Technische Universität München
Fakultät für Elektrotechnik und Informationstechnik

Towards a better understanding of eye movements in natural contexts

Ioannis Agtzidis

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. Bernhard Wolfrum
Prüfer der Dissertation:	1. TUM Junior Fellow Dr.-Ing. Michael Dorr
	2. Prof. Dr.-Ing. Werner Hemmert
	3. Prof. Dr. med. Rebekka Lencer

Die Dissertation wurde am 15.06.2020 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 08.09.2020 angenommen.

Abstract

For many of our everyday activities vision is the most important sense and eye movements form a significant part of it. Therefore the research of eye movements allows us to better understand many functions of our visual system and offers us a window to the corresponding brain functionality. People understood the significance of eye movements a long time ago and they have been systematically studied for more than 100 years. As new technologies are invented and commercialized new opportunities arise for more thorough research of eye movements in closer to natural environments. But these new environments pose new challenges that did not exist with the experimental setups of the early days.

In this thesis we provide the foundation for the automatic analysis of eye movement functionality in more unconstrained scenarios that are closer to our natural environment and include videos of everyday scenes presented either on a monitor or a head-mounted display. More specifically, we provide three hand-labeled ground-truth eye movement data sets that span many hours, contrary to minutes, which had been so far the standard. Then we improve the quality of automatic eye movement detection by providing two new algorithms that work with monitor-based experiments and achieve state-of-the-art performance. We also extend the field of application of many pre-existing algorithms to work with head-mounted displays and we propose a new algorithm too. At the end we conclude with two applications of our infrastructure in new domains. The first offers a better understanding of the relationship between the brain and smooth pursuit eye movements and the second investigates how transferable results are across different experiments with varying levels of naturalness.

Abstrakt

Augenbewegungen formen einen zentralen Bestandteil des menschlichen Sehens, das wohl der wichtigste Sinn für viele unserer Alltagsaktivitäten ist. Die Forschung zu Augenbewegungen ermöglicht uns daher ein besseres Verständnis des Sehens und damit auch allgemein der Hirnfunktion. Schon seit über 100 Jahren werden Augenbewegungen deswegen systematisch untersucht, doch erst kürzlich wurden neue Technologien entwickelt und kommerziell zur Verfügung gestellt, die Studien unter weitgehend natürlichen Bedingungen ermöglichen. Diese neuen Technologien erzeugen wiederum eigene Herausforderungen bei der Datenaufnahme und -analyse.

In der vorliegenden Arbeit legen wir die Grundlagen für die automatische Analyse von mit natürlichem Bildmaterial aufgenommenen Blickrichtungsdaten. Die Stimuli umfassten dabei Videos von Alltagsszenen, die auf einem Monitor oder in einem kopfgetragenen Display angezeigt wurden. Wir erstellten drei "ground truth" Datensätze, die von Experten manuell annotiert wurden und die statt weniger Minuten, wie bisher üblich, mehrere Stunden an Material umfassen. Darauf aufbauend verbesserten wir die Güte automatischer Klassifikation von Blickrichtungsdaten durch die Entwicklung zweier neuer Algorithmen für monitor-basierte Experimentaldaten. Für Daten, die mithilfe von kopfgetragenen Displays aufgenommen wurden, entwickelten wir neben einem gänzlich neuen Algorithmus auch Verfahren, wie bereits für monitor-basierte Experimente existierende Algorithmen optimal auf diesen neuen Datentyp angewendet werden können. Zum Schluss stellen wir beispielhaft zwei Anwendungsfälle für unsere entwickelten Methoden vor. In einer ersten Studie untersuchten wir den Zusammenhang von MRT-gemessener Aktivität in verschiedenen Hirnarealen mit dem Vorliegen verschiedener Augenbewegungstypen. In einer zweiten Studie variierten wir den Komplexitätsgrad von typischerweise in Psychophysik-Studien verwendeten Stimuli hin zu einem höheren Natürlichkeitsgrad, um die öklogische Validität von unter Laborbedingungen gefundenen Resultaten zu untersuchen.

Acknowledgements

Many people contributed for the completion of this thesis in various forms. First of all I would like to thank my family and especially my mother and my late grandfather for their support throughout the years. Then I would like to thank my friends both in Greece and Germany for the great times that we had together and most importantly Anna for the great experiences that we shared together.

I would also like to thank my collaborators. I worked with Mikhail Startsev for the whole duration of my PhD sharing the same office and throughout this time we had some very interesting research related discussions. He was also the person that I collaborated with the most and he always had new ideas that improved our work. Rebekka Lencer and Inga Meyhöfer offered me advice and a more medical oriented perspective regarding the analysis of fMRI data. Also I would like to thank Alexander Goettker from Karl's Gegenfurtner's lab for the excellent cooperation. Finally, I would like to thank Michael Dorr, my great supervisor, who was supportive for the whole duration of my PhD, which I feel would have been impossible without him.

Contents

Abstract i

Abstrakt ii

Acknowledgements iii

I Setting the scene 1

1 Introduction 3

1.1 Thesis organization 5

1.1.1 Previous publications 6

2 Basics 7

2.1 The role of eye movements 7

2.2 Eye movement types 8

2.2.1 Eye movement types definitions 9

2.3 Smooth pursuit 10

2.4 Head-mounted eye tracking 12

2.5 Evaluation metrics 14

II Foundational data sets 17

3 Labeling tool 19

3.1 Data format 19

3.2 Labeling interface 21

3.3 Handling of 360-degree data 22

4 Hand-labeled data sets 27

4.1 GazeCom data set 27

4.1.1 Data set description 27

4.1.2 Labeling procedure 28

4.1.3 Inter-rater agreement 29

4.1.4 Hand labeling statistics 29

4.1.5 Basic statistics 31

4.2 Hollywood2 data set 33

4.2.1 Data set description 33

4.2.2 Labeling procedure 34

4.2.3 Inter-rater agreement 35

4.2.4 Hand labeling statistics 35

4.2.5 Basic statistics 37

4.3 360-degree data set 38

4.3.1 Data set collection 38

4.3.2 Manual annotation 43

4.3.3 Basic statistics 46

4.3.4 Discussion 46

III Improving automated gaze trace segmentation in unstructured environments 51

5 Eye movement segmentation in monitor-based experiments 53

5.1 Smooth pursuit detection based on multiple observers 53

5.1.1 Prefiltering 54

5.1.2 Clustering 54

5.2 Deep learning eye movement segmentation 56

5.3 Algorithm evaluation 57

5.3.1 Literature algorithms 57

5.3.2 Algorithm Optimization 59

5.3.3 Results 60

6 Eye movement detection with 360-degree stimuli 65

6.1 Conversion of monitor-based algorithms 66

6.1.1 Equirectangular to Cartesian space 66

6.1.2 Application to existing algorithms 68

6.1.3 Conversion of data 71

6.1.4 Conversion evaluation 74

6.2 I-S^{5}T algorithm 77

6.3 Overall evaluation 80

IV Applications in dynamic natural contexts 83

7 Understanding smooth pursuit brain activations in dynamic natural scenes 85

7.1 Methods 86

7.1.1 Data set 86

7.1.2 Motion estimation in the stimulus 87

7.1.3 Eye movement classification 87

7.1.4 fMRI analysis 88

7.1.5 Additional validation regressors 91

7.2 Results 91

7.2.1 Eye movement statistics 92

7.2.2 SP- and saccade-related activations 93

7.2.3 SP-saccade related activations 94

7.2.4 Accounting for movie motion 95

7.3 Discussion 97

7.4 Chapter conclusion 102

8 Saccade and smooth pursuit initiation interactions 103

8.1 Methods 105

8.1.1 Selection of baseline trajectories 105

8.1.2 Experimental design 106

8.1.3 Experimental setup 108

8.1.4 Participants 108

8.1.5 Data analysis 109

8.2 Results 110

8.2.1 Saccadic eye movements 110

8.2.2 Pursuit eye movements 112

8.2.3 Effect of object size 113

8.3 Discussion 114

8.3.1 Effect of scene complexity on saccadic eye movements 115

8.3.2 Saccade pursuit interaction 117

8.3.3 Effect of scene complexity on pursuit eye movements 117

8.4 Chapter conclusion 118

9 Conclusion **119**

Bibliography **120**

Part I

Setting the scene

Chapter 1

Introduction

In our everyday life we continuously explore our visual environment by shifting our point of regard several times per second. This behavior creates the percept of a high-resolution world based only on a very small area of the retina with high photoreceptor density, the fovea. The constant movement of the eyes creates a path, the gaze trace, that represents the areas of our surroundings that have been attended. Even though the gaze trace in itself can offer rich information about the functioning of the human visual system, its segmentation into eye movements offers even richer information. Eye movements enable us to better understand the fundamentals of visual processing, to relate the different visual areas of the brain, and additionally to perform more detailed analyses. For example raw gaze traces are enough for creating attention maps (also called saliency maps) but the constituent eye movements enable us to understand how the attention is allocated, along with the timing of events preceding a gaze shift to a target.

The eye movement behavior has been so far mainly studied in experimental setups that have varying degrees of fidelity to the natural everyday viewing behavior. The most common constraint is the use of a monitor as a presentation medium where the experiment's participants are placed in front of it. This setup either allows for a small amount of head motion or restricts it by stabilizing the head with a chin rest. This choice simplifies the experiment and its subsequent analysis but it excludes head motion, which forms a significant aspect of natural viewing behavior. Another constraint is the use of computer-generated stimuli that mostly contain up to a couple of artificial potential targets, such as dots. Again this choice allows for the accurate measurement of specific eye movement characteristics, such as their latency, but it provides an uncluttered environment that is not reminiscent of the rich visual environment that we live in. Finally, the majority of the experiments are constrained

to static stimuli in the form of natural or computer-generated images. When static stimuli are used only saccadic and fixational eye movements are performed and they have almost linearly separable speed profiles. Also, these eye movements are performed in isolation without any interactions with all the other eye movement types that we utilize in our everyday visual behavior (for all the eye movement types that are mentioned in this thesis refer to Section 2.2.1).

In this thesis, we tackle many of the challenges that are associated with the use of more natural and unstructured stimuli in eye movement research and demonstrate in practical applications how such technical solutions can open up new avenues to answering research questions about neuroscientific phenomena and the generalizability of research outcomes in new domains. One of the biggest challenges for eye movement research in natural scenes is the difficulty of defining a "ground-truth" against which things can be compared. Our contribution towards this problem is the creation of three large eye movement data sets that contain hand-labeled eye movements based on clear definitions. The data sets span a diverse set of viewing conditions that include free viewing of dynamic everyday natural scenes (ex. cars driving on a street) and Hollywood videos on a monitor screen as well as free viewing of 360-degree content in a head-mounted display (HMD) that allows for free head motion. In all data sets apart from the usual fixational and saccadic eye movements we also labeled the smooth pursuit eye movement type, which has been often overlooked due to the technical challenges that its analysis is posing. In the 360-degree data set apart from the previously mentioned eye movements we also account for "eye movements" that arose due to free head motion and provide a formal taxonomy for all of these. Overall the data sets comprise around 7 hours of recordings, which are the biggest to date and allow for the development and evaluation of machine learning and more importantly deep learning applications.

Another challenge is automatic labeling of eye movements when working with dynamic natural environments because (i) most of the pre-existing algorithms classify fixations and saccades only, and (ii) most of them were developed for monitor-based experiments. Firstly, we developed two new algorithms for eye movement classification in monitor-based experiments that include the smooth pursuit label, which is more challenging since its characteristics overlap with those of fixations and saccades. Secondly, we converted five popular pre-existing eye movement classification algorithms in order to work with HMD gathered data, provided a method for using pre-existing algorithms without any modifications with HMDs by converting the underlying data instead of the algorithms, and finally we have developed a new algorithm that combines eye and head motion together to return richer labels.

1.1 Thesis organization

The thesis is organized into five main parts and here we will describe in more detail the content of each part. In Part I we set the scene and we provide information about the state of eye movement research with a higher emphasis towards smooth pursuit and the challenge of head-mounted eye tracking, which allows gaze recordings with unrestrained head motion.

In Part II we talk about the scaffolding that we built in order to be able to research eye movements in more "natural" conditions. We start with Chapter 3 where we present our tool for manual eye movement labeling, which can handle gaze data recorded both in monitor-based and HMD-based experiments. In Chapter 4 we describe in detail the process that was followed during the hand-labeling of eye movements in three large data sets. The first data set comprises of the complete annotation of eye movements for the naturalistic free viewing GazeCom [Dorr et al., 2010] data set. Apart from the more common fixations and saccades our annotation also includes labeled smooth pursuit, which can occur in the presence of motion. The second data set is a partial annotation of the Hollywood2 [Mathe and Sminchisescu, 2012] data set and includes the same eye movement labels. The last data set consists of 360-degree content that allowed for unconstrained head motion. Because no publicly available gaze data sets existed we designed a new HMD experiment, we collected gaze data, and manually annotated part of it. Here special care was given to the different frames of reference, which can either be world-fixed or head-fixed, and based on which one is chosen the gaze signal demonstrates different patterns. To overcome this we (i) expanded our labeling tool in order to be able to display the same information for both frames of reference and (ii) we used a two label scheme to characterize each gaze sample.

In Part III we move from manual to automatic eye movement labeling and we present new algorithms that achieve state-of-the-art performance both for monitor- and HMD-based experiments. Chapter 5 contains algorithms that were designed for experiments that present the stimulus on a monitor. In Section 5.1 we present a tool that automatically detects fixations, saccades and smooth pursuits. Its fixation and saccade detectors are based on literature algorithms but its SP detector is based on the clustering effect of gaze samples towards salient moving objects when multiple observers watch the same video clip. By taking advantage of the clustering effect our tool can detect SP with high precision. Then we further improve the eye movement detection performance for all three eye movement types by using a deep neural network architecture, which is explained in Section 5.2. In Chapter 6 we

move towards more immersive scenarios and in Section 6.1 we convert 5 popular eye movement detection algorithms in order to work correctly with gaze recordings that come from HMD experiments with equirectangular stimuli. Because such a conversion is not always feasible, in the same section, we present an alternative that converts the equirectangular data into an almost linear space where the original algorithms can be applied with minimal artifacts. All of the so far presented 360-degree algorithms do not account for the presence of the two different frames of reference. For this reason in Section 6.2 we design a new algorithm that detects eye movements based on both the field-of-view and eye+head frames of reference. This algorithm uses simple speed thresholds that were optimized based on the hand-labeled part of the 360-degree data set and can detect all eye movement types that are used for information retrieval and head movement compensation.

Based on the rich content of the hand-labeled data sets and the automatic detection algorithms, in Part IV we present the application of these in two different domains. Chapter 7 investigates which areas of the brain are activated during SP while humans are free-viewing the Hollywood movie 'Forrest Gump' and Chapter 8 examines how the saccade-SP initiation interactions are influenced by different experimental complexities that vary from simple moving dots to free viewing of naturalistic scenes. Finally, we conclude this thesis in Chapter 9.

1.1.1 Previous publications

The work presented in this thesis has been published in 9 conference and journal papers. At the beginning of each part we reference the relevant papers that the subsequent content is based on and we acknowledge the respective co-authors when their work is used.

Chapter 2

Basics

In this chapter we will talk about the role of eye movements and why they exist in the first place. We will then present the eye movement types that were studied in the context of this thesis and we will analyze in more detail the smooth pursuit eye movement because it is more challenging both to define and detect. Then we will move towards wearable eye tracking and how different frames of reference influence the gaze signal and the detection characteristics.

2.1 The role of eye movements

Eye movements exist in order to direct the point of regard to the most interesting areas and therefore maximize the information gain. From the physiological perspective they arise from the structure of our eyes, which is visualized in Figure 2.1. The light enters through the cornea at the front of the eye and is focused through the lens onto the retina at the back of the eye. The retina is lined with photoreceptor cells that absorb light at different wavelengths and emit signals that are transferred to the brain through the optic nerve [Findlay and Gilchrist, 2003]. The photoreceptor cells are divided into rods and cones with each category having different characteristics. Rods are achromatic cells, which can function in low light conditions and form the majority of photoreceptors (approx. 90 million) [Curcio et al., 1990]. Cones are larger than rods and function in well-lit conditions. They can be divided in S, M, and L types with each type being sensitive to different wavelengths and because of this humans have color vision. Even though their count is low ($\sim$5 million) they are concentrated in a very small area of the retina, which is called fovea (Figure 2.1). The fovea is the area of our retina with the highest resolution and therefore when we

need to acquire detailed information about an object we rotate our eyes in unison to align the object projection through the lens onto the fovea.

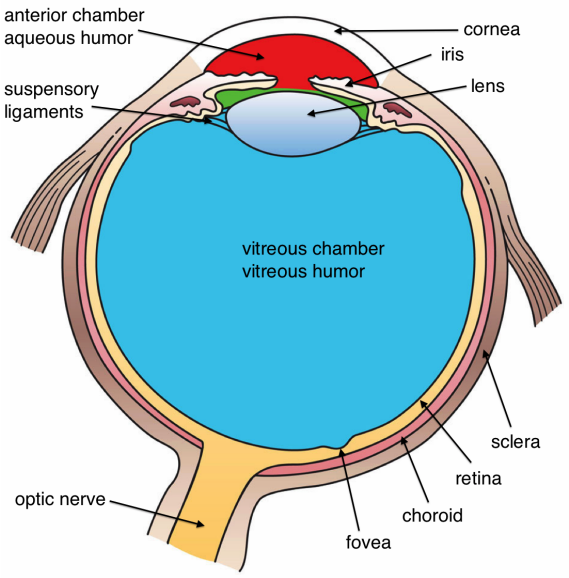


Figure 2.1: Physiology of the human eye. Courtesy of Wikimedia.

