and the corresponding significance levels in Figure 5.11). Interestingly, some objects show a significantly flat fit, suggesting that their *independence* of difference in viewing direction is not just coincidence and that observers attend to the same features in different views.

5.5.5 Material Dependence

In the same way that we analyzed dependence on viewing direction, we now examine dependence on material. We ask if the difference in viewing behavior for different materials is in any way significantly different. For this test, we generate 7 different sets of gaze density maps

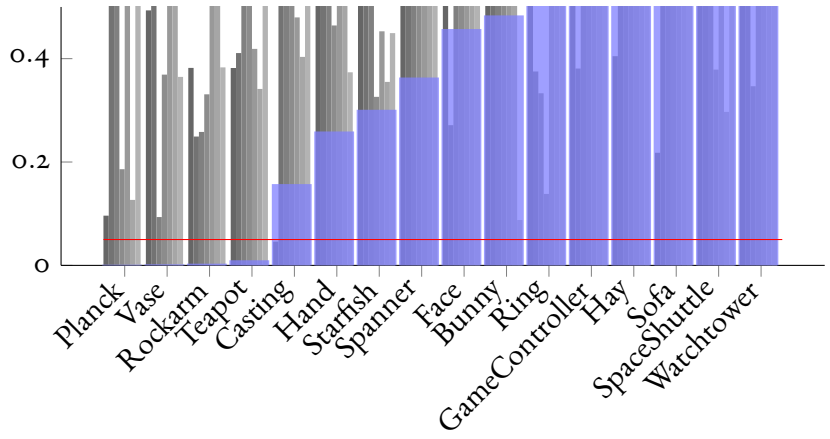


Figure 5.12: Significance test of whether a split along the material condition leads to large difference in the gaze density maps than an arbitrary split. The small bars show the results for the individual viewing directions and the large bar shows the results combining all views.

$\mathbf{g}^0(m), \mathbf{g}^1(m), \ldots, \mathbf{g}^6(m)$, where m takes on only two different values, and we consider all viewing directions combined or each separately. For each of the 7 sets there is only one difference that can be computed. Without considering viewing direction this is the difference of two sets stemming from 35 observers each, and in the other case it is the sets from 5 observers. As above, we are generating a random distribution for the difference values, by either considering all data of different directions and randomly splitting it into two sets of 35 observers each, or considering the data from one viewing direction and randomly splitting into 5 each. Then we can directly compute the rank of true value in the distribution of randomly generated ones to provide the significance.

Figure 5.12 shows the result of this test. We find that 4 models exhibit a significant difference between gaze density maps for the different materials when the viewing direction is ignored. In all other cases the dependence on material is not statistically significant.

5.5.6 Stable Features

We wonder whether any surface features are consistently attended to by the observers across the different conditions. Based on our analysis so far, we drop the dependence on material as a condition and only consider viewing direction. This means, for each object we consider the 7 gaze density maps $\mathbf{g}(\varphi)$ consisting of the data from 10 observers each.

We may consider a region on the surface and ask whether it has been attended from 3 or more viewing directions. This can be estimated using the Matusita affinity $\|\mathbf{s}(\mathbf{g}(\varphi), \ldots)\|_1$ for the set of gaze density maps of the different viewing conditions, restricted to the region of

Figure 5.13: The geometric mean of all conditions combined, illustrating regions on the surface attended to in all views and for both materials. The selected shapes have the highest similarity measure in our data set.

Figure 5.14: Visualizing features that are stable across a variation in viewing direction. Top row shows the combination of heat maps that result from considering the geometric means of 3 adjacent viewing directions, i. e., features that have been attended to consistently within 30°. Results in the lower row are based on requiring that features appear consistently across 60° viewing angle. Note how features retained for the larger insensitivity to viewing direction are exclusively of semantic nature.

interest. Note that the vector $\mathbf{s}$ contains large values exactly for those regions that have high affinity. So we might as well inspect the affinity vector over all of the surface.

First, we compute the affinity for the set of all viewing directions, showing which surface regions are attended to from all directions. The 4 objects with overall largest affinity are depicted in Figure 5.13, for most objects the resulting affinity is zero. The range of views covering 90 degrees is apparently too wide for features to be consistently attended to.

Consequently, we reduce the desired range of views to either 30 or 60 degrees. This means we compute the affinity vector for sets of 3 or 5 views. For visualization purposes we combine the resulting affinity vectors. The result is depicted in Figure 5.14, showing that stability across 30 degrees works quite well, yet stability for 60 degrees leaves only very few regions consistently attended. From a visual inspection of the visualization we would speculate that stable features contain more semantic information, such as the eyes of the Bunny, the Face, and the windows for the watchtower, or the points of symmetry for the ring and the starfish.

5.6 Computational Model of Gaze Density

The idea of geometric saliency has been used in applications probably because of the existence of computational models, i. e., the possibility to guess the gaze density map for a given 3D shape based on the geometry alone. Here we try to develop such a computational model based on the data we have collected and using the currently popular convolutional neural networks (CNN). This computational model can then be used in applications.

The dependence of salient features on viewing direction suggests to predict saliency based on view-dependent information unlike common geometric saliency models, which are independent of viewing conditions[167,244].

In particular, the prediction of saliency is based on the surface normals (relative to the view coordinate system) as well as the depth information of objects. With this information as input, we train two different models for gaze density estimation.

1. All shapes in different viewing directions are used to predict the gaze density of a shape for a new viewing direction.

2. Shapes in different viewing directions are used to predict the gaze density of a new shape.

In the first model the computational model only needs to predict a new viewing direction, having information on the viewing behavior for the shape from other viewing directions. The second model analyzes generalization towards unknown shapes.

5.6.1 CNN Model and Training

For both models we train a simple 5-layer CNN consisting of three convolutional layers followed by a fully connected and upsampling layer. The network layout and further details are given in Figure 5.15.

The training input images contain the normals and depth map of the sample objects. The first three channels of the input image represent the surface normals at each (visible) point of the object, the last channel represents the depth value of the underlying surface points. The output of the network is the predicted gaze density map for different viewing directions of an object.

As a loss function we use the mean squared error (MSE) on the predicted gaze density. We employ a dropout layer and early stopping to prevent overfitting. The overall validation loss is decreasing during the first 50 training epochs due to the rather small amount of training data. The complete dataset with 224 samples (of 16 objects with 14 view directions each) is evaluated with 4-fold cross-validation. For the first trained model, the dataset was split according to the view directions. In each cross validation run, 3 view directions were used as the test set (48

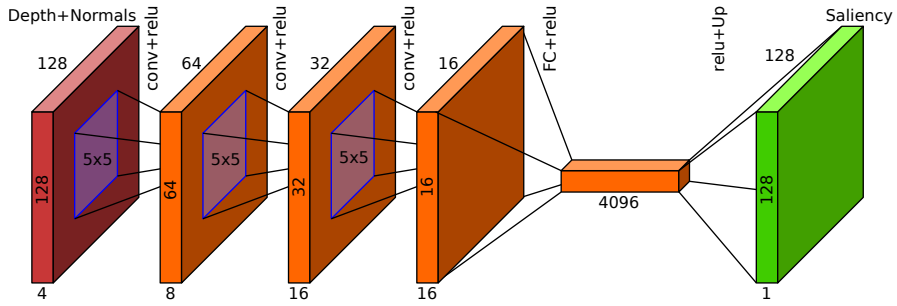


Figure 5.15: Saliency prediction network architecture. The input to the network is an image of size $128 \times 128 \times 4$. The network consists of 3 convolutional layers with filter kernels of size 5×5 with stride 2 and padding 2 and ReLU as activation functions. This follows a fully connected layer and a subsequent upsampling layer to produce a resulting saliency map of size $128 \times 128 \times 1$.

samples in total). The remaining samples were used as the training set (176 samples). For the second model, the dataset was split according to object categories. In this case, the samples of 4 objects were used as the test set (56 samples in total) in each cross validation run. This left a total of 168 samples as the training set.

5.6.2 Gaze Density Map Prediction

The first prediction model is able to predict gaze density maps for previously unseen viewing directions. However, given the small amount of available data, it is difficult to prevent the model from overfitting. Even though we employ multiple measures to prevent overfitting, it remains unclear how well the model generalizes to completely different unseen viewing directions. Some exemplary test input images and the resulting predicted gaze density maps are depicted in Figure 5.16.

The second prediction model is trained only on a subset of the objects (12 out of 16) and is able to predict gaze density maps for the 4 unseen objects (of each cross-validation fold). This situation is similar to other generic computational models for the prediction of gaze density, such as mesh saliency[167]. We wish to compare the trained CNN to mesh saliency, however, comparing the MSE would be unfair, as our model is specifically trained to minimize this error, while mesh saliency only promises to provide qualitative results. Consequently, we base the comparison only on the relative ordering of the values. Specifically, we use Kendall's rank correlation coefficient[139], which measures the correlation between two variables in the range $-1 \leq \tau \leq 1$. We rank the estimated gaze density maps with the ground truth gaze density maps and compare them to the rank of the calculated saliency maps with respect to the ground truth.

The mean τ coefficient for the CNN-predicted gaze density is 0.40 (with all p-values below

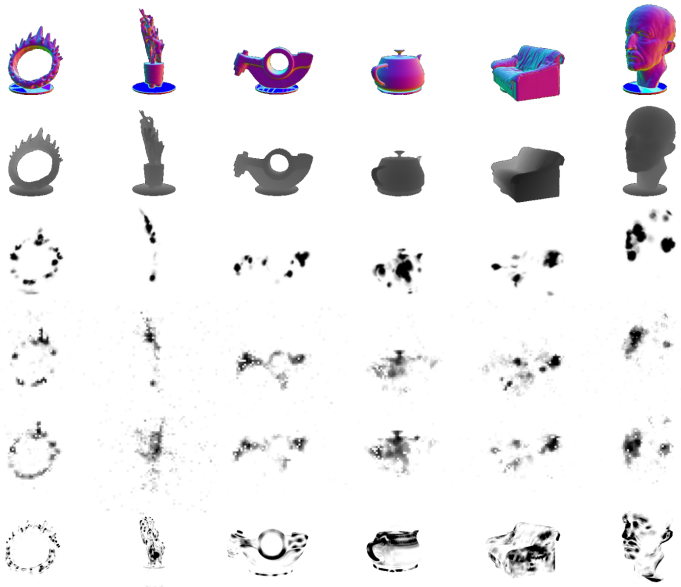


Figure 5.16: CNN prediction results for different objects (from top to bottom): normal and depth channels of the input image, ground truth gaze density map, predicted gaze density map from the trained model for unseen viewing direction prediction, predicted gaze density map for unseen objects, calculated saliency map.

0.01), while mesh saliency yields a mean τ coefficient of 0.13 (with all but 4 p-values below 0.01). These results indicate that both computational models are positively correlated with the ground truth gaze density maps in a significant way, yet the correlation for our CNN-model is much higher.

The resulting MSE (averaged over the cross-validation) for the first model (unseen view direction prediction) is 189.5. For the second model the training results in an averaged MSE score of 249.5.

The fact that using the shape to train the model for new views improves the prediction suggests that certain shape features cannot be learned from the geometry alone–they are likely higher level features of the shapes. We expect that more refined neural networks would result in better computational models for gaze density prediction.

5.7 Discussion

Our analysis as well as the computational model suggest some characteristics of salient features on 3D shapes, at least for a majority of object stimuli:

- There is no significant dependence of fixations on the two materials used for the stimuli.

- Salient features exhibit a tendency to be view-dependent and the ones that are stable across a wide range of views appear to be features with semantic meaning.

Both observations have consequences and deserve some discussion. The independence of fixation on the moderate gloss of the surface may seem natural, but it contradicts the idea that local contrast is the strongest low-level feature in the image function. It rather suggests that saccade targets on geometry are independent of contrast, either governed by the occulomotor system alone, or dependent on other features of the scene. On the other hand, the materials used in our experiment only differ slightly, and it would be interesting to understand the extent to which the fixations are stable across different materials and under various lighting conditions.

The dependence of salient features on viewing angle is also intuitive. The better performance of our simple CNN-model compared to mesh saliency could be due to the dependence of salient features on view direction. Not using information on the view direction should lead to reduced predictive power. We would speculate that the success of computational models is based on a bias in the commonly used shapes: relevant features almost always have larger curvature variation and thus appear as part of the salient features predicted by the model. It would be interesting to modify features with semantic meaning such that computational models fail to predict them and then see if they are still dominant in a human subject experiment. Yet how to quantify semantics still remains a topic for future study.

Note that our analysis is based on the whole viewing sequences without considering temporal changes. It would be interesting to see whether saliency of objects changes over time.

While we have made a significant effort in our experiment, involving more than 70 participants and using custom-built hardware, the data would still benefit from being based on a larger corpus. To this end, we believe that automation would help to avoid errors in setting up the individual conditions for each observer and may also increase the geometric accuracy of the presented stimulus.

We have decided to use a mapping from the 4D space of pairs of pupil positions directly to 3D. While this has led to data with few significant outliers, it did create a tendency for the fixations to have depth values that are too small. In the specific setup, we could have also intersected eye rays computed for each eye individually against the geometry of the shape. This, however, makes it more difficult to estimate which of several possible intersections along a silhouette region are the right match. It would be interesting to combine all available information, yet, we are unsure how to do this. Further studies about characteristics of eye movements in space would offer useful guidelines in this regard.

Part III

Comparing Eye Movements during Encoding to Eye Movements during Recall

Where words are restrained, the eyes often talk a great deal.
Samuel Richardson

6 The Looking-at-nothing Phenomenon and Data Acquisition

This chapter introduces the well-established looking-at-nothing (LAN) phenomenon in psychology, which studies the eye movements while looking at an empty space. It describes the phenomenon that humans usually move their eyes spontaneously as if the image were in front of them, when retrieving images from memory. Studies show that such eye movements correlate strongly with the spatial layout of the recalled content and function as memory cues facilitating the retrieval procedure. In order to study this phenomenon and explore the information contained in the eye movements, we collect a dataset and record eye movements while a photo is immediately recalled from memory. Here we briefly review the background on eye movements during mental imagery and provide the details on data acquisition. Basic statistics of the dataset are reported at the end.

6.1 The Looking-at-nothing Phenomenon

Research on eye movements during visual imagery has a long history. Early work[121,194,212] at the beginning of the twentieth century have linked eye movements during visual imagery to the corresponding mental images, as indicated by the intense activity of eye movements. Neisser[197] argued that eye movements are actively associated with the construction of a visual image, and Hebb[94] suggested that eye movements during imagery and memory retrieval are

necessary to assemble and organize "part images" into a whole visualized image.

A large amount of recent studies [25,127,128,129,159,160,223,236,247] have shown that humans spontaneously move their eyes when recalling a scene from memory and that such eye movement patterns closely resemble the spatial arrangement of the elements of the recalled scene. This effect has been demonstrated for participants who encode visual scenes and later recall those scenes while looking at an empty screen [25,126,127,129,159,247] as well as for participants who encode verbal information in association with a spatial cue and later recall such information while looking at an empty screen [223,236]. Moreover, it has been shown that participants who listen to scene descriptions also make eye movements which correspond to spatial positions from the described scene [126,127,247]. When recalling text information, imagery eye movements are associated with the situation model of the described scene and not with fixated text locations during a preceding reading phase [130].

Previous studies aimed to evaluate the functionality of eye movements to blank space, in particular whether they play any functional role during memory retrieval. Results suggest that eye movements while looking at nothing act as "spatial indices" [7,94,159,197,223] that may provide assistance in remembering the spacial layout of a scene [25,94,159,197], but whether such eye movements can facilitate memory retrieval as additional cues is still arguable [71,222]. In support of such a functional role, impaired episodic memory performance has been reported in experiments where participants are restricted to look at a fixation cross during recall [126,129,159,160]. Moreover, participants' gaze direction was manipulated towards positions on an empty screen that either overlapped or not with the original locations of the to-be-retrieved visuospatial information in [129]. Results demonstrated that the likelihood of successfully remembering increased when there was an encoding-retrieval overlap in gaze locations. Corresponding results were reported in a follow-up study, where similar gaze manipulations were used for participants who recalled verbal information that had previously been encoded in association with a particular space [236].

Imagery movements have been reported for other muscles beyond the ocular ones, and a consistent finding is that muscle activation during imagery is but a fraction of what it is during action-related activation [121,196]. Eye movements are the only muscular activity that during imagery largely replicate activity during actual perception or action.

6.2 Data Collection

We adopted the encoding-recall experimental paradigm from [129,159]. We instructed observers to encode and then to recall a set of 100 images in front of an empty screen. Fig. 6.1a shows an exemplary structure of one trial. Using a video-based eye tracker, we recorded observers' eye movements during both encoding and recall.

Participants Twenty-eight naive participants took part in the study (9 females, mean age 26±4). All participants had normal or corrected-to-normal vision and none of them had colour deficiencies. They gave their written informed consent before the experiment and their time was compensated.

Apparatus The experiment was conducted in a dark and quiet room. Participants were seated in front of a 24-inch display (resolution: 1920 × 1200 pixels; physical size: 0.52m × 0.32m; distance: 0.7m) on which image stimuli were presented. Each participant viewed and recalled 100 images binocularly, however, only the dominant eye movement was recorded with an EyeLink 1000 in remote mode at 1000Hz. Gaze point on the screen was calibrated using a standard 9-point calibration and a chin and forehead rest was used to help participants stabilize their positions.

Visual Stimuli We randomly selected 100 images from the MIT data set[132], including both indoor and outdoor scenes. In practice, an eye-tracking experiment session must have a limited amount of time. Therefore, we decided to use only 100 images (a subset of the original data set) in this study considering our experimental design. All images were presented at the center of display in their original size and the largest dimension has 1024 pixels.

Experimental Procedure We followed the standard LAN paradigm[129,159] to collect eye movements during both encoding and recall as shown in Figure 6.1a.

Observers were instructed to look closely at each image and try to remember its content. The instruction for recall was to think about the image and generate a visual representation of the content which you find interesting. Each observer had an initial round of 10 practice trials. We assumed that they had then learned the task, and after that there were no further instructions regarding the task.

The total 100 trials were divided into five blocks and for each participant, images were presented in a randomized order. At the beginning of each block of 20 trials, we calibrated the eye tracker on the display. We repeated the calibration procedure unless it achieved a good accuracy (below 0.5°) in the following validation. We had 200 eye movement sequences from each participant viewing and recall 100 images.

Memory Task At the end of the experiment, we randomly selected five from the viewed 100 images and presented them together with five new unseen images to the participants in a randomized order. We asked participants to report whether one image has been seen before. All participants could easily determine which image had been seen before, except for one single mistake made by one participant. This indicates that participants still had the image contents in their memory.

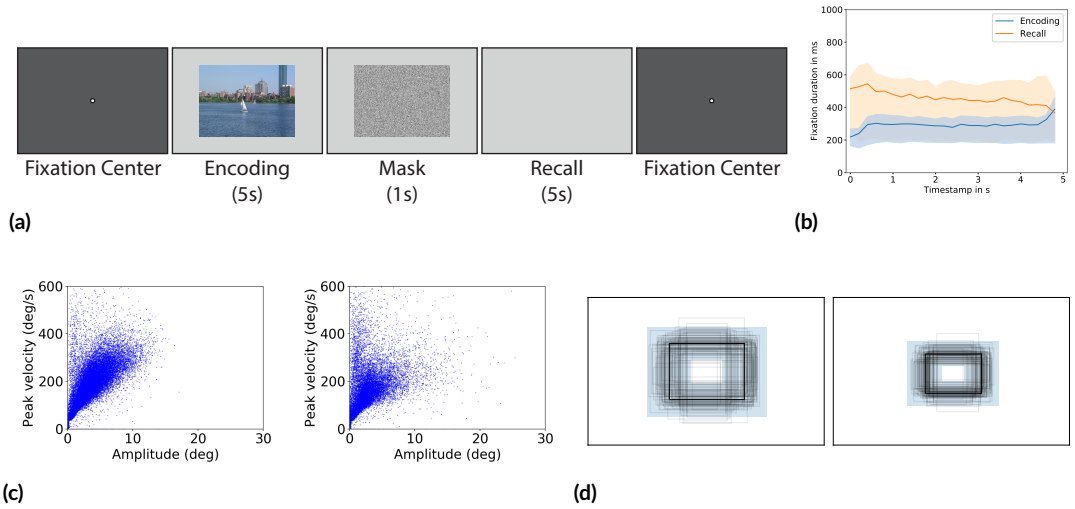


Figure 6.1: Experimental paradigm and eye movements statistics. (a) Paradigm of one single trial in the experiment. After an initial fixation (of 500ms) at the center of the display, one image stimulus was presented for 5*s*, followed by a noise mask. Briefly after that, observers were asked to recall the image they had just seen for another 5*s*. (b) Average of fixation duration in encoding (blue) and recall (orange). X-axis indicates the time and y-axis corresponds to the duration of each fixation. The black curves in the middle of each plot correspond to the average duration within the five seconds and the light coloured areas indicate the center intervals of 50 percent. (c) Main sequence diagrams during encoding (left) and recalling (right) where peak velocity is plotted as a function of the amplitude of saccades. (d) The spacial coverage of all fixations over the 100 stimuli. For each image, one bounding box of all fixations during encoding is drawn in the left and one for fixations during recall in the right. Outermost box indicates the screen size. The black box visualizes the averaged size of regions covered in each condition and the light orange areas depict one standard deviation of all coverage areas.

6.3 Eye Movement Statistics

6.3.1 Data Analysis

Each eye movement sequence, either during encoding or during recall, consisted of raw pupil positions on the screen. Collected at a sampling rate of 1000 Hz, we had slightly less than 5000 raw data in each sequence, very often with missing data due to blink. Saccades and fixations were detected from the raw data using the velocity based algorithm provided by *SR Research* (the velocity threshold = $30°/s$ and the acceleration threshold = $8000°/s^2$). Only 95% fixations from eye movements during thinking about an image were located inside of the stimuli domain (SD=8%), while 99%($SD = 1.9\%$) fixations during looking at the images were within the stimuli boundaries.

6.3.2 Characteristics of Eye Movements on Real and Mental Images

Three measurements were used to analyze the eye movement sequences: the main sequence graphs of saccades, the fixation count and duration, and the overall spatial coverage. Figure 6.1c shows the main sequence graphs of saccades during encoding (left) and recall (right), where peak velocity is plotted as a function of amplitude. Imagery saccades are shorter and slower. Fixation duration is plotted over time in Figure 6.1b. On average each eye movement sequence during encoding has 16 (SD=2.8) fixations, which has an averaged duration of 278.0*ms* (SD=73.4 ms), and each sequence during recall has 11 (SD=3.6) fixations with an averaged duration of 452.2*ms* (SD=308.0 ms). Recall sequences have less fixations than encoding sequences ($t(5.6e3) = 62.77, p < .001$, Welch's t-test) and recall fixations are longer than encoding fixations ($t(6.7e4) = -57.29, p < .001$, Welch's t-test). As indicated in previous studies[127,129], information retrieval from memory might account for the longer durations of fixations during mental imagery.

Similar to previous findings[77,126,127], we also observed a shrinkage of eye movements on mental images, as shown in Fig. 6.1d. This goes in line with the observation that eye movements during LAN are distorted due to the lack of reference frame. Unlike fixations during encoding, which are perfectly aligned with the corresponding visual content, fixations during recall very often do *not* coincide with the intended elements in the original image. Consequently, this makes it difficult to estimate the intended locations from the recall fixations alone.

7

The Mental Image Revealed by Gaze Tracking

Based on the "looking-at-nothing" phenomenon, we know that humans involuntarily move their eyes when retrieving an image from memory. This motion is often similar to actually observing the image. In this chapter, we suggest to exploit this behavior as a new modality in human computer interaction, using the motion of the eyes as a descriptor of the image. Interaction requires the user's eyes to be tracked, but no voluntary physical activity. Using the collected dataset from the controlled experiment, we develop matching techniques using machine learning to investigate if images can be discriminated based on the gaze patterns recorded while users merely recall an image. Our results indicate that image retrieval is possible with an accuracy significantly above chance. We also find that such a retrieval approach can be generalized to unseen images. Furthermore, we show that these results generalize to images not used during training of the classifier and extends to uncontrolled settings in a realistic scenario.

7.1 Introduction

Imagine that you are thinking about your vacation photos from last summer. After 5 seconds a photo appears in front of you, from that vacation, very similar to the moment you just recalled.

We believe such a seemingly magical task is possible through the use of eye movements during mental imagery. The close similarity between eye movements during perception and those during imagery appear to open up the possibility to use imagery eye movements for computational image retrieval: Pick the image where the similarity between perceptual eye movements and imagery eye movements is the largest. In view of the very robust findings on imagery eye movements, computational image retrieval based on those eye movements may at first seem like a trivial task, but there are four reported issues that comprise challenges.

- Even though eye movements when recalling an image play a functional role, they do *not* reinstate the eye movements made during encoding of the image[128,236].
- It has been shown that eye movements during imagery might be potentially driven by covert attention, which may explain the functionality of eye movements while looking at nothing[235].
- What is also known is that recalling an image with closed eyes might be preferable for some of us. This effect has not been fully understood[183,263], and it has posted a challenge in practice to record eye movements using available video-based eye trackers.
- Many previous studies consistently reported that eye movements while looking at nothing contain a large variation due to the lack of reference frame. In particular, the area spanned by eye movements during recall is scaled down comparing to the area spanned by eye movements during encoding while the visual image is onsite[25,77,126,127,128]. A clue to why some people scale down their images was given by[126,128], in which all participants were tested for working memory capacity and the *Object-Spatial Imagery and Verbal Questionnaire* (OSIVQ)[18] was used as an assessment for individual differences in object imagery, spatial imagery and verbal cognitive style. Scaling down imagery eye movements was most pronounced with participants who had high scores on spatial imagery. Eye movements during visual imagery tasks are employed to reduce cognitive resources associated with the processing of spatial information, and a weaker spatial imagery ability increases the need for those eye movements.

This image retrieval technique would only work if there is a strong similarity between the eye-movements on real versus imagined images across different viewers. Previous studies mainly used simple grid-based stimuli[129,159,236], where eye movements were discretized unnaturally in low resolution. A single picture of more complex scene was used in[126,127] but no other visual stimuli were included. It remains unclear what the degree of similarity is when encoding and recall a large amount of *natural images*. In view of the fact that such a similarity exists but its strength was not quantified, we evaluate different retrieval scenarios. In all cases

we essentially ask: how well can images be computationally discriminated from other images based on only gaze data.

The scenarios differ based on what gaze data is used as a query and what gaze data is used for matching. We consider different combinations which can be generally divided into two scenarios: in Scenario 1, we follow the idea of exploring available information contained in the data; in Scenario 2, we test a realistic setting in applications, which allows for the possibility of extension to new users.

We develop two types of retrieval algorithms for these scenarios. Restricting ourselves to using spatial histograms of the data, we consider an extended version of earth movers distance (EMD) and, at least for the scenarios that provide enough data, we also use deep neural nets. In general, we find that retrieval is feasible, albeit the data from LAN is challenging, and the resulting performance varies significantly across scenarios and observers.

Based on the promising results in a lab setting we make a first step towards a real application (see Section 7.5): we sent several participants with a mobile eye tracker to a staged 'museum' exhibiting paintings. After their tour, we ask them to recall some of the images while looking at a blank whiteboard. We find that the median rank in a classification approach is small, showing that the idea is promising for practical use.

Despite the encouraging results, our proposed new modality still faces challenges that require further investigations.

7.2 Background & Related Work

Gaze Interaction The well-known 'Midas Touch' problem[120] describes the difficulty of distinguishing between attentive gaze during exploration and communicative gaze that is used for instance for selection. Even so, gaze still offers an attractive option as input, since it is comparatively fast and accurate. Especially in hand-free interaction systems, gaze data provides an additional input modality. As an input method, gaze is used to indicate users' intention[119,158], and to type words[177] and passwords[28]. Other types of applications are based on the characteristic of eye movements captured by dwell time[195] or smooth pursuit[68]. For a more comprehensive review of eye-based interaction, we refer readers to work by Majaranta and Bulling[176].

Most similar to our endeavor are attempts at learning something about the image or the observer by simply processing the gaze in a free viewing task. Fixations have been used to select images from a collection to indicate the desired attributes of a search target[231]; or gaze data is exploited to identify fine-grained differences between object classes[135]. Comparing to previous work, where users are mostly required to *consciously* move their eyes, the proposed interaction system relies more on *unconscious* eye movements.

Brain-computer Interfaces The idea of using brain activity for direct communication with the computer is intriguing. Early attempts based on electroencephalogram (EEG) go back to 1970s[261], and first consolidations of various research efforts[275] argue that *reading the mind* will be difficult and the focus of attention should be on individuals training to steer the EEG for providing certain commands. Today EEG is used in applications such as identity authentication[190] and studies about brain responses[78]. Despite the bandwidth of communication using EEG still seems to be very limited, the trend appears to be combining continuous EEG data with other modalities[161,279].

Likewise, there have been attempts at reading a person's mental state using functional magnetic resonance imaging (fMRI)[91,133]. The task of reconstructing mental images has generated much interest and shows promising results[60,138,198,201,240]. These results are similar to what we want to achieve: guess the image one is recalling from memory.

Many of EEG based applications are restricted to binary classification. Participants are instructed to imagine moving their left or right hand[17], which needs to be repeated many times to give a sufficient signal-to-noise ratio[192]. Yet it opens up a possibility of communication for patients with aggravating conditions[161,279]. Participants in studies using fMRI are required to keep still so that acquired images can be aligned. Such restriction of movements during fMRI scanning poses another challenge (besides costs) in terms of practical usage. This chapter is based on tracking the gaze patterns during mental imagery, which is more feasible for practical purposes.

Photo Management and Retrieval Managing photos and image collections can be difficult for users (cf. Kirk et al.[148]), oftentimes requiring context-based organization and sorting. Especially retrieval of individual images can pose a significant challenge and requires advanced interfaces and interactions (cf. e.g. Kim et al.[145], Shneiderman et al.[242]). Sketches are probably the oldest way to communicate mental images. Unlike in image recall, gaze patterns are significantly distorted and incomplete compared to the real visual appearance of an object. Information in a sketch is much richer and more coherent than the data we have to deal with. Nevertheless sketches have been shown to work well for retrieving images[65,230,278]. In contrast to sketching, we explore how users can retrieve an image without explicit interaction, which can be beneficial in situations where sketching is not available (e.g. because the hands are occupied).

7.3 Retrieval

Retrieval is based on two steps: first the raw eye tracking data is turned into a more compact representation; then, we try to assign a query representation to one of the 100 image classes. For the second part, we consider different approaches, either measuring the distance between

pairs of data representations using an appropriate distance metric, or using deep neural nets. Depending on what gaze data is used for query and matching, we have different retrieval scenarios. In particular, we consider the following combinations:

1. Within the same condition: the query gaze data and the gaze data used for matching are both taken from the observers while

   (a) looking at the images or

   (b) looking at nothing.

   In these situations we always consider matching across *different observers*. This means that when querying with data from a specific observer, his/her data is *not* included for matching.

2. Across conditions: the query data is taken from an observer looking at nothing and it is matched against gaze data from looking at the images. Here we differentiate between matching the query data

   (a) against the gaze data from all other observers or

   (b) against the gaze data from the same observer.

Scenario 1 is meant to establish an idea about the available information contained in the data. Scenario 2, we believe, is a realistic setting in applications, where we want to use gaze data from imagining an image for retrieval while collecting viewing data from other users.

7.3.1 Data Representation

We have experimented with a number of different representations and were unable to identify a representation that clearly outperforms others in terms of retrieval performance. For reasons of a clear exposition we use *spatial histograms*, as they consistently performed well, are simple to explain, and are a natural input for neural nets.

A spatial histogram is defined on a regular lattice with $k = m \times n$ cells. We denote the centers of the gird cells as $\{\mathbf{c}_i\}, 0 \leq i < k$. To each cell we associate a weight w_i. The weights are computed based on the *number* of *raw* gaze samples (returned from the eye tracker) that fall into each cell, i.e., that are closer to $\mathbf{c}_i$ than to any other center $\mathbf{c}_j, j \neq i$. Because of the grid structure, we may consider the set of weights $\mathbf{w}$ as intensities of a discrete image. This image is sometimes called *heat map* in the literature on eye tracking.

We are aware that it is common to first extract fixations from the raw gaze data before further processing. We have opted not to do this, as (1) fixations are not defined in a way that could be universally accepted, (2) the spatial histogram is dominated by the fixations

anyways, and (3) saccade data may or may not be useful in classification. In an initial test using MultiMatch[54], a tool developed in the eye tracking community for comparing gaze sequences based on fixation data, we found that the retrieval tasks we are interested in performed only at chance level.

It has been observed that the sequence information could be useful[135] but we could not reproduce this observation. This makes sense, as it has been observed that the order of where people look during imagery is not identical to the order during looking at the image[126].

7.3.2 Distance Measure

The nature of the data suggests that simple measures of distance are unable to capture the desired correspondences. Indeed, we found that simple Euclidean distances on the weight vectors of spatial histograms, i.e. $d(\mathbf{w}, \mathbf{w}') = \|\mathbf{w} - \mathbf{w}'\|$ are not promising. The reasons for this are that the fixations in one gaze sequence are commonly a subset of the fixations in the other sequence. Moreover, the data for mental images is spatially distorted and contains elements that are unrelated to the original stimulus.

Research on fixations and image saliency has investigated various ways to measure the agreement (or disagreement) of real fixations with computational predictions[131]. Bylinskii et al.[30] observe that the Earth Mover's Distance (EMD)[227] tolerates some positional inaccuracy. More generally, EMD is well known to be robust to deformation and outliers for comparing images.

In our context the weights $\mathbf{w}$ and $\mathbf{w}'$ are representative for the gaze data. The *flows* $\mathbf{F} = \{f_{ij}\}$ describe how much of the weight w_i is matched to w'_j. Based on the idea that $\mathbf{w}'$ is potentially a subset of $\mathbf{w}$ we scale the weight vectors so that

$$1 = \|\mathbf{w}\|_1 \geq \|\mathbf{w}'\|_1 = \sigma \tag{7.1}$$

The flow is constrained to relate all of the weights in $\mathbf{w}'$, but not more than w_i, to i:

$$\sum_i f_{ij} = w'_j, \quad \sum_j f_{ij} \leq w_i. \tag{7.2}$$

This implies $\sum_{ij} f_{ij} = \|\mathbf{w}'\|$. Among the flows satisfying these constraints (i.e., the flows that completely match $\mathbf{w}'$ to a part of $\mathbf{w}$) one wants to minimize the flow. Therefore we need to specify how far apart the elements i and j are by a distance *metric* $\{d_{ij}\}$. The resulting minimization is:

$$\underset{\mathbf{F}}{\operatorname{argmin}} \sum_i \sum_j f_{ij} d_{ij}. \tag{7.3}$$

In many cases it is natural to use Euclidean or squared Euclidean distance between the cell

centers $\mathbf{c}_i$ and $\mathbf{c}_j$ to define d_{ij}. On the other hand, we cannot be sure that the space in which the recall sequences 'live' is aligned with other such spaces, or the fixed reference frame of the images. In line of this, learning has been used to optimize the metric $\{d_{ij}\}$ for EMD [48,266]. Our idea in this context is to restrict the potential mapping to be affine. This means we define the distances to be

$$d_{ij} = \|\mathbf{c}_i - \mathbf{T}\mathbf{c}_j\|_2^2 \tag{7.4}$$

where $\mathbf{T}$ is an arbitrary affine transformation. We optimize for the flows $\mathbf{F}$ and the transformation $\mathbf{T}$ in alternating fashion. With fixed transformation $\mathbf{T}$ this is the standard EMD problem and can be efficiently computed for the size of data we have to deal with. To optimize $\mathbf{T}$ we consider the matching pairs of viewing and recall sequences. Based on the given flows, computing an affine transformation for the *squared* distances is a linear problem. This procedure typically converges to a local minimum close to the starting point [40], which is desirable in our application, as we expect the transformation to be close to identity. In any case, the minimization indirectly defines the resulting distance as

$$d(\mathbf{w}, \mathbf{w}') = \frac{1}{\sigma} \sum_i \sum_j f_{ij} d_{ij} \tag{7.5}$$

7.3.3 KNN Classification.

In order to have a baseline for comparison, we used a simple k-nearest neighbor (kNN) classifier based on euclidean distance. The basic idea of kNN classifier is each query is exhaustively compared to all training samples and classified as the class of the k-closest ones. k was set to 27 in our experiment and the class label was determined as the distance weighted labels of the nearest 27 samples. The overall accuracy was determined by leave-one-out cross validation. Each time we dropped out one dataset for testing and used all remaining datasets for training. This method was used for image retrieval based on eye movements during encoding and during recall separately.

7.3.4 Distance-based Retrieval

Based on a distance measure d between spatial histograms represented by their weight vectors $\mathbf{w}$, $\mathbf{w}'$, we can perform retrieval, assuming $\mathbf{w}'$ represents the query data.

The simplest case is scenario 2b (described in Section 7.3), where retrieval is restricted to a single observer. So we have 100 spatial histograms $\mathbf{w}_i$ representing the gaze sequences while looking at the images, and a single histogram $\mathbf{w}'$ representing a recall sequence as a query. Computing the 100 distances $d(\mathbf{w}_i, \mathbf{w}')$ allows ranking the images.

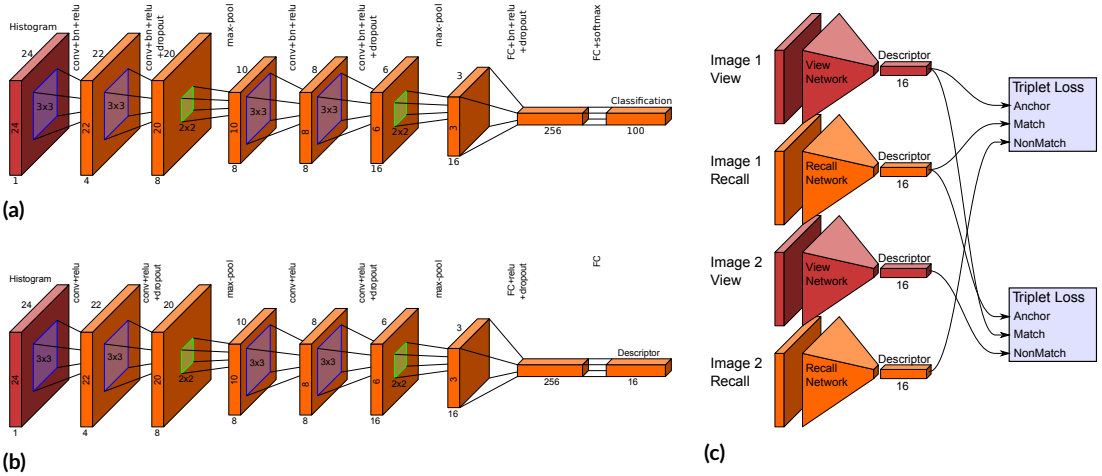


Figure 7.1: (a) CNN architectures employed for histogram based matching. Numbers around layers indicate width and height, as well as the number of channels (written below). (b) Descriptor learning setup with two triplet losses. For two different images, encoding and recall histograms were fed through their respective networks (truncated pyramids) to produce 16-dimensional descriptors. Two triplet losses were used to force matching descriptors to be closer to each other (based on Euclidean distance) than non-matching descriptors.

In the other three scenarios query data from a single observer is matched to the gaze data from other observers. In these scenarios we base our approach on *leave-one-out* cross validation, meaning we always use all the data from the remaining 27 observers for matching. Let $\mathbf{w}_i^k$ be the spatial histogram for image i provided by observer k. For each image we compute the smallest distance of the query $\mathbf{w}'$ to each image across all observers:

$$d_i(\mathbf{w}') = \min_k d(\mathbf{w}_i^k, \mathbf{w}'). \tag{7.6}$$

Then the ranking is based on $d_i(\mathbf{w}')$. Note that the first rank in this case is the same as using the nearest neighbor in the space of spatial histograms with the distance measure defined above.

7.3.5 Convolutional Neural Networks

The data we have collected encompasses 2700 gaze data histograms–we felt this number justifies trying to learn a classifier using the currently popular *Convolutional Neural Networks* (CNN). We design the architectures to have few parameters to reduce overfitting on the data, which is still rather small comparing to the typical number of parameters in CNNs.

Network Layout The basic setup for the CNN is similar to the ones used for image classification and visualized in Figure 7.1a. Each encoding (or, respectively, recall) sequence

of eye movements was accumulated into a histogram which served as a one channel input image to the CNN. Each convolution filters the previous layer with learned 3×3 filter kernels and produces a new image with as many channels as filter kernels are used. As is common, we combine each convolution with *Batch Normalization* (BN) and *ReLU* non-linearity. After two blocks of convolution, we perform *Max Pooling* to reduce spatial size. After two blocks of max pooling, the spatial size is down to 3×3 elements at which point we flatten and employ regular fully connected layers. The first of these two fully connected layers is again using BN and ReLU. The last one, however, directly feeds into a softmax layer to produce a probability distribution over the 100 image classes. We designed our network in such an architecture (Figure 7.1a) that it only had few parameters so that overfitting on the small dataset was reduced. To improve generalization, we employ dropout layers throughout the network with a dropout probability of 30% in the convolutional layers and 20% in the fully connected layer.

Application to Scenario 1 The network can be directly used to perform retrieval within the same condition (scenarios 1a and 1b, described in Section 7.3), as the query data is of the same type as the data used for generating the network. All weights were initialized using Xavier initialization[79] and, where applicable, initial filters were orthogonal to each other. We trained the networks using Cross Entropy Loss for 50 epochs using the Adam parameter update scheme[147] with $\beta_{1,2} = 0.95$ and a batch size of 100. We started with a base learning rate of 0.02 and annealed it over time by halving it every 10 epochs. The individual layers had their learning rate scaled based on the number of inputs, analog to the scaling in Xavier initialization. The intermediate layers had no weight decay as they were followed by batch normalization layers. The output layer had a mild L2 weight decay of 1e-8.

To test classification accuracy (in the first two cases), we performed leave one (subject) out cross validation. That is, we trained the network 28 times, each time withholding the data from one of the 28 subjects from the training pool. Testing was then performed only on the data of the withheld subject, thus testing how the network generalizes to new subjects. Overall accuracy was reported as the average over all tests.

Extension for Scenario 2 For working with histograms coming from different processes, it seems better to learn independent encodings. Instead of mapping the histograms directly to their respective image index, and thus casting image retrieval as a classification task, we rather perform image retrieval by learning proper encodings and then comparing them based on distance. In other words, we train an embedding of the histograms from the two conditions into a low dimensional descriptor space such that matching pairs are close together, and non-matching pairs are far apart.

The network architecture (see Figure 7.1b) is similar to the classification architecture detailed above. The difference is in the lack of BN layers, as they could not be used in this

training setup, the removal of the softmax output, and the reduction from 100 outputs to just 16. We simultaneously train two instances of this architecture, one for each condition. Both map histograms to 16 dimensional descriptors. These descriptors were then used to match a query histogram to a set of previously recorded reference histograms and, by extension, find the closest encoding histogram for a given recall histogram. The difference to the previous network was in the smaller output layer (16 outputs instead of 100) and the removal of the batch normalization layers. Initialization of weights and scaling of the learning rate was handled as in the previous network.

The *Triplet Loss*[239], employed during training, forces the networks to produce descriptors which, for matching pairs, are more similar (based on Euclidean distance) than for non-matching pairs. We used two triplet losses in tandem with four descriptor computations to balance the training (see Figure 7.1c). We trained each fold for 1000 epochs with an initial base learning rate of 0.001 that we scaled by 0.75 every 100 epochs. We used the same Adam optimizer with $\beta_{1,2} = 0.95$ and a batch size of 20. All layers had a small L2 weight decay of 1e-8. We also experimented with hinge loss, which simply forced matching descriptors to be close and non-matching descriptors apart, but found that triplet loss, while training more slowly, generalized better.

A ranking for a query histogram can be computed by generating its representation in the low dimensional descriptor space and then using Euclidean distance. The procedure for selecting the best match is identical to the one explained above for the distances computed based on EMD.

Extension to Unseen Images. Following the previous formulation of matching histograms. We tested how well the learned descriptors can be generalized to unseen images. To this end, we performed 10 fold cross validation to show that the matching network generalizes to new images. The 100 images were split into 10 groups of 10 images each. Training was performed 10 times, each time withholding one group of images from the training data and using it for testing. Reported numbers were the average over the 10 folds.

7.4 Results

Using the two retrieval methods we now present results for different scenarios. They are described based on the rank of the queried image among all images. Based on the ranks we form a curve that describes the percentage of the 100 different queries (one for each image) to be matched if we consider the first x ranks ($x \leq 100$). This is essentially a Receiver Operating Characteristic (ROC) curve. The leave-one-out cross validation generates one curve for each observer, so 28 curves in total. We consider the mean curve as the overall retrieval result and also provide the 50% interval in the visualization (see Figure 7.2, Figure 7.3, and Figure 7.4).

For a good retrieval algorithm, the ROC curve climbs up steeply at the beginning and quickly reaches the highest point. This would mean that for most queries the correct image match is among the top ranked retrieval results. The overall retrieval accuracy is then measured by the standard area under the curve (AUC). If retrieval was based on chance, the expected rank for retrieval would be 50 among the 100 images. The chance for being ranked first would be 1%, the chance to be among the first 5 would be 5%, and so on. This means the ROC curve for chance retrieval would be the bisector line with AUC = 0.5.

Both methods depend on the size of the histogram, and the distance measure provides the additional parameter σ. We also checked how enabling the affine transformation influences the results. In general, the results in terms of the AUC are not varying much with these parameters and we make the following recommendations:

- Choosing k too small slightly degrades the results and we suggest using histograms of size $k \geq 16 \times 16$.
- Allowing a partial match improves the result, i.e. choosing $\sigma = 1$ is suboptimal. We have settled for $\sigma = 0.5$.
- Allowing a global affine transformation shows improvements in retrieval rate for some of the scenarios. This means learning a global transformation matrix could potentially adjust the deformed recall sequences.

7.4.1 Scenario 1

The tentatively easiest case is retrieving an image based on the gaze data while looking at the image, matched against similarly obtained gaze data. Indeed, we find that the AUC for kNN is 94.5% (Figure 7.2, $U = 9.6e3, n_1 = n_2 = 100, p < .001$ two-tailed, Mann-Whitney rank test) and the distance measure is 96.3% (Figure 7.2, U=9.73e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). In particular, using kNN the top ranked images are correct in 60.6% of the cases, and the correct match is among the top 3 in 77.9%, and 52.2% and 72.4% using the distance measure. These results are largely independent of the choice of parameters and the use of the affine transformation.

The CNN performs only slightly better with AUC = 97.5% (Figure 7.2, U=9.86e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test), and achieved top-1 and top-3 ranks of 61.3% and 79.1%.

This demonstrates that visual images can be easily discriminated based on eye movements from observers exposed to those images, at least for a database of 100 images. Even a simple classifier can correctly identify the visual stimulus more than 60% of the time.

If recall gaze data is matched against recall gaze data, the results are significantly worse. The performance for the kNN classifier dropped down to chance level (Figure 7.2, AUC =

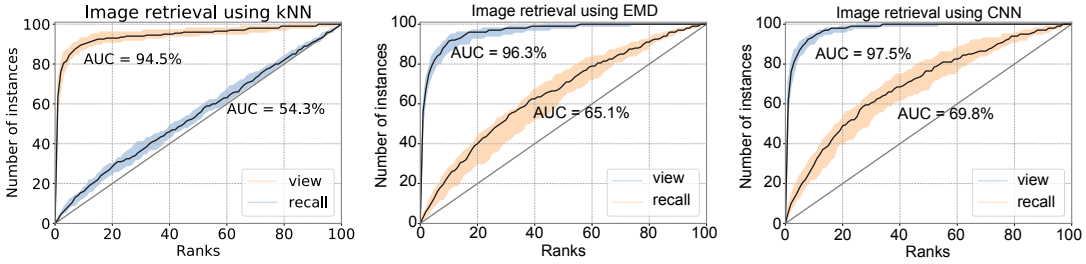


Figure 7.2: Retrieval performance using kNN(left), EMD(middle) and CNN(right). The 50 percent center intervals of all ROC curves over all observers are depicted in light colors, viewing based retrieval in blue and recall based retrieval in orange.