2.2 Eye movement types

The quantitative study of eye movements started more than a century ago when a French ophthalmologist observed through a mirror the human eye movements during reading [Javal, 1878]. [Huey, 1908] provided a review of the eye movement literature of the time, which mostly consisted of reading research. He also reviewed the methods that were used for measuring eye movement characteristics at this time with some of them being fairly intrusive (e.g. Figure 1, page 26 of the original paper). However the translation in English of the seminal work of Yarbus [Yarbus, 1967] reheated the interest of the scientific community in eye movements by presenting new recording methods (i.e. suction cups) and by demonstrating that eye movements are task dependent. All of the previous researchers along with many researchers of their time used static images or written texts in their work and they only observed two types of eye movements: fixations and saccades. They observed that the eyes moved very fast 2-3 times per second in order to reorient the point of regard and theses "jumps" are now called saccades with the time of relative ocular stability between saccades being called fixations. Fixations comprise the majority of the viewing time because the information processing happens during these periods.

The constraint of the eye movement research into fixations and saccades only arises

from the use of static stimuli, which have some very strong advantages: (i) ease of data analysis due to simpler stimuli and the simple criteria (e.g. speed threshold) that can separate fixations from saccades (ii) the ability to precisely measure oculomotor functions (e.g. delays in an antisaccade test). But these experiments have a significant disadvantage: they cannot model the dynamic environment that humans have evolved to live in.

To overcome the previous problem researchers have used videos instead of static stimuli. Because videos introduce motion they also evoke smooth pursuit eye movements in order to follow moving targets and comprehend the environment. Moreover, if the head is allowed to move freely, as we do in our everyday life, the eyes start to perform other functions such as compensatory movements (e.g. vestibulo-ocular reflex). These compensatory movements have the effect of changing the gaze signal in HMD-based experiments and thus making not only the analysis but also the definition of eye movements more challenging. Below we tackle this problem by providing clear definitions of all the eye movement types that are used in this thesis [Startsev et al., 2019b, Agtzidis et al., 2019].

2.2.1 Eye movement types definitions

Because this thesis provides data sets with hand-labeled eye movements and algorithms that automate the labeling process it is of paramount importance to clearly define each term that we use. This becomes even more important because the eye movement community seems to disagree about the definitions of eye movements even when we talk about the most common fixations and saccades [Hessels et al., 2018]. Our eye movement definitions are based on two of our previously published manuscripts [Startsev et al., 2019b, Agtzidis et al., 2019] and below is the full list of those that are mentioned in this thesis.

Fixation: A period of time where no movement of the eye inside the head is triggered by retinal input. This can include reflexive eye motions that compensate for head motion or slow gaze signal drifts that arise from changes in pupil dilation [Drewes et al., 2014].

Saccade: High-speed ballistic movement of the eye to shift the point of regard, thus bringing a new (part of an) object onto the fovea (including adjusting the gaze position to match the tracked object via catch-up saccades during pursuit, or similar).

Post-saccadic oscillations (PSOs): As the name suggests they appear at the end

of saccades and their shape varies depending on the amplitude of the saccade [Hooge et al., 2015]. Also people have argued that PSOs are not an actual eye movement but a relative movement of the pupil inside the iris that is detected by video-based eye trackers [Nyström et al., 2013]. Therefore in the context of this work we are not treating PSOs as a separate eye movement but as part of the saccades.

Smooth pursuit: A period of time during which the eyes are in motion inside the head and a moving (in world coordinates, relative to the observer) target is being foveated. The motion of the target can either arise due to its own movement or camera motion.

Noise: Even though noise is not an actual eye movement type, we accumulate blinks, drifts, tracking loss, and physiologically implausible gaze signals under this one name.

Vestibulo-ocular reflex (VOR): A period of time when the eyes are compensating for head motion and stabilizing the foveated area.

Optokinetic nystagmus (OKN) or nystagmus: Sawtooth-like eye movement patterns, composed of fast saccadic parts alternating with slow stabilization parts. We labeled all such patterns as OKN, though it has to be noted that some of these labels correspond to nystagmus, e.g. when a person is observing a blank part of the synthetic stimulus while simultaneously turning the head, so the reflexive movement is not actually triggered by the visual input.

OKN+VOR: This is a combination of the two previous categories: The eye signal exhibits a sawtooth pattern during head rotation.

Head pursuit: A period of time where a pursuit of a moving target is performed only via head motion, with the gaze direction within the head relatively constant.

Throughout this thesis the first five eye movements (fixation to noise) are often mentioned as *primary* because they can be performed if the head moves freely or if it is fixed. The rest of the eye movements (last four) are mentioned as *secondary* because they require free head motion and the purpose of the eye movements is to compensate or to counteract the head movement.

2.3 Smooth pursuit

Even though we have presented what smooth pursuit means in the context of this thesis, it is worth providing a more extensive overview. Smooth pursuit started to be

investigated more than a century ago and roughly at the same time as fixations and saccades [Dodge, 1904]. Also people realized very early that smooth eye movements are important for the perception of motion [Dodge, 1904, Lord and Wright, 1949] and started to investigate its characteristics [Westheimer, 1954, Robinson, 1965] and its interplay with other oculomotor functions [Fox and Dodge, 1929, Dodge et al., 1930, Rashbass, 1961]. The study of the relationship between brain areas and the dynamics of smooth pursuit also started at a relatively early point [Sharpe et al., 1979, Katsanis and Iacono, 1991] and became more thorough as new imaging techniques became available [Petit and Haxby, 1999, Lencer and Trillenberg, 2008].

As personal computers started to become ubiquitous at the turn of the century researchers started to investigate ways that could make the analysis of gaze data easier. By this time eye movement research was dominated by static stimuli (including reading) and thus the engineers of the time did not notice SP or considered its analysis as something niche. Therefore the first algorithms for automatic eye movement classification did not detect SP at all [Sauter et al., 1991, Goldberg and Schryver, 1995, Salvucci and Anderson, 1998, Salvucci and Goldberg, 2000]. Oftentimes when dynamic natural scenes were used there was no distinction between fixations and SP because either SP was defined as a fixation on a moving target [Steil et al., 2018] or they were implicitly grouped together [Mathe and Sminchisescu, 2012, Wang et al., 2018] assuming that SP constituted a small fraction of the overall viewing time. However, both of these assumptions are not valid because the SP percentage is on average high and varies from 10 to 24 % depending on the stimulus type (for more details see the basic statistics of Chapter 4) and because different neural mechanisms drive different eye movements [Luna et al., 1998, Beauchamp et al., 2001, Kimmig et al., 2008].

When SP was specifically analyzed more often than not the input modality was artificially created stimuli where up to a couple of moving targets were present [Heinen and Watamaniuk, 1998, Schütz et al., 2011]. This type of stimuli apart from allowing to investigate specific eye movement attributes, such as the delay of attending a new moving object, does not require per se the detection of SP in the signal since a simple distance-based metric would suffice for identifying the followed target. However, the biggest shortcoming of the previous approach is the diminished "naturalness" of the experiment, which excludes the decision making and planning processes involved in the gazing of objects of interest in dynamic natural scenes. To overcome this shortcoming researchers recently have used videos of natural scenes [Dorr et al., 2010, Mital et al., 2011], head-mounted eye trackers [Martens and Fox, 2007, Giannopoulos et al., 2015], as well as head-mounted displays (HMD) [David et al.,

2018, Sitzmann et al., 2018]. However, in these dynamic environments SP is much more challenging to detect both manually but also algorithmically. For monitor-based experiments the algorithms that detect SP based on simple criteria such as velocity or dispersion [Komogortsev and Karpov, 2013] usually return poor results (see Section 5.3.3) due to SP's overlapping characteristics with fixations and saccades (see Sections 4.1.5, 4.2.5, and 4.3.3). More elaborate algorithms have been developed [Berg et al., 2009, Larsson et al., 2015, Dar et al., 2019], which perform very well on average but their SP detection performance has a much larger room for improvement in comparison to fixation and saccade detection. In this thesis (Chapter 5) we present two new eye movement classification algorithms that become the new state-of-the-art in automatic eye movement labeling. Finally, the data sets for the evaluation of these algorithms under naturalistic conditions are very few and they only span a couple of minutes [Larsson et al., 2013, Andersson et al., 2017]. To overcome this limitation in Sections 4.1 and 4.2 we present two large monitor-based ground-truth data sets that span in total more than 6 hours of recordings. In Section 4.3 we present a new data set that was recorded in an HMD and allowed free head motion together with a hand-labeled ground-truth subset of it that spans roughly 30 minutes.

2.4 Head-mounted eye tracking

In recent years due to the dramatic decrease in price and set-up complexity of remote gaze tracking hardware[1], eye tracking has been increasingly applied in more challenging domains. Currently, eye tracking is being integrated into consumer-oriented virtual and augmented reality (VR/AR) devices[2,3,4] and should be widely available in the following years. Previously we mentioned that head-mounted eye tracking can provide an almost uninterrupted viewing experience and thus a better representation of natural eye movement behavior. But before being widely applied in research many of the intricacies that arise from the transition to free head movement have to be handled.

The biggest source of confusion when free head motion is allowed arises from the different frames of reference where data can be reported. These frames of reference are summarized in Figure 2.2. In the simplest and most widespread experimental

[1] `https://gaming.tobii.com/product/tobii-eye-tracker-4c/`
[2] `https://www.vive.com/eu/pro-eye/`
[3] `https://www.getfove.com/`
[4] `https://www.microsoft.com/en-us/hololens/hardware`

setup a monitor is used for the stimulus presentation and the head is either fixed (e.g. with a chin rest) or allowed to move freely, usually within a limited bounding box because a remote eye tracker is utilized. The lack of confusion here arises from the coincidence of the stimulus coordinate system with that of the monitor, which simplifies both the understanding and the analysis of the data. Therefore any potential head motion is disregarded and a pair of x, y values suffices to describe the state of the experiment, since the monitor is a 2D plane.

In the most liberal scenario the participant wears a pair of (AR) eye-tracking glasses and moves in the real world, or wears a VR headset and moves in a virtual world. In this scenario two 3D coordinate systems are needed to describe the experiment: (i) a world coordinate system that is fixed permanently at a location in the world and (ii) a coordinate system that is attached to the participant's head with its origin usually placed in the midpoint between the two eyes. Then the location of the participant in space and the head's orientation can be described through the translation and rotation of the head coordinate system in relation to the world coordinate system. In this type of experiments the eye tracker is attached to the participant's head and reports the gaze vectors in head coordinates. Because of this the gaze signal contains eye movements such as VOR that compensate for head motion. But if the gaze vectors are then reported in the world coordinate system through the head to world transformation the signal is now a combination of eye and head motion and therefore different than before. Due to this difference, the combination of these two different signals can be used for the assignment of primary (fixation, saccade, SP) and secondary (VOR, OKN, head pursuit) labels to each gaze sample.

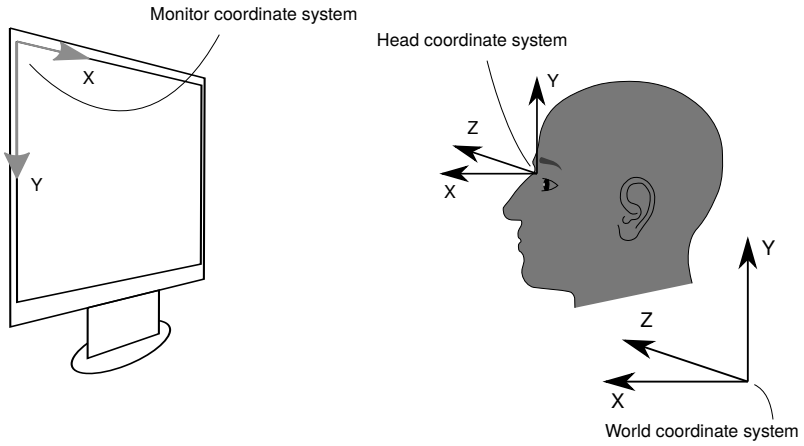


Figure 2.2: Visualization of the different frames of reference that are present in eye tracking experiments.

In the context of this thesis as a first step towards completely unconstrained eye movement analysis we exclude the translation component from the previous setup

and make the origin of the head and world coordinate systems to coincide. We also use as stimulus monoptic 360-degree videos displayed in an HMD, which allows for a simplified data representation. The world gaze vectors can be represented in the stimulus coordinate system (equirectangular video) just with a pair of values and can be directly visualized with pre-existing tools. Similarly the head orientation can be represented with a pair of values in the stimulus space together with a third value representing the roll component. The combination of the head orientation together with the world gaze representation lets us compute the gaze representation in the head coordinate system, which is equivalent to the rotation of eye ball within its socket. This process is described in more detail in Section 6.1.

2.5 Evaluation metrics

In the past many algorithms have been developed for the automatic segmentation of the gaze trace into eye movements and have achieved some kind of state-of-the-art performance. In this thesis we also develop algorithms for eye movement classification and diverse ground-truth data sets for the evaluation of these algorithms. When we report the performance of an algorithm we provide a number that is derived from an "objective" evaluation function that compares the algorithmic output against the ground truth. The objective function can rely on a single metric or a combination of metrics, but there exists no single metric that is globally acceptable and can describe all aspects of an algorithm. For example the MIT300 [Bylinskii et al., 2016] saliency benchmark uses 8 different metrics to evaluate the performance of the submitted algorithms.

The eye-tracking community is no different from any other scientific field and many different evaluation metrics have been proposed. For example, [Komogortsev and Karpov, 2013] have proposed behavior metrics but they are oftentimes difficult to apply and interpret when dynamic natural stimuli are used. The easiest to understand evaluation metrics are based on sample level statistics due to the discrete nature of the eye-tracking signal that arises from the sampling frequency of the eye tracker. Some of the most commonly used statistics are precision, recall, and Cohen's Kappa [Cohen, 1960]. Precision (Equation 2.1) represents the proportion of correctly labeled samples among all retrieved samples. Recall (Equation 2.2) is the proportion of correctly retrieved samples among all samples of the same category. A balanced representation of these two is the F1 score (Equation 2.3), which is their harmonic mean. Cohen's Kappa measures the agreement between raters by accounting for the chance agreement between the two (introduced in Equation 6.9).

$$precision = \frac{TP}{TP + FP} \quad (2.1) \qquad recall = \frac{TP}{TP + FN} \quad (2.2)$$

where TP = true positives, FP = false positives, FN = false negatives

$$F_\beta = (1 + \beta^2)\frac{precision * recall}{\beta^2 * precision + recall} \quad (2.3)$$

for $\beta = 1$ we get the $F1$ score

The previous statistics can use "events" instead of independent samples but then their application is not always straightforward. First, the term "event" and secondly, the criteria for event matching have to be defined. Throughout this thesis we use the terms "event"and "episode" interchangeably and both refer to a period of time where all the gaze sample class labels (either in human annotations or in the output of an algorithmic detector) are identical. Thus, any gaze recording is subdivided into non-overlapping eye movement events (episodes). [Hoppe and Bulling, 2016] were matching a ground truth event to the majority vote of the algorithm's samples within the temporal window of the earlier. [Hooge et al., 2017] were matching an event in the ground truth with the earliest algorithmically detected event that intersects with it and only one-to-one matching was allowed. [Zemblys et al., 2018a] changed the matching criterion and the intervals with the longest intersection were matched. [Startsev et al., 2019a] further controlled the quality of matching by specifically defining the minimum amount of overlap between two intervals. For this purpose they used the intersection over union [Everingham et al., 2010, Everingham et al., 2015] with a standard value of 0.5.

A more recent approach [Startsev et al., 2019c] uses all the previous metrics during evaluation but not each one independently. Instead, it creates different baselines that are either randomly drawn from an event pool or are based on the similarity between subjects and compares the algorithm against them. As the authors report often times the "oblivious" baselines return higher scores than the classification algorithms, which indicates the weak generalization ability of the latter.

In this thesis we will use the F1 score to report the performance of the algorithms since it encompasses both precision and recall in one number and it is easier to

interpret than Cohen's kappa. Together with sample-level F1 scores for each eye movement we will also report event-level F1 scores that are matched with the scheme of [Hooge et al., 2017].

Part II

Foundational data sets

The objective of this thesis is to work towards an understanding of eye movements as they are performed in unconstrained environments. However, the majority of the publicly available information about eye movements regards either static stimuli or dynamic stimuli in the form of moving dots. Therefore, here we provide the largest set to date of hand-annotated eye movements in many diverse scenarios.

In Chapter 3 we present the tool that was used for the manual annotation of eye movements. Its initial version could handle monitor based stimuli only and was presented in [Agtzidis et al., 2016a]. A significantly expanded version of this tool with improved performance and the ability to handle head-free experiments with 360-degree equirectangular input was published in [Agtzidis et al., 2019]. This tool is used in Chapter 4 in order to hand annotate three large dynamic natural data sets.

The first data set contains the labeled eye movements of the 50 participants as they were watching the GazeCom data set [Dorr et al., 2010] that contains short clips of everyday scenes and its full details have been published in [Startsev et al., 2019b]. The second data set comprises the eye movements of 16 participants as they were watching Hollywood movie excerpts [Mathe and Sminchisescu, 2012] and its details have been published in [Agtzidis et al., 2020b]. The last data set contains the labeled eye movements of 13 participants as they were watching clips of everyday scenes in a head-mounted display that allowed free head motion and was presented in [Agtzidis et al., 2019].

Chapter 3

Labeling tool

Before presenting the hand-labeled eye movement data sets we will explain in this chapter the infrastructure that was developed and used during labeling. We will start with the presentation of the used format for data representation and then explain the structure of the labeling interface and the labeling process. At the end we explain how the tool handles data recorded with 360-degree equirectangular stimuli.

3.1 Data format

For the data representation we used an extendable data file format, which has been used extensively in the data mining community and more specifically was introduced and used in WEKA [Hall et al., 2009]. The ARFF (Attribute-Relation File Format) file is a text file that describes a list of instances sharing a set of attributes.

In ARFF, all keywords start with a "@" symbol and the following names are case-insensitive; all lines starting with "%" are considered comments. Any file comprises two sections for a header and the data. The header starts with "@relation", which defines the relation name. After this, the attributes can be declared through the "@attribute" keyword followed by the name and type of the attribute.

The data section starts with the "@data" keyword. The further lines describe the instances with one instance per line and comma-separated attributes. The attributes should follow the same order used for their declaration in the header section.

In our implementation, we introduce minor deviations from the regular ARFF format, but we maintain compatibility with the standard. Since an eye-tracking experi-

ment require extra information about the experimental setup, we introduce "special" comments (maintaining ARFF compatibility) starting with the "%@metadata" keyword and followed by a key-value pair.