54.3%, $U = 5.5e3, n_1 = n_2 = 100, p = 0.245$ two-tailed, Mann-Whitney rank test). This failure reflects the repeated finding that eye movements during mental imagery suffer from spatial distortion [25,126,240], so we should not use the Euclidean distance measures on imagery data, unless we could somehow remap the distorted gaze data onto the (unknown) correct positions.

Using matching based on the distance measure we find AUC = 65.1% (U=7.08e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). The top-1 and top-3 ranks are 3.4% and 8.2% respectively. For the CNN we get AUC = 69.8% (Figure 7.2, U=7.08e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). In 5.9% of the cases the system could correctly identify the image, and the top-3 rank dropped to 13.8% (chance would have been 1% and 3%). Notably, the variance of all ROC curves of retrieval based on recall gaze data is higher (SD = 7.13%, compared to SD = 2.42% for retrieval based on viewing gaze data). We interpret this poor performance to reflect the previously noted variation among observers' recall behaviour.

7.4.2 Scenario 2

We perform cross-condition matching in scenario 2 by matching a recall query to viewing gaze data. When matching is performed against all other observers' gaze data, the performance drops to AUC = 68.4% as shown in Figure 7.3 on the left (U=6.937e3, n_1=n_2=100, P<10^{-5} two-tailed, Mann-Whitney rank test), with top-1 of 5.8% and top-3 of 12.8%. When matching is against the gaze data from the same observer, we achieved an AUC of 73.7% (U=7.47e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test) based on EMD distances (right plot in Figure 7.3). Top-1 and top-3 matches are 10.0% and 19.6% respectively.

Surprisingly, in this scenario the performance using CNN is not better than using EMD. Matching against data from all other observers gives AUC = 68.1% (U=6.91e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). Similar to the results based on EMD distance, the retrieval performance improves for matching against the data of the observer: AUC = 73.4% (U=7.44e3, n_1=n_2=100, p<.001 two-tailed, Mann-Whitney rank test). We suspect this

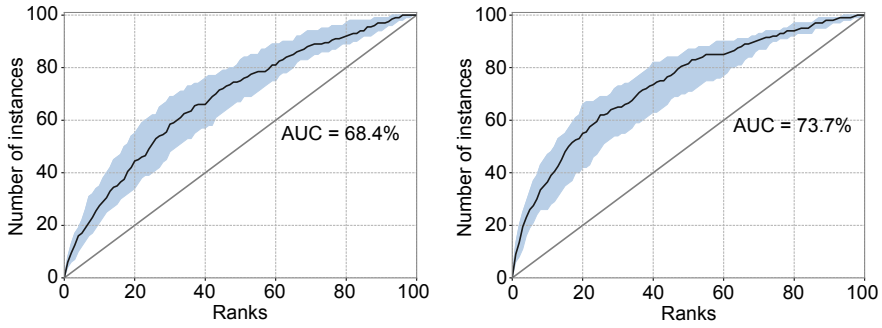


Figure 7.3: Retrieval performance in scenario 2. The left plot is matching result against viewing from all other observers and the right plot shows matching against data from the same observer. The 50 percent center intervals of all ROC curves over all observers are depicted in light colors.

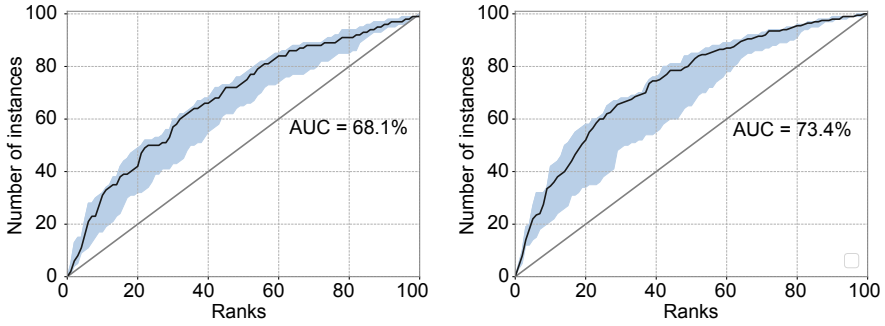


Figure 7.4: Retrieval performance in scenario 2 using CNN. The left plot is matching result against viewing from all other observers and the right plot shows matching against data from the same observer. The 50 percent center intervals of all ROC curves over all observers are depicted in light colors.

is due to the difficulty of the data in general and, in particular, due to the well known problem of CNN when dealing with global affine transformations. If there was a way to compensate for such global transformations in CNN, the results might be better. It would be interesting to continue with this idea in future work.

7.4.3 Extension to Unseen Images

The matching networks allowed us to see whether retrieval based on eye movements during mental imagery can generalize to new image content that is not in the training dataset used by the networks. In other words, whether the system can generalize to unseen images (with eye movement data). Matching was restricted to the pairs for each observer and a 10-fold cross validation over the images was used to determine the retrieval accuracy, e.g., for each fold training on 90 images and testing on the remaining 10. An average AUC of 66.4% was achieved over the 10 folds ($U = 78.5, n_1 = n_2 = 10, p = 0.250$ two-tailed, Mann-Whitney

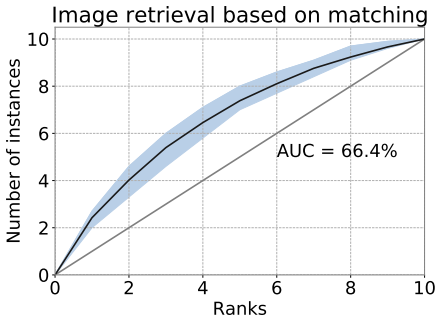


Figure 7.5: Generalization to unknown stimuli. ROC curve of retrieval based on matched eye movements during recalling. Blue area depicts the 50 percent intervals of all ROC curves of each subject.

rank test) (Figure 7.5). The top-1 and top-3 ranks were 23.5% and 40.2% respectively. These results indicate that the learned space can be used to describe eye movements related to unseen images, and allows the generalization about new stimuli with a reasonable retrieval accuracy.

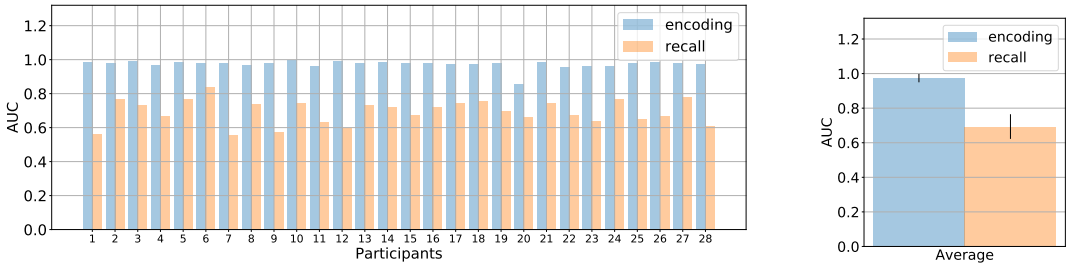
7.4.4 Dependence on Individual Participants' Behavior

Here we examined the variation of the retrieval accuracy among individual testing set. We looked at the achieved AUCs from previous leave-one-out cross validation using CNN (Figure 7.6 (a)). We noticed that imagery vs imagery retrieval worked better for some observers who actively move their eyes during mental imagery. Our intuition is that the current neural network design works well when observers resemble their perceptional eye movements during recall but performs poorly when it comes to severe distortion, i.e. when eye movements during recall are largely shifted, scaled or translated, or when observers don't move their eyes much during recall. On average, recall-based retrieval had a larger variation (SD=0.07) then encoding-based retrieval (SD=0.02) as depicted in Figure 7.6 (b).

7.4.5 Dependence on Stimuli Content

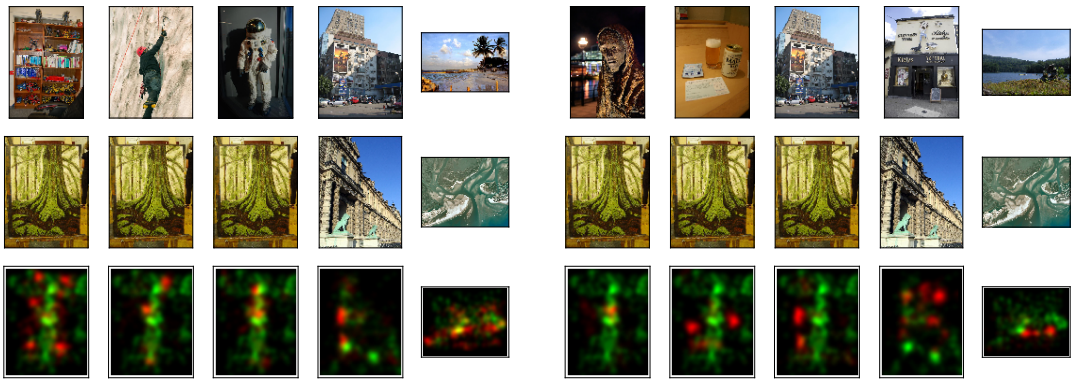
Images used in the experiment had 1024 pixels in the largest dimension and histograms of 24×24 are essentially the downsampled representations. This poses a natural limitation on the total number of images that could be computationally discriminated. As histograms encode information of where we look and how long we look, two distinct images inevitably lead to two similar histograms. To further investigate this aspect, we visually inspected the most and least misclassified image pairs when CNN was used as the retrieval method.

For encoding-based retrieval, the most confusing image pairs (Figure 7.7 (a)) had noticeably higher similarities especially in the image layout, whereas the least confusing pairs (Figure 7.7 (b)) shared less similarity. Interestingly, similar results were observed in recall-based



(a) (b)

Figure 7.6: Area under the curve of each leave-one-out test using CNN. (a) AUC when data of each subject is used for testing. Classification based on encoding eye movements are shown in blue and classification based on recall eye movements are shown in orange. The mean values and standard errors (one standard deviations) of the AUCs are shown in (b).



(a) (b)

Figure 7.7: Top-5 image pairs that are most often confused (a) and never confused (b) in encoding eye movements based retrieval using CNN. (a) Image pairs are sorted by the number of confusions from left to right: the first column shows the two images for which the eye movements during encoding are confused most often. In each pair, the image in first row is classified as the image in the second row. Overlaid histograms aggregated over all eye movements of the complete dataset are shown in the third row. Histograms of images in the first row are show in red and histograms of the images in the second row in green. (b) Examples of the most distinct pairs of images, i.e., images in the first row are never classified as images in the second row. Similar to (a), overlaid histograms are plotted in the third row.

retrieval and Figure 7.8 shows the most confusing image pairs. Despite the inaccuracy of eye movements during mental imagery, the similarity of image content still seems to be a source of confusion for the classification.

Figure 7.8: Top-5 most confused image pairs (in each column) in recall-based retrieval. Image in the first row is most often misclassified as the image below. The frequency of each pair's confusion goes down from left to right. Notably that images of each pair have very similar content. For example, the two images in the second column have dominant features in the right half: the boy in the top image and the text and display in the bottom one. In the third pair, a dog and a house are placed similarly to the positions where the boy and the adults are in the bottom image.

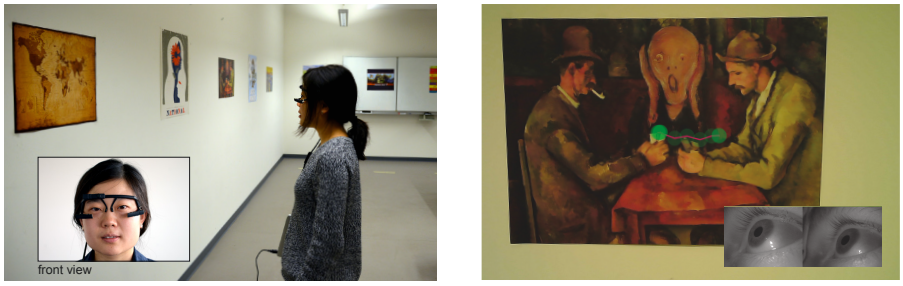
7.4.6 Discussion

Identifying an image based on gaze data during looking at the image works well, based on a variety of different approaches. It is clear, however, that this result is highly dependent on the set of images being used. If the 100 images evoke similar gaze patterns, they could not be discriminated based on gaze data. Our results indicate that gaze on the images we chose is different enough to allow for computational image retrieval. This is important, because if discrimination had been difficult based on gaze data from viewing, it would have been unlikely that gaze sequences during recall contained enough information for any task.

The results for retrieval based on gaze data from recall using data for recall to match indicate how severely distorted the recall data is. The performance significantly decreases and indicates that the task we believe is most important in applications, namely matching gaze data from recall against gaze data from viewing, may be very difficult.

The two scenarios we consider for matching recall data against viewing data show quite different results. Matching the recall data of an observer against their own viewing data works much better than matching against the viewing data from other observers. This indicates that viewing is idiosyncratic. The fact that observers agree more with themselves than with others is also consistent with the findings that fixations during recall are reenactments[7,223].

To our knowledge, this is the first quantitative assessment of the amount of information in the recall gaze that can be used to identify images.



(a) (b)

Figure 7.9: (a) The setup in the 'museum'-experiment. Participants view 20 images wearing an eye tracker. Later, they were asked to think about the images while looking at the (blank) whiteboard. Gaze patterns from this recall are used for identifying the image. (b) Example of video stream (scene and eye cameras). Green dots indicate fixations. Red lines represent the trajectory to the previous viewed position.

7.5 Real-world Application

We have established that the gaze pattern while only recalling an image could be used to identify the image using standard techniques from vision and learning. The data, however, was collected under artificial conditions. We are interested in performing a similar experiment, albeit this time under more realistic conditions.

A useful application could be retrieving one or more of the images seen from a museum visit. To explore if this is possible, we hung 20 posters in a seminar room (see Figure 7.9a), simulating a museum*. The images had slightly varying sizes at around $0.6m \times 1.0m$ on both portrait and landscape orientations. Their centroids were at a height of $1.6m$. For the recall phase, an (empty) whiteboard was used.

We recruited 5 participants (mean age = 32, 1 female) for the museum visit; only 1 had participated in the earlier experiment. Each of them was outfitted with a Pupil mobile eye tracker with reported accuracy of 0.6 degrees[137], equipped with two eye cameras at 120Hz and one front-facing camera at 30Hz. The eye tracker was calibrated on a screen prior to viewing the images. No chin rest was used during the calibration but participants were asked to keep their head still. As in the controlled experiment, we used a 9-point calibration and a display (1920×1200 pixels) placed at $0.7m$ distance. This also approximates the viewing distance between the visitor and the museum items. After calibration, positions of pupil center in eye cameras are mapped into the front-facing camera, yielding the resulting *gaze positions*. As long as the eye tracker stays fixed on the head, the calibrated mapping is valid. Participants were asked to inform us if they notice any displacement of the eye tracker.

*We initially planned to do this in cooperation with any of the large museums close by, but legal and administrative issues connected to the video camera in the eye tracker caused more complication then we felt the merely symbolic character was worth.

After calibration the cameras were continuously recording. Participants were asked to view all the images, in no particular order or time frame (but without skipping or ignoring any). Furthermore, participants were asked to obey markers on the floor indicating minimal distance to the images (similar to the rope in a museum). After viewing the images, participants were asked to recall images they liked, in any order, and as many as they wanted to recall. Each recall sequence started with instructions by the experimenter and ended with participants signaling completion.

We manually partitioned the resulting eye tracking data based on the video stream from the front facing camera. 20 viewing sequences are extracted for each participant and Figure 7.9b shows several frames of the data. The viewing duration varied greatly among participants, from few seconds to more than a minute per image. All participants recalled at least 5 images, with 10 being the highest number. Viewing and recall sequences are represented as variable-length gaze positions mapped in the front-facing camera. In total, each dataset from one participant contains gaze positions of 20 viewing sequences and five or more recall sequences.

For the analysis it turned out to be relevant what part of the viewing and recall sequences we would use. We considered the first 5, 10, 15, 20 seconds of the viewing sequence and the first 5, 8, 10, 12 seconds of the recall sequence. We represented the resulting raw data using k-means clustering with $k = 10$ and computed distances using EMD. The EMD based distance measure is able to compensate for global rigid transformation, which seems very important for the setting. The resulting ranks are above chance in all cases, yet different parameter settings worked best for different participants. The results did improve with optimization for a transformation. There was no significant advantage in including rotation or scaling, but translation was important.

The best median ranks for 5 recalled images were, in order from best to worst, 2, 2, 3, 4, 5 (without translation we found 2, 3, 3, 4, 7). So if duration can be adapted to each participant the results are surprisingly good. Applying any combination of duration to all 5 participants in the same way leads to worse overall results. Optimization for translation pays off in this situation as well, as the results get more stable against the choice of parameters: for a wide variety of combinations we find median ranks of 4 and 5. All the reported median ranks have to be seen in context of the median of a purely random retrieval being 10.

None of our participants had reported big movements during the experimental session and no secondary calibration was conducted. The camera rate of the eye tracker also seems to be irrelevant in our setting as long as meaningful eye movements are recorded. There is no special requirement for high-speed eye camera since saccades and other type of micro eye movements are not included in the analysis.

7.6 Discussion

Our study aimed at investigating the possibility of using computational models to retrieve natural images based on eye movements during encoding and mental imagery. We found that image retrieval based on eye movements when looking at the images can achieve very high accuracy even with a simple method. In other words, the gaze patterns when observers inspecting the images contain unique signatures which allow us to easily retrieve the corresponding images. However, a successful retrieval based on eye movements when LAN requires more complicated machine learning algorithms. Using a convolutional neural networks (CNN), we demonstrated that eye movements during mental imagery can be used to retrieve images with a reasonable performance. Furthermore, our results show that the intrinsic similarity between eye movements during encoding and those during recall allows us to extend the retrieval framework to new photos.

Using physiological measurements for the retrieval of mental imagery is not new. Others have aimed at reading a person's mental state[91,133] or reconstructing the imagery they have in mind[60,198,240], by making use of brain activities measured by functional magnetic resonance imaging (fMRI). Eye movements during looking at images have also been previously used in practice. In human-computer interaction, fixations were used as a mouse cursor for selection[231]. In Karessli et al.[135], gaze data were used to indicate where observers allocate their attention in a fine-grained object classification task, and subsequently incorporated into a classification framework as it provided the information on where to look for the distinction. Surprisingly, eye movements have not been used to retrieve the mental imagery before, despite substantial previous work demonstrating that eye movements during recall in front of an empty space are abundant and play a functional role in memory retrieval[25,127,128,129,159,160,223,236,247].

We have replicated previous research showing that gaze patterns from observers encoding images differ from the patterns while observers think about the same images[25,77,126,127,128]: imagery eye movements cover a smaller area compared to the area covered by encoding eye movements (Figure 6.1d), and imagery fixations are longer (Figure 6.1b). For this reason, the discriminatory information contained in imagery eye movements is poorer. More specifically, successful discrimination does not work with straight-forward k-nearest-neighbor search (kNN) and its Euclidean distance measure, while classification with convolutional neural networks (CNN) more easily discriminates the content. In contrast, gaze patterns from encoding the photographical content contain sufficient information that allow both the kNN and the CNN methods to perform successful discrimination. Additionally, our analysis results indicate that the retrieval performance varies among individuals and similarities of image content will pose a major challenge of extending to larger databases.

7.6.1 Eye Movements Representations.

In this study, eye movements were represented with a 24×24 2D histogram of gaze positions (also known as a gridded gaze density map). This is a natural choice, as most imagery research has shown that the location of fixations but not fixation order is preserved during imagery. However, other types of information might well be suppressed. Besides the simple 2D histograms, we experimented with several different representations which are 1). binary histogram excluding the time spent in each cell; 2). histograms with a third dimension of time; 3). concatenated histograms of differences between consecutive eye positions. None of them performed significantly better. It is therefore unclear how much better the image discrimination would be if order information were to be included in the representation of eye movement sequences. We suspect observers may have difficulties to recall anything else than locations of objects in the limited five seconds time. But it is also possible that other details of the mental imagery (e.g. color and texture) are hidden in finer gaze patterns. They might require longer time to recall. It remains to be an interesting direction in future work to explore different representations for other types of information.

7.6.2 Improvement on Performance.

We noticed that image retrieval based on eye movements from recall worked better for some observers than for others. Johansson et al.[128] could show that people with a poorer spatial imagery ability produced eye movements during imagery with a larger dispersion, more similar to those during encoding of the physical image. People with a good spatial imagery ability can recall photos without making extensive eye movements, and scale, shift and translate their gaze area. From this perspective, computational retrieval of image content using eye movements suits people with a poorer spatial imagery ability better.

The down-scaled eye movements during mental imagery make the retrieval task more difficult as clusters are formed in close range around the image center. Maybe we could instruct observers to extensively look around and actively make larger eye movements in the full extent of the tracking space, in other words, to imagine an up-scaled photo. Asking participants to make an effort is not unreasonable – and in many brain-computer interface applications, it is the norm. It is known that a successful brain scan using fMRI requires the participants to lie still during imaging. Electroencephalogram (EEG) provides a way to communicate with less physical constrains, especially for patients with aggravating conditions[161,279]. However, the measured signals can only be used to distinguish in binary cases[17]. Even though it requires a long training time[192], blink or other body movements can still lead to low signal-to-noise ratio[50]. In comparison, making eye movements during mental imagery is much more natural and intuitive, and has a potential to be a product that many people could use.

Surprisingly, in the third case (the generalization result), the resulting AUC (Figure 7.5)

was somewhat weaker than in the second case (the imagery result) in Figure 7.2. We were uncertain about the reason of this performance drop but we suspected this is likely caused by the increased difficulty of having to deal with variations in the encoding *and* imagery data simultaneously. We decided not to include more components than 16 in the descriptor vector, as they would just increase the networks likelihood to overfit.

7.6.3 Long-term Memory

The current study only focuses on the effect of short-term memory where image recall was performed immediately after the inspection of the original images. As indicated by a study[159], recall/imagery eye movements are likely to be less accurate as memory deteriorates over time, and consequently the retrieval performance would decrease. We have not explored this and it remains an interesting future direction.

7.6.4 Scalability

Retrieval depends on image content. Clearly, the more photos are added to the database, the more likely that several different photos will give rise to the same eye movement behavior, which inevitably affects performance. This is because image representations are essentially downsampled to gaze patterns, and similar representations lead to increased ambiguity. This would pose a limit to how many photos can be used with this retrieval method. In similar studies performed on fMRI data, Chadwick et al.[35] showed brain activity patterns can be used to distinguish among imagining only three film events and in Cowen et al.[46] fMRI signals were used to reconstruct face images but only 30 images were used.

Moreover, it is very likely that if all participants would look around in the full extent of the monitor, when imagining the photo, the retrieval performance of the system would increase significantly. Maybe it would be possible to instruct participants to make more extensive eye movements than they would normally do, in order to help the computer find the right image? It is not unreasonable to expect a small effort on the part of the participant–which in many brain-computer interface applications is the norm.

7.6.5 Temporal Stability

Our recall data were recorded immediately after image viewing. A longer delay might reduce the discrimination performance as the memory deteriorates. In future work, it would be interesting to explore the influence of memory decay and its effect on image retrieval from long-term memory.

7.6.6 Alternative Sensing Modalities

Our proposed technique relies on sensing the motion of the eyes over time. To gather this data, we used a video-based eye tracker, which relies on users' eye to be open, even when they look at a neutral surface during recall. In contrast, other sensing techniques, e.g. electrooculography (EOG), allow to sense eye movement when users' eyes are closed. This would also allow users to create a neutral background by simply closing their eyes, which might even increase the vividness of mental imagery[183,263]. However, the lack of reference with closed eyes might introduce different types of distortions. We believe that exploring alternative sensing modalities such as EOG as replacement or additional data source will allow our concept to become more viable for everyday interactions. We aim to explore additional sensing modalities as well as their deployment in less controlled environments in the future.

7.6.7 Towards Real-world Applications

This chapter focuses on the evaluation of the proposed new interaction model along with the development of computational tools. We used image retrieval as an indicator for the success of understanding user's intention. From a practical point of view, it would be interesting to compare our method to existing techniques such as manual selection or speech-based interface.

How to accurately track eye movements using mobile eye trackers poses another challenge. In our museum visiting experiment, we did notice that the calibration accuracy gradually got worse. As the shifts of the eye tracker over time are rigid translations, our comparison methods should be able to compensate them. Such limitations rely on the further improvements of mobile eye tracking.