An example ARFF file is provided in Listing 3.1. At the beginning the relation describes the data present in the file. In this example the data was recorded in a 360-degree headset. Then follow the metadata that explain the experimental setup with regard to the used headset and the video stimuli. The following attributes explain the gathered data, which include two types of labels from a hand-labeler.

Listing 3.1: Sample ARFF file.

```
@RELATION gaze_360

%@METADATA distance_mm 0.00
%@METADATA fov_height_deg 100.00
%@METADATA fov_height_px 1440
%@METADATA fov_width_deg 100.00
%@METADATA fov_width_px 1280
%@METADATA height_mm 0.00
%@METADATA height_px 1920
%@METADATA width_mm 0.00
%@METADATA width_px 3840

@ATTRIBUTE time INTEGER
@ATTRIBUTE x NUMERIC
@ATTRIBUTE y NUMERIC
@ATTRIBUTE confidence NUMERIC
@ATTRIBUTE x_head NUMERIC
@ATTRIBUTE y_head NUMERIC
@ATTRIBUTE angle_deg_head NUMERIC
@ATTRIBUTE handlabeler_1_pl {unassigned,fixation,saccade,SP,noise}
@ATTRIBUTE handlabeler_1_sl {unassigned,OKN,VOR,OKN+VOR,noise,
   head_pursuit}

@DATA
0,2012.91,1192.18,1.00,1899.00,1060.30,1.49,fixation,unassigned
8000,2012.87,1192.16,1.00,1899.01,1060.35,1.49,fixation,unassigned
...
553000,1928.15,1038.05,1.00,1895.76,1058.68,1.43,saccade,unassigned
564000,1916.92,1029.23,1.00,1895.36,1058.45,1.41,saccade,unassigned
571000,1910.23,1019.49,1.00,1895.06,1058.26,1.40,saccade,unassigned
578000,1909.20,1019.19,1.00,1894.69,1058.02,1.39,fixation,VOR
587000,1907.08,1019.64,1.00,1894.25,1057.72,1.38,fixation,VOR
```

3.2 Labeling interface

The labeling interface was developed in C++ with Qt 5 providing the graphical interface. It has been tested in Ubuntu 18.04 and the source code is publicly available under an open-source license[1]. As its name suggests the tool can be used for the visualization and labeling of eye-tracking data and does not contain eye movement classification functionality. If the output of an algorithm needs to be visualized, the algorithm has to be run in advance and its output is added as an extra attribute in the ARFF file.

The labeling interface provides the necessary information through four panels and an example is provided in Figure 3.1. The top-left panel displays the video itself overlaid with the gaze samples from a 200 ms temporal window centered at the current time (red circles represent past samples and gray circles future samples). The two right panels display the x and y coordinates of the gaze data and the bottom-left panel its speed. The speed panel applies low-pass smoothing in the computed speed because without smoothing it usually returns noisy results. The smoothing is achieved by computing the speed between consecutive samples that span 100 ms in time but the temporal window can be changed by the user. Moreover, the background of all three panels is color-coded based on the visualized attribute. When the attribute is of nominal type then the nominal values are displayed in the top-left corner of the panels, otherwise a set of default values is used as in the example screenshot below.

Also the three panels handle most of the user interaction with regard to the color-coded panels. The following actions can be performed through the panels:

- Right-clicking and dragging moves the current position in time (backwards or forwards according to the direction of the mouse movement).
- Scrolling the mouse wheel changes the temporal scale, i.e. increases or decreases (according to the scroll direction) the temporal window represented by the plots on the right and bottom-left panels.
- Left-clicking and dragging a border expands or shrinks the adjacent intervals. If a border is moved further than the interval duration then the interval is deleted.
- Holding the left-click on an interval and pressing a number on the keyboard changes the label of the interval. The legend provides information on the correspondence between numbers and the assigned labels.

[1] `https://gin.g-node.org/ioannis.agtzidis/gta-vi`

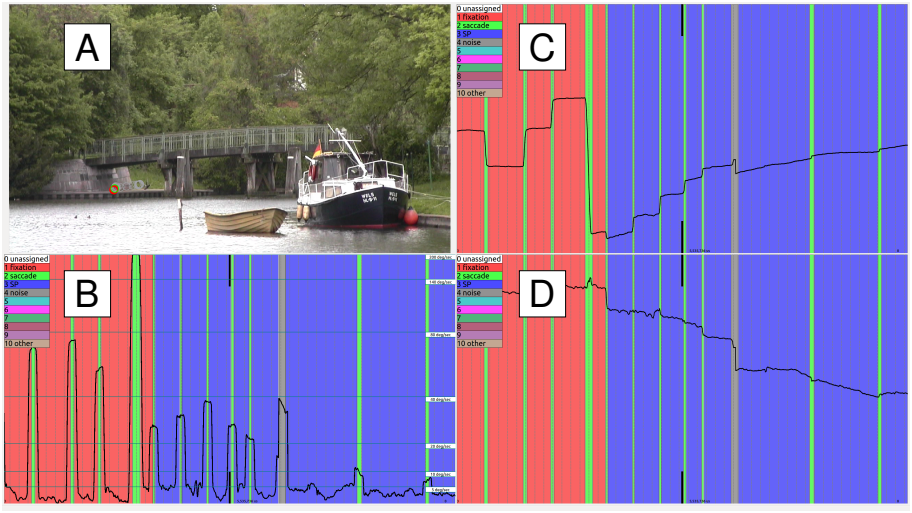


Figure 3.1: Screenshot of the labeling tool. The video with overlaid gaze traces is playing in the top left (panel A), while x- and y-coordinates are plotted as a function of time on the right (panels C, D). The bottom left panel (B) displays the gaze speed in the same time window as the gaze panels. The longer duration of high speed values during saccades in the speed panel is due to the low-pass filtering.

- The sequence of a left-click, a number key press and finally pressing the *Insert* key inserts a new interval of the selected type spanning a temporal window of +/- 40 ms around the current time; this interval can then be adjusted as above.
- The sequence of a left-click and pressing the *Delete* key unassigns the label of the selected interval.

The interaction is completed with some standard keyboard shortcuts:

- Pressing *Space* key starts playing or pauses the video.
- Pressing *Ctrl-Z* reverts the last change.
- Pressing *Ctrl-Shift-Z* acts as "redo" (reapplies the last canceled change).

3.3 Handling of 360-degree data

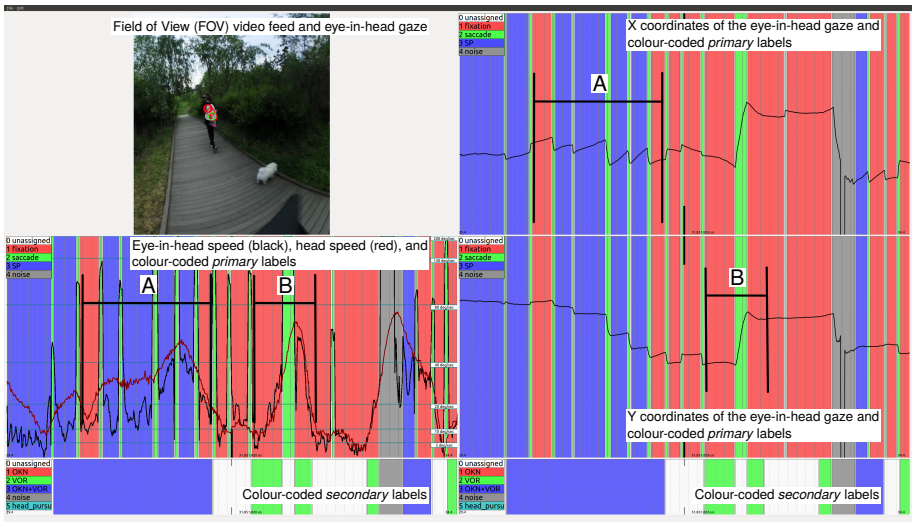
In order to be able to label 360-degree data with our tool we had to make some changes regarding the labeling process and how the different coordinate systems (head-fixed and world) of Section 2.4 could be handled. For the different coordinate

systems we provide two modes of operation that we call *field of view* and *eye+head* and for both of these we provide the option to assign secondary labels.

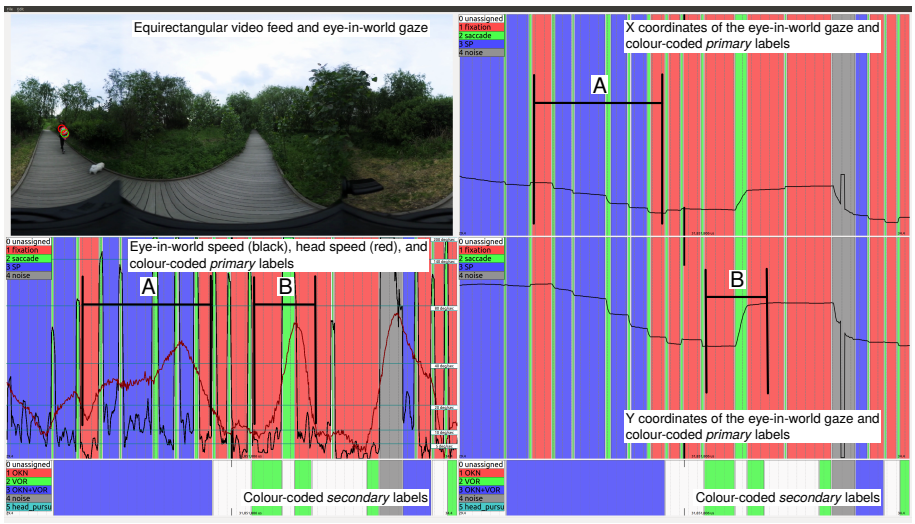
In the *field of view (FOV)* mode (Figure 3.2a), the annotator is presented with the view of the scene that is defined by the corresponding head rotation of the subject (the size of the visualized video patch roughly corresponds to the field of view that the participant had in the VR headset). This view corresponds to the frame of reference that moves together with the participant's head and allows us to see the actual visual stimulus that was perceived by the participant and to analyze the eye-within-head gaze behavior.

In the *eye+head (E+H)* mode (Figure 3.2b), the full equirectangular video frame is presented to the annotator. Visualizing gaze locations in this view enables the annotator to see the combination of the head and eye movement, which corresponds to the overall gaze in the frame of reference of the world (or the 360-degree camera, to be more precise).

In both operation modes the x and y gaze coordinates as well as the gaze speed are plotted over time as before. However, the coordinate systems used for these plots differ between the two modes: In the FOV mode, the gaze coordinates and the speed of gaze are reported in the *head*-centered coordinate system, whereas in the E+H mode, the coordinates and the speed are reported in the *world* coordinate system. This way, the FOV representation provides the annotator with the eye motion information within the eye socket, while the E+H representation is responsible for highlighting the absolute movement of the foveated objects, which is necessary for determining the precise label type, e.g. distinguishing between fixations and pursuits. The difference between the two modes is evident in the marked areas of Figure 3.2.



(a) FOV mode



(b) E+H mode

Figure 3.2: Schematic of field-of-view (a) and eye+head (b) operation modes of the hand-labeling tool together with panel description. Colored intervals correspond to different primary (on three large panels) and secondary (bottom panels) labels. Differences in gaze coordinates and speed patterns (e.g. intervals A and B) allow for improved annotation.

For 360-degree data we also display the head speed (red line in the speed panel) and two more panels at the bottom of the interface for the visualization of the secondary labels, which are used for the description of the gaze samples in more detail (e.g. fixation together with VOR). The secondary panels have the same behavior as the interactive panels (explained in the previous section) but they can also add a new interval that temporally matches the respective primary interval by double left-clicking on the mouse. The value of the added interval is again here selected through a number key press.

Chapter 4

Hand-labeled data sets

With the labeling tool of the previous chapter and by using the eye movement definitions of Section 2.2.1 we have hand-labeled three large eye movement data sets that span approx. 7 hours of recordings. Two of the labeled data sets represent monitor-based experiments and they are based on previously published gaze data sets. The last data set represents an HMD-based experiment that was recorded and labeled in the context of this thesis. All the data presented in this chapter are made publicly available with an open-source license[1,2,3].

4.1 GazeCom data set

4.1.1 Data set description

Because the GazeCom [Dorr et al., 2010] data set is one of the three data sets on which we built our work, we briefly describe its set-up and provide some basic information here. The data set comprises 18 short naturalistic video clips (20 s each), depicting everyday scenes. These include beach scenes, pedestrian and car-filled streets, boats, animals, etc. There is little to no camera motion in the recorded clips (11 out of 18 clips lack it completely, four have slow panning camera motion, and the camera was slightly shaking in the other three), and the scenes themselves contain both rigid (e.g. cars) and non-rigid (e.g. human or animal) motion at a variety of speeds. These clips thereby form a set of dynamic and relatively naturalistic stimuli.

[1] https://gin.g-node.org/ioannis.agtzidis/gazecom_annotations
[2] https://gin.g-node.org/ioannis.agtzidis/hollywood2_em
[3] https://gin.g-node.org/ioannis.agtzidis/360_em_dataset

All video clips were presented at 1280×720 pixels, 29.97 frames per second, at a distance of 45 cm from the observers. The frames covered an area of 48×27 degrees of visual angle. The gaze of 54 participants was recorded at 250 Hz with an SR Research EyeLink II eye tracker. Even though the eye tracker allowed for small head motion, a chin rest was used to stabilize the participants' heads. Some recordings were discarded by the authors of the data set due to frequent (over 5%) tracking loss, leaving 844 recordings in the published data set (46.9 per clip on average). These data total 4.5 h of gaze tracking recordings, all of which we annotate and analyze in the context of this work.

4.1.2 Labeling procedure

Before the annotators started labeling the data set the gaze samples were automatically pre-labeled using our implementation of the saccade and fixation detection algorithms of [Dorr et al., 2010] and an early version of the clustering algorithm of [Agtzidis et al., 2016b] for the annotation of SP. The purpose of pre-labeling the samples was to speed-up the labeling process since the annotators would mainly have to adjust the borders and change the labels of the already present intervals instead of adding all the intervals manually[4].

The algorithmic labeling of the gaze samples prior to manual annotation allowed us to roughly double the speed of the labeling process: For an expert annotator, the labeling time decreased from ca. 10 to ca. 4 mins on average per single ca. 20 s recording. This speed-up becomes more important considering that the GazeCom data set comprises of 4.5 h of gaze recordings and that multiple passes were performed during its labeling.

The labeling procedure involved novice and expert annotators. The novice annotators were undergraduate students of the Technical University of Munich and they were compensated for their work. Initially they received information about the different types of eye movements that they were going to label along with some representative examples. Throughout the duration of their work they were free to ask for clarifications about ambiguous cases. In the first pass they were presented with the pre-labeled suggestions and they were instructed to change them accordingly. As an added quality assurance measure some of their first annotations were visually inspected and feedback was provided for cases where the eye movement

[4]An evaluation of this method showed that the manual annotators were not biased by the suggested intervals and this is specifically evident in the contribution of different eye movements in the final SP amount (Figure 4.1).

definitions of Section 2.2.1 were violated. Because their level of expertise and understanding of eye movements had changed at the end of the first annotation round the novice annotators performed a second annotation pass. In the second pass they were presented with their own annotations and they were instructed to change them wherever they felt it was needed. The third (expert) annotator (the author of this thesis) re-examined all the recordings in the data set and his main purpose was to resolve disagreements between the first two annotators but he was also able to change the labels wherever he felt it was appropriate.

4.1.3 Inter-rater agreement

In Table 4.1 we present how well the labels of the three annotators agreed in terms of sample-level F1 scores. The event-level F1 scores are omitted because they are quantitatively similar to the ones presented here. We can see the annotation of fixations and saccades returns high agreement scores among annotators and different passes of the same annotator. However, the agreement scores for SP are substantially lower across the board. The final annotator, who was resolving conflicts between the two final passes of the novice annotators, tended to agree the most with the first of the two. Interestingly, the SP agreement scores between the two passes of the novices are similar to the inter-rater agreement (excluding the 1_{final} vs. $final$). These low scores demonstrate the difficulty of labeling smooth pursuit in naturalistic stimuli even when clear definitions are provided.

Table 4.1: Agreement between the initial (1_{ini} and 2_{ini}) and final (1_{final} and 2_{final}) annotations of the two non-expert annotators, and all annotator pairs in the form of sample-level F1 scores. The "final" label refers to the annotations of the third (expert) rater, who consolidated the labels of 1_{fianl} and 2_{final}.

EM type	1_{ini} vs. 1_{final}	2_{ini} vs. 2_{final}	1_{final} vs. 2_{final}	1_{final} vs. final	2_{final} vs. final
Fixation	0.950	0.977	0.933	0.975	0.949
Saccade	0.904	0.951	0.863	0.937	0.883
SP	0.787	0.796	0.629	0.904	0.697

4.1.4 Hand labeling statistics

Labeling the full Gazecom data set lasted the equivalent of several months of full-time work (including the two passes through the whole data set for each of the

two novice annotators). On average for all three annotators, labeling one GazeCom recording (usually ca. 20 s) took between 5 and 6 minutes, which is equivalent to a labeling time of 15–18 s for each second of the recorded gaze signal. The labeling process also benefited from pre-labeling the gaze signal, which more than doubled the labeling speed.

Figure 4.1 visualizes the changes in labels between the suggested (algorithmically pre-labeled) and the final hand-labeled eye movement samples, in the form of a confusion matrix. The matrix cells sum to 1.0 and therefore each individual cell contains the overall share (e.g. 6.5% of the samples were suggested as part of a fixation but were changed to SP). The colors in the matrix (see color bar of the figure) represent how much each suggested label contributed to the final "ground-truth" label, i.e. the color values are normalized per column and if the suggestions were perfect only the diagonal would be brightly colored. We observe that the majority of the samples for fixations and saccades were correctly suggested by the algorithms (over 90% of final labels came from the corresponding cell in the suggestions). However, ca. 59% of the final SP labels were suggested as fixations and only 27% was correctly suggested. Also for the noise label most of the final labels were initially unassigned and around 30% came from saccades that were forming part of blinks but the eye tracker returned a signal similar to a rapid eye movement because video-oculography [Holmqvist et al., 2011, p. 177] was used for the recording of the GazeCom data set.