With the development in wearable devices, we believe tracking the motion of the eyes would be a natural by-product. Combination with other interaction modalities, such as the possibility offered by recent work in speech interface[134], offers a rich source of information. With additional sources of information, we believe that our interface would provide an improved interaction between users and software agents.

8

A Methodological Investigation of the Spontaneously Prioritized Image Content

While looking in front of nothing, we retrieve certain details from memory that are encoded before. Based on this insight, we study the information contained in the looking-at-nothing (LAN) eye movements on the level of scene elements. In this chapter, we propose a new way to experimentally probe what content of a scene is prioritized in the recall from memory, based on LAN eye movements. More precisely, we propose a novel elastic matching algorithm to find out which encoding fixations have been subsequently recalled. The obtained results correlate highly to the subjective evaluations of visual importance, based on data collected in a separate experiment, indicating that the encoded fixations correspond to important image content. We then show examples of how this method can be used to investigate visual memory of bottom-up features (such as colour) versus objects with semantic meaning (such as faces). Other meaningful memory effects are also demonstrated in the results.

8.1 Introduction

From change blindness studies, we know that in some situations participants look straight at a task-relevant object, and still no working memory trace can be registered[258]. In the well-known gorilla experiment in psychology, observers reported no recognition of the gorilla even when it had been fixated for a significant amount of time[189]. Clearly, not all fixations

are the same, and only some result in a later reported memory of the fixated item, which is theoretically important, but also has practical implications for a variety of research fields in the vision sciences and beyond.

The main goal in this chapter is to distinguish the fixations that are more likely to be remembered from those that are less likely to be remembered, based on eye movements during looking at nothing. In this paradigm, we use the gaze positions during recall as a *selection criterion* of the fixations during exploration: we take the matching of a fixation in the exploration phase with a fixation from recall as an indication that the corresponding scene content has been prioritized in recall from episodic visual memory.

This method takes its basis in the well-established finding from the LAN paradigm that our eyes move while recalling an image. This was first shown by[194], and later research established that in the absence of other visual features (i.e. while looking at nothing), the motion of the eyes is reminiscent of the gaze pattern while looking at the original stimulus[127,159]. In particular, it has been shown that the fixations during recall of the stimulus reveal the *location* of objects[71,180], but not necessarily reinstate the sequences[86,181]. Participants with a good spatial imagery ability make fewer imagery eye movements, while participants with a poor ability make more and wider eye movements[126]. A number of studies[19,126,129,159,211,236,237] have shown that eye movements during recall play a *functional* role in memory retrieval. Laeng & Teodorescu[160] showed that inhibiting eye motion, by asking observers to maintain fixation on a central point during exploration, led to reduced eye motion during recall, and inhibiting eye motion during recall led to degraded recall performance.[53] confirmed that inhibiting eye motion during recall decreases memory performance. Moreover, attending to regions that have been previously looked at before has been linked to imagery vividness[160], change detection performance[208], memory accuracy[159,236] and recognition accuracy[36], further suggesting that eye movements during looking at nothing correlate to what has been encoded in memory.

There is, however, a fundamental limitation to the LAN paradigm: without the stimulus being present, there is no physical frame of reference other than the boundaries of the screen. The locations of fixations during recall exhibit a significant local displacement, i.e. the spatial reproduction of fixation positions contains error. This deformation of the imagery space has been consistently reported in imagery literature and involves shrinking, translation and not making any eye-movements at all[126,127,159]. To overcome this obstacle, instead of using natural images, most previous studies employed single face images[36,100] or grid-based stimuli[129,159,181,236], for which area-of-interest (AOI) methods are sufficient to find the correspondences between encoding fixations and recall fixations. However, for complex stimuli such as photographic images, visual features are irregularly distributed and rigid AOI methods very often fail to handle the displacements in recall locations, forcing researchers to perform time-consuming manual coding, often using spoken language as a mediator[127].

In this chapter, we propose a new method to computationally distinguish between fixated regions in the original image that are spontaneously recalled* from visual episodic memory vs. those that are not. The fixated content which is recalled might coincide with the relative importance of different scene regions, appearing in the recall fixations after it has been prioritized in episodic memory.

In order to match fixations during recall to fixations during exploration, we therefore need to compute a mapping. After applying the mapping, we retain fixations from the exploration sequence that are close enough to a fixation in the relocated recall. A threshold on the distance between fixations in the exploration sequence from fixations in the recall allows us to steer the distance criteria and to control the amount of image content being considered as recalled (more detailed in Section 8.4). In this study, we address the following two questions, employing both quantitative and qualitative evaluations: First, will the prioritized image content coincide with separate judgments of the most important position in the image? Second, how would salience, faces, and gaze guidance contribute to the prioritization in recall from short-term visual episodic memory?

8.2 Analysis of Eye Movements during Exploration and Recall

The dataset of eye movements during exploration and recall as described in Chapter 6 is used in this study. Here we present additional results focusing on the similarity between each pairs of encoding and recall eye movement sequences.

Data from Encoding

The fixations from all participants for a single image can be summarized in a spatial histogram (a so-called gaze density map), commonly plotted as a *heat map* for the image. We computed the spatial histograms for the data from our experiment and for the publicly available MIT-data, which includes eye movement data during encoding. Following Judd et al.[131], we removed the first centre fixation from each sequence and applied a Gaussian filter with a kernel size equivalent to 1 degree of visual angle. The heat maps resulting from the publicly available data and the heat maps from the exploration phases in our experiment are very similar (mean Pearson's correlation coefficient (CC) = 0.766, SD=0.115).

Data from Recall

Compared to encoding sequences, recall sequences have fewer but longer fixations (see Figure 6.1b). For some photos, the correspondence between fixations during encoding and recall

*We will use the term 'recall' for 'retrieval of memory content' during looking at nothing even though the recalled information is not spoken but only exhibited through gaze.

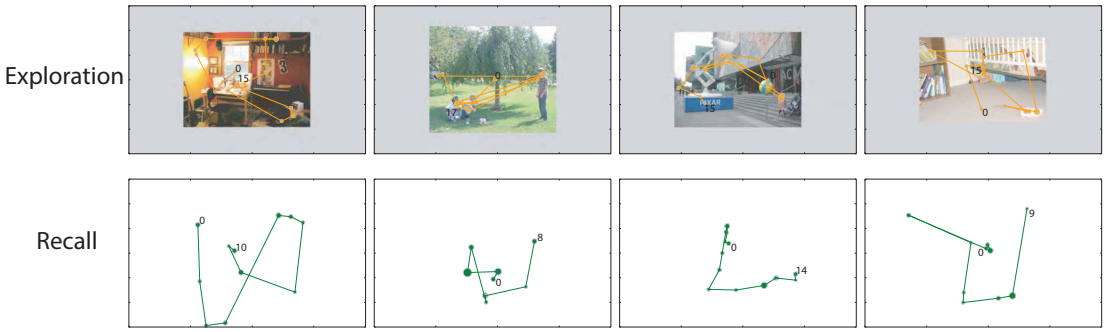


Figure 8.1: Pairs of exploration and recall fixation sequences from four observers. Fixation sequences during exploration are shown in the first row overlaid on top of the image stimuli. Corresponding fixations during recall are shown in the second row. The temporal order of each sequence is indicated by the numbers and consecutive fixations are connected by line segments. Notice that fixations in recall are distorted relative to image features and there is often no clear correspondences between fixations during exploration and recall.

is very clear. However, for the majority, while fixations from recall roughly resemble the maps from the encoding phase, they are more constrained towards the centre and typically fail to *exactly* correspond with features in the image that participants would have remembered during recall. We also found that the temporal order is in general not preserved (see Figure 8.1 for some examples). This is consistent with previous results in imagery research[126].

Some observers tended to stall during recall. They stopped moving their eyes, leading to fixations that are unlikely corresponding to image content. For 5 out of the 28 participants, the number of fixations in recall is less than half compared to encoding. One participant reported after the experiment that he changed his strategy through the experiment and only recalled the single most interesting element of the image.

8.2.1 Quality of Individual Recall Data

Participants developed certain behaviours or strategies during the experiment that distort the data, despite that their eye movements during recall are generally minimal. For example, a low number of fixations clustered around the centre, or fixations that are randomly spread over the stimulus area cannot be used to reliably identify locations of scene elements corresponding to one of the fixations during encoding. We evaluated the quality of the recall data, which also indicates the level of difficulties in matching. We expected that systematic undesirable gaze behaviour during recall would show in gaze density map aggregated over all 100 images during recall. For comparison we took the aggregated gaze density map during encoding and computed the correlation coefficient (CC) of the two gaze density maps (the median value is $CC = 0.856$). A few examples and the resulting CC scores are illustrated in Figure 8.2.

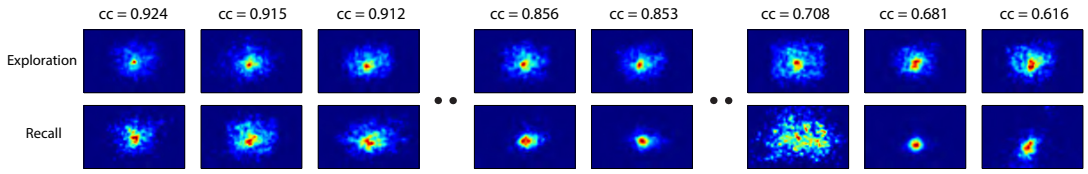


Figure 8.2: The similarity between the encoding and recall gaze density maps for each of the 28 participants aggregated over the 100 stimulus images. Participants are ranked according to correlation between the two cumulative gaze density maps. Two different types of inconsistencies appear in the least similar pairs: the distributions of recall fixations are either peaked at one point or spread randomly.

8.3 Collection of Visual Importance Data

During recall, some objects in the memorized photo images are prioritized over others in the choice of recall fixation positions. Are these prioritized items the important objects? In order to evaluate the prioritization during recall, we made another measurement, where we asked a group of new observers to identify the most important scene element by clicking at its position after being briefly exposed to a stimulus. In this, we follow studies that show how the gist of a scene can be perceived in a single glance within as little as hundred milliseconds[16,32,205,216]. Clicking has previously been used to determine important areas[204].

8.3.1 Method

Participants A group of 21 participants (6 female, mean age = 24) were recruited and their participation was compensated. All participants had normal or corrected-to-normal vision and none of them had taken part in previous data collection. Consent forms were signed before the experiment which allow us to use their data.

Apparatus The apparatus setup was identical to that used in previous data collection. Observers' eye movements were tracked with a standard EyeLink 1000 in remote mode.

Stimuli and Design The same set of 100 images were used as visual stimuli. In order to help observers better locate scene elements, grids were overlaid over the images as inconspicuously as possible. Furthermore, smoothed, semi-transparent images were displayed during clicking. High spatial frequencies were removed from the images and remaining low frequencies functioned as a reminder of the visual content. A Gaussian kernel with radius of 10 was used for smoothing and the alpha transparent blending value was set to 0.3 (an example is shown in Figure 8.3).

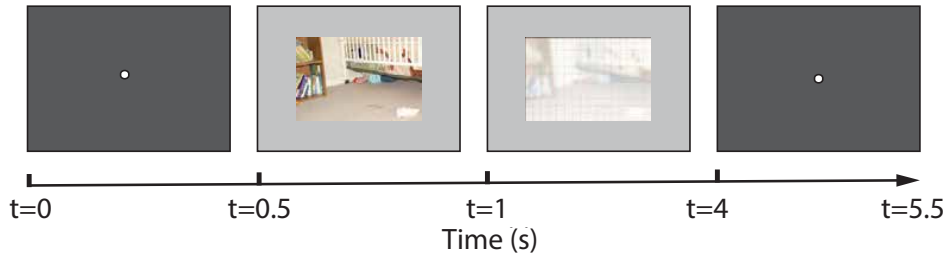


Figure 8.3: Recording paradigm used in data collection 2. After a brief exposure of image stimulus, observers were asked to click at the most important scene element.

Procedure The procedure for each trial is illustrated in Figure 8.3. Similarly to the previous data acquisition procedure, each trial started with a black screen with a white fixation circle at its centre, followed by a brief encoding phase where an image stimulus was displayed for 500 ms. Observers were instructed to 'select the most important part of the image'. Subsequently, observers were asked to 'click at that selected position'. The amount of time given for clicking was 3s in each trial.

8.3.2 Data Processing

We use 2093 out of 2100 trials, for which observers clicked within the given time. The average click latency was 1.63*s* measured from the moment when the image got blurred. Figure 8.4 shows three examples of the data where heat maps for click data are smoothed with a Gaussian kernel of 2° of visual angle. These heat maps are called the *clicking maps* in later sections.

8.4 Mapping Recall Fixations onto Encoding Fixations

In the first data collection, participants were asked to immediately recall the previously seen image in front of an empty screen. We will now present an algorithm that determines which elements in the original image correspond to the LAN (recall) fixations. Due to the lack of reference frame while looking at nothing, it is known that fixations during recall contain large spatial errors. In contrast, fixations during encoding are mostly perfectly aligned with image elements. Hence, matching recall fixations onto encoding fixations would provide us with the alignment between recall fixations and the original image objects that have been recalled.

8.4.1 Relocation Mapping

The problem of establishing the mapping is that spatial locations of the fixations in recall are distorted relative to the locations of features in the image. Due to a lack of reference frame,

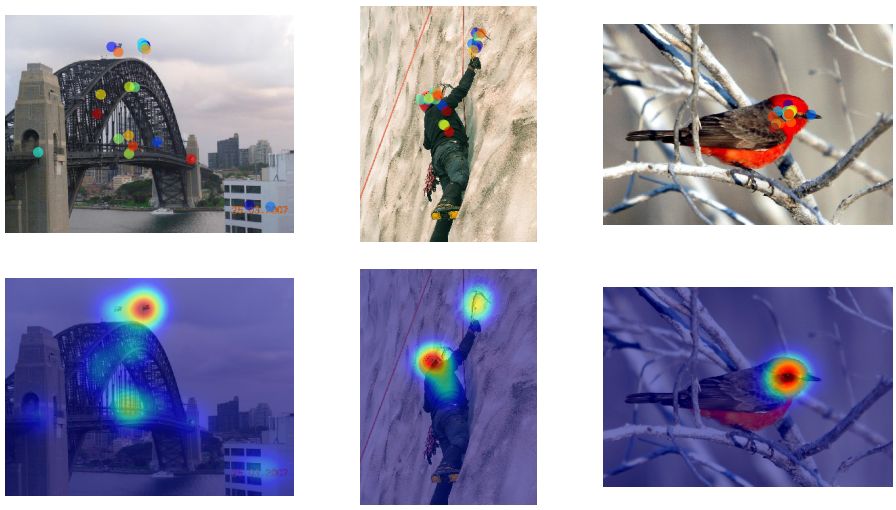


Figure 8.4: Examples of clicking results where agreement of clicking positions among all participants increases from left to right. First row shows clicked positions after the instruction to click what is important. Below it, heat maps of the same click data.

fixations during looking at nothing contain a large deformation, and such inaccuracy in each fixation may accumulate over time. Therefore, we view the operation to relocate the recall fixations to their proper position as a composition of a global rigid transformation (due to a lack of reference frame) and local deformation (due to local inaccuracy and propagated errors). We speculate that without this relocation, several recalled elements remain unmatched, causing significant but unwarranted changes to the mapping.

Figure 8.5 provides an illustration: The six points in a) depict a sequence of fixations during encoding, of which only five are being recalled. The observed set of fixations during recall projected onto the same reference frame is shown in b). We may think of the locations as being displaced by the rigid transformation in d) and the (local) deformation in c). The deformation contains no global rigid transformation in the sense that moving or rotating the fixations would not make the sum of the squared distances to the matched fixations in encoding any smaller. Note that matching fixations in recall and encoding cannot be found by simply considering their distances, as such matching will not take into account the relative structure within recall fixations (see Figure 8.1 for some examples).

When data are corrupted by global transformation and deformation, it is common to approximate the effects by minimizing the squared distances between matched data points. This is known as a chicken and egg problem: The estimated mapping is intended to improve the matching, while the matching is needed to estimate the mapping. The common

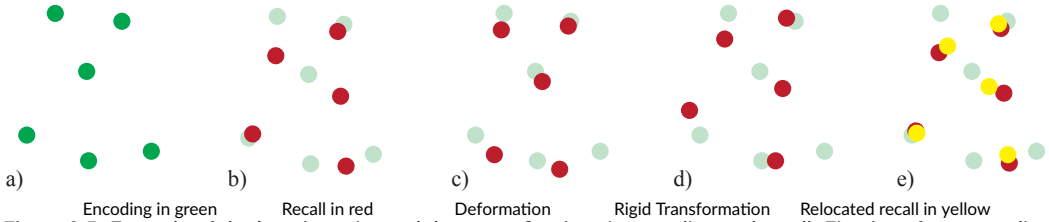


Figure 8.5: Example of the location mismatch between fixations in encoding and recall. Fixations from encoding are shown in green (a); recall fixations are subset of the fixations during encoding and are shown in red (b). Their displacement contains deformation (c) and global rigid transformation (d). After the relocation mapping, recall fixations are transformed to the yellow locations (e). Refer to Section 8.4.1 for more details.

solution to this dilemma is to start with a first guess about the matching, then compute the transformation, and based on the transformation make a better guess about the matching, and so on until the method converges. A common approach of this type is Iterated Closest Point (ICP)[15], which is used to compute a global rigid transformation to match two partially overlapping point sets. This allows to cope with the global transformation, caused by the lack of reference frame while looking at nothing, and the accumulated sequential distortions in each fixation. The approach we suggest for matching the recall fixations to the fixations in encoding is inspired by ICP, yet extends it in two important ways: First, rather than using a closest point matching, it matches fixations in recall to *consensus* locations in the encoding data. Consensus locations are computed by considering several possible scenarios (details below). Second, rather than computing a global rigid transformation, it is based on a elastic mapping that has a controllable overall deviation from a global rigid transformation. In such a way, a global transformation offers a reference frame for fixations during looking at nothing and a local transformation provides the flexibility to adjust the variant local distortions.

The first modification is motivated by the situation in Figure 8.5b, namely that matching the recall fixations to the *closest* fixations in encoding may lead to unintended assignments, especially when considering the large variations among participants. These mismatched assignments will steer the estimated mapping away from the desired solution. Given the data, we believe several situations have to be accommodated:

1. One fixation in recall maps to exactly one fixation from encoding.

2. One fixation in recall maps to several close fixations from encoding, i.e. the observer recalls just one scene element that drew several fixations during encoding.

3. A fixation in recall may have no corresponding fixation in encoding (i.e. because the fixation is unrelated to the process of recalling, the spatial error is too large to be rectified, or non-fixated scene element is recalled).

Motivated by the fact that there are fewer recall fixations than encoding fixations, we suggest to accommodate all cases by computing a consensus location for each recall fixation. A weighted average of the fixation locations in encoding is computed and the weights decay exponentially with distance.

This has the effect that if only one fixation in the encoding sequence is close while the others are far away (case 1: one-to-one matching), the closest location will receive a large weight, while the others get relatively small weights, so that the consensus location will be the matching fixation.

In case several fixations in encoding are close (case 2: one-to-many matching), all of them receive equal weights, and the consensus location is the centre of these fixations. For one-to-many matching, a balanced solution would give each fixation an equal weight. This could correspond to situations where participants carefully inspect the eyes, nose and mouth of a face, which correspond to many fixations during encoding, but during looking at nothing only the face area is recalled.

If no fixation is found for matching (case 3: one-to-none matching), all fixations receive little weight, and the consensus location is roughly where the recall fixation already is.

Figure 8.5 e) shows the consensus locations for the situation in b) and recall fixations are placed at the consensus locations after applying the relocation mapping. The mathematical modeling of the computation of consensus locations is explained in 8.4.2. It can be controlled by a parameter w_p, measured in visual angle, which could be interpreted as the distance of fixations that contribute significantly to the weighted averaging procedure. We set the parameter to $w_p = 2^\circ$ and discuss this choice in Section 8.4.4.

The consensus locations for recall fixations are used to compute a relocation mapping $D : \mathbb{R}^2 \mapsto \mathbb{R}^2$ from the current positions of the recall fixations to the desired ones. Allowing elastic transformation overcomes the problem that the best we could achieve with computing a rigid transformation is to be left with the elastic matching, i.e. the situation depicted in Figure 8.5 d). However, some global rigidity needs to be preserved, i.e. the positions should not deform arbitrarily, as fixations while looking at nothing are not arbitrary and correlated to the mental imagery during recall. This is important, as we would otherwise always match all recall fixations to some fixations in encoding. For example, if we allowed arbitrary scale, it would always be possible to scale the set of recall fixations to a single point and then match it to one of the fixations in encoding. To avoid such degenerated solutions where recall fixations are mapped to the centroid of all encoding fixations, it is important to restrict the mapping to preserve the global structure. The computation of the mapping based on consensus locations and the current positions of the recall fixations is explained in 8.4.2. The global rigidity of the mapping can be controlled by a parameter w_d, measured in visual angle. Roughly speaking, w_d describes the distance of points that may be transformed by two rigid transformations that differ significantly. If w_d is large, fixations are transformed by very sim-

ilar rigid transformations, restricting the deformation; if it is small, fixations are transformed by independent rigid transformations, allowing large deformation. We set the parameter to $w_d = 10°$ and discuss this choice in Section 8.4.4.

The necessary steps to compute the relocation mapping are given in pseudo code in Section 8.4.2. Once the mapping is computed, we simply use distance as the sole criterion for testimony: a fixation in encoding has been recalled, if there is a mapped fixation in recall that is closer than the matching radius ε. Intuitively, this allows us to measure what has been recalled while looking at nothing.

8.4.2 Computing the Relocation Mapping

Let $\mathbf{p}_i \in \mathbb{R}^2$ be the positions of the fixations in the encoding sequence and $\mathbf{r}_j \in \mathbb{R}^2$ the positions of fixations in the recall sequence (in a common coordinate system). We wish to compute a relocation $D : \mathbb{R}^2 \mapsto \mathbb{R}^2$ that is applied to the recall locations $\mathbf{r}_j$ with the aim to align the data with the fixation positions during encoding.

We need two ingredients for computing the relocation: partial matching and elastic mapping. For the elastic mapping, we suggest Moving Least Squares (MLS)[171]. Here, we use this framework applied to rigid transformations, i.e. local rigid transformations are fitted using weighted least squares. This approach has become popular in geometric modelling where it is usually derived as minimizing the deviation of the mapping from being locally isometric[37,232,245].

We model the relocation D as a rigid transformation that varies smoothly over space:

$$D(\mathbf{x}) = \mathbf{R}_\mathbf{x}\mathbf{x} + \mathbf{t}_\mathbf{x}. \tag{8.1}$$

The subscript $\mathbf{x}$ indicates that rotation and translation vary (smoothly) with the location $\mathbf{x}$ in the plane. They are computed by solving a spatially weighted orthogonal Procrustes problem[83], where the weights depend on the distances of the points to $\mathbf{x}$. Assume the desired position for $\mathbf{r}_j$ is the position $\mathbf{q}_j$, then $\mathbf{R}_\mathbf{x}, \mathbf{t}_\mathbf{x}$ are computed by solving

$$\underset{\mathbf{R}_\mathbf{x}{}^\top\mathbf{R}_\mathbf{x}=\mathbf{I},\mathbf{t}_\mathbf{x}}{\operatorname{argmin}} \sum_j \vartheta(\|\mathbf{x}-\mathbf{r}_j\|)\|\mathbf{R}_\mathbf{x}\mathbf{r}_j + \mathbf{t}_\mathbf{x} - \mathbf{q}_j\|_2^2. \tag{8.2}$$

Here, the weight function ϑ should be smoothly decaying with increasing distance. We use the common choice

$$\vartheta_d(x) = e^{-\frac{x^2}{w_d^2}}, \tag{8.3}$$

which gives us control over the amount of elasticity in the mapping with the parameter w_d. The minimization can be solved directly using the singular value decomposition (SVD), see

[246] for an accessible derivation.

Note that for computing the mapping we simply assume the desired positions $\mathbf{q}_j$ were given.