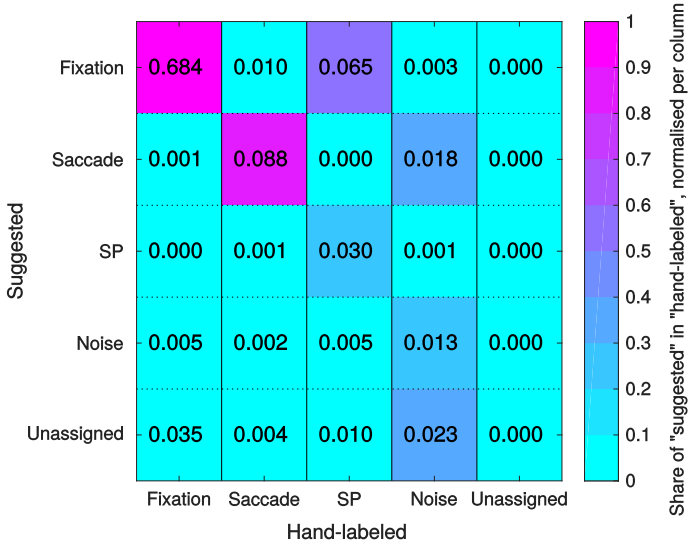


Figure 4.1: Confusion matrix for the pre-labeled and manually annotated eye movement samples. Rows correspond to the suggested eye movement labels, columns – to the final hand-labeled classes. The color bar on the right does not match the numbers in the cells because cell color reflects the share of samples in the final hand-labeling that were originally pre-labeled as the respective suggested classes (i.e. per-column normalization is employed; cf. the color bar on the right).

Figure 4.1 already reflects the proportions of samples that each eye movement type was allocated during the pre-labeling or manual labeling phase (these numbers can be obtained by summing either the matrix rows or columns, respectively). In Table 4.2 we provide these numbers together with the corresponding number of episodes (periods of time where consecutive samples have the same label). From the table we see that the percentages and number of episodes for fixations and saccades have not changed substantially. On the contrary, the percentage of SP has increased more than threefold (from 3.3% to 11%) and the corresponding number of episodes from ca. 3000 to ca. 4500. To conclude, we can say that the suggested labels were changed substantially and most of the changes appeared in the SP and noise labels.

Table 4.2: The overall percentage of gaze samples and number of episodes of all eye movement types in the algorithmically suggested ("pre-labeled") labels and the final set of labels produced in our annotation procedure.

EM type	Suggested label		Final expert label	
	Share	Episodes	Share	Episodes
Fixation	76.2%	39,293	72.6%	38,629
Saccade	10.7%	40,233	10.5%	39,217
SP	3.3%	2879	11%	4631
Noise	2.5%	6319	5.9%	3493
Unassigned	7.3%	27,165	0%	0

4.1.5 Basic statistics

Overall, the hand-labeled GazeCom data set contains 38,629 fixations, 39,217 saccades, and 4631 SP episodes (Table 4.2). While the number of SP episodes may seem small there are more pursuit than saccade samples (11% vs. 10.5%).

Further down we visualize some basic and commonly used (e.g. [Salvucci and Goldberg, 2000, Komogortsev and Karpov, 2013, Santini et al., 2016, Zemblys et al., 2018b, Startsev et al., 2019a]) statistics of the ground-truth fixations, saccades, and pursuits.

Figure 4.2a visualizes the distribution of the overall speed, duration, and amplitude of the events of each eye movement class. Notably, some average saccade speeds (Figure 4.2a) were very low due to the inclusion of PSOs in our definition, but overall very fast compared to the other two classes. Smooth pursuits, on the other hand, are expectedly faster than fixations on average, but there is a substantial overlap between the two classes in terms of their average speed, as it is evident

from the quartile lines (vertical dashed lines) in the figure. This overlap makes the distinction between drifting fixations and SP more challenging, at least when purely speed-based thresholding is concerned.

Examining event durations (Figure 4.2b), saccades are again clearly separable from the other two types as their maximum duration does not exceed 100 ms. By contrast, 75% of fixation and SP intervals lasted longer than 200 ms, and their overall duration distributions almost perfectly overlap.

Finally, the amplitude distributions of the three eye movement types (related to the dispersion feature used by classifiers) is presented in Figure 4.2c. Here, a fair separation between fixations and saccades would be possible with a single threshold, but the distribution of SP amplitudes in this data set significantly overlaps with both of the other distributions.

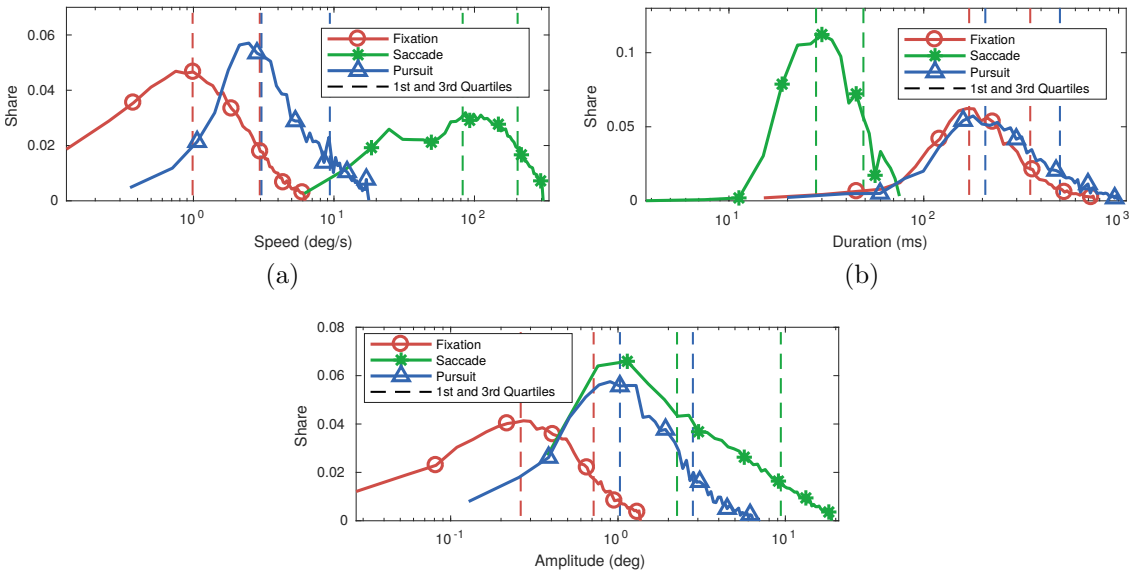
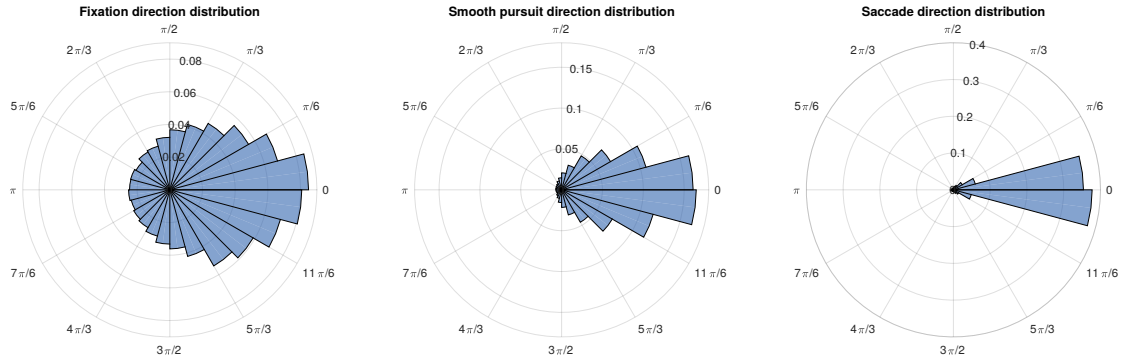


Figure 4.2: Overall per-episode speed, duration, and amplitude distributions for fixations, saccades, and smooth pursuits. These are the (normalized) histograms, which were computed for each eye movement type independently with 50 equal-sized bins covering each respective speed range. These were then plotted here in log-scale (see x-axis), with the y-axis representing the share of episodes in each of the bins. The dashed vertical lines visualize the quartiles (first and third) of the respective distributions. Note that since the horizontal axis is in log-scale, it is difficult to visually compare the areas under different parts of the curves. For example, for fixations (red solid line), 50% of the labeled episodes (between the first and third quartile lines) had an overall speed between 1 and 3°, as indicated by the left and right vertical red lines, respectively.

Figure 4.3 visualizes the angular deviation of the sample-to-sample gaze direction



(a) (b) (c)

Figure 4.3: Directional deviation distributions for fixations (a), pursuits (b), and saccades (c) presented as circular histograms. To decrease the effect of noisy recordings we compute the directional deviation in 16 ms windows that move one sample at a time. The height of each bar represents the share of the velocity vectors with the given angular deviation from the overall direction of their corresponding episode. Zero deviation angle means perfect alignment with the overall direction of the respective episode.

from the overall episode direction in the form of a circular histogram. Values close to zero mean that the gaze course in the given movement is almost co-directional with the overall episode direction. From the figure we observe that each eye movement type has a distinctly different pattern from the other two. Fixations are the most evenly distributed (Figure 4.3a) but they are not exactly uniform around the circle due to the fact that minuscule gaze shifts or eye-tracking drifts would result in biasing the distribution. SPs (Figure 4.3b) have a more pronounced peak around zero and this is to be expected due to the existence of a pursued moving target. Finally, saccades (Figure 4.3c) have the most pronounced peak and the majority of the values are found in the two bins around zero. These distinct directional patterns among fixations, saccades, and SPs demonstrate that they are valid features for eye movement classification and researchers have already demonstrated this [Larsson et al., 2016, Startsev et al., 2019a].

4.2 Hollywood2 data set

4.2.1 Data set description

The second hand-labeled data set is based on the gaze recordings of the Hollywood2 data set [Mathe and Sminchisescu, 2012] and provides the labels for a subset of it

(approx. 130 minutes). The Hollywood2 data set was recorded, as its name suggests, with Hollywood movies (movie excerpts, to be precise) as stimuli. It contains ca. 20 hours of gaze recordings. The purpose of the data set was action recognition through eye movements, and the pool of 16 eye-tracking experiment participants was split into two groups. The task of the "active" subgroup (12 subjects) was to assign one of the 12 action classes to each video clip while their eye movements were recorded. The "free viewing" subgroup (4 subjects) had no task and was simply watching the video clips. The participants' head was stabilized with a chin rest and the eye movements were recorded monocularly for the dominant eye at 500 Hz with an SMI iView X HiSpeed 1250 eye tracker. A relatively high eye tracking accuracy of 0.75 degrees was achieved via a 13-point calibration procedure at the beginning of each recording block, plus a validation step at the end – if the validation accuracy fell outside these limits, the data were discarded.

4.2.2 Labeling procedure

Again, here we used the labeling tool of Chapter 3 for the labeling of eye movements. On each gaze recording two human annotators worked sequentially. The first labeler was a paid student at the Technical University of Munich, working part-time (8 h/week for 22 weeks). She obtained basic knowledge about eye movements from following a relevant course and additional clarifications from the authors. She was also provided with representative examples for the eye movement definitions from the previous section in action in the context of the labeling interface. During the full duration of the labeling process, experts were available to answer any of her questions. Randomly chosen annotated files were periodically visually inspected by the authors, and feedback was provided.

To speed up the annotation process, the gaze files were pre-segmented with the I-VVT algorithm [Komogortsev and Karpov, 2013] with default parameters before being processed by the first annotator. By providing the automatically detected intervals, even if those were poorly aligned with the actual eye movements, the task of the annotator was simplified to mainly merging intervals and correcting their temporal location, instead of constantly adding new intervals one by one and then correcting their borders. By using the I-VVT algorithm instead of a more elaborate approach (see Section 5.3), the labeler could not leave the suggested labels uncorrected: The outputs of I-VVT on our data were very noisy, meaning that the first annotator had to carefully inspect the full file, thus decreasing the potential of labels being biased by the algorithmic pre-segmentation. For the amount of changed

samples per eye movement type see Table 4.4.

The second annotator (the author of this thesis) then performed the final pass over all the gaze files. The second labeler could freely modify the gaze event intervals wherever it was deemed necessary. The labels yielded by this labeler we consider as final.

4.2.3 Inter-rater agreement

We here report the agreement between the novice and expert annotators similarly to Section 4.1.3. We can see that the agreement between the two annotators is very high, with the final annotator making only minor changes. These agreement levels are roughly the same as those between the first and final annotator of the GazeCom data set.

Table 4.3: Agreement between the first and final annotations corresponding to the novice and expert annotators. Values correspond to sample-level F1 scores.

EM type	1 vs. final
Fixation	0.943
Saccade	0.911
SP	0.890

4.2.4 Hand labeling statistics

The hand-labeling process required approximately 230 hours of labor for labeling 130 minutes of gaze data. The time was roughly split into 170 hours for the novice labeler and 60 for the expert. Based on these durations each second of gaze recording required 106 s of human time to label. This value is higher than the number reported in GazeCom (Section 4.1.4: 15 to 18 s) even when we consider all five labeling passes. This discrepancy can be attributed to two main factors. First, for the pre-labeling of Hollywood2 we used a much simpler algorithm and the annotator had to change more samples (Figure 4.4) in comparison to GazeCom (Figure 4.1). Second, in Hollywood movies camera motion is very common and can follow very complex patterns (combination of panning and zooming), which makes the distinction between fixations and SP more challenging than in the GazeCom data set where camera motion is very limited.

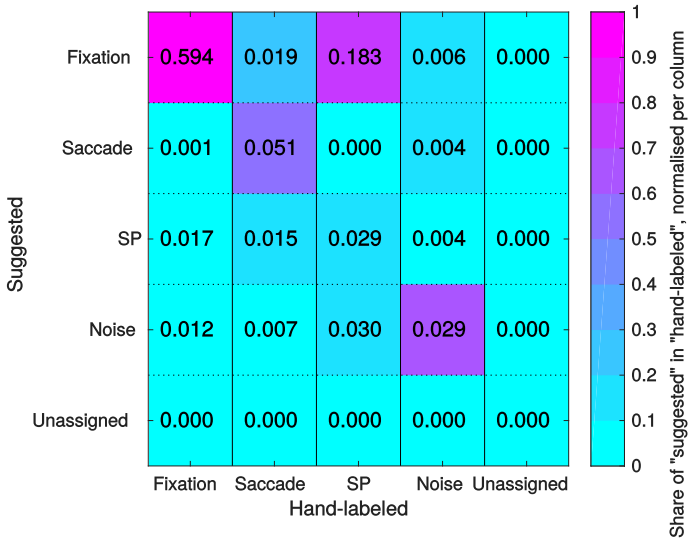


Figure 4.4: Confusion matrix for the pre-labeled and manually annotated eye movement samples. Rows correspond to the suggested eye movement labels, columns – to the final hand-labeled classes. The color bar on the right does not match the numbers in the cells because cell color reflects the share of samples in the final hand-labeling that were originally pre-labeled as the respective suggested classes (i.e. per-column normalization is employed; cf. the color bar on the right).

Table 4.4: The overall percentage of gaze samples and number of episodes of all eye movement types in the algorithmically suggested ("pre-labeled") labels and the final set of labels produced in our annotation procedure.

EM type	Suggested label		Final expert label	
	Share	Episodes	Share	Episodes
Fixation	80.2%	96,643	62.4%	14,643
Saccade	5.6%	19,807	9.1%	15,082
SP	6.4%	99,460	24.2%	5649
Noise	7.7%	21,708	4.3%	1045
Unassigned	0%	0	0%	0

In Table 4.4 we provide the percentages of the suggested and final labels, which is equivalent to the summation of Figure 4.4 row and column-wise. As can be seen the annotators had to change the provided suggestions substantially and especially for SP, which increased from 6.4 % to 24.2 %. Also we can observe that the I-VVT algorithm is very prone to return very short intervals as it is evidenced by the second column of the table. These results are in stark contrast with Table 4.2, where a more elaborate suggestion algorithm was used.

4.2.5 Basic statistics

Overall, the hand-labeled subset of the Hollywood2 data set contains 14,643 fixations, 15,082 saccades, 5649 SP episodes. Here, the number of SP episodes is lower than the other two eye movement classes, as in GazeCom, but here its overall percentage is much higher than in Gazecom (24.2 % vs. 11 %). Based on the negligible difference in the SP share between the "active" and the "free viewing" groups in the current data set (24% vs. 24.3%), the difference in the amount of SP between the GazeCom and Hollywood2 data sets likely originates from the different stimuli types (Hollywood movie clips vs. naturalistic videos), and not from the task of the observers (free-viewing vs. action recognition).

In Figure 4.5 we visualize again the distribution of speed, duration, and amplitude for each eye movement class in Hollywood2. These distributions are almost identical to the equivalent distributions of the GazeCom data set (Figure 4.2). The directional deviation of Figure 4.6 also follows a similar as before.