We compute them as the distance weighted centroid of encoding fixations:

$$\mathbf{q}_j = \frac{\sum_i \vartheta_p(\|D(\mathbf{r}_j) - \mathbf{p}_i\|)\mathbf{p}_i}{\sum_i \vartheta_p(\|D(\mathbf{r}_j) - \mathbf{p}_i\|)}, \tag{8.4}$$

where $\vartheta_p(d)$ quickly decreases such that points further away are receiving relatively insignificant contribution.

Algorithm 1: Relocation mapping

Data: Fixations locations in encoding $\mathbf{p}_1, \ldots, \mathbf{p}_n$ and recall $\mathbf{r}_1, \ldots, \mathbf{r}_m$, convergence criteria λ

Result: Locations of mapped recall fixations $\mathbf{r}'_1, \ldots, \mathbf{r}'_m$

$\forall (i,j) \in [1,m]^2 : w_{ij} \leftarrow \exp(-\|\mathbf{r}_i - \mathbf{r}_j\|^2 / w_d^2)$

for $j \in [1,m]$ do

  $\mathbf{r}'_j \leftarrow \mathbf{r}_j$

  $\bar{\mathbf{r}}_i \leftarrow \left(\sum_{j=1}^m w_{ij}\mathbf{r}_j\right) / \left(\sum_{j=1}^m w_{ij}\right)$

  $\mathbf{R}_i = (w_{i1}(\mathbf{r}_1 - \bar{\mathbf{r}}_i), \ldots, w_{im}(\mathbf{r}_m - \bar{\mathbf{r}}_i))$

repeat

  for $i \in [1,m]$ do

    $\mathbf{q}_i \leftarrow \frac{\sum_{j=1}^n \exp(-\|\mathbf{r}'_i - \mathbf{p}_j\|^2 / w_p^2)\mathbf{p}_i}{\sum_{j=1}^n \exp(-\|\mathbf{r}'_i - \mathbf{p}_j\|^2 / w_p^2)}$

  $\sigma \leftarrow 0$

  for $i \in [1,m]$ do

    $\bar{\mathbf{q}} \leftarrow \left(\sum_{j=1}^m w_{ij}\mathbf{q}_j\right) / \left(\sum_{j=1}^m w_{ij}\right)$

    $\mathbf{Q} = (\mathbf{q}_1 - \bar{\mathbf{q}}, \ldots, \mathbf{q}_m - \bar{\mathbf{q}})$

    $\mathbf{U}\mathbf{V}^\mathsf{T} \leftarrow \mathsf{SVD}\left(\mathbf{R}_i\mathbf{Q}^\mathsf{T}\right)$

    $\mathbf{s} \leftarrow \mathbf{V}\mathbf{U}^\mathsf{T}(\mathbf{r}_i - \bar{\mathbf{r}}_i) + \bar{\mathbf{q}}$

    $\sigma \leftarrow \sigma + \|\mathbf{r}'_i - \mathbf{s}\|$

    $\mathbf{r}'_i \leftarrow \mathbf{s}$

until $\sigma < \lambda$

Note that we are considering the distances of the fixations $\mathbf{p}_i$ to the *relocated* locations of the fixations in the recall sequence. This means that setting the target locations depends

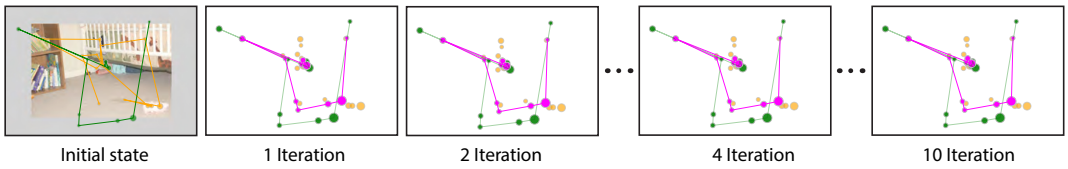


Figure 8.6: Iterations of the relocation mapping. The initial state shows two fixation sets of encoding (orange) and recall (green). The size of each circle corresponds to the duration of each fixation. Pink circles are the deformed recall fixations after each iteration. Note that the matching converges quickly after a few iterations.

on the relocation mapping and computing the relocation mapping depends on the target locations. Consequently, we alternate the two steps as shown in Algorithm 1. We start this process with D being the identity. Then we compute the desired positions $\mathbf{q}_j$ as explained above. The procedure converges after very few iterations (see Figure 8.6). Note that the relocation D needs to be evaluated only in the location $\mathbf{r}_j$. This means for the next step we only need to compute $\mathbf{R}_\mathbf{x}$ and $\mathbf{t}_\mathbf{x}$ for $\mathbf{x} = \mathbf{r}_j$.

8.4.3 Quantitative Results of the Mapping

In the previous section, we presented an algorithm that relocates recall fixations in the stimulus space. Next, we will match each relocated recall fixation to the original encoding fixations. This mapping will depend on the choice of the matching radius ε. First, we present global statistics to show the effect of this procedure.

Note that as there are fewer fixations in recall, after we have established the mapping, we select those encoding fixations that were matched to recall fixations. The *leftover* encoding fixations correspond to scene content that was never expressed in recall fixations, i.e. the encoding fixations that were forgotten. The number of leftover encoding fixation depend on the matching radius ε, which gives us a measure of the reduction rate (number of leftover encoding fixations divided by the original number of encoding fixations).

In order to understand whether relocation of recall fixations is important to achieve the results we present, we also calculate the rate of leftover encoding fixations based on the original, un-relocated recall fixations. We expect that even for relatively small matching radii ε, a significant number of fixations remain after a meaningful matching. The graph in Figure 8.7(a) depicts the number of matched encoding fixations as a function of the matching radius. As expected, matching based on the relocated recall fixations lead to better preservation of encoding fixations, when using a small ε. When we do not relocate recall fixations, fewer encoding fixations are matched. For $\varepsilon = 1°$ nearly 20% of encoding fixations are matched to relocated recall fixations, whereas only 5% are matched to original, un-relocated recall fixations. Note that this effect is quite similar across different images and observers. For values larger than

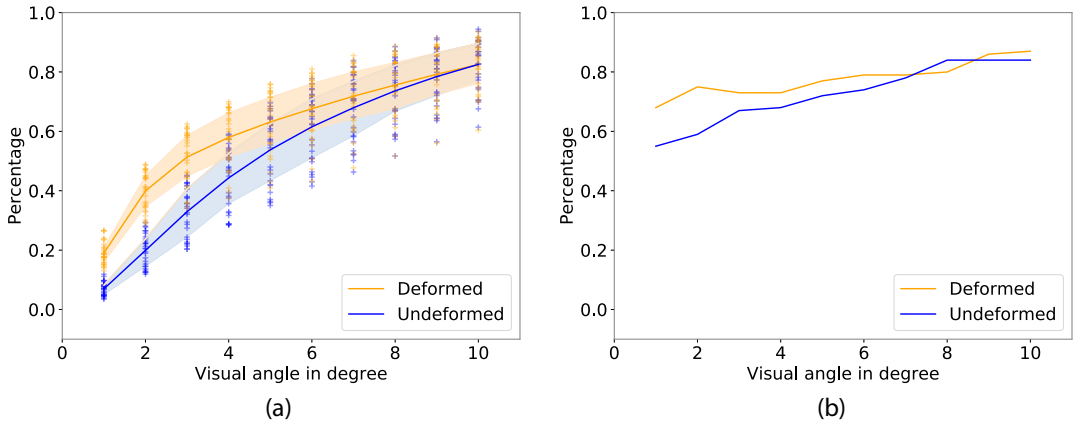


Figure 8.7: In (a), the reduction rate as a function of matching radius $\varepsilon \in [1°, 10°]$. The original recall fixations are shown in blue. Because they match with fewer of the encoding fixations, reduction rate is higher for the original recall fixations. Recall fixations that have been relocated are more likely to be mapped onto an encoding fixation (yellow), and therefore there will be fewer leftover encoding fixations and a lower reduction rate for relocated recall fixations, in particular when ε is small. (b) This figure shows for what proportion of the 100 images the peak of the gaze density map for encoding fixations is within $4°$ of the peak of the gaze density map for the matched encoding fixations based on the relocated recall fixations, as a function of ε. For comparison, also the matching based on the un-relocated recall fixations, for which fewer gaze density peaks are within $4°$ of the peak for encoding fixations. For small ε, peak position in more images are kept the same, which demonstrates the effectiveness of the relocation mapping.

$\varepsilon = 10°$, we effectively match all encoding fixations. In other words, $10°$ is so big that every encoding fixation will be matched to a recall fixation, which contradicts our earlier observation that the number of recall fixations is smaller than the number of encoding fixations, i.e. some elements are forgotten / not prioritized and the ε should be low enough that the method shows that.

We also analysed whether the matching procedure shifts the gaze density map, as opposed to just uniformly reducing the number of fixations without affecting the resulting distribution. For this we identify in each image the region with the most fixations during encoding, based on the smoothed spatial histogram. We do the same for the matched encoding fixations, which are fewer, and consider that the region has shifted if the difference in spatial location exceeds $4°$. The right graph in Figure 8.7 shows the resulting difference as a function of the matching radius ε. As expected, for large matching radii, the distribution of encoding fixations that could be matched against relocated recall fixations have a distribution very similar to the distribution of encoding fixations that are matched against the original un-relocated recall fixations. For smaller radii, there is a shift in the gaze density distribution in about 30% of the images when matching is made against relocated recall fixations. Matching against the

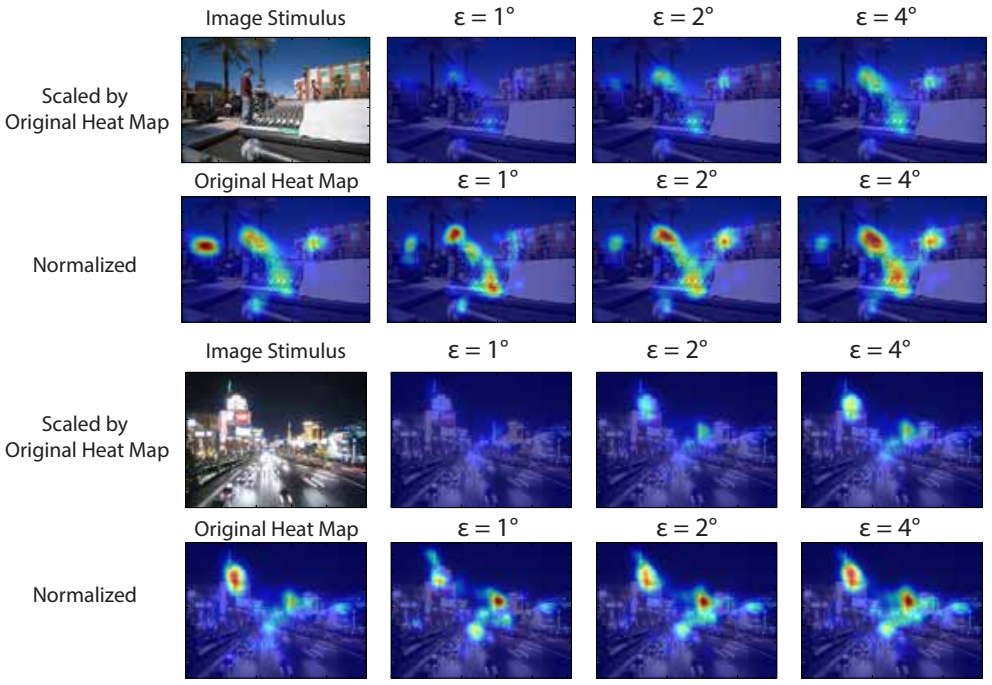


Figure 8.8: Matched encoding maps using different thresholds. We show examples of $\varepsilon = 1^\circ, 2^\circ, 4^\circ$. Original heat map shows the result based on all fixations during encoding of the image. The colour coding of the scaled maps is based on the highest value appearing in all heat maps, which is naturally the one generated without matching. They are used to show the resulting maps under the same scale. The normalized map is colour coded in its own range.

original recall fixations leads to distribution shifts in 40% of the images.

Figure 8.8 illustrates the same effect of varying the threshold on the resulting gaze density maps for two of the images, for the values $\varepsilon = 1^\circ, 2^\circ, 4^\circ$. We can observe that some regions that attracted a lot of encoding fixations appear to have been less recalled, for example the sign to the left in the top image. Other regions are more fixated during recall, such as the objects in the foreground in the top image or the tower to the right of the road in the bottom image.

8.4.4 Mapping Effect

The results of the deformation process are stable across a wide range of parameters (see Figure 8.9). We found that the radius for matching recall fixations to encoding fixations w_p should be chosen in the range of $2^\circ - 4^\circ$ of visual angle. This means we expect that matching fixations are usually not separated by more than twice this amount. To limit the deformation of the mapping we have tried values of $w_d \in [4^\circ, 16^\circ]$. Based on experimentation we have settled for $w_p = 2^\circ$ and $w_d = 10^\circ$. Figure 8.10 shows the results of three examples using

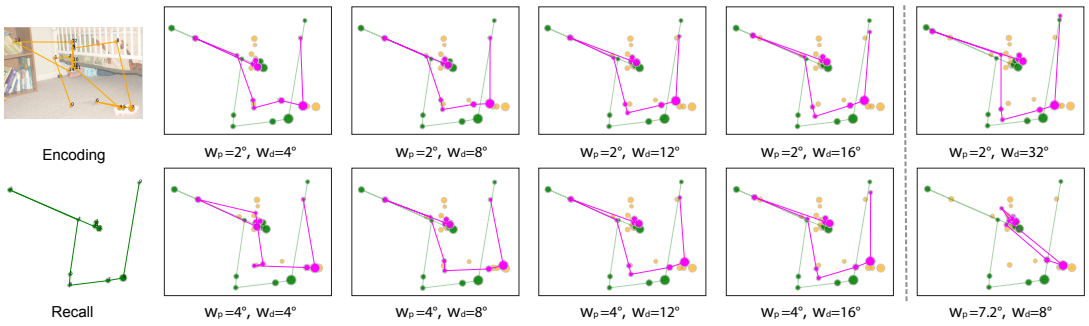


Figure 8.9: Matching results using different parameters. Original recall fixations (in green) are matched to fixations during encoding (in yellow) and relocated recall fixations are shown in pink. Matching between recall and encoding is stable across all parameters, except in the rightmost column, which shows cases of undesired mapping for extreme parameter settings. More encoding fixations are considered for matching with a larger w_p while the relocation mapping is more rigid with a larger w_d.

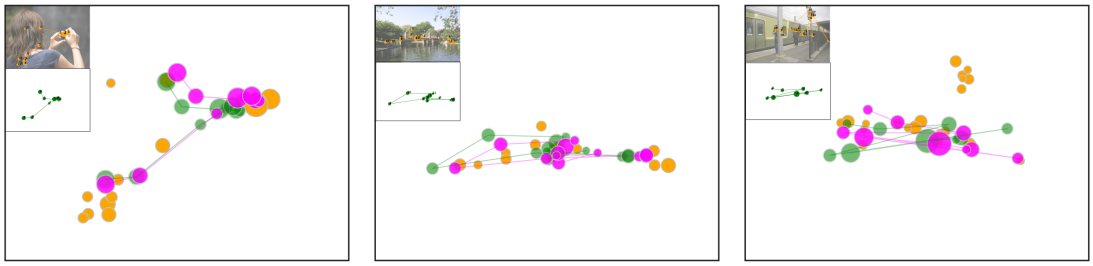


Figure 8.10: The relocation result on three different examples. Parameters w_p is set to 2° and w_d to 10°. Fixations during encoding are depicted in yellow and recall fixations in green. Mapped recall fixations are shown in pink. Encoding fixation sequences overlaid with the stimulus and the corresponding original recall fixation sequences are illustrated at the top-left corners.

this parameter setting.

The outcomes of the process for relocating recall fixations are stable across a wide range of parameters, as Figure 8.9 illustrates. We found that the radius for matching recall fixations to encoding fixations w_p should be chosen in the range of $2^\circ - 4^\circ$ of visual angle. This means we expect that matching fixations are usually not separated by more than twice this amount. To limit the amount of the relocation in the mapping we have tried values of $w_d \in [4^\circ, 16^\circ]$. We applied the relocation process to each pair (encoding/recall) of fixation sequences for all images. Then we applied a matching using $\varepsilon = 1^\circ$, such that an encoding fixation was matched only if there was a corresponding relocated recall fixation within 1° visual angle. This was a rather strict setting, leaving many encoding fixations unmatched. We optioned for a small ε for stronger effects.

Figure 8.10 shows three examples of relocated recall fixations based on the corresponding encoding fixations, using this parameter setting.

8.5 Performance of the Proposed Method

8.5.1 Do mapped recall fixations prioritise subjectively important scene elements?

We expect the matched encoding fixations based on relocated recall ones to fall on positions in the scene that were judged important by the clicking data. To investigate this, we first compared the similarity between clicking maps (such as Figure 8.4) and encoding maps (i.e. gaze density maps from encoding fixations) versus matched encoding maps (i.e. gaze density maps from matched encoding fixations). Area under ROC curve[131] and correlation coefficients (CC) measures[30,105] are used. The averaged similarity scores are plotted on the left in Figure 8.11 where encoding maps (blue bars) and the the matched encoding maps (orange bars) were compared to clicking maps. Additionally, we also compared the clicking data to gaze density maps when randomly paired recall sequences are used in mapping (green bars). Even though large standard deviations are presented in the results, the decreases in similarity when random sequences were used for matching are significant comparing to the matched encoding results, regardless of the measurement ($t(198) = 8.75, p \ll 0.01$ using ROC and $t(198) = 19.84, p \ll 0.01$ using CC, two-tailed student t-test). As shown on the right in Figure 8.11, we then conducted similar comparisons between clicking maps versus the recall maps (i.e. gaze density maps from original un-relocated recall fixations) (purple bars) versus the relocated recall maps (i.e. gaze density maps from relocated recall fixations) (yellow bars). Light green bars on the right show the similarities between clicking maps and the gaze density maps of the relocated recall fixations when randomly selected encoding sequences are used for relocation mapping. We observed significant decreases in matching from randomly paired sequences comparing to matching between the corresponding pairs ($t(198) = 4.69, p \ll 0.01$ using ROC and $t(198) = 13.02, p \ll 0.01$ using CC, two-tailed student t-test). The larger similarity between clicking and recall fixations that have been relocated (yellow bars on the right) as well as the matched encoding fixations (orange bars on the left) suggests that the proposed matching algorithm produces a meaningful outcome, in the sense that the areas clicked as important are largely equal to most of the spontaneously prioritized areas during visual imagery / recall.

Figure 8.12 shows all three data for three example photos. In all cases, the peak - the most prioritized area - is the same. It can be clearly seen that during encoding there is more gaze outside the top peak, than during recall, and that during clicking, the activity outside the peak has fallen almost to zero. The first example photo shows that the low entropy in both the click map and the matched encoding map would correlate better than clicking with the

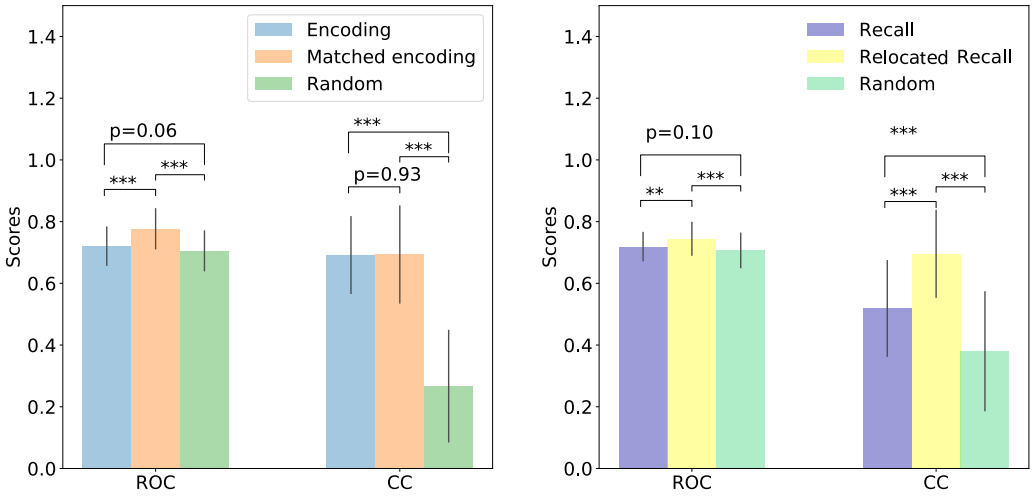


Figure 8.11: Averaged ROC and CC similarity scores. All comparisons are made against click data for the most important position in the image. On the left, the blue bars compare gaze density maps during encoding (i.e. the encoding maps) to clicking (i.e. the clicking maps), while the orange bars compare the matched encoding fixations (i.e. the matched encoding maps) to clicking, and finally, the green bars compare the matched encoding maps based on random recall sequences to clicking. On the right, the blue bars instead compare gaze density maps during recall (i.e. the recall maps) to clicking, while orange bars compare the relocated and mapped recall results (i.e. the relocated recall maps) to clicking, and the green bars show the relocated recall maps when random encoding sequences are used for matching. Student t-test was used with sample size $n = 100$. $**$ is $p < 0.005$ and $***$ is $p < 0.0005$. Together the results indicate the proposed relocation algorithm is meaningful as the resulting gaze density maps have a higher correlation to clicking (importance), depicted by the orange and yellow bars.

encoding maps. But the differences of correlations in the other two examples go down with an increased entropy in both encoding and matched encoding maps.

8.5.2 Applying the Method to Study What Scene Content is Prioritised during Recall

Our analysis was based on comparing the heat maps on images from all encoding fixations, the encoding maps, against all the encoding fixations that are matched to the relocated recall fixations, the matched encoding maps, averaged over all observers. For each image, the difference between the encoding map and the matched encoding map defines the leftover encoding fixations, made on objects that were not fixated during recall, and thus not prioritized during recall. In the subsequent examples, we present tentative effects of prioritization during recall, that can be seen using our method, and which could be investigated as hypotheses in future experimental studies. Recall fixations in these examples always refer to relocated recall

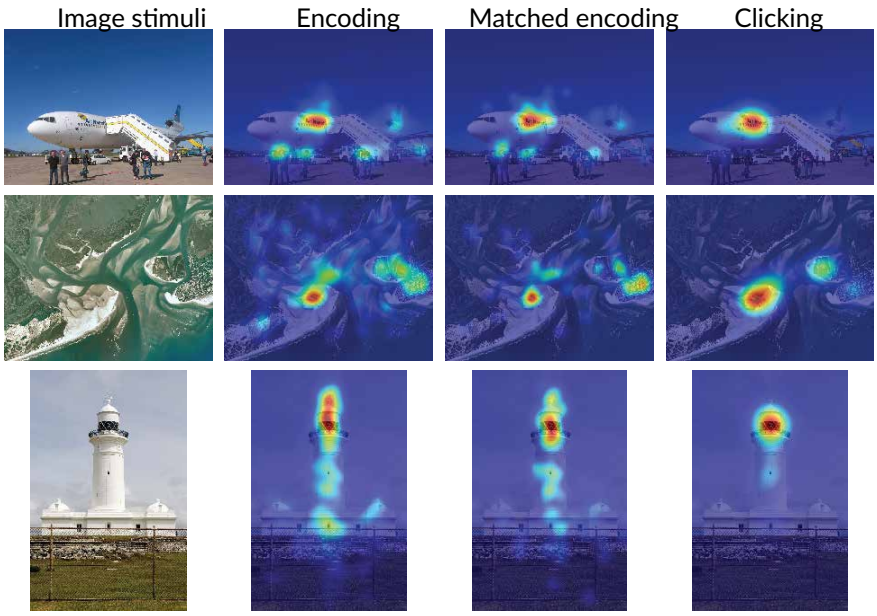


Figure 8.12: Examples of photographic images, gaze density maps for original encoding, for matched encoding based on relocated recall, and clicking data. Three examples are chosen such that the differences of the ROC scores are the largest, the median and the smallest.

fixations, not the original ones.

8.5.3 Example 1: Low-level Features

Low-level features have been shown to contribute to saccade target selection[9,143,155]. Consequently, low-level features contribute significantly to fixation-based attentional models. Figure 8.13 shows three examples of un-prioritized / forgotten scene elements. A white box in the lower right corner over a black background in the first image draws a lot of encoding fixations, possibly due to its high contrast. This box is less dominant in the resulting heat maps as shown in the last columns. The wall sign with the number 3 has a similar fate. In the second example, the red car behind the trees is large gone in the matched encoding fixations, in favour of fixations on the train. In the third example, the colourful ball is much less fixated during recall, but the little girl became relatively more fixated.

To quantify this effect in our non-experimental stimuli, we therefore calculated the feature values for luminance, chromaticity, contrast and edge-content following Tatler et al.[252]. As shown on the left in Figure 8.14, no differences in low-level feature values could be found between positions of the leftover fixations during encoding and the matched (the remembered) fixations. The equal distribution of low-level features in leftover and matched encoding fixations may suggest that the prioritization in episodic memory is not influenced by low-level features.

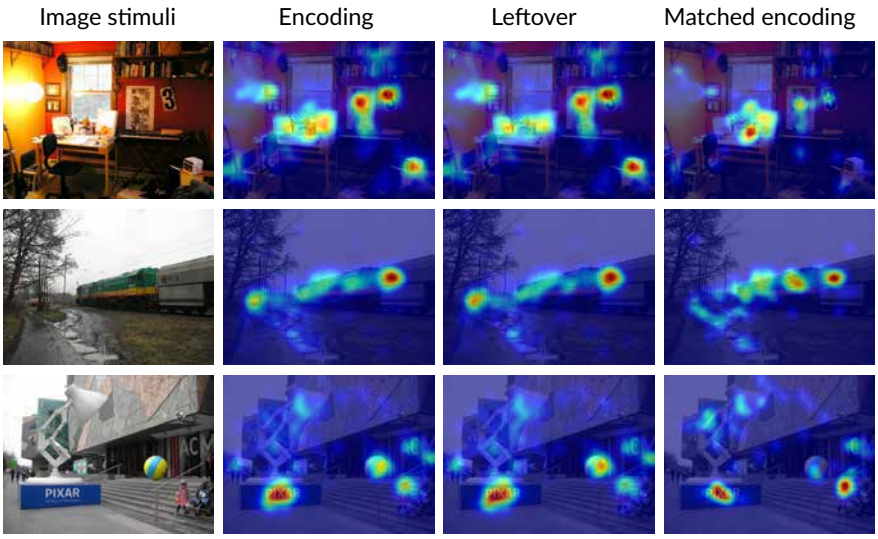


Figure 8.13: Examples of how low level features are not prioritized in recall. The second column shows the encoding map of the image. Leftover map shows the *map of non-prioritised objects* (very loosely "map of forgotten things"), while matched encoding fixations are used to generate the *matched encoding map* (very loosely "map of remembered things"). Each map is normalized in its own range. Fixations triggered by low-level features are largely absent in the *matched encoding* maps.

8.5.4 Example 2: High-level Features

Many high-level (or semantic) features have been known to attract fixations during image inspection, most prominently faces but also signs and other artifacts and out-of-context objects. We have high-level examples in our stimuli in the form of text, signs, and people. Figure 8.15 shows two examples. However, the variability in these uncontrolled photos made the comparison non-significant (Figure 8.14 right).

To further examine the role of high-level features, we excluded the half of the participants whose recall fixations were less correlated to their encoding fixations, and then repeated the analysis for the remaining participants - with highly correlated fixations. The results showed a significant effect of not prioritizing text ($t(54) = 2.83$, $p \ll 0.05$, two-tailed student t-test), indicating that text elements may not have a high recall priority once their meaning is decoded.

Additionally, we considered all features in each category regardless of their size. Size matters for attention, and may affect the prioritisation. In order to investigate the effect of size, we used the many differently sized faces in our diverse stimulus. Faces in photos are known to be fixation magnets, and hence also definite candidates for well-remembered objects in future controlled studies.

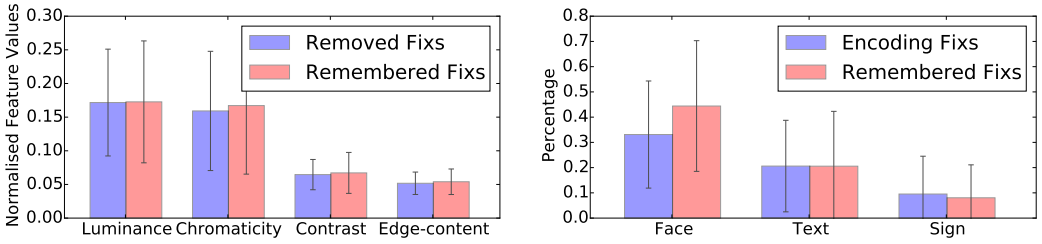


Figure 8.14: Left: The comparison of low-level feature values at fixated positions. Compared low-level features are luminance, chromaticity, contrast and edge-content. No difference between leftover fixations and matched encoding fixations (i.e. the remembered fixations) can be observed. Right: The percentage of fixations on faces, texts and signs in the encoding phase, and the corresponding number of matched encoding fixations (i.e. the remembered fixations) on the same objects. For faces, the p-value is 0.09.

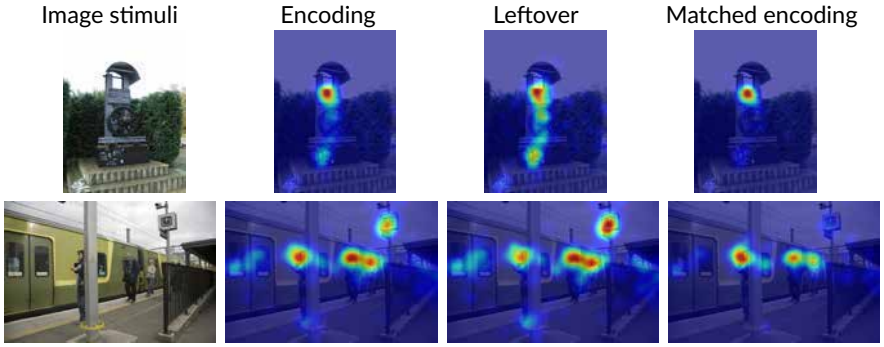


Figure 8.15: Text in the first example is entirely un-fixated during recall as shown in the last column, as well as the sign in the second example, where the people are fixated during recall, at the cost of other objects in the photos. Illegible or irrelevant text (upper row) as well as signs without particular relevance for the scene (lower row) are almost entirely un-fixated during recall.

8.5.5 Example 3: Big enough People are Remembered

There was an effect of size of people and faces on recall fixations. If a person was big enough, he or she would be fixated during recall, as in Figure 8.16, usually at the cost of other things in the image not being fixated during the 5 seconds that were available in the LAN recording. However, if people or faces were small in the photo, like in the top two rows of Figure 8.16, they are largely un-fixated during recall. Figure 8.17 summarizes this size effect for all people and faces in the photos. In order to be spontaneously fixated during recall of a photo, your body should fill around $2^\circ \times 2^\circ$ or more of the photo.

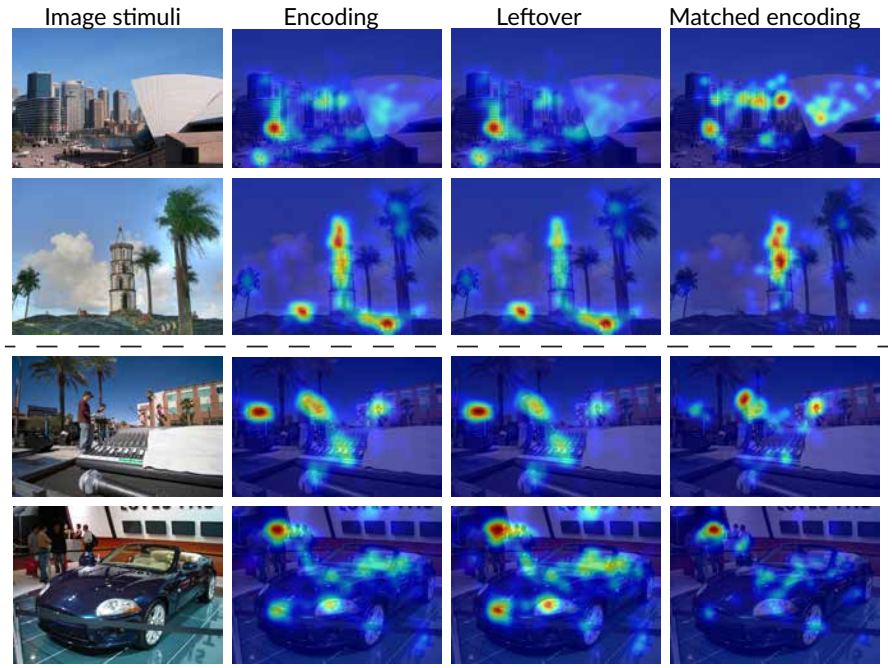


Figure 8.16: People in the scene are fixated during encoding but their reappearance during the recall fixation depends on their size in the photo. In the upper rows, the small people are not prioritized in recall, while in the lower rows they are big enough to be remembered. The meaning of the colour coded images is as before.

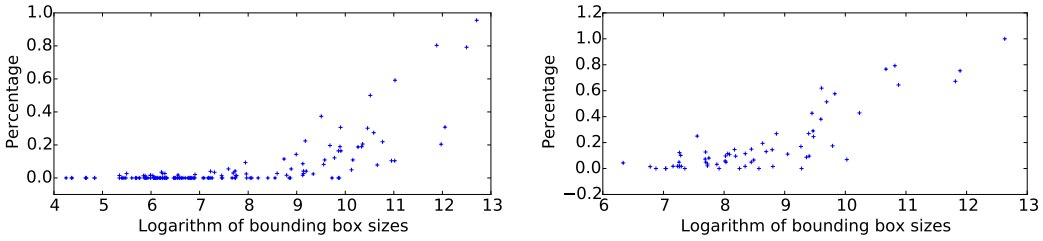


Figure 8.17: Left: the probability of a person being fixated during recall as a function of the size of the person in the photo. Small-sized people are forgotten, but at a certain extent, people are remembered. Right: The probability of a face being fixated during recall as a function of the size of the person in the photo. Faces covering a small area are forgotten, while larger faces are commonly recalled.

8.5.6 Further Examples of Memory Effects that could be Studied with this Method

We argue that our paradigm is general. With more controlled experimental designs, it can be used to study other memory prioritization effects.

Similar objects in a scene tend to be all inspected during exploration, independent of their

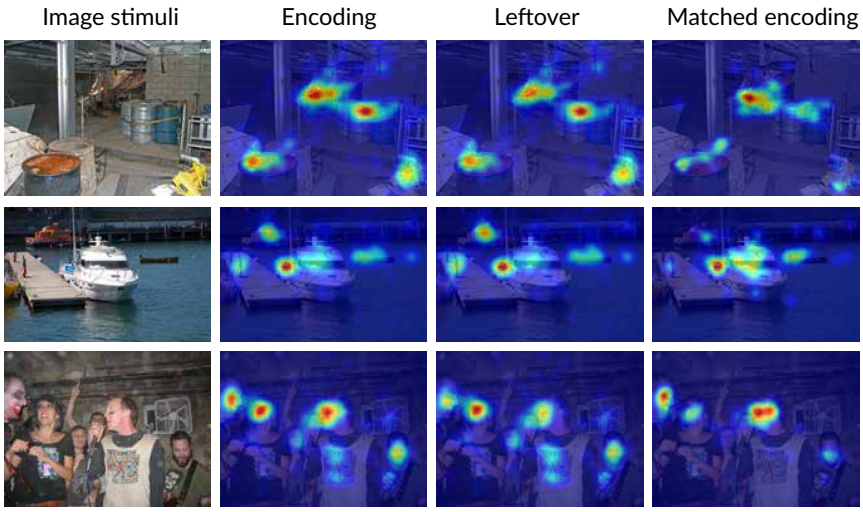


Figure 8.18: Effects on elements with similar meaning. The meaning of the colour coded images is as before. In the top row, similar distributions of encoding fixations among similar items are not preserved after matching against the recall fixations. The smaller boat in the middle row loses in relative priority during recall, while still being dominated by the larger boat. In the last row, the face that stands out among other faces increases in its dominance in the matched encoding map.

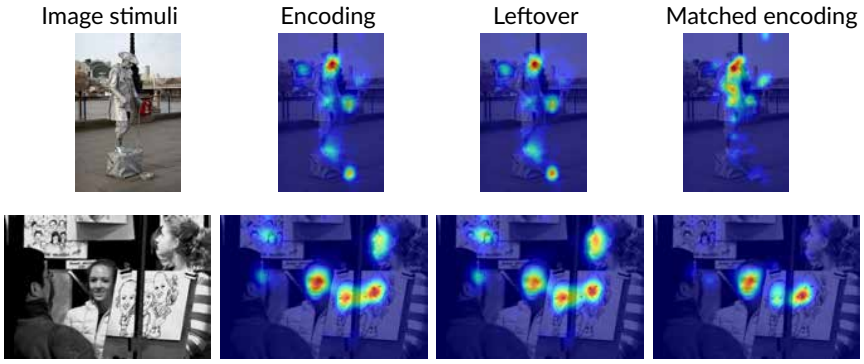


Figure 8.19: The objects looked at by the person in these images are also looked at by our participants during encoding, but much less fixated at during recall.

visual representation. However, the distribution of fixations among similar items does not seem to be preserved during recall and several such examples are shown in Figure 8.18.

For instance, when looking at an image with a person looking at something, we often follow that overt gaze to look at the same item, as it may provide additional information about their intention or the local environment[13,43,74,75,76,85,111]. This gaze deictic effect is notoriously hard to model computationally[220,277]. Figure 8.19 shows two examples of how such gaze following during encoding is not followed by fixations on the same object during recall.

The instances of horizons and mirrors in the 100 stimulus images were items that attracted a lot of fixations during encoding, but appear of little interest when the image is being recalled

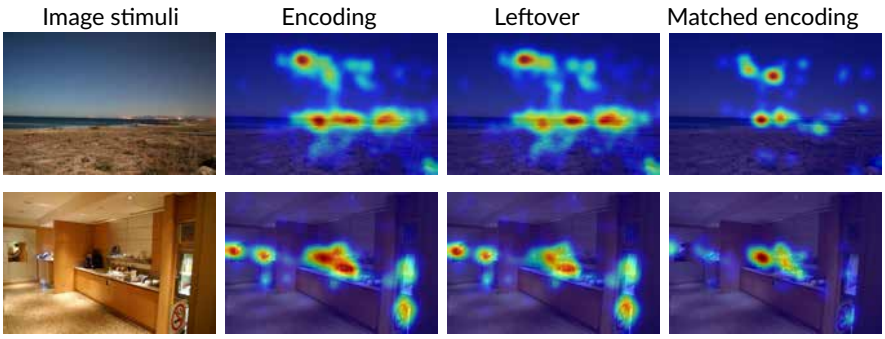


Figure 8.20: Observers scan the horizon during encoding, yet during recall only fixate part of it. Similarly, the mirror is a frequently fixated object during encoding, but it does not gain many fixations during recall.

(see Figure 8.20).

We had too few instances of overt directed gaze, horizons and mirrors in our data set. To produce meaningful quantitative data, future experiments may be designed to investigate recall of these objects thoroughly.

8.6 Discussion

We have presented a novel algorithmic method to study how items within a recently inspected photographic image are prioritized during spontaneous LAN recall. No slow mediation via language or sketching is necessary to find out what is being prioritized in recall from visual episodic memory. Making effortless eye movements during recall does not even need to be instructed, but happens spontaneously for the majority of participants.

The proposed elastic matching algorithm requires no specific parameter setting or fine-tuning. The stable results achieved across a wide range of parameters demonstrates the effectiveness of the matching algorithm as well as the underlying correspondences between fixations during encoding and subsequent recall.

Importantly, when asked to click the most important item in the image, participants click positions consistently coincided with the peak of the gaze density map made from the matched encoding fixations based on relocated recall fixations. This suggests that the peaks in that matched encoding maps are in fact the most prioritized items for recall from episodic memory. Notice that eye movements are significantly faster than clicking, and therefore eye-movement data provide peaks on several items throughout the same trial duration.

The proposed method assumes that what is being prioritized in episodic memory has been fixated before, or in other words, that items not fixated during encoding cannot be recalled. This is contradicted by Underwood et al. [260], who reported that of all the objects that were *not* fixated while participants watched video recordings of moving vehicles, nearly 20% were

nevertheless recalled. When asked to describe pictures, participants often talk about objects that were never fixated before[84]. This seems to be a natural limitation not only to the LAN-based paradigm proposed in this chapter, but also of the method of eye-tracking in general.

Note that our concept of importance differs from the definition of meaningfulness in Henderson & Hayes[97]. Their meaning maps capture the distribution of semantic information across scenes. Results in Henderson & Hayes[97] show that such maps provide a major guidance for attention. Meaningfulness was rated by participants comparing patches of image regions, which elegantly avoids potential interference of the context of the image as a whole. In this way, meaningfulness is mainly controlled by participants' long-term knowledge and experience[95,99]. In contrast, what we investigated here is the meaningfulness (importance) in spontaneous prioritization from episodic visual memory as expressed in LAN fixations.

In fact, visual attention and visual working memory are entangled: attention controls encoding in working memory by selecting what is the relevant information, and visual working memory biases the focus of attention[57,104,207,233]. Several studies suggest that visual attention and visual working memory share the same representations[5,69,91,103], and it has been debated whether they should be regarded as one cognitive function since they also share some of the same mechanisms (see Olivers[206] for a review). Our tentative results suggest that the prioritization in visual attention differs from the prioritization in visual episodic memory.

It appears that the recall performance improves with repeated trials. Similar effects have been observed in other recall experiments[136,179]. Foulsham & Kingstone[73] show that repeated viewing of fixated image content improves scene recognition accuracy. It is not known whether prioritization after repeated viewing of one image differs from that resulting from single viewing.

The paradigm we present offers a computational method to investigate prioritization in *short-term* visual episodic memory. The studies of LAN fixations during *long-time* recall provide inconsistent results, specifically as to whether locations of fixations still encode the locations of features in images[159,181,271]. If our method were to be used in studies of long-term memory, we would expect the amount of noise in the data to increase and, therefore, finding the match between recall and encoding fixations to be more difficult. Future research would have to tell. If the method would work for long-term memory, our paradigm could be used for experiments related to image memorability[29,113,141], which explores the memorability of visual objects, and the results would likely become even more dependent on personal factors. Witness psychology would be one obvious application area. Finally, we must acknowledge that the very fixations during looking at nothing have a fundamentally different function from fixations on a scene such as a photo. They are significantly longer, they do not obey the focal and ambient viewing effects[109,273], and they land on slightly different locations.

8.6.1 Tentative Future Research with this Method

As a form of feasibility investigation in our randomly selected stimuli, we could show that participants make recall fixations on faces/people in the photos, depending on the size of them. People who take up little space, mostly because they are far away when the photo was taken, are much less fixated in recall than close-by, large people (Figure 8.17. This we think is evidence that our method could be used in controlled studies of face and people perception as a function of gender, emotions, social contexts and gaze deixis in the stimuli.

Semantic importance in a general sense has repeatedly reported to attract fixations[98,254] during viewing. In contrast to those studies, hard-to-interpret elements in the scene tend to be less fixated in the LAN recall (Figure 8.13). Further studies are needed of this difference between encoding and LAN recall. Our results, together with the presented examples, suggest that image content that better represent the scene gist are prioritized higher in visual episodic memory.

Image salience models (e.g. Itti et al.[118]) suggests that fixations during encoding are (partially) driven by low-level features, especially during the short period after stimuli onset (although this is disputed, e.g. Nyström & Holmqvist[204]). However, our results show that low-level features are equally distributed between the matched encoding fixations vs. the leftover encoding fixations (Figure 8.13). This suggests that prioritization in recall from visual episodic memory is less driven by low-level features.

Research on LAN and mental imagery could benefit from the mapping function we developed, as it offers means to correct for the spatial distortion of eye movements during recall. More real-world scenes can be studied in such experimental paradigm, and the mapping function allows us to overcome the limitation of grid-based Area of Interest (AOI) analysis[159,180,236] as well as the MultiMatch method[54], which lacks the ability to model non-rigid transformation.

Our data suggest that the relative importance of scene elements during recall in some cases appear to be prioritized differently from the selection of fixation targets. Not all fixations during encoding are spontaneously recalled from episodic memory. In the fields of computer visualization and human-computer interaction, it is often a goal to make specific design elements clearly visible and easily memorable. Researchers create maps of the relative 'importance' of different regions of the stimulus, commonly based on eye tracking, accumulating fixations from several observers[22,26,45,131,132,277]. The underlying reasoning is that most visual information is processed while fixating[186], and saccade target selection is likely optimized to attend to the most important visual elements first[116,210], which may not hold for recall.

9 Conclusion

With rapid developments of hardware, the ubiquitous accessibility to eye movement recording has been increased dramatically over the last decades. Interdisciplinary researches are enabled across the fields of cognitive science, behaviour psychology, neuroscience, as well as applied science. As the core component of the human biological system, eye movements offer the unique possibility to study how human perceive the world and how the mind reacts.

In this thesis, we study eye movement from two distinct perspectives with the aim of focusing on one of its roles at one time. More precisely, we conduct eye tracking experiments on physical 3D shapes (Chapter 4-5), which demonstrate the existence of saliency for real objects. The analysis results indicate that consistent features across different views contain more semantic information but there is no significant fixation dependence on the tested materials. We would be curious to see how other computational models of visual saliency perform on our data, especially with advanced machine learning techniques. However, collection of human data is expensive and the small amount of available data poses another challenge for data-hungry deep learning approaches. In the future, we would also like to explore how the collected visual saliency data of 3D objects could be used in practice, for example simplifying the models in a meaningful way such that time and material cost are reduced in fabrication.

In Chapter 6-8, we conduct studies following the well-established 'looking-at-nothing' paradigm. Chapter 7 presents the study on the similarity between eye movements during encoding and eye movements during recall from memory. Using the shared characteristics of eye movements between these two conditions, we apply it in an image retrieval task using

only eye movements as queries. Our study provides evidence that eye-movement based image retrieval is computationally feasible, even in a more real-world like scenario. Without additional training and practice, naive participants could use our method to retrieve an image. We cannot estimate the feasibility when significantly more images are included in the database, however, the presented results achieve a clear performance benefit over brain activity based measurements. Fewer restrictions are posed on the participants' side, and if we ask them to exercise a small effort to actively imagine a larger image, the recall-based retrieval is very likely to achieve a comparable accuracy to when encoding eye movements are used. Chapter 8 proposes a matching function for encoding vs. recall fixations, and reveals the fixated content that is spontaneously prioritized in recall from episodic memory. Examples of such prioritization show interesting results that differ from the prioritization during image encoding. The proposed matching function can be further used to study eye movements during mental imagery in general, and it also extends the possibility of using the 'looking-at-nothing' paradigm in applications.

The first part of this thesis presents several methods developed for general eye tracking in three dimensional space. In Chapter 2, we develop a system to accurately track eye movements in 3D space. Experimental validation shows that the achieved accuracy is about 0.8° of visual angle, which is on par with the standard eye tracking accuracy on screen. The proposed approach in particular offers the building blocks for studies presented in Part II, and we hope it could enable researchers to extend their data collection without significantly changing their current work flow. Chapter 3 presents the study on the bias of gaze estimation based on vergence. It shows that the estimated point of interest from intersection of eye rays has large error in depth. The mean of the linearly estimated points of vergence is biased and depends on the horizontal vs. vertical noise distribution of the fixation positions. It is generally hard to interpret results for depth from binocular vision and our suggestion is to take the average of fixations of the same target to minimize the uncertainty, in both calibration and experiment phase.

This thesis studies the role of eye movements in both inwards and outwards flows of information processing, but it barely scratches the surface. Besides, we know that eye movements function as a combination of various roles and it is difficult to disentangle them from each other. This remains as a severe challenge for future work. We believe interesting research directions include integration of eye movement recordings and data from other modalities (e.g. speech, gesture, EEG etc.), in both fundamental studies and applied scenarios. Recent developments in machine learning, particularly deep learning, provide powerful computational tools to leverage the advantages of big data. Such computational methods would offer means to better understand the human mind, use the understanding to further build applications that combine different roles of eye movements, develop intelligent tools to assist daily life, and further create a larger impact on society.