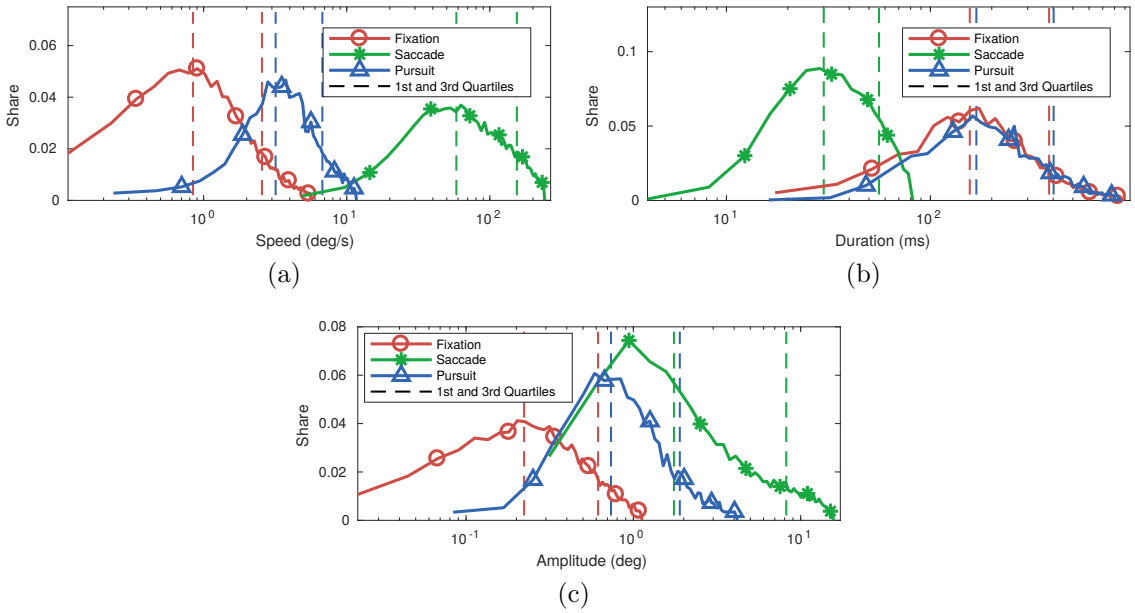
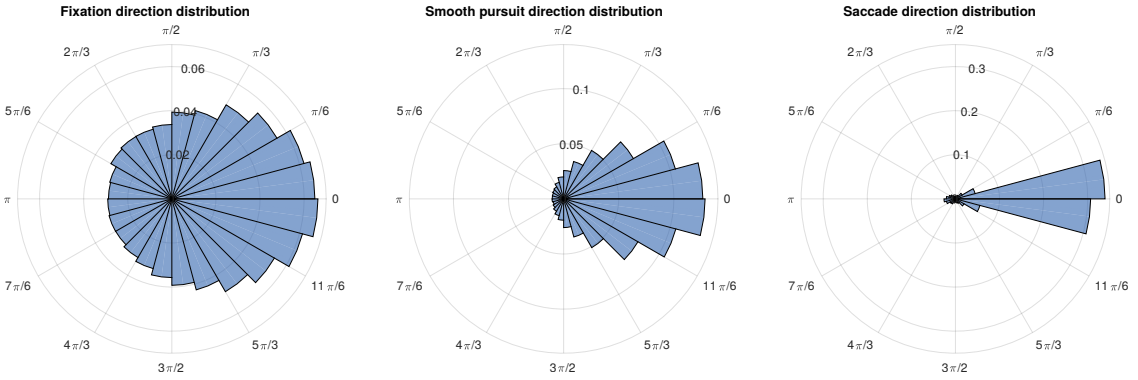


Figure 4.5: Overall per-episode speed, duration, and amplitude distributions for fixations, saccades, and smooth pursuits. These are the (normalized) histograms, which were computed for each eye movement type independently with 50 equal-sized bins covering each respective speed range. These were then plotted here in log-scale (see x-axis), with the y-axis representing the share of episodes in each of the bins. The dashed vertical lines visualize the quartiles (first and third) of the respective distributions. Note that since the horizontal axis is in log-scale, it is difficult to visually compare the areas under different parts of the curves.



(a) (b) (c)

Figure 4.6: Directional deviation distributions for fixations (a), pursuits (b), and saccades (c) presented as circular histograms. To decrease the effect of noisy recordings we compute the directional deviation in 16 ms windows that move one sample at a time. The height of each bar represents the share of the velocity vectors with the given angular deviation from the overall direction of their corresponding episode. Zero deviation angle means perfect alignment with the overall direction of the respective episode.

4.3 360-degree data set

The previous two data sets provided us with ground-truth eye movement labels for a diverse set of stimuli that were displayed on a monitor. Here we move a step closer to full immersion by hand-labeling the eye movements in a free viewing data set that was recorded in an HMD with monocular 360-degree equirectangular videos as stimuli.

4.3.1 Data set collection

Here, we had to gather our own data set because so-far the publicly available 360-degree video-based gaze data sets provide only scanpaths. The scanpaths comprise of a time series where each point represents many gaze samples (usually spanning 100 ms) of the gaze signal. However, gathering a data set of eye-tracking recordings for 360-degree videos differs from the common monitor-based experiments. The experimental set-up, the selection criteria for the used stimuli, as well as the way drifts are handled during recordings are all influenced by the new stimulus type. Our choices and the full data collection procedure are described in more detail below.

Hardware and software

For data gathering we used the FOVE[5] virtual reality headset with an integrated 120 Hz eye tracker. For video presentation we used the integrated media player of SteamVR[6], which supports 360-degree content including the equirectangular video format. A small custom C++ program was used to handle the eye-tracking recordings and store them to disk. The data that we stored for each recording include (i) x and y coordinates of the gaze point on the full 360-degree video surface in equirectangular coordinates, (ii) the same x and y coordinates of the head direction, as well as its tilt. This allowed us to disentangle the eye motion from the head motion (computing the eye-in-head motion) and to reconstruct the gaze position in each participant's field of view. We also stored (as metadata) the dimensions of the headset's field of view (in degrees and in pixels).

In order to make a recording, the experimenter would start the C++ program from the command line with the size of the equirectangular video as parameters. After this the video was loaded in the SteamVR media player and was moved to its first frame after pausing it. All this time our program waits for the Ctrl-A keyboard shortcut to start recording. When it detects the shortcut it emits a play signal to the Steam VR media player and at the same time it starts polling the FOVE SDK for new eye and head data. After the end of the video the same shortcut stops recording.

For the presented videos we kept their original sound. In all clips but two it corresponded to the environment noises (the two exceptions had silence and an overlaid soundtrack). Sound has a bearing on eye movements during monitor-based video viewing [Coutrot et al., 2012], and should affect the viewers even more in virtual environments as noises may induce head rotation towards video regions that would otherwise never be in the field of view.

In our experimental set-up the participants were sitting on a swivel chair with the headset and headphone cables suspended from a hook above the subject's head (Figure 4.7). This allowed the subjects to swivel on the chair freely, without interference from the cords, which could have otherwise led them to avoid head rotation. In addition to the discomfort of feeling the attached cables, unless those are suspended from above, their stiffness would have likely caused the displacement of the headset relative to the observer's head during the experiment, thus lowering the quality of eye-tracking recordings. The experimenter was sitting in front of the participant with

[5] `https://www.getfove.com`
[6] `https://store.steampowered.com/steamvr`

the computer monitor turned towards the experimenter so that the HMD tracking quality and eye camera feeds could be controlled throughout the experiment.

Figure 4.7: Experimental setup for 360-degree gaze recordings.

Stimuli

The assembled video collection includes 14 naturalistic clips from YouTube and one synthetically generated video. All the naturalistic data are licensed under the Creative Commons license[7]. We give attribution to the original creators of the content by providing the YouTube IDs of the original videos together with our data set. The selected clips represent a diverse set of different categories of scene content and context, e.g. static camera, walking, cycling, or driving, as well as such properties as the content representing an indoors or an outdoors scene, the environment being crowded or empty, urban or mostly natural. Representative sample scenes are provided in Figure 4.8. The duration of the complete videos varied greatly, and we decided to use a maximum of one minute per stimulus. For each of these clips, we extracted a continuous part of the original recording that contained no scene cuts to preserve the immersion. The details for each video (name, scene categories, duration) are listed in Table 4.5.

In addition, we generated one stimulus clip synthetically for a more controlled scenario. The circular gaze target we used for this part of the experiment followed the recommendations of [Thaler et al., 2013] in order to improve fixation stability. It

[7] `https://creativecommons.org/licenses/by/3.0/legalcode`

Figure 4.8: Sample scenes taken from our 360-degree data set. They are presented in their original equirectangular format and on the top-left corner of each scene we display the name of the video that it was taken from. The full list of the videos is presented in Table 4.5.

measured two degrees of visual angle in diameter and was displayed in white on a black background. For simplicity, we neglected the idiosyncrasies of the equirectangular format for the stimulus generation here, as the target always stayed close to the equator of the video, meaning that shape distortions would be negligible.

The synthetic clip we generated consisted of five phases. Each phase started with a short instruction set (displayed for ca. 7 s), after which the fixation gaze target appeared. The first four phases lasted 10 s after the stimulus appeared and were designed (together with their respective instructions) to induce (i) eye movements that are typically seen in controlled lab settings: fixations, saccades, and smooth pursuits, all without excessive head motion, (ii) VOR with voluntary head motion while maintaining a fixation on a stationary target, (iii) "natural" long pursuit, without any additional instructions (an arbitrary combination of body or head rotation, VOR, and smooth pursuit), where the target moved with a constant speed of 15°, covering 150°, and (iv) a special combination of VOR and smooth pursuit, when the eyes are relatively stationary inside the head, but the gaze keeps track of a moving target. We refer to the latter type of eye-head coordination as "*head*

Table 4.5: Used Video Stimuli

Video Name	Categories	Duration
01_park	static camera, nature, empty	1:00
02_festival	static camera, urban, busy	1:00
03_drone	drone flight, urban, very high	1:00
04_turtle_rescue	static camera, nature, busy	0:38
05_cycling	cycling, urban, busy	1:00
06_forest	walking, nature, empty	1:00
07_football	static camera, nature, busy	1:00
08_courtyard	static camera, urban, busy	1:00
09_expo	static camera, indoors, busy	1:00
10_eiffel_tower	static camera, urban, busy	0:57
11_chicago	walking, urban, busy	1:00
12_driving	car driving, urban, busy	1:00
13_drone_low	drone flight, urban, empty	1:00
14_cats	static camera, urban, busy	0:43
15_synthetic	moving dot	1:25

pursuit". During the fifth phase, OKN was induced by targets rapidly moving for a short period of time (at 50 degrees symmetrically around the center of the video), disappearing, and then repeating the motion, covering 25° on each pass. Both left-to-right and right-to-left moving targets were displayed with a 2.5 s pause between the sequences of same-direction target movement (5 s each).

Experimental procedure

In order to be able to detect and potentially compensate for eye tracking quality degradation, we added a stationary fixation target at the beginning (for 2 s) and the end (for 5 s) of each video clip. Overall, the 15 videos had a cumulative duration of ca. 17 minutes including these fixation targets. The recording process was split into three sessions for each participant. During the first and the second sessions, 7 naturalistic videos were presented in succession. The last session only included the synthetic video. The participants could have an arbitrary-length break between the sessions. The eye tracker was calibrated through the headset's built-in routine shortly before every recording session. We then empirically and informally validated the calibration using the FOVE sample Unity project[8] where the participant's gaze is visualized. If the quality was deemed insufficient, the calibration procedure was repeated. We accounted for eye-tracking drifts between recordings of the same session

[8] `https://github.com/FoveHMD/FoveUnitySample`

by performing a one-point re-calibration with the fixation target at the beginning of each video.

The naturalistic videos were presented in a pseudo-random order (same for all subjects); the synthetic clip was presented last not to prompt the observers to think about the way they moved their eyes before it was necessary. If the participant at any point was feeling unwell, the recording was interrupted. Afterwards, a new calibration was performed, the unfinished video was skipped, and the recording procedure was continued from the next clip.

Overall, we recorded gaze data of 13 subjects (10 m / 3 f; 27.2 ± 4.6 y; for optical correction status see the data set). The number of recordings per stimulus video clip was between 11 and 13 (12.3 on average) because two of the participants felt unwell at different points during the experiment. Overall, our recordings amount to ca. 3.5 h of eye-tracking data in total.

4.3.2 Manual annotation

When working with 360-degree equirectangular videos, the natural visualization of the recording space is the camera (or the observer's head) placed at the center of a sphere that is covered by the video frame pixels. Computationally, this directly matches the equirectangular video representation, where the x and y coordinates on the video surface are linearly mapped to the spherical coordinates of this sphere (longitude and latitude, respectively). Since the field of view is limited (up to 100° in our HMD), the observers will use head rotation (as in everyday life) to explore their surroundings, so this aspect of the viewing behavior needs to be accounted for both in the definitions of the eye movements and the annotation procedure.

Definitions

In order to fully describe the interplay of the movement of the head and the eyes, we cannot assign just a single eye movement label to every gaze sample, since the underlying process may differ when eye-head coordination is involved. Therefore, we used two labels for each gaze sample, which we refer as *primary* and *secondary* labels. Following the recommendations of [Hessels et al., 2018], we defined the eye movements that we annotated in Section 2.2.1 to avoid potential confusion in terminology. We did not include post-saccadic oscillations or microsaccades in our

annotations as the headset's eye tracker frequency and precision did not permit their confident localization by the annotator.

The *primary* label was necessarily assigned to *all* gaze samples and characterized fixations, saccades, SP, and noise.

The *secondary* label was *not* assigned to all the gaze samples and was used to describe in more detail how the primary eye movements were executed and were mostly a consequence of head motion (except for OKN). The secondary label could take one of the following values: vestibulo-ocular reflex (VOR), optokinetic nystagmus (OKN) or nystagmus, VOR + OKN, head pursuit, unassigned.

Labeling procedure

To thoroughly describe the labeling process, we focus primarily on the information that was available to the manual annotator during this process. At first, we used a two-stage annotation pipeline, with stages corresponding to different frames of reference (for the visualized gaze speed and coordinates), sets of assigned labels, and projections used for the scene content display. We refer to these stages (or modes of operation) as *field of view* and *eye+head*.

In the *field of view (FOV)* mode, the annotator is presented with the view of the scene that is defined by the corresponding head rotation of the subject (the size of the visualized video patch roughly corresponds to the field of view that the participant had in the VR headset). This view corresponds to the frame of reference that moves together with the participant's head and allows us to see the actual visual stimulus that was perceived by the participant and to analyze the eye-within-head gaze behavior.

In the *eye+head (E+H)* mode, the full equirectangular video frame is presented to the annotator. Visualizing gaze locations in this view enables the annotator to see the combination of the head and eye movement, which corresponds to the overall gaze in the frame of reference of the world (or the 360-degree camera, to be more precise).

In both operation modes, the currently considered gaze sample as well as previous and future gaze locations (up to 100 ms) are overlaid onto the displayed video surface. In addition, the plots of the x and y gaze coordinates over time, as well as the plot of both the eye and the head speeds are presented (see Figure 3.2a and 3.2b for the FOV and E+H mode examples). The coordinate systems used for these plots,

however, differ between the two modes: In the FOV mode, the gaze coordinates and the speed of gaze are reported in the *head*-centered coordinate system, whereas in the E+H mode, the coordinates and the speed in the *world* coordinate system are visualized. Figure 2.2 visualizes and summarizes these two coordinates systems with the only difference in this data set being that their start coincides. This way, the FOV representation provides the annotator with the eye motion information within the eye socket, while the E+H representation is responsible for highlighting the absolute movement of the foveated objects, which is necessary for determining the precise label type, e.g. distinguishing between fixations and pursuits.

The manual annotator began (i.e. *the first stage*) with the FOV operation mode and assigned all primary eye movement labels without taking head motion into account: Ballistic eye-in-head motion would correspond to saccades, relatively stationary (in the coordinate system of the head) gaze direction – to fixations, smoothly shifting gaze position – to pursuits (provided that a correspondingly moving target existed in the scene), etc. To speed up the process, we pre-labeled saccades with the I-VT algorithm of [Salvucci and Goldberg, 2000], applied in the FOV coordinates (instead of the coordinates of the full equirectangular video) with a speed threshold of 140°. The labeler then went through each recording, correcting saccade limits or inserting missed ones, assigning fixation, SP, and noise labels, inserting new events where necessary. OKN was labeled in this stage as well because the sawtooth pattern of the eye coordinates was more visible without the head motion effects.

After the annotator felt confident about the first labeling stage results, *the second stage* would begin: The annotator went through the video again, this time – in the E+H operation mode. On the second pass, the previously assigned primary labels were visible and needed to be re-examined in the context of the eye-head coordination, with respective additions of the secondary labels:

- *SP to fixation*: If the primary SP label of the first stage corresponded to the foveation of a stationary (in world coordinates) target, the label was changed to fixation, and a matching VOR episode was added to the secondary labels. If the SP episode in question belonged to an OKN episode, the respective part of the latter was re-assigned to the OKN+VOR class.
- *Fixation + head pursuit*: If the primary fixation label of the first stage (i.e. little to no movement of the eye within its socket) corresponded to following a moving (in world coordinates) target, the secondary "head pursuit" label was added.
- If the primary SP label was maintained in the second stage in the presence of head motion, a VOR episode was added to the secondary labels.

The annotation was performed by the author of this thesis, who first annotated five minutes of pilot data in order to familiarize himself with the procedure and the interface. Labeling a single recording (of about a minute of gaze data) took between 45 min and 1 h. In total, our annotations cover about 16% of the data (two recordings per stimulus clip) and amount to ca. 33 min.

4.3.3 Basic statistics

Overall, the hand-labeled part contains 33 min of manually annotated data that are split into a training and a test set in order to enable easier development and evaluation of automatic labeling algorithms. Each set consists of 15 recordings with one hand-labeled recording for each of the data set videos (for all the included videos see Table 4.5). To further make the two sets more separable the recordings in each one of these come from unique non-shared subjects (subjects 2 to 9 in the training set; subjects 10 to 14 in the test set). In the collected data set (train and test sets together) 75.2% of the samples are labeled as fixations (4035 events), 10.4% as saccades (3837 events), 9.8% as SP (518 events), and 4.6% as noise (524 events) for the primary eye movement classes. These percentages are very close to the percentages of the GazeCom data set (Section 4.1.5), which is a free viewing data set too, and demonstrate that SP is a vital eye movement for the comprehension of our environment even when the head is allowed to move freely. The secondary eye movement labels include 27.6% VOR (1728 events), 15.8% of a combination of OKN+VOR (286 events), 0.8% OKN without VOR (19 events), and 1.5% head pursuit (52 events) with the rest being left unassigned. It is worth mentioning that the head pursuits did not occur only during the synthetic clip viewing, where in one of its tasks the participants were instructed to follow a target with their head while trying to hold their eyes steady, but also during the free viewing of naturalistic stimuli. Also 48 % of the time the head was moving with more than 10°/s and due to its almost continuous movement we observe the high percentage of VOR, which counteracts the head movement and stabilizes the point of regard.