A

Analytical analysis of bias

Our variable errors contain independent horizontal and vertical components. Each of them follows the zero-mean Gaussian distribution, which is a symmetric function around 0, i.e. $p(a) = p(-a)$.

To simplify the scenario, we set vertical errors to zero and only consider errors in the horizontal direction. Assume left eye and right eye have identical Gaussian distributions, we have $p(\eta_l, \eta_r) = p(\eta_r, \eta_l)$. The probability of observing one pair of errors is the same when it corresponds to left and right eyes or vice versa.

Based on previous two symmetries, we have the same probability for 8 error pairs as shown in Figure A.1. As η_l and η_r follow two independent Gaussian distribution of the same variance, we have a joint normal distribution. Each circle centered at origin is a contour line where samples have the same probability. Given an error vector (η_l, η_r), we can find the other seven pairs by the following operations using symmetry:

- $p(\eta_l, \eta_r) = p(-\eta_l, \eta_r)$ by changing the sign of left eye,
- $p(\eta_l, \eta_r) = p(\eta_l, -\eta_r)$ by changing the sign of right eye,
- $p(\eta_l, \eta_r) = p(-\eta_l, -\eta_r)$ by changing the sign of both eyes,
- $p(\eta_l, \eta_r) = p(\eta_r, \eta_l)$ by swapping between left and right eyes,
- $p(\eta_l, \eta_r) = p(-\eta_r, \eta_l)$ by swapping and changing one sign,

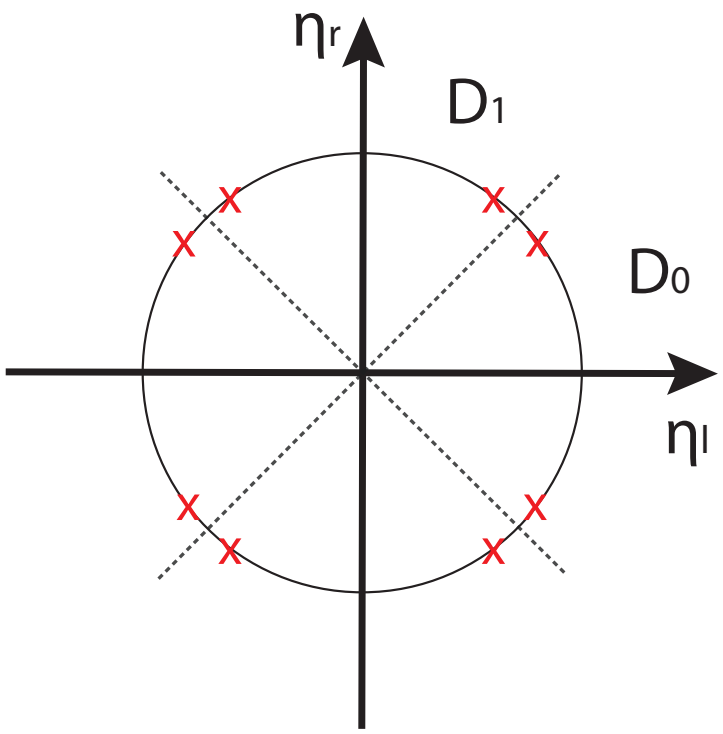


Figure A.1: Symmetry of error pairs. x and y axes correspond to the horizontal errors of left eye η_l and right eye η_r. Eight crosses on the circle have the same probability.

- $p(\eta_l, \eta_r) = p(\eta_r, -\eta_l)$ by swapping and changing one sign,
- $p(\eta_l, \eta_r) = p(-\eta_r, -\eta_l)$ by swapping and changing both signs.

Eight samples are symmetric across four lines which are $\eta_l = 0, \eta_r = 0, \eta_l = \eta_r, \eta_l = -\eta_r$. Together these four lines divide the whole variable domain into eight domains $\{D_i, i = 0, \cdots, 7\}$. We define function Φ_i that maps samples from D_0 to D_i. For example $\Phi_1(\eta_l, \eta_r) \coloneqq (\eta_r, \eta_l)$ which maps sample in D_1 to D_0 as shown in Figure A.1. As Φ_i is a linear map, we know $\det(d\Phi_i) = \det(\Phi_i) = 1$.

Given a pair of horizontal errors (η_l, η_r), we know the intersection point $\mathbf{x}$ can be computed as

$$\mathbf{x}_{\eta_l, \eta_r} = (\mathbf{E}'_l(1) + \mathbf{E}'_r(1))^{-1}(\mathbf{E}'_l(1)\mathbf{p}_l + \mathbf{E}'_r(1)\mathbf{p}_r) \tag{A.1}$$

and its probability is $p(\eta_l, \eta_r)$. Let us define a function f that $f(\eta_l, \eta_r) \coloneqq \mathbf{x}_{\eta_l, \eta_r}(\eta_l, \eta_r) \in D_i$ and $f_i = f \circ \Phi_i$. The expected value of $\mathbf{x}$ in the whole domain is

$$\begin{aligned}
\mathbf{x} &= \iint f(\eta_l, \eta_r) p(\eta_l, \eta_r) d\eta_l d\eta_r \\
&= \sum_{i=0}^{7} \iint_{D_i} f(\eta_l, \eta_r) p(\eta_l, \eta_r) d\eta_l d\eta_r \\
&= \sum_{i=0}^{7} \iint_{\Phi_i(D_0)} f(\eta_l, \eta_r) p(\eta_l, \eta_r) d\eta_l d\eta_r \\
&= \sum_{i=0}^{7} \iint_{D_0} (f \circ \Phi_i)(p \circ \Phi_i) \det(d\Phi_i) d\eta_l d\eta_r \\
&= \sum_{i=0}^{7} \iint_{D_0} f_i p d\eta_l d\eta_r \\
&= 8 \iint_{D_0} \sum_{i=0}^{7} \frac{1}{8} f_i p d\eta_l d\eta_r
\end{aligned} \tag{A.2}$$

Since probability on the same contour line is the same, $p \circ \Phi_i = p$. The expected intersection point $\mathbf{x}$ can be computed by computing the average of f_i. Therefore, we only show the simplified computation of one pair of error vectors in Section Qualitative analysis analysis of bias.

References

[1] Abbott, W. W. & Faisal, A. A. (2012). Ultra-low-cost 3D gaze estimation: an intuitive high information throughput compliment to direct brain-machine interfaces. *Journal of Neural Engineering*, 9, 1–11.

[2] Abrams, R. A., Meyer, D. E., & Kornblum, S. (1989). Speed and accuracy of saccadic eye movements: Characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3).

[3] Agarwal, S., Mierle, K., & Others (2015). Ceres solver. http://ceres-solver.org.

[4] Aherne, F. J., Thacker, N. A., & Rockett, P. I. (1998). The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4), 363–368.

[5] Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology*, 23(15), 1427 – 1431.

[6] Alexa, M. & Wardetzky, M. (2011). Discrete laplacians on general polygonal meshes. *ACM Trans. Graph.*, 30(4), 102:1–102:10.

[7] Altmann, G. T. (2004). Language-mediated eye movements in the absence of a visual world: The 'blank screen paradigm'. *Cognition*, 93(2), B79–B87.

[8] Atchison, D. A., Smith, G., & Smith, G. (2000). *Optics of the human eye*. Butterworth-Heinemann Oxford.

[9] Badcock, D. R., Hess, R. F., & Dobbins, K. (1996). Localization of element clusters: Multiple cues. *Vision Research*, 36(10), 1467 – 1472.

[10] Banitalebi-Dehkordi, A., Nasiopoulos, E., Pourazad, M. T., & Nasiopoulos, P. (2018). Benchmark 3d eye-tracking dataset for visual saliency prediction on stereoscopic 3d video. *arXiv preprint arXiv:1803.04845*.

[11] Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.

[12] Barz, M., Bulling, A., & Daiber, F. (2015). *Computational Modelling and Prediction of Gaze Estimation Error for Head-mounted Eye Trackers*. Technical report, German Research Center for Artificial Intelligence (DFKI).

[13] Bayliss, A. P., Murphy, E., Naughtin, C. K., Kritikos, A., Schilbach, L., & Becker, S. I. (2013). "Gaze leading": Initiating simulated joint attention influences eye movements and choice behavior. *Journal of Experimental Psychology: General*, 142(1), 76–92.

[14] Berryhill, M. E., Phuong, L., Picasso, L., Cabeza, R., & Olson, I. R. (2007). Parietal lobe and episodic memory: Bilateral damage causes impaired free recall of autobiographical memory. *Journal of Neuroscience*, 27(52), 14415–14423.

[15] Besl, P. J. & McKay, N. D. (1992). Method for registration of 3-d shapes. In *Proc. SPIE*, volume 1611 (pp. 586–606).

[16] Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103(3), 597–600.

[17] Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & r. Muller, K. (2008). Optimizing spatial filters for robust eeg single-trial analysis. *IEEE Signal Processing Magazine*, 25(1), 41–56.

[18] Blazhenkova, O. & Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. *Applied cognitive psychology*, 23(5), 638–663.

[19] Bochynska, A. & Laeng, B. (2015). Tracking down the path of memory: eye scanpaths facilitate retrieval of visuospatial information. *Cognitive Processing*, 16(1), 159–163.

[20] Borji, A. & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1), 185–207.

[21] Borji, A. & Itti, L. (2015). Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*.

[22] Borji, A., Lennartz, A., & Pomplun, M. (2015). What do eyes reveal about the mind?: Algorithmic inference of search targets from fixations. *Neurocomputing*, 149, Part B, 788 – 799.

[23] Borji, A., Tavakoli, H. R., Sihite, D. N., & Itti, L. (2013). Analysis of Scores, Datasets, and Models in Visual Saliency Prediction. In *2013 IEEE International Conference on Computer Vision* (pp. 921–928).: IEEE.

[24] Bradski, G. (2000). OpenCV. *Dr. Dobb's Journal of Software Tools*.

[25] Brandt, S. A. & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9(1), 27–38.

[26] Bruce, N. D. B. & Tsotsos, J. K. (2005). Saliency based on information maximization. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05 (pp. 155–162). Cambridge, MA, USA: MIT Press.

[27] Bulbul, A., Capin, T., Lavouè, G., & Preda, M. (2011). Assessing visual quality of 3-d polygonal models. *IEEE Signal Processing Magazine*, 28(6), 80–90.

[28] Bulling, A., Alt, F., & Schmidt, A. (2012). Increasing the security of gaze-based cued-recall graphical passwords using saliency masks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12 (pp. 3011–3020). New York, NY, USA: ACM.

[29] Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, 116, Part B, 165–178. Computational Models of Visual Attention.

[30] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3), 740–757.

[31] Cassin, B., Rubin, M. L., & Solomon, S. (1984). *Dictionary of eye terminology*, volume 10. Triad Publishing Company Gainsville.

[32] Castelhano, M. S. & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human perception and performance*, 34(3), 660.

[33] Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2008). Predicting human gaze using low-level saliency combined with face detection. In *Advances in neural information processing systems* (pp. 241–248).

[34] Cerrolaza, J. J., Villanueva, A., & Cabeza, R. (2012). Study of polynomial mapping functions in video-oculography eye trackers. *ACM Trans. Comput.-Hum. Interact.*, 19(2), 10:1–10:25.

[35] Chadwick, M. J., Hassabis, D., Weiskopf, N., & Maguire, E. A. (2010). Decoding individual episodic memory traces in the human hippocampus. *Current Biology*, 20(6), 544 – 547.

[36] Chan, J. P., Kamino, D., Binns, M. A., & Ryan, J. D. (2011). Can changes in eye movement scanning alter the age-related deficit in recognition memory? *Frontiers in psychology*, 2, 92.

[37] Chao, I., Pinkall, U., Sanan, P., & Schröder, P. (2010). A simple geometric model for elastic deformations. *ACM Trans. Graph.*, 29(4), 38:1–38:6.

[38] Chen, X., Saparov, A., Pang, B., & Funkhouser, T. (2012). Schelling points on 3d surface meshes. *ACM Trans. Graph.*, 31(4), 29:1–29:12.

[39] Chen, X., Starke, S. D., Baber, C., & Howes, A. (2017). A cognitive model of how people make decisions through interaction with visual displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17 (pp. 1205–1216). New York, NY, USA: ACM.

[40] Cohen, S. & Guibas, L. (1999). The earth mover's distance under transformation sets. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2 (pp. 1076–1083).

[41] Collewijn, H., Erkelens, C. J., & Steinman, R. M. (1995). Voluntary binocular gaze-shifts in the plane of regard: Dynamics of version and vergence. *Vision Research*, 35(23), 3335 – 3358.

[42] Collewijn, H., Erkelens, C. J., & Steinman, R. M. (1997). Trajectories of the human binocular fixation point during conjugate and non-conjugate gaze-shifts. *Vision Research*, 37(8), 1049 – 1069.

[43] Corkum, V. & Moore, C. (1998). The origins of joint visual attention in infants. *Developmental psychology*, 34(1), 28–38.

[44] Cournia, N., Smith, J. D., & Duchowski, A. T. (2003). Gaze-vs. hand-based pointing in virtual environments. In *CHI'03 extended abstracts on Human factors in computing systems* (pp. 772–773).: ACM.

[45] Coutrot, A. & Guyader, N. (2014). How saliency, faces, and sound influence gaze in dynamic social scenes. *Journal of Vision*, 14(8), 5.

[46] Cowen, A. S., Chun, M. M., & Kuhl, B. A. (2014). Neural portraits of perception: reconstructing face images from evoked brain activity. *Neuroimage*, 94, 12–22.

[47] Cui, X., Jeter, C. B., Yang, D., Montague, P. R., & Eagleman, D. M. (2007). Vividness of mental imagery: Individual variability can be measured objectively. *Vision Research*, 47(4), 474 – 478.

[48] Cuturi, M. & Avis, D. (2014). Ground metric learning. *J. Mach. Learn. Res.*, 15(1), 533–564.

[49] Dalziel, C. C. (1981). Effect of vision training on patients who fail sheard's criterion. *American journal of optometry and physiological optics*, 58(1), 21–23.

[50] Davis, J. J. J., Lin, C.-T., Gillett, G., & Kozma, R. (2017). An integrative approach to analyze eeg signals and human brain dynamics in different cognitive states. *Journal of Artificial Intelligence and Soft Computing Research*, 7(4), 287–299.

[51] De Graef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, 52(4), 317–329.

[52] de Meij, L., Telleman, M., Luijten, M., Polling, J. R., & Gutter, M. (2017). An optimal measurement of fixation disparity using ogle's apparatus. *Strabismus*, 25(3), 128–133.

[53] de Vito, S., Buonocore, A., Bonnefon, J.-F., & Della Sala, S. (2014). Eye movements disrupt spatial but not visual mental imagery. *Cognitive processing*, 15(4), 543–549.

[54] Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4), 1079–1100.

[55] Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, 23(5), 423 – 434.

[56] Ding, Y., Naber, M., Gayet, S., Van der Stigchel, S., & Paffen, C. L. E. (2018). Assessing the generalizability of eye dominance across binocular rivalry, onset rivalry, and continuous flash suppressionDing et al. *Journal of Vision*, 18(6), 6–6.

[57] Downing, P. E. (2000). Interactions between visual working memory and selective attention. *Psychological Science*, 11(6), 467–473. PMID: 11202491.

[58] Drewes, J. (2014). Smaller is better: drift in gaze measurements due to pupil dynamics. *PLoS ONE*, 9(10), 1–6.

[59] Dryden, I. L., Koloydenko, A., & Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3), 1102–1123.

[60] Du, C., Du, C., & He, H. (2017). Sharing deep generative representation for perceived image reconstruction from human brain activity. In *Neural Networks (IJCNN), 2017 International Joint Conference on* (pp. 1049–1056).: IEEE.

[61] Duchowski, A. T., Pelfrey, B., House, D. H., & Wang, R. (2011). Measuring gaze depth with an eye tracker during stereoscopic display. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization - APGV '11* (pp.15). New York, New York, USA: ACM Press.

[62] Dutagaci, H., Cheung, C. P., & Godil, A. (2011). Evaluation of 3d interest point detection techniques. In *Proceedings of the 4th Eurographics Conference on 3D Object Retrieval*, 3DOR '11 (pp. 57–64). Aire-la-Ville, Switzerland, Switzerland: Eurographics Association.

[63] Dutagaci, H., Cheung, C. P., & Godil, A. (2012). Evaluation of 3d interest point detection techniques via human-generated ground truth. *The Visual Computer*, 28(9), 901–917.

[64] Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3), 589–599.

[65] Eitz, M., Hildebrand, K., Boubekeur, T., & Alexa, M. (2011). Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Transactions on Visualization and Computer Graphics*, 17(11), 1624–1636.

[66] Erkelens, C. J., Steinman, R. M., & Collewijn, H. (1989). Ocular vergence under natural conditions. ii. gaze shifts between real targets differing in distance and direction. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 236(1285), 441–465.

[67] Essig, K., Pomplun, M., & Ritter, H. (2006). A neural network for 3D gaze recording with binocular eye trackers. *International Journal of Parallel, Emergent and Distributed Systems*, 21(2), 79–95.

[68] Esteves, A., Velloso, E., Bulling, A., & Gellersen, H. (2015). Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (pp. 457–466).: ACM.

[69] Farah, M. J. (1985). Psychophysical evidence for a shared representational medium for mental images and percepts. *Journal of Experimental Psychology: General*, 114(1), 91.

[70] Feixas, M., Sbert, M., & González, F. (2009). A unified information-theoretic framework for viewpoint selection and mesh saliency. *ACM Trans. Appl. Percept.*, 6(1), 1:1–1:23.

[71] Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Sciences*, 12(11), 405 – 410.

[72] Förstner, W. & Moonen, B. (2003). A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium* (pp. 299–309). Springer.

[73] Foulsham, T. & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: An experimental test of scanpath theory. *Journal of Experimental Psychology: General*, 142(1), 41.

[74] Foulsham, T. & Lock, M. (2014). How the eyes tell lies: Social gaze during a preference task. *Cognitive Science*, 39(7), 1704–1726.

[75] Freeth, M., Foulsham, T., & Kingstone, A. (2013). What affects social attention? social presence, eye contact and autistic traits. *PLOS ONE*, 8(1), 1–10.

[76] Gallup, A. C., Hale, J. J., Sumpter, D. J. T., Garnier, S., Kacelnik, A., Krebs, J. R., & Couzin, I. D. (2012). Visual attention and the acquisition of information in human crowds. *Proceedings of the National Academy of Sciences*, 109(19), 7245–7250.

[77] Gbadamosi, J. & Zangemeister, W. H. (2001). Visual imagery in hemianopic patients. *Journal of Cognitive Neuroscience*, 13(7), 855–866.

[78] Glatz, C., Krupenia, S. S., Bülthoff, H. H., & Chuang, L. L. (2018). Use the right sound for the right job: Verbal commands and auditory icons for a task-management system favor different information processes in the brain. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 472).: ACM.

[79] Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Y. W. Teh & D. M. Titterington (Eds.), *AISTATS*, volume 9 of *JMLR Proceedings* (pp. 249–256).: JMLR.org.

[80] Godwin, H. J., Menneer, T., Cave, K. R., Thaibsyah, M., & Donnelly, N. (2015). The effects of increasing target prevalence on information processing during visual search. *Psychonomic bulletin & review*, 22(2), 469–475.

[81] Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Sciences*, 17(11), 585 – 593.

[82] Gottschalk, S., Lin, M. C., & Manocha, D. (1996). Obbtree: A hierarchical structure for rapid interference detection. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96 (pp. 171–180). New York, NY, USA: ACM.

[83] Gower, J. C. & Dijksterhuis, G. B. (2004). *Procrustes problems*, volume 30. Oxford University Press on Demand.

[84] Griffin, Z. M. & Spieler, D. H. (2006). Observing the what and when of language production for different age groups by monitoring speakers' eye movements. *Brain and Language*, 99(3), 272–288.

[85] Gullberg, M. & Holmqvist, K. (2006). What speakers do and what addressees look at: Visual attention to gestures in human interaction live and on video. *Pragmatics and Cognition*, 14(1), 53–82.

[86] Gurtner, L. M., Bischof, W. F., & Mast, F. W. (2019). Recurrence quantification analysis of eye movements during mental imagery. *Journal of Vision*, 19(1), 17–17.

[87] Gutierrez Mlot, E., Bahmani, H., Wahl, S., & Kasneci, E. (2016). 3d gaze estimation using eye vergence. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies*, BIOSTEC 2016 (pp. 125–131). Portugal: SCITEPRESS - Science and Technology Publications, Lda.

[88] Häkkinen, J., Kawai, T., Takatalo, J., Mitsuya, R., & Nyman, G. (2010). What do people look at when they watch stereoscopic movies? In A. J. Woods, N. S. Holliman, & N. A. Dodgson (Eds.), *Stereoscopic Displays and Applications XXI*.

[89] Hammer, J. H., Maurus, M., & Beyerer, J. (2013). Real-time 3d gaze analysis in mobile applications. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, ETSA '13 (pp. 75–78). New York, NY, USA: ACM.

[90] Hanhart, P. & Ebrahimi, T. (2014). EYEC3D: 3D video eye tracking dataset. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)* (pp. 55–56).: IEEE.

[91] Harrison, S. A. & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 458(7238), 632.

[92] Hassabis, D. & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, 11(7), 299 – 306.

[93] Hayhoe, M. & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4), 188 – 194.

[94] Hebb, D. O. (1968). Concerning imagery. *Psychological review*, 75(6), 466.

[95] Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11), 498 – 504.

[96] Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. *Eye movements: A window on mind and brain*, (pp. 537–562).

[97] Henderson, J. M. & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, 1(10), 743.

[98] Henderson, J. M. & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, 50(1), 243–271.

[99] Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, 16(5), 850–856.

[100] Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition*, 33(1), 98–106.

[101] Heng, L., Li, B., & Pollefeys, M. (2013). Camodocal: Automatic intrinsic and extrinsic calibration of a rig with multiple generic cameras and odometry. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on* (pp. 1793–1800).: IEEE.

[102] Hennessey, C. & Lawrence, P. (2009). Noncontact binocular eye-gaze tracking for point-of-gaze estimation in three dimensions. *IEEE Transactions on Biomedical Engineering*, 56(3), 790–799.

[103] Hollingworth, A. & Hwang, S. (2013). The relationship between visual working memory and attention: retention of precise colour information in the absence of effects on perceptual selection. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 368(1628).

[104] Hollingworth, A., Matsukura, M., & Luck, S. J. (2013). Visual working memory modulates rapid eye movements to simple onset targets. *Psychological Science*, 24(5), 790–796. PMID: 23508739.

[105] Holmqvist, K. & Andersson, R. (2017). *Eye Tracking. A comprehensive guide to methods, paradigms, and measures*. Lund Eye-tracking Research Institute.

[106] Holmqvist, K., Blignaut, P., & Niehorster, D. (2018). Small eye movements cannot be reliably measured by current video-based eye-trackers. *Under preparation for Behavior Research Methods*.

[107] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

[108] Hooge, I. T., Hessels, R. S., & Nyström, M. (2019). Do pupil-based binocular video eye trackers reliably measure vergence? *Vision Research*, 156, 1 – 9.

[109] Horrey, W. J. & Wickens, C. D. (2004). Focal and ambient visual contributions and driver visual scanning in lane keeping and hazard detection. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(19), 2325–2329.

[110] Howard, I. P. (2012). *Preceiving in Depth*. Oxford Psychology Series. Oxford University Press.

[111] Hutton, S. B. & Nolte, S. (2011). The effect of gaze cues on attention to print advertisements. *Applied Cognitive Psychology*, 25(6), 887–892.

[112] Huynh-Thu, Q. & Schiatti, L. (2011). Examination of 3D visual attention in stereoscopic video content. In B. E. Rogowitz & T. N. Pappas (Eds.), *IS&T/SPIE Electronic Imaging* (pp. 78650J–78650J–15).: International Society for Optics and Photonics.

[113] Isola, P., Xiao, J., Parikh, D., Torralba, A., & Oliva, A. (2014). What makes a photograph memorable? *IEEE transactions on pattern analysis and machine intelligence*, 36(7), 1469–1482.

[114] Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6), 1093–1123.

[115] Itti, L. & Borji, A. (2015). Computational models: Bottom-up and top-down aspects. *arXiv preprint arXiv:1510.07748*.

[116] Itti, L. & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10), 1489–1506.