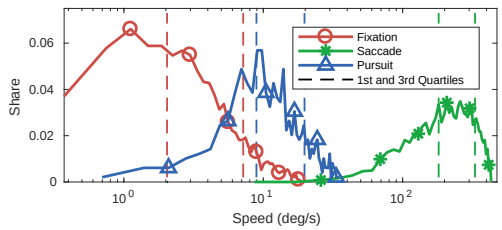
4.3.4 Discussion

In Figure 4.9 we visualize the speed, duration, and amplitude distributions for the primary eye movement types as they appear in the E+H mode, which is equivalent to their projection in the world coordinate systems. The E+H mode representation enables us to visualize the eye movements without the influence of the head motion,

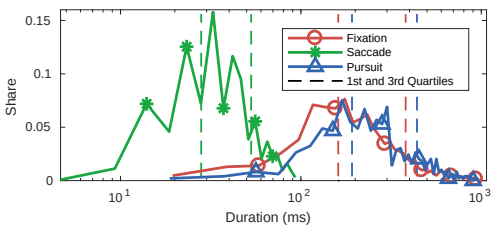
which is not possible in the FOV mode. Here, the relationship among the eye movements is no different than their relationship in the other two monitor-based data sets (Figures 4.2 and 4.5), bare for the noisier plot lines that appear sometimes due to the lower event count. But differences start to appear when we compare a specific eye movement type across the three data sets.

The amplitude and speed of saccades, which are linearly dependent on one another with the so-called "main sequence" [Bahill et al., 1975], are much higher in this data set than in the other two. From the vertical green lines in Figure 4.9a we can see the first speed quartile for this data set has roughly the same value as the third quartile in Figures 4.2a and 4.5a. Similarly, here the saccade amplitudes are substantially larger than in the other two data sets with their durations being comparable among the three. These differences can be partially attributed to the much wider field of view in the HMD (roughly double), which allows for bigger gaze shifts. Also, smooth pursuits exhibit on average higher speeds and amplitudes in the current data set too. But this difference most probably can be attributed to the video content and not to the different experimental conditions because our SP definition requires a moving target to be followed. For fixations the differences among the three data sets are located in the amplitude and speed plots and in the current data set fixation intervals with mean speed of a couple of degrees per second are not uncommon (first quartile at 2°/s).

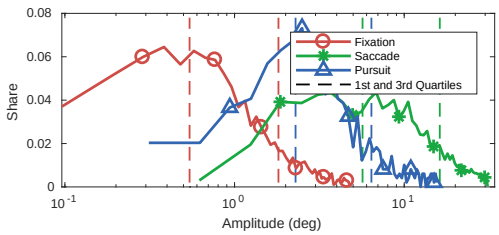
As before in Figure 4.10 we present the deviation of samples in each interval from the overall interval direction. Here the pattern is the same as in Figures 4.3 and 4.6 with the fixation deviation being more spread, the saccade deviation being very concentrated around zero, and SP falling somewhere in between.



(a)

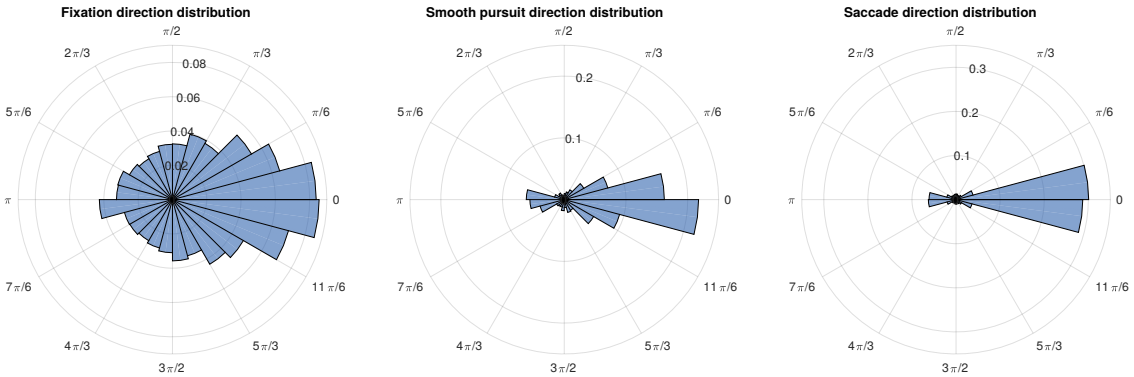


(b)



(c)

Figure 4.9: Overall per-episode speed, duration, and amplitude distributions for fixations, saccades, and smooth pursuits. These are the (normalized) histograms, which were computed for each eye movement type independently with 50 equal-sized bins covering each respective speed range. These were then plotted here in log-scale (see x-axis), with the y-axis representing the share of episodes in each of the bins. The dashed vertical lines visualize the quartiles (first and third) of the respective distributions. Note that since the horizontal axis is in log-scale, it is difficult to visually compare the areas under different parts of the curves. Additionally the presented amplitudes were computed in the 3D space based on the equations of Section 6.1.2.



(a) (b) (c)

Figure 4.10: Directional deviation distributions for fixations (a), pursuits (b), and saccades (c) presented as circular histograms. To decrease the effect of noisy recordings we compute the directional deviation in 16 ms windows that move one sample at a time. The height of each bar represents the share of the velocity vectors with the given angular deviation from the overall direction of their corresponding episode. Zero deviation angle means perfect alignment with the overall direction of the respective episode.

Part III

Improving automated gaze trace segmentation in unstructured environments

In this part of the thesis we move from the tedious process of manually annotating eye movements to algorithmically detecting them. In Chapter 5 we propose two new algorithms that achieve state-of-the-art performance for monitor-based experiments. The first algorithms detects smooth pursuit by using a clustering algorithm and was published in [Agtzidis et al., 2016b] with its open-source implementation developed by Mikhail Startsev [Startsev et al., 2019b]. The second algorithm uses a deep network architecture for the detection of fixations and saccades along with smooth pursuit and was published in [Startsev et al., 2019a]. The development of the network architecture was the work of Mikhail Startsev.

Then Chapter 6 moves towards automatic eye movement detection in 360-degree content. This chapter contains techniques that allow for the conversion of pre-existing algorithms in the 360-degree domain together with techniques that allow for the conversion of the gaze data instead of the algorithms. These techniques have been published in [Agtzidis and Dorr, 2019]. Then we introduce a new algorithm that was developed specifically for content that allows free head motion [Agtzidis et al., 2019].

Chapter 5

Eye movement segmentation in monitor-based experiments

5.1 Smooth pursuit detection based on multiple observers

As people explore their surroundings they tend to attend to the same areas. This observation allows predicting the gaze behavior of future participants from previously recorded gaze patterns through computational saliency models both in static [Itti and Koch, 2000, Kienzle et al., 2009, Kümmerer et al., 2016] and dynamic stimuli [Zhong et al., 2013, Startsev and Dorr, 2018]. Also this shared behavior can be influenced either by a given task [Yarbus, 1967, Rothkopf et al., 2007] or the scene's content [Võ et al., 2019, Healey and Enns, 2011].

Our algorithm takes advantage of the similar behavior among multiple observers when they watch the same scene, which is common in eye-tracking experiments, in order to improve the detection of SP. In brief it first removes fixations and saccades from the gaze recordings with the remaining samples being considered as SP candidates. Then the SP candidates across all observers are handled together and a portion of these is marked as actual SP based on a clustering method that works as proxy to gaze trace similarity. We will refer to this algorithm as "sp_tool", which is the name of its publicly available implementation[1].

[1] `https://github.com/MikhailStartsev/sp_tool`

5.1.1 Prefiltering

The separation of fixations and saccades from other eye movements can reach high quality levels, as will be demonstrated in Section 5.3, even with relatively simple algorithms. The easier separability arises from their well separated basic statistics that have already been presented in Sections 4.1.5, 4.2.5, and 4.3.3. Therefore, in the prefiltering step we remove confidently detected saccades, blinks, and fixations. The prefiltering starts by detecting saccades with the detector of [Dorr et al., 2010] that utilizes two speed thresholds and returns high quality results. Then we process intervals of tracking loss and expand them to include saccades that start or end within 25 ms of their border. Then the expanded intervals (including the saccades) are marked as blinks. Fixations are processed in the intersaccadic space. If the gaze shift within the intersaccadic interval is below 1.41° the whole interval is marked as fixation. If not, a sliding with a duration of 100 ms is applied and fixations are detected when the mean gaze speed falls below 2°/s.

5.1.2 Clustering

After the pre-filtering step, all the remaining gaze samples from all the observers are pooled together and are considered as *pursuit candidates*. Our algorithm processes these candidates and creates clusters with a variation of the DBSCAN clustering algorithm [Ester et al., 1996]. It is worth mentioning that this step differs from the pre-filtering step, which processes the gaze trace of each observer independently.

The algorithm finds clusters of a predefined density and does not assume any shape for the cluster (e.g. Gaussian mixture models) nor searches for a representative point in the form of a centroid (e.g. k-means [MacQueen, 1967]). This particular attribute is very appealing for our use case because pursued targets in dynamic natural scenes can move at arbitrary directions with varying speeds.

More specifically DBSCAN divides the pursuit candidates into three categories based on the number of neighbors that each point has within a user defined distance (parameter ϵ). (i) Core points are the points that have at least a certain number of neighbors (parameter $minPts$, user defined) within their neighborhood. (ii) Border points are the points that are not core points but have at least one core point as neighbor. (iii) Outliers are the points that do not fulfill any of the previous criteria and are marked as noise by our algorithm. The core and border points are considered as pursuit samples.

An intricate point in applying DBSCAN in the eye movement domain is finding a way to compare distances in time and space. Our solution modifies the original algorithm and uses two distance thresholds for the neighborhood definition. The x and y parameters are grouped together and a threshold $\epsilon_{xy} = 4°$ of visual angle is used. Time t is considered independently with a threshold $\epsilon_t = 80\,ms$. The net effect of these two parameters is a neighborhood that has the shape of a cylinder with its axis spanning the time domain. The number of neighbors threshold $minPts$ was set to 160. All the parameter values presented were optimized with the GazeCom data set (Section 5.3.2).

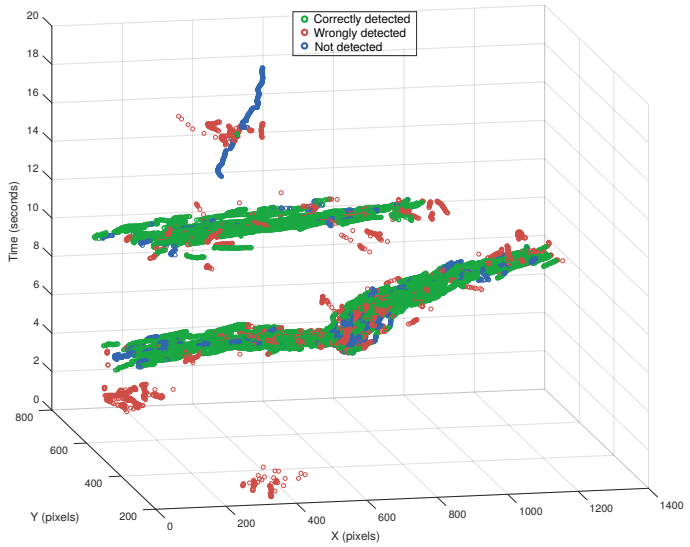


Figure 5.1: Visualization of clustering-based pursuit classification in one video of our data set (*ducks_boat*). Data points for all observers are presented. Correctly detected smooth pursuit samples (in green) as well as detection errors (in red – false detections, in blue – missed samples) of our SP detection algorithm in the sp_tool framework.

Figure 5.1 visualizes the SP detection results of our algorithm in the *ducks_boat* clip of the GazeCom data set. In this clip two ducks are flying across a lake as well as across the camera field of view and appear in the gaze recordings as the two big (predominantly green) clusters in the figure. Here, the green points represent SP gaze samples that were detected by our algorithm (true positives) while the red points represent other eye movements labeled as SP (false positives) and the blue points represent missed SP samples (false negatives). We can see that most of the pursuit intervals were correctly detected by our algorithm although some weaknesses appear. Samples at the edge of the big clusters are missed due to a drop in cluster density. Also the two small clusters at the bottom of the figure were wrongly labeled

as SP because a dense group of non-SP samples passed through the prefiltering step. The elongated blue line at the top of the figure represents a single observer following a target who was missed by our algorithm.

5.2 Deep learning eye movement segmentation

The basic eye movement characteristics of the hand-labeled data sets have already demonstrated that high quality separation of all three eye movement types with simple thresholds is not feasible. Because of this, a deep learning approach would be able to learn the more complex relationships among the different gaze features that each eye movement type contains. This algorithm utilizes a series of 1D convolutional layers followed by a one layer bidirectional long short-term memory network (see Figure 5.2). Another defining point of this algorithm is the classification of bigger gaze intervals instead of single samples [Hoppe and Bulling, 2016, Zemblys et al., 2018a].

Features for classification

Because of the small size of the used architecture the network could not learn the complex relationships between the raw x and y coordinates and therefore we provided it with additional pre-computed features. These features included speed, acceleration, and direction of gaze on five different temporal scales (4, 8, 16, 32, and 64 ms) in order to make them noise robust and let the network use the most efficient combination of them. The use of speed as a classification feature has been used extensively in other algorithms [Sauter et al., 1991, Salvucci and Goldberg, 2000, Komogortsev et al., 2010, Komogortsev and Karpov, 2013] and its usefulness was demonstrated with the basic characteristics of our data sets. Acceleration has also been used in literature [Collewijn and Tamminga, 1984, Nyström and Holmqvist, 2010, Behrens et al., 2010, Larsson et al., 2013] contrary to the gaze direction. Gaze direction can aid in eye movement classification because SP is performed mostly on the horizontal plane while drifts predominantly on the vertical plane [Ko et al., 2016].

Additionally we conducted experiments with different feature combinations in order to further enhance performance.

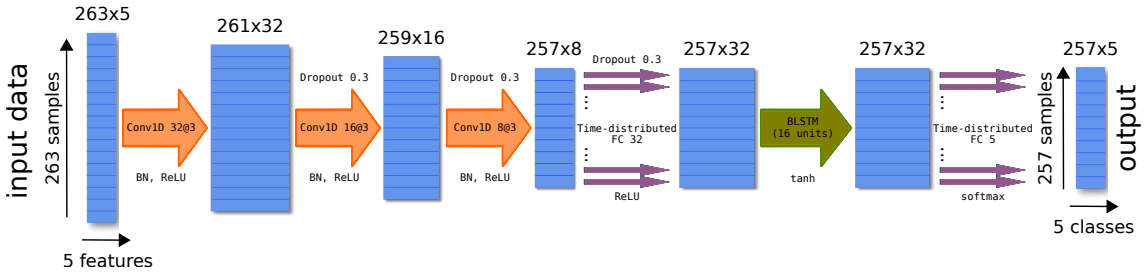


Figure 5.2: The architecture of our 1D CNN-BLSTM network. Courtesy of [Startsev et al., 2019a].

Network architecture

For the eye movement classification network a simple architecture consisting of one-dimensional temporal convolution layers together with a Bidirectional LSTM was used. The network is able to label fixations, saccades, SP, noise, and "unknown". Since we are labeling 1D time sequence the parameter count is relatively low (ca. 100000) in comparison with other deep architectures that are used for processing images or videos. As can be seen from Figure 5.2 the network contains three convolutional layers followed by a fully-connected layer. No pooling is used because the network labels chunks of the gaze signal that contain many samples. Then a fully connected layer follows before a final BLSTM layer.

5.3 Algorithm evaluation

In this section we present the evaluation results of the two proposed algorithms along with 12 literature algorithms. All the algorithms were compared against the two monitor-based hand-labeled data sets of Chapter 4. Our algorithms together with three literature algorithms were optimized for the GazeCom data set and were then tested in the Hollywood2 data set in order to obtain more objective performance estimations. Here we start with the presentation of the literature algorithms, followed by the optimization procedure, and finally with the evaluation results.

5.3.1 Literature algorithms

Several eye movement classification algorithms have been proposed throughout the years with varying degrees of complexity. The majority of these do not provide a publicly available implementation that would enable easier evaluation and comparison with newly proposed algorithms. But some algorithms have a publicly available

implementation that is provided either by the original authors or other researchers and here we are going to present 12 of these.

Eight of the evaluated algorithms are included in the publicly available and open-source toolbox of [Komogortsev, 2014]. The toolbox is implemented in Matlab and contains five algorithms that detect fixations and saccades only (I-VT, I-DT, I-HMM, I-KF, I-MST) with rest detecting SP too (I-VVT, I-VDT, I-VMP). This toolbox along with other implementations use the term *velocity* when they actually use the length of the velocity vector. Therefore in this thesis we refrain from using the term *velocity* and we use the more accurate *speed*. I-VT and I-DT [Salvucci and Goldberg, 2000] are the simplest algorithms in the toolbox and they use a simple speed or dispersion threshold for eye movement classification. Saccades are detected when the measured characteristic is above threshold with the rest of the samples labeled as fixations. I-HMM [Salvucci and Anderson, 1998] uses the transition probabilities among the different eye movements via a Hidden Markov Model to improve upon the I-VT. I-KF [Sauter et al., 1991, Komogortsev et al., 2010] uses a single threshold but the used statistics are more elaborate. Initially the gaze speed is predicted with a Kalman filter, which is then compared to the actual eye speed through a Chi-square test. A threshold on this test is used for the separation of fixations from saccades. I-MST [Goldberg and Schryver, 1995] requires the user to provide the maximum saccade duration, which is then used for the creation of a spanning tree. Based on the point-to-point distances fixations are separated from saccades based on a simple threshold.

The rest of the algorithms in the toolbox were developed for dynamic stimuli and can also detect smooth pursuit. The I-VVT algorithm is an extension of the I-VT algorithm and uses two speed thresholds instead of one. The samples above the high speed threshold are marked as saccades, the samples between the thresholds as SP, and those below the low threshold as fixations. I-VDT [San Agustin, 2010] is a combination of the I-VVT and I-DT algorithms and it replaces the low speed threshold of the earlier with the dispersion threshold of the latter. I-VMP [Komogortsev and Karpov, 2013] detects saccades with a high speed threshold similar to the previous two algorithms. Then a window-based approach is used for the separation of fixations from SP with a threshold that uses motion characteristics such as the magnitude and direction of movement.