[117] Itti, L. & Koch, C. (2001). Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3), 194–203.

[118] Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.

[119] Jacob, R. & Stellmach, S. (2016). What you look at is what you get: gaze-based user interfaces. *interactions*, 23(5), 62–65.

[120] Jacob, R. J. (1990). What you look at is what you get: eye movement-based interaction techniques. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 11–18).: ACM.

[121] Jacobson, E. (1932). Electrophysiology of mental activities. *The American Journal of Psychology*, 44(4), 677–694.

[122] Jansen, L., Onat, S., & König, P. (2009). Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1), 1–19.

[123] Jaschinski, W. & Schroth, V. (2008). Ocular prevalence: difference between crossed and uncrossed disparities of stereo objects. *Strabismus*, 16(4), 159–164.

[124] Jaschinski, W., Švede, A., & Jainta, S. (2008). Relation between fixation disparity and the asymmetry between convergent and divergent disparity step responses. *Vision Research*, 48(2), 253–263.

[125] Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[126] Johansson, R., Holsanova, J., Dewhurst, R., & Holmqvist, K. (2012). Eye movements during scene recollection have a functional role, but they are not reinstatements of those produced during encoding. *Journal of Experimental Psychology: Human Perception and Performance*, 38(5), 1289–1314.

[127] Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science*, 30(6), 1053–1079.

[128] Johansson, R., Holsanova, J., & Holmqvist, K. (2011). The dispersion of eye movements during visual imagery is related to individual differences in spatial imagery ability. In *Proceedings of the Cognitive Science Society*, volume 33.

[129] Johansson, R. & Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. *Psychological Science*, 25(1), 236–242.

[130] Johansson, R., Oren, F., & Holmqvist, K. (2018). Gaze patterns reveal how situation models and text representations contribute to episodic text memory. *Cognition*, 175, 53–68.

[131] Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*.

[132] Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 2106–2113).: IEEE.

[133] Kamitani, Y. & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature neuroscience*, 8(5), 679.

[134] Kapur, A., Kapur, S., & Maes, P. (2018). Alterego: A personalized wearable silent speech interface. In *23rd International Conference on Intelligent User Interfaces*, IUI '18 (pp. 43–53). New York, NY, USA: ACM.

[135] Karessli, N., Akata, Z., Schiele, B., Bulling, A., et al. (2017). Gaze embeddings for zero-shot image classification. In *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[136] Kaspar, K. & Koenig, P. (2011). Viewing behavior and the impact of low-level image properties across repeated presentations of complex scenes. *Journal of Vision*, 11(13), 26–26.

[137] Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct* (pp. 1151–1160). New York, New York, USA: ACM Press.

[138] Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452, 352–355.

[139] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2), 81–93.

[140] Kensler, A. & Shirley, P. (2006). Optimizing ray-triangle intersection via automated search. In *2006 IEEE Symposium on Interactive Ray Tracing* (pp. 33–38).

[141] Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2390–2398).

[142] Ki, J. & Kwon, Y.-M. (2008). 3D Gaze Estimation and Interaction. In *2008 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video* (pp. 373–376).: IEEE.

[143] Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of vision*, 9(5), 7–7.

[144] Kienzle, W., Wichmann, F. A., Franz, M. O., & Schölkopf, B. (2007). A nonparametric approach to bottom-up visual saliency. In *Advances in neural information processing systems* (pp. 689–696).

[145] Kim, S. J., Ng, H., Winkler, S., Song, P., & Fu, C.-W. (2012). Brush-and-drag: A multi-touch interface for photo triaging. In *Proceedings of the 14th International Conference on Human-computer Interaction with Mobile Devices and Services*, MobileHCI '12 (pp. 59–68). New York, NY, USA: ACM.

[146] Kim, Y., Varshney, A., Jacobs, D. W., & Guimbretière, F. (2010). Mesh saliency and human eye fixations. *ACM Trans. Appl. Percept.*, 7(2), 12:1–12:13.

[147] Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

[148] Kirk, D., Sellen, A., Rother, C., & Wood, K. (2006). Understanding photowork. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06 (pp. 761–770). New York, NY, USA: ACM.

[149] Kirkby, J. A., Blythe, H. I., Drieghe, D., Benson, V., & Liversedge, S. P. (2013). Investigating eye movement acquisition and analysis technologies as a causal factor in differential prevalence of crossed and uncrossed fixation disparity during reading and dot scanning. *Behavior Research Methods*, 45(3), 664–678.

[150] Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4), 219–27.

[151] Koenderink, J. J. (1998). Pictorial relief. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 356(1740), 1071–1086.

[152] Koenderink, J. J. (1999). Virtual Psychophysics. *Perception*, 28, 669–674.

[153] Kovesi, P. (2015). Good colour maps: How to design them. *CoRR*, abs/1509.03700.

[154] Kowler, E. (2011). Eye movements: The past 25years. *Vision Research*, 51(13), 1457 – 1483. Vision Research 50th Anniversary Issue: Part 2.

[155] Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision*, 13(2), 201–214.

[156] Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*.

[157] Kümmerer, M., Wallis, T. S., & Bethge, M. (2016). Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.

[158] Kytö, M., Ens, B., Piumsomboon, T., Lee, G. A., & Billinghurst, M. (2018). Pinpointing: Precise head-and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp.81).: ACM.

[159] Laeng, B., Bloem, I. M., D'Ascenzo, S., & Tommasi, L. (2014). Scrutinizing visual images: The role of gaze in mental imagery and memory. *Cognition*, 131(2), 263 – 283.

[160] Laeng, B. & Teodorescu, D.-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science*, 26(2), 207–231.

[161] Lance, B. J., Kerick, S. E., Ries, A. J., Oie, K. S., & McDowell, K. (2012). Brain-computer interface technologies in the coming decades. *Proceedings of the IEEE*, 100(Special Centennial Issue), 1585–1599.

[162] Land, M. F. & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25), 3559 – 3565.

[163] Lang, C., Nguyen, T. V., Katti, H., Yadati, K., Kankanhalli, M., & Yan, S. (2012). Depth Matters: Influence of Depth Cues on Visual Saliency. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision – ECCV 2012*, Lecture Notes in Computer Science (pp. 101–115). Springer.

[164] Lau, M., Dev, K., Shi, W., Dorsey, J., & Rushmeier, H. (2016). Tactile mesh saliency. *ACM Trans. Graph.*, 35(4), 52:1–52:11.

[165] Lavoué, G., Cordier, F., Seo, H., & Larabi, M.-C. (2018). Visual attention for rendered 3d shapes. *Computer Graphics Forum*, 37(2), 191–203.

[166] Lavoué, G. & Corsini, M. (2010). A comparison of perceptually-based metrics for objective evaluation of geometry processing. *IEEE Transactions on Multimedia*, 12(7), 636–649.

[167] Lee, C. H., Varshney, A., & Jacobs, D. W. (2005). Mesh saliency. *ACM Trans. Graph.*, 24(3), 659–666.

[168] Lee, J. W., Cho, C. W., Shin, K. Y., Lee, E. C., & Park, K. R. (2012a). 3D gaze tracking method using Purkinje images on eye optical model and pupil. *Optics and Lasers in Engineering*, 50, 736–751.

[169] Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2012b). Disentangling visual imagery and perception of real-world objects. *NeuroImage*, 59(4), 4064 – 4073.

[170] Lehmann, E. L. & Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. New York: Springer, third edition.

[171] Levin, D. (1998). The approximation power of moving least-squares. *Math. Comput.*, 67(224), 1517–1531.

[172] Li, L.-J., Su, H., Fei-Fei, L., & Xing, E. P. (2010). Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems* (pp. 1378–1386).

[173] Lindlbauer, D., Mueller, J., & Alexa, M. (2016). Changing the appearance of physical interfaces through controlled transparency. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16 (pp. 425–435). New York, NY, USA: ACM.

[174] Liversedge, S. P., Rayner, K., White, S. J., Findlay, J. M., & McSorley, E. (2006). Binocular coordination of the eyes during reading. *Current Biology*, 16(17), 1726–1729.

[175] Maggia, C., Guyader, N., & Guérin-Dugué, A. (2013). Using natural versus artificial stimuli to perform calibration for 3D gaze tracking. In *SPIE, Human Vision and Electronic Imaging XVIII* (pp. 865113). Bulingame, CA, United States.

[176] Majaranta, P. & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing* (pp. 39–65). Springer.

[177] Majaranta, P. & Räihä, K.-J. (2002). Twenty years of eye typing: Systems and design issues. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ETRA '02 (pp. 15–22). New York, NY, USA: ACM.

[178] Mansouryar, M., Steil, J., Sugano, Y., & Bulling, A. (2016). 3d gaze estimation from 2d pupil positions on monocular head-mounted eye trackers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ETRA '16 (pp. 197–200). New York, NY, USA: ACM.

[179] Mäntylä, T. & Holm, L. (2006). Gaze control and recollective experience in face recognition. *Visual Cognition*, 14(3), 365–386.

[180] Martarelli, C. S., Chiquet, S., Laeng, B., & Mast, F. W. (2017). Using space to represent categories: insights from gaze position. *Psychological Research*, 81(4), 721–729.

[181] Martarelli, C. S. & Mast, F. W. (2013). Eye movements during long-term pictorial recall. *Psychological Research*, 77(3), 303–309.

[182] Martinez-Conde, S., Macknik, S. L., & Hubel, D. H. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, 5, 229 EP –. Review Article.

[183] Mastroberardino, S. & Vredeveldt, A. (2014). Eye-closure increases children's memory accuracy for visual material. *Frontiers in Psychology*, 5, 241.

[184] Mathe, S. & Sminchisescu, C. (2012). Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision – ECCV 2012* (pp. 842–856).: Springer.

[185] Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(1).

[186] Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81(12), 899–917.

[187] Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. *Annals of the Institute of Statistical Mathematics*, 19(1), 181.

[188] Maurus, M., Hammer, J. H., & Beyerer, J. (2014). Realistic heatmap visualization for interactive analysis of 3d gaze data. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14 (pp. 295–298). New York, NY, USA: ACM.

[189] Memmert, D. (2006). The effects of eye movements, age, and expertise on inattentional blindness. *Consciousness and Cognition*, 15(3), 620 – 627.

[190] Merrill, N. & Chuang, J. (2018). From scanning brains to reading minds: Talking to engineers about brain-computer interface. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 323).: ACM.

[191] Meyer, M., Barr, A., Lee, H., & Desbrun, M. (2002). Generalized barycentric coordinates on irregular polygons. *Journal of Graphics Tools*, 7(1), 13–22.

[192] Millán, J. d. R., Rupp, R., Mueller-Putz, G., Murray-Smith, R., Giugliemma, C., Tangermann, M., Vidaurre, C., Cincotti, F., Kubler, A., Leeb, R., Neuper, C., Mueller, K., & Mattia, D. (2010). Combining brain–computer interfaces and assistive technologies: State-of-the-art and challenges. *Frontiers in Neuroscience*, 4, 161.

[193] Monty, R. A., Fisher, D. F., & Senders, J. W. (2017). *Eye movements: cognition and visual perception*. Routledge.

[194] Moore, C. S. (1903). Control of the memory image. *The Psychological Review: Monograph Supplements*, 4(1), 277–306.

[195] Mott, M. E., Williams, S., Wobbrock, J. O., & Morris, M. R. (2017). Improving dwell-based gaze typing with dynamic, cascading dwell times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 2558–2570).: ACM.

[196] Mulder, T., Zijlstra, S., Zijlstra, W., & Hochstenbach, J. (2004). The role of motor imagery in learning a totally novel movement. *Experimental Brain Research*, 154(2), 211–217.

[197] Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

[198] Nemrodov, D., Niemeier, M., Patel, A., & Nestor, A. (2018). The neural dynamics of facial identity processing: insights from eeg-based pattern analysis and image reconstruction. *eNeuro*.

[199] Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2018). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, 50(1), 213–227.

[200] Niehorster, D. C., Zemblys, R., Hein, O., Beelders, T., & Holmqvist, K. (in prep). Noise in eye-tracker data: what it is and what shapes it. *Behavior Research Methods*.

[201] Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19), 1641 – 1646.

[202] Nuthmann, A. & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8), 20.

[203] Nuthmann, A. & Kliegl, R. (2009). An examination of binocular reading fixations based on sentence corpus data. *Journal of Vision*, 9(5), 31–31.

[204] Nyström, M. & Holmqvist, K. (2008). Semantic override of low-level features in image viewing – both initially and overall. *Journal of Eye Movement Research*, 2(2), 2:1–2:11.

[205] Oliva, A. & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. In S. Martinez-Conde, S. Macknik, L. Martinez, J.-M. Alonso, & P. Tse (Eds.), *Visual Perception*, volume 155 of *Progress in Brain Research* (pp. 23 – 36). Elsevier.

[206] Olivers, C. N. (2008). Interactions between visual working memory and visual attention. *Frontiers in Bioscience*, 13(3), 1182–1191.

[207] Olivers, C. N., Meijer, F., & Theeuwes, J. (2006). Feature-based memory-driven attentional capture: visual working memory content affects visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, 32(5), 1243.

[208] Olsen, R. K., Chiew, M., Buchsbaum, B. R., & Ryan, J. D. (2014). The relationship between delay period eye movements and visuospatial memory. *Journal of Vision*, 14(1), 8–8.

[209] Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ETRA '10 (pp. 141–144). New York, NY, USA: ACM.

[210] Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1), 107–23.

[211] Pathman, T. & Ghetti, S. (2015). Eye movements provide an index of veridical memory for temporal order. *PLOS ONE*, 10(5), 1–17.

[212] Perky, C. W. (1910). An experimental study of imagination. *The American Journal of Psychology*, 21(3), 422–452.

[213] Pfeiffer, T., Latoschik, M. E., & Wachsmuth, I. (2008). Evaluation of binocular eye trackers and algorithms for 3d gaze interaction in virtual reality environments.

[214] Pfeiffer, T. & Renner, P. (2014). Eyesee3d: A low-cost approach for analyzing mobile 3d eye tracking data using computer vision and augmented reality technology. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 369–376).: ACM.

[215] Pfeiffer, T., Renner, P., & Pfeiffer-Lessmann, N. (2016). Eyesee3d 2.0: Model-based real-time analysis of mobile eye-tracking in static and dynamic three-dimensional scenes. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ETRA '16 (pp. 189–196). New York, NY, USA: ACM.

[216] Potter, M. C. & Levy, E. I. (1969). Recognition memory for a rapid sequence of pictures. *Journal of experimental psychology*, 81(1), 10.

[217] Ramanathan, S., Katti, H., Sebe, N., Kankanhalli, M., & Chua, T.-S. (2010). An Eye Fixation Database for Saliency Detection in Images. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer Vision – ECCV 2010* (pp. 30–43).: Springer.

[218] Ramasamy, C., House, D. H., Duchowski, A. T., & Daugherty, B. (2009). Using eye tracking to analyze stereoscopic filmmaking. In *Posters on - SIGGRAPH '09* (pp.1). New York, New York, USA: ACM Press.

[219] Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8), 1457–1506.

[220] Recasens, A., Khosla, A., Vondrick, C., & Torralba, A. (2015). Where are they looking? In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28* (pp. 199–207). Curran Associates, Inc.

[221] Rencher, A. C. (2003). *Methods of multivariate analysis*, volume 492. John Wiley & Sons.

[222] Richardson, D. C., Altmann, G. T., Spivey, M. J., & Hoover, M. A. (2009). Much ado about eye movements to nothing: a response to ferreira et al.: Taking a new look at looking at nothing. *Trends in Cognitive Sciences*, 13(6), 235–236.

[223] Richardson, D. C. & Spivey, M. J. (2000). Representation, space and hollywood squares: Looking at things that aren't there anymore. *Cognition*, 76(3), 269–295.

[224] Ringer, R. V., Throneburg, Z., Johnson, A. P., Kramer, A. F., & Loschky, L. C. (2016). Impairing the useful field of view in natural scenes: Tunnel vision versus general interference. *Journal of Vision*, 16(2), 7–7.

[225] Ritter, J. (1990). An efficient bounding sphere. In A. S. Glassner (Ed.), *Graphics Gems* (pp. 301–303). San Diego, CA, USA: Academic Press Professional, Inc.

[226] Robinson, D. A. (1965). The mechanics of human smooth pursuit eye movement. *The Journal of Physiology*, 1801(3), 569–591.

[227] Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), 99–121.

[228] Sachs, L. (2012). *Applied statistics: a handbook of techniques*. Springer Science & Business Media.

[229] Saladin, J. J. (1986). Convergence insufficiency, fixation disparity, and control systems analysis. *American journal of optometry and physiological optics*, 63(8), 645–653.

[230] Sangkloy, P., Burnell, N., Ham, C., & Hays, J. (2016). The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 35(4), 119:1–119:12.

[231] Sattar, H., Bulling, A., & Fritz, M. (2017). Predicting the category and attributes of visual search targets using deep gaze pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2740–2748).

[232] Schaefer, S., McPhail, T., & Warren, J. (2006). Image deformation using moving least squares. *ACM Trans. Graph.*, 25(3), 533–540.

[233] Schmidt, B. K., Vogel, E. K., Woodman, G. F., & Luck, S. J. (2002). Voluntary and automatic attentional control of visual working memory. *Perception & Psychophysics*, 64(5), 754–763.

[234] Schneider, P. & Eberly, D. H. (2002). *Geometric tools for computer graphics*. Elsevier.

[235] Scholz, A., Klichowicz, A., & Krems, J. F. (2017). Covert shifts of attention can account for the functional role of "eye movements to nothing". *Memory & Cognition*, (pp. 1–14).

[236] Scholz, A., Mehlhorn, K., & Krems, J. F. (2016). Listen up, eye movements play a role in verbal memory retrieval. *Psychological research*, 80(1), 149–158.

[237] Scholz, A., von Helversen, B., & Rieskamp, J. (2015). Eye movements reveal memory processes during similarity- and rule-based decision making. *Cognition*, 136, 228 – 246.

[238] Schönberger, J. L. & Frahm, J.-M. (2016). Structure-from-motion revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4104–4113).

[239] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015* (pp. 815–823).

[240] Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2017). Deep image reconstruction from human brain activity. *bioRxiv*.

[241] Shilane, P. & Funkhouser, T. (2007). Distinctive regions of 3d surfaces. *ACM Trans. Graph.*, 26(2).

[242] Shneiderman, B., Bederson, B. B., & Drucker, S. M. (2006). Find that photo!: Interface strategies to annotate, browse, and share. *Commun. ACM*, 49(4), 69–71.

[243] Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., & Wetzstein, G. (2017). How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*.

[244] Song, R., Liu, Y., Martin, R. R., & Rosin, P. L. (2014). Mesh saliency via spectral processing. *ACM Trans. Graph.*, 33(1), 6:1–6:17.

[245] Sorkine, O. & Alexa, M. (2007). As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, SGP '07 (pp. 109–116). Aire-la-Ville, Switzerland, Switzerland: Eurographics Association.

[246] Sorkine-Hornung, O. & Rabinovich, M. (2016). Least-squares rigid motion using svd. Technical note.

[247] Spivey, M. J. & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological research*, 65(4), 235–241.

[248] Stellmach, S., Nacke, L., & Dachselt, R. (2010). 3d attentional maps: aggregated gaze visualizations in three-dimensional virtual environments. In *Proceedings of the international conference on advanced visual interfaces* (pp. 345–348).: ACM.

[249] Švede, A., Hoormann, J., Jainta, S., & Jaschinski, W. (2011). Subjective fixation disparity affected by dynamic asymmetry, resting vergence, and nonius bias. *Investigative ophthalmology & visual science*, 52(7), 4356–4361.

[250] Švede, A., Treija, E., Jaschinski, W., & Krūmiņa, G. (2015). Monocular versus binocular calibrations in evaluating fixation disparity with a video-based eye-tracker. *Perception*, 44(8-9), 1110–1128.

[251] Tasse, F. P., Kosinka, J., & Dodgson, N. (2015). Cluster-based point set saliency. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 163–171).

[252] Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5), 643 – 659.

[253] Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5), 5–5.

[254] Theeuwes, J., Kramer, A. F., Hahn, S., & Irwin, D. E. (1998). Our eyes do not always go where we want them to go: Capture of the eyes by new objects. *Psychological Science*, 9(5), 379.

[255] Tian, Y. & Styan, G. P. (2001). Rank equalities for idempotent and involutary matrices. *Linear Algebra and its Applications*, 335(1), 101 – 117.

[256] Toet, A. (2011). Computational versus psychophysical bottom-up image saliency: a comparative evaluation study. *IEEE transactions on pattern analysis and machine intelligence*, 33(11), 2131–46.

[257] Toussaint, G. T. (1974). Some properties of matusita's measure of affinity of several distributions. *Annals of the Institute of Statistical Mathematics*, 26(1), 389–394.

[258] Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, 3(1), 86–94.

[259] Tulving, E. (1985). *Elements of Episodic Memory*. Oxford University Press.

[260] Underwood, G. M., Chapman, P., Berger, Z., & Crundall, D. (2003). Driving experience, attentional focusing, and the recall of recently inspected events. *Transportation Research Part F: Traffic Psychology and Behaviour*, 6(4), 289–304.

[261] Vidal, J. J. (1973). Toward direct brain-computer communication. *Annual review of Biophysics and Bioengineering*, 2(1), 157–180.

[262] Vidal, M., Bulling, A., & Gellersen, H. (2012). Detection of smooth pursuits using eye movement shape features. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12 (pp. 177–180). New York, NY, USA: ACM.

[263] Vredeveldt, A., Tredoux, C. G., Kempen, K., & Nortje, A. (2015). Eye remember what happened: Eye-closure improves recall of events but not face recognition. *Applied Cognitive Psychology*, 29(2), 169–180. ACP-14-0042.R2.

[264] Wagner, M., Ehrenstein, W. H., & Papathomas, T. V. (2009). Vergence in reverspective: Percept-driven versus data-driven eye movement control. *Neuroscience Letters*, 449(2), 142 – 146.

[265] Wang, D., Mulvey, F. B., Pelz, J. B., & Holmqvist, K. (2017a). A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods*, 49(3), 947–959.

[266] Wang, F. & Guibas, L. J. (2012). Supervised earth mover's distance learning and its computer vision applications. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part I*, ECCV'12 (pp. 442–455). Berlin, Heidelberg: Springer-Verlag.

[267] Wang, R. I., Pelfrey, B., Duchowski, A. T., & House, D. H. (2014). Online 3d gaze localization on stereoscopic displays. *ACM Trans. Appl. Percept.*, 11(1), 3:1–3:21.

[268] Wang, W., Shen, J., Yu, Y., & Ma, K.-L. (2017b). Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Transactions on Visualization and Computer Graphics*, 23(8), 2014–2027.

[269] Wang, X., Lindlbauer, D., Lessig, C., & Alexa, M. (2017c). Accuracy of monocular gaze tracking on 3d geometry. In M. Burch, L. Chuang, B. Fisher, A. Schmidt, & D. Weiskopf (Eds.), *Eye Tracking and Visualization* (pp. 169–184). Cham: Springer International Publishing.

[270] Wang, X., Lindlbauer, D., Lessig, C., Maertens, M., & Alexa, M. (2016). Measuring the visual salience of 3d printed objects. *IEEE Computer Graphics and Applications*, 36(4), 46–55.

[271] Wantz, A. L., Martarelli, C. S., & Mast, F. W. (2016). When looking back to nothing goes back to nothing. *Cognitive Processing*, 17(1), 105–114.

[272] West, T. G. (1991). *In the mind's eye: Visual thinkers, gifted people with learning difficulties, computer images, and the ironies of creativity*. Prometheus Books.

[273] Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(3), 449–455.

[274] Wismeijer, D. A., van Ee, R., & Erkelens, C. J. (2008). Depth cues, rather than perceived depth, govern vergence. *Experimental Brain Research*, 184(1), 61–70.

[275] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6), 767 – 791.

[276] Wu, J., Shen, X., Zhu, W., & Liu, L. (2013). Mesh saliency with global rarity. *Graphical Models*, 75(5), 255 – 264.

[277] Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of vision*, 14(1), 28–28.

[278] Yu, Q., Liu, F., SonG, Y.-Z., Xiang, T., Hospedales, T., & Loy, C. C. (2016). Sketch me that shoe. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 799–807).

[279] Zander, T. O. & Kothe, C. (2011). Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *Journal of Neural Engineering*, 8(2), 025005.