[Dorr et al., 2010] provides two algorithms for fixation and saccade detection in dynamic contexts. For saccade detection it employs two speed thresholds. The high speed threshold is used for the initiation of a saccade, which is then expanded forwards and backwards in time until the low speed threshold is reached. For the

fixation detection it takes the dynamic nature of the content into account and tries to avoid mistakenly label SP as part of it. Fixations are detected with a sliding window that combines a dispersion and a speed threshold. When the requirements are not fulfilled the samples are left unassigned because there existed too much motion in the gaze signal. In our evaluation further down we consider these samples as SP.

The toolbox of [Walther and Koch, 2006] contains the implementation of the algorithms by [Berg et al., 2009], which were designed for dynamic stimuli. It initially low-pass filters the gaze signal and then it computes its principal components (PCA) in different temporal scales. The combination of the ratio of the principal axes together with the gaze speed are used for the separation of saccades from SP. The rest of the samples are finally labeled as fixations.

[Larsson et al., 2015] proposed an algorithm that works in the intersaccadic space and separates fixations from saccades. A publicly available re-implementation of it was developed as part of this thesis[2] and in order to provide a complete toolbox the saccade detector of [Dorr et al., 2010] was included in it. In the intersaccadic space a sliding window is used for the classification of fixations and SP based on four criteria. Samples are labeled as fixations when none of the criteria are fulfilled and SP when all the criteria are fulfilled. In cases where one to three criteria are fulfilled the window is labeled based on its similarity with its neighboring already labeled windows.

[Dar et al., 2019] proposed the REMoDNaV algorithm, which is an elaborate speed-based detection algorithm and an extension of [Nyström and Holmqvist, 2010]. It uses a multistep preprocessing pipeline with the aim of suppressing the recording noise. Then the gaze trace is split into chunks that are separated by periods of high gaze speed. These periods of high gaze speed together with detected saccades as returned from the algorithm of [Nyström and Holmqvist, 2010] are marked as saccades. At the end the intersaccadic intervals are low passed filtered and again an adaptive speed threshold is utilized for the separation of fixations from SP.

5.3.2 Algorithm Optimization

Out of the 14 evaluated algorithms we optimized the two proposed algorithms of this thesis together with the I-VDT, I-VVT, and I-VMP as provided from the toolbox of [Komogortsev, 2014]. All the algorithms were optimized with the ground-truth

[2] `https://www.michaeldorr.de/smoothpursuit/larsson_reimplementation.zip`

labels of the GazeCom data set and never saw the labels Hollywood2, which acted as the test set.

For the clustering algorithm we randomly sampled the multi-dimensional parameter space of our fixation and pursuit detectors, which enabled us to understand the performance range of our detector and pick the values that returned the best average F1 score by considering both the sample- and event-level F1 scores across all eye movements types (fixations, saccades, SP). For more detailed information about the used metrics refer to Section 2.5. The optimal parameters were presented together with the algorithm in Section 5.1.

For the deep learning algorithm we ran experiments with different feature sets and context sizes in order to identify the best architecture and most informative features. The optimization was performed on the GazeCom data set, from which we sampled 50,000 windows with replacement after splitting it into training (90 %) and validation (10 %) sets. The model that achieved the highest average F1 score had a context size of 257 samples and used only the speed and direction features.

The I-VDT, I-VVT, and I-VMP were optimized through a grid search for all of their parameters. Overall, the best parameter set for I-VDT was $80°/s$ for the speed threshold and 0.7° for the dispersion threshold. For I-VVT, the low speed threshold of $80°/s$ and the high threshold of $90°/s$ were chosen. For I-VMP, the high speed threshold parameter was fixed to the same value as in the best parameter combination of I-VVT ($90°/s$), and the best parameters for the window duration and the "magnitude of motion" thresholds were 400 ms and 0.6 respectively.

5.3.3 Results

Evaluation on GazeCom

In Table 5.1 we display the scores of the 14 evaluated algorithms in descending average F1 score order. From this table some interesting observations can be made. Our algorithms occupy the first two spots with the 1D CNN-BLSTM algorithm achieving the highest scores for five out of the six metrics. As expected the older and simpler eye movement detection algorithms that do not detect SP occupy the last spots. One exception is the I-VVT algorithm, which was designed to detect SP, but during the optimization process attained the highest average score when it did not detect it. This is evident by the almost zero sample- and event-level F1 scores

but also from its two optimal "velocity" thresholds that are very close together at $80°/s$ and $90°/s$ respectively.

Table 5.1: GazeCom evaluation results as F1 scores for *sample-level* and *episode-level* detection (sorted by the average of all columns). Table adapted from [Startsev et al., 2019a].

		Sample-level F1			Event-level F1		
Model	average F1	Fixation	Saccade	SP	Fixation	Saccade	SP
1D CNN-BLSTM^{+**} (ours)	**0.830**	**0.939**	**0.893**	**0.703**	0.898	**0.947**	**0.596**
sp_tool** (ours)	0.769	0.886	0.864	0.646	0.810	0.884	0.527
[Larsson et al., 2015]	0.730	0.912	0.861	0.459	0.873	0.884	0.392
I-VMP**	0.718	0.909	0.680	0.581	0.792	0.815	0.531
[Berg et al., 2009]	0.695	0.883	0.697	0.422	0.886	0.856	0.424
REMoDNaV	0.690	0.823	0.692	0.480	0.858	0.898	0.391
[Dorr et al., 2010]	0.680	0.919	0.829	0.381*	**0.902**	0.854	0.193*
I-VDT**	0.606	0.882	0.676	0.321	0.823	0.781	0.152
I-KF	0.563	0.892	0.736	–	0.877	0.876	–
I-HMM	0.546	0.891	0.712	–	0.817	0.857	–
I-VVT**	0.531	0.890	0.686	0.000	0.778	0.816	0.013
I-VT	0.528	0.891	0.705	–	0.761	0.810	–
I-MST	0.497	0.875	0.570	–	0.767	0.773	–
I-DT	0.480	0.877	0.478	–	0.759	0.765	–

CNN-BLSTM results marked with $^{+}$ are for context window size of with 1 s (257 samples) and speed and direction features. The * signs mark the numbers where the label was assumed from context and not actually assigned by the algorithm. Performance estimates for models marked with ** can be potentially optimistic because they were either trained or optimized for this data set. In each column, the highest value is **boldified**.

The newer algorithms that can detect SP occupy the higher places in the table. On average all of them can detect fixations very well and can achieve almost human-level agreement with the final hand-labeler based on the comparison between the scores in the current table with the hand-labeler agreement scores of Tables 4.1 and 4.3. Also for saccades they can reach human-level agreement based on the event-level F1 score but they tend to not identify well the beginning and the end of the saccadic intervals, which appears on the table as lower sample-level F1 scores. This discrepancy can potentially arise from the difficulty in defining PSOs [Hooge et al., 2015] and whether they are included in the saccadic intervals or not. Contrary the SP detection performance is substantially lower in comparison to the other two eye movement types both in sample and event-level terms. On this front our algorithms significantly improve the detection of SP in comparison with the state-of-the-art algorithms and the deep learning algorithm starts to approach human-level performance.

However, we should mention that some of the algorithms presented in Table 5.1 (including ours) might have an unfair advantage in comparison to the rest of the algorithms due to their training or optimization with the GazeCom data set (see previous section). In order to get a more objective performance estimate we apply again these algorithms on the previously unseen Hollywood2 data set and the results

are presented below.

Evaluation on Hollywood2

Table 5.2 contains the evaluation results for the Hollywood2 data set. The overall pattern for the eye movement detection quality stays the same as before with the fixations being the best detected followed by saccades and then by SP. The GazeCom optimized algorithms drop in performance with the most notable being the I-VMP algorithm, which drops from 71.8 % to 46.7 % average F1 score and it is a clear example of overfitting. The other previously optimized algorithms achieve a couple of percentage lower average F1 scores with our algorithms' performance decreasing by 4.3 % and 6.6 % respectively. This small drop indicates that they have avoided overfitting and can generalize well in new unseen contexts.

Table 5.2: Hollywood2 evaluation results as F1 scores for *sample-level* and *episode-level* detection (sorted by the average of all columns).

		Sample-level F1			Event-level F1		
Model	average F1	Fixation	Saccade	SP	Fixation	Saccade	SP
1D CNN-BLSTM^{+**} (ours)	**0.787**	**0.872**	**0.827**	**0.680**	0.808	**0.946**	0.588
REMoDNaV	0.748	0.779	0.755	0.622	0.784	0.931	**0.615**
sp_tool** (ours)	0.703	0.819	0.815	0.616	0.587	0.900	0.483
[Dorr et al., 2010]	0.685	0.832	0.796	0.373*	0.821	0.884	0.403*
[Larsson et al., 2015]	0.647	0.796	0.803	0.317	0.807	0.886	0.274
[Berg et al., 2009]	0.601	0.824	0.729	0.137	**0.845**	0.826	0.243
I-VDT**	0.570	0.828	0.665	0.412	0.657	0.601	0.258
I-KF	0.523	0.816	0.770	–	0.748	0.803	–
I-HMM	0.480	0.811	0.720	–	0.646	0.700	–
I-DT	0.473	0.803	0.486	–	0.744	0.802	–
I-VMP**	0.467	0.811	0.672	0.045	0.586	0.624	0.067
I-VVT**	0.448	0.809	0.672	0.002	0.536	0.622	0.050
I-VT	0.432	0.810	0.705	–	0.520	0.555	–
I-MST	0.385	0.793	0.349	–	0.590	0.576	–

CNN-BLSTM results marked with $^{+}$ are for context window size of 257 samples and speed and direction features. The * signs mark the numbers where the label was assumed from context and not actually assigned by the algorithm. Algorithms marked with ** were trained or optimized for the GazeCom data set and tested independently here. In each column, the highest value is **boldified**.

The 1D CNN-BLST algorithm still achieves the top performance in the Hollywood2 data set but the sp_tool algorithm drops in the third position behind REMoDNaV. The drop of the sp_tool in the third position can be attributed to the increased performance of REMoDNaV in this data set, but also to an inherent weakness of our algorithm: the segmentation of long intervals into shorter ones. This last attribute can be observed by its substantially lower fixation and SP event-level scores in comparison to sample-level scores in both Tables 5.1 and 5.2 and from Figure 5.1. In this figure we can observe that many short duration intervals appear

mostly next to the two flying ducks (predominantly green clusters) but also in the wrongly detected cluster (pure red clusters).

Chapter 6

Eye movement detection with 360-degree stimuli

In the previous chapter we presented two new algorithms for eye movement detection and we evaluated them against 12 publicly available literature algorithms, which all together form a small part of the multitude of all the proposed eye movement detection algorithms. The caveat of all the aforementioned algorithms is that they cannot be applied directly to 360-degree equirectangularly projected spaces because they were developed by assuming monitor-based experiments (visualized in Figure 2.2). But this important assumption simplified the overall algorithm design in several aspects, and we are going to explain some of these simplifications here. The video coordinate system and the monitor coordinate system coincide and they are usually placed at the top-left corner of the monitor, as can be seen in Figure 2.2. By having a single frame of reference we avoid difficult to understand conversions between different frames of reference. The area of application is almost uniform with nearly the same pixels per degree density for the whole monitor even though there is a small deviation between its edges and its center. The last and most important factor that makes understanding and visualization easier is the fact that we work in the 2D Cartesian space.

As we move towards the 360-degree equirectangular space things become more complicated and the previous assumptions do not apply anymore. In the new space the experiment coordinate system differs from the video coordinate system. The participant lives inside a sphere surrounded by the video and as can be seen from Figures 6.1a and 6.1b the correspondence between the equirectangular coordinates and the experiment coordinates is not trivial. Also the frame of reference of the experiment can be either fixed or can move together with the participant's head.

Moreover, the 2D equirectangular plane that is used for the data representation does not have uniform pixel to sphere surface density (Table 6.1 and Figure 6.3) and distortions start to appear as we move away from the horizontal midline. However, if we use the 3D Cartesian space of Figure 6.2 the previous limitations can be overcome but understanding and visualization of concepts become more difficult.

Thus, in this chapter we are going to tackle the previous problems by presenting the different frames of reference and the correspondence between the equirectangular and the 3D spaces. Based on these we are going to present algorithms that work directly in the 3D Cartesian space but we will also demonstrate a method that allows using monitor-based algorithms on converted equirectangular data.

6.1 Conversion of monitor-based algorithms

In this section we provide information on how we converted 5 popular algorithms (a subset of the evaluation algorithms of Section 5.3), namely the I-VT, I-DT, the saccade and fixation detectors of [Dorr et al., 2010], and the algorithm of [Larsson et al., 2015], in order to work in the equirectangular space. Because the conversion of pre-existing algorithms is not always possible we also provide a method that converts the data instead of the algorithm, which then can be used as input to the original monitor-based algorithms.

6.1.1 Equirectangular to Cartesian space

Before moving to the actual algorithm conversion we have to understand the connection between the equirectangular frame of reference and the 3D Cartesian frame of reference. As already mentioned, when we display the video in the HMD environment we live inside a sphere surrounded by the video. If we cut the sphere horizontally in the middle and look at it from the top we get Figure 6.1a with the Y axis pointing towards us. The x_{eq} equirectangular coordinate is the equivalent of angle ϕ in spherical coordinates and varies from 0° to 360°. Consequently, the 3D vector $(1, 0, 0)$ is the point where the left and right sides of the video meet; this is usually behind the participant at start time. The vector $(-1, 0, 0)$ represents the center of the video and the initial viewing position.

Now if we cut the previous figure along the Z - Y plane and look at it from the right we get the view of Figure 6.1b with the X axis pointing away from us. Here, the

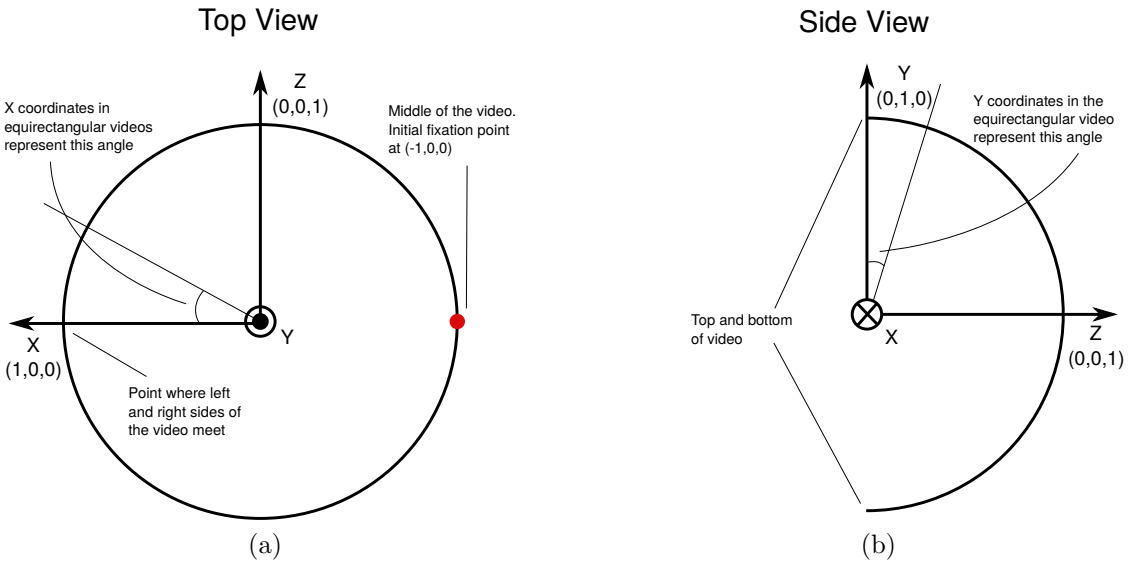


Figure 6.1: **(a)** Top view of 3D Cartesian coordinate system of reference for the HMD experiment with the Y axis pointing towards us. The surrounding circle visualizes the intersection of the sphere of the 360-degree video with the X-Z plane. The 3D Cartesian X axis points to the left with the middle of the video represented by the red dot at $(-1, 0, 0)$. The x_{eq} equirectangular coordinate represents the angle of the projected unit vector on the X-Z plane with the X axis. **(b)** Side view of the 3D Cartesian coordinate system with the X axis pointing away from us. The y_{eq} coordinate represents the angle of a unit vector with the Y axis with the angle ranging from $[0, \pi]$ and is visualized through a half circle. The top and bottom of the equirectangular projection are at the $(0, 1, 0)$ and $(0, -1, 0)$ respectively.

y_{eq} equirectangular coordinate is equivalent to the angle θ in spherical coordinates and varies from 0° to 180°; hence, we have a half circle.

With the definition of x_{eq} and y_{eq} coordinates in place we can now define the conversions between equirectangular, spherical, and 3D Cartesian coordinates in the following manner.

Equirectangular to spherical:

$$\begin{aligned} \phi &= x_{eq} * 2\pi / width_{eq} \\ \theta &= y_{eq} * \pi / height_{eq} \end{aligned} \tag{6.1}$$

Spherical to Cartesian:

$$\begin{aligned} x &= \sin\theta * cos\phi \\ y &= \cos\theta \\ z &= \sin\theta * \sin\phi \end{aligned} \tag{6.2}$$

6.1.2 Application to existing algorithms

For the actual algorithm conversion we can now use the 3D Cartesian unit vectors from equation 6.2 for calculating the angular distance, speed, and dispersion, which are the basic building blocks for all five algorithms that we want to convert. The angular distance and speed are calculated between two unit vectors in the same fashion as in [Diaz et al., 2013, Duchowski et al., 2002].

$$distance(\vec{v}_1, \vec{v}_2) = arccos(\vec{v}_1 \cdot \vec{v}_2) \tag{6.3}$$

$$speed(\vec{v}_1, \vec{v}_2) = \frac{distance(\vec{v}_1, \vec{v}_2)}{t_2 - t_1} \tag{6.4}$$

In the literature the dispersion typically is defined as a function of the size of the bounding box of the sample set defined on the video coordinate system [Salvucci and Goldberg, 2000, Larsson et al., 2015]. In the 360-degree scenario, however, this definition loses its meaning since the two principal axes of the dispersion can be in any position and orientation on the sphere. Therefore we define the dispersion as the angle of the cone that starts from the center of the coordinate system and contains all the data points at its intersection with the sphere. This is visualized in Figure 6.2 and is defined below. A similar approach to ours is followed by the PUPIL headset in its fixation detector [Barz, 2015].

$$\begin{gathered} dispersion(\mathbb{A}) = \{\underset{i,j}{argmax}\{distance(\vec{v}_i, \vec{v}_j)\} \mid i, j \in \mathbb{A}\} \\ \text{where } \mathbb{A} = \{1, 2, ..., n\} \end{gathered} \tag{6.5}$$

By using equation 6.4 we can now directly convert the I-VT algorithm [Komogortsev et al., 2010], which uses a single threshold for velocity (or more specifically, its absolute magnitude, speed) to the 360-degree equirectangular video domain. The original algorithm labels saccades when the gaze speed is above threshold and fixations when it is below threshold. Because we have free head movement often the eyes will perform compensatory eye movements at intermediate speeds. Therefore, in the 360-degree video domain the algorithm is only reliable for saccade detection and all the samples below the velocity threshold are left unassigned.

The saccade detector described in [Dorr et al., 2010] is more elaborate than I-VT

and uses two speed thresholds. A saccade detection is initiated when the speed is above the high threshold and is then expanded forward and backwards in time until the speed reaches the lower threshold. With this approach, it avoids erroneous detection of saccadic samples especially in 360-degree videos where the head can move freely and reach high angular speeds without performing a saccade. To convert this algorithm we again use the definition of speed from equation 6.4.

The I-DT fixation detector from [Salvucci and Goldberg, 2000] calculates the dispersion of gaze samples in windows of fixed length. If the dispersion is below threshold the samples within the window are labeled as part of a fixation otherwise the window moves to the next gaze sample. Here, we use the dispersion definition of equation 6.5 for 360-degree videos as a replacement for the original definition.

Our second fixation detector is described in [Dorr et al., 2010] and uses a combination of dispersion and speed thresholds. Initially it starts from a fixed duration window and checks whether the dispersion and mean speed are below their respective thresholds. If both criteria hold true the window duration is extended forward in time and the dispersion threshold increases logarithmically in relation to the new window duration. The expansion phase stops when one of the conditions becomes false. At this point all the samples within the previously valid window are labeled as fixations. The conversion of this algorithm is straightforward and uses equations 6.4 and 6.5 for speed and dispersion calculation in the new domain.

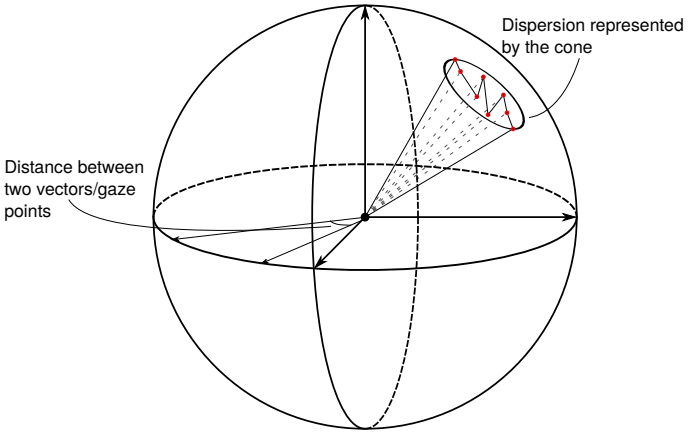


Figure 6.2: Visualization of the 3D Cartesian coordinate system within the HMD presentation sphere. The angle between two vectors represents the distance in the 3D space. The angle of the cone that contains a gaze segment represent the dispersion in the 3D space. The path length is the sum of distances between consecutive gaze points.

The algorithm described in [Larsson et al., 2015] works on windows in intersaccadic intervals and labels fixations and smooth pursuit (SP). The algorithm uses four cri-

teria, namely dispersion, consistent direction, positional displacement and spatial range for the classification of eye movements. When none of the previous criteria are satisfied the window samples are labeled as fixations. Contrary to fixations, the window samples are labeled as SP when all the criteria are fulfilled. For the remaining cases where between one and three criteria are satisfied the algorithm looks at other labeled segments in the same intersaccadic interval to make its decision.

A publicly available implementation of this algorithm [Larsson et al., 2015] was developed as part of this thesis[1] and includes the saccade detector from [Dorr et al., 2010] that we have already described above. For the conversion of this algorithm apart from using equations 6.3 and 6.5 we have to define the sample-to-sample direction, which is used in the preliminary segmentation described in section 2.2 of the original paper [Larsson et al., 2015]. The sample-to-sample direction within a window in the 3D Cartesian space is defined as

$$\begin{aligned} \vec{d_i} &= \vec{v}_{i+1} - \vec{v}_i \\ &\text{where } i = \{1, 2, \ldots, N-1\} \end{aligned} \tag{6.6}$$

with the rest of the calculations for the Rayleigh test staying the same.

In the evaluation of spatial features described in section 2.3 of the original paper [Larsson et al., 2015], the parameter p_D represents the dispersion within each segment and is calculated as the ratio between the explained variance of the second and first principal components. In the 3D space we keep it as is. The parameter p_{CD} is a measure for the consistency of movement direction and is calculated as the ratio between d_{ED} and the length of the first principal axis, where d_{ED} is the distance of the first and last samples of the segment. In the 3D Cartesian space d_{ED} is the angular distance of the first and last vectors of an interval as defined in equation 6.3. In the 3D Cartesian space the length of the first principal axis is equivalent to the maximum dispersion of the segment and is taken from equation 6.5. The positional displacement parameter p_{PD} is the fraction of d_{ED} to d_{TL} with d_{TL} being the gaze path length. The gaze path length is calculated as the sum of consecutive vector distances within a window by using equation 6.3. The parameter p_R is the spatial range, which in the original algorithm is the length of the diagonal of the bounding box for all the sample of a window. In the 3D Cartesian space it is equivalent to the dispersion of the segment and is calculated through equation 6.5.

[1] `https://www.michaeldorr.de/smoothpursuit/larsson_reimplementation.zip`

The last modification to the [Larsson et al., 2015] algorithm comes in section 2.4 of the original paper for the classification of eye movements when between 1-3 criteria are satisfied and criterion 3 is true. In this case we use the mean gaze direction between two segments, which in the 3D Cartesian space is calculated as the angle (equation 6.3) between the mean direction vectors of each segment after vector normalization.

6.1.3 Conversion of data

The approach to convert an algorithm to the 3D Cartesian space as described above may be either deemed too time consuming in comparison to the amount of data that we want to process or impossible due to closed source or very convoluted code. Because of these limitations we describe here a method of how to transform equirectangular gaze recordings to a new space where the original algorithms can be applied without any modification.

Cartesian to equirectangular space

Before presenting the reprojection process we first have to understand how we can move from the 3D Cartesian space back to equirectangular space. First we convert a normal vector $\vec{v} = (x, y, z)$ to its spherical representation with equation 6.7, which is the inverse of equation 6.2.

$$\begin{aligned} & \phi = arctan2(z, x) \\ \text{if}(\phi < 0): \ & \phi = \phi + 2\pi \\ & \theta = arccos(y) \end{aligned} \tag{6.7}$$

The two spherical angles are directly connected to the equirectangular coordinates and are calculated from equation 6.8, which is the inverse of equation 6.1.

$$\begin{aligned} x_{eq} &= width_{eq} * \phi / 2\pi \\ y_{eq} &= height_{eq} * \theta / \pi \end{aligned} \tag{6.8}$$

Gaze data reprojection

As we have already mentioned the equirectangular projection for video and gaze representation comes with the drawback of higher distortions as we move away from the equatorial line (horizontal line passing through the middle of the video). We visualize these distortions by projecting the intersection of a cone with the sphere (Figure 6.2) to the equirectangular representation of Figure 6.3. The angle of the cone is 20° and it is centered at 0°, 22.5°, 45°, 67.5°, and 90° away from the equatorial line. The gray area represents the equirectangularly projected cone with the black circle representing a hypothetical undistorted cone.

Table 6.1: Distortions of equirectangular projection increase exponentially the further away we move from the equator. Distortions are zero at the equator and infinite at the top and bottom of the projection.

Distance from equator	Distortion
0°	0.00
22.5°	0.08
45°	0.41
67.5°	1.61
90°	Inf

Table 6.1 summarizes the distortions at different distances from the equatorial line. The distortions represent how much longer a line segment appears in the equirectangular projection than in the HMD. For example if a horizontal line segment at 45 degrees from the equator covers 10 degrees of visual angle in the headset it would be represented by a projected line segment of 14.1 degrees on the equirectangular space. From the table we see that the equirectangular representation has no distortions in the middle and infinite distortions at the top and bottom. The infinite distortions arise from the fact that a single vector on the Y axis, which covers zero degrees in the 3D Cartesian space, acquires a whole line of pixels (360°) in the equirectangular representation. From this observation, it becomes obvious that pre-existing algorithms would fail close to the top and bottom areas of Figure 6.3. Intuitively these algorithms should not experience significant detection quality deterioration in the areas close to the equator.

If we do not want to change the algorithms as we did in the previous section we can take advantage of the central area of 45° vertically that has low distortions of no more than 8%. Therefore if all the gaze samples for a given period of time are exclusively within this horizontal stripe we can then use all the pre-existing algorithms in their original forms. We can achieve those preconditions by splitting

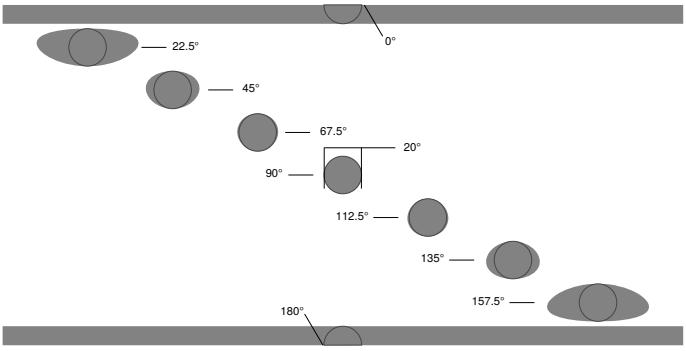


Figure 6.3: Visualization of distortions at the points of Table 6.1. The gray areas are the projection of the intersection of a cone with the presentation sphere. The black circles visualize a hypothetical undistorted circle.

the gaze input into time periods of no more than 45° vertical spread, which can then be moved to the central part.

During the centering of gaze samples we have to be careful not to simply displace vertically the gaze coordinates in the equirectangular representation because we will keep all the distortions, but to perform a rotation in the 3D Cartesian space. Another point of attention should be the limitation of monitor-based algorithms to handle movement along the vertical edges of the equirectangular video, which are seamlessly presented during the HMD viewing. These transition cannot be efficiently processed and therefore we choose to move samples not only vertically but also horizontally to the middle of the video. By doing so the possibility of the gaze moving along the vertical edges of the equirectangular frame decreases dramatically.

For the rotation matrix computation we use the extrinsic Y-Z-X Tait-Bryan angles of the mean gaze vector to the coordinate system of reference. The Y rotation is computed with the right-hand rule from the projected vector on the X-Z plane of Figure 6.1a and is equivalent to ϕ in equation 6.7. Then the Z rotation is computed similarly to the angle of Figure 6.1b but with the X and Z axes exchanged and this is equivalent to θ in equation 6.7. The rotation around the X axis is provided as head tilt in the recorded data. Then we find the rotation matrix between the middle of the equirectangular video, which is represented by the $(-1, 0, 0)$ vector, to the coordinate system of reference in the same fashion. The combination of these two rotation matrices gives us the overall rotation from the mean gaze vector to the center of the video. In the last step we apply the already calculated rotation matrix to all of the gaze samples and project them back to the equirectangular space by using equations 6.7, 6.8. Now the transformed data is ready to be handled by any algorithm that was developed for monitor-based experiments.

6.1.4 Conversion evaluation

For the evaluation of the conversion methods we used the full 360-degree data set of Section 4.3 and not only the hand-labeled part of it. The evaluation on the non-labeled gaze recordings was not a limiting factor because we wanted to assess how the algorithms fared with different levels of distortion. For that reason, we evaluated the original algorithm, its converted version, and the original algorithm with converted input data in the eye+head scenario (i.e. overall gaze). During testing we split the input space into four pairs of distinct areas, which represent horizontal stripes in the equirectangular video of 22.5° height. These pairs were placed symmetrically on both sides of the equator line, and depending on its distance from this line, each pair represented a different level of distortion.

Table 6.2: Percentage of samples for the four areas representing different distortion levels.

Range	% Samples
{67.5°, 112.5°}	84.97
{45°, 67.5°} ∪ {112.5°,135°}	10.29
{22.5°, 45°} ∪ {135°, 157.5°}	3.58
{0°, 22.5°} ∪ {157.5°, 180°}	1.16

We evaluated the algorithm proposed by [Larsson et al., 2015] (with the saccade detector by [Dorr et al., 2010]) because it is the most complete package by providing labels for fixations, saccades, SP, and noise, which includes blinks (with a separate algorithm). Here, for the evaluation we deviate from the so far used F1 metrics because we are interested in the overall sample-level agreement between the different versions of the algorithm and we use Cohen's Kappa as defined in equation 6.9. Each of the four eye movements is evaluated separately in a binary classification problem with positive labels for the detection of the given eye movement. P_o represents the proportion of agreement between the two versions of the algorithm and P_c the proportion of chance agreement. Kappa ranges from $[-1, 1]$ with higher scores representing higher agreement.

$$K = \frac{P_o - P_c}{1 - P_c} \tag{6.9}$$

For the evaluation of the two versions of the algorithm we kept all the parameters (16 in total) the same but η_{maxFix}. The last parameter represents the dispersion of the interval, which is measured as the diagonal of the bounding box containing the

Table 6.3: Cohen's Kappa scores for four different areas with varying distortion levels. Evaluating the original algorithm against the converted algorithm against the original algorithm with converted data.

	Cohen's Kappa between **original and converted algorithms**			
Range	Fix.	Sacc.	SP	Noise
{67.5°, 112.5°}	0.817	0.794	0.714	0.468
{45°, 67.5°} ∪ {112.5°,135°}	0.802	0.796	0.638	0.557
{22.5°, 45°} ∪ {135°, 157.5°}	0.753	0.763	0.558	0.479
{0°, 22.5°} ∪ {157.5°, 180°}	0.325	0.416	0.084	0.355

	Cohen's Kappa between **original and converted data**			
Range	Fix.	Sacc.	SP	Noise
{67.5°, 112.5°}	0.976	0.975	0.959	0.934
{45°, 67.5°} ∪ {112.5°,135°}	0.915	0.938	0.800	0.881
{22.5°, 45°} ∪ {135°, 157.5°}	0.817	0.883	0.638	0.788
{0°, 22.5°} ∪ {157.5°, 180°}	0.361	0.451	0.099	0.431

	Cohen's Kappa between **converted algorithm and converted data**			
Range	Fix.	Sacc.	SP	Noise
{67.5°, 112.5°}	0.815	0.778	0.705	0.387
{45°, 67.5°} ∪ {112.5°,135°}	0.810	0.780	0.635	0.453
{22.5°, 45°} ∪ {135°, 157.5°}	0.830	0.792	0.654	0.426
{0°, 22.5°} ∪ {157.5°, 180°}	0.816	0.789	0.627	0.357

interval samples in the original algorithm. In the converted algorithm the dispersion is calculated from equation 6.5 as the angle of a cone. The maximum difference between these two definitions of dispersion is achieved for a square bounding box where its diagonal is $\sqrt{2}$ times bigger than the radius of the circle that is tangent to all of its sides. Therefore we used an η_{maxFix} of 2.7° for the original version and 1.9° for the converted version.

Table 6.2 reports the percentage of time spent in each of the four areas, with Table 6.3 reporting the Kappa values for agreement between three different combinations of algorithms and data for each eye movement type. These combinations use the original version of the algorithm with the original data, the converted version of the algorithm with the original data, and the original algorithm on converted data.

When we compare the two versions of the algorithm applied to the original data,

we see that the highest agreement is achieved in the middle part of the video that spans the 67.5° to 112.5° range and holds roughly 85% of all samples. In this area the Kappa scores are 0.82 for fixations, 0.79 for saccades, and 0.73 for smooth pursuit. Even though these scores are high, we do not reach 1, which represents perfect agreement. The not *perfect* agreement may be attributed in part to the small distortions that appear at the borders of the middle area. More importantly, though, there is also no perfect agreement between the initial and converted criteria (16 in total) such as the slightly different definitions of dispersion. These two factors become more obvious for the SP detection, which is more sensitive to disturbances because often its characteristics (such as speed and dispersion) overlap with those of fixations and saccades. For now we will not comment on noise and leave its explanation for later.

The second area, which spans from 45° to 67.5° and from 112.5° to 135° from the top of the video, accounts for roughly 11% of the total samples. In this area distortions range from 8% to 45%, which is not enough to influence the overall detection quality for fixations and saccades since their Kappa scores remain static. The flat Kappa scores are probably the result of the clear separation for the characteristics of fixations and saccades, but we see a substantial drop in the agreement for SP.

In the third area, which spans from 22.5° to 45° and from 135° to 157.5°, we have roughly 4% of the total gaze sample. Here we have high distortions that range from 41% to 161% and therefore the uncorrected artifacts of the original algorithms lead to a strongly reduced agreement between algorithms for all three detected eye movement types.

In the last area, which covers the top and the bottom areas of the equirectangular projection, its extreme distortions result in a significant drop in the agreement for fixations and saccades with almost no agreement (at chance level) for SP detection. However, this area holds only a small share of the overall samples (1%).

Now when we look at the results of applying the original algorithm on the original and converted data we see that the agreement is almost perfect for the middle area. Still the small deviation can be attributed to the small distortions at the border of this area and the temporal border effects that are introduced by the segment-wise data conversion. As expected, the higher distortions for the off-center areas in the original data lead to roughly the same disagreement scores at the top and bottom, meaning that the approach of converting the data also reduces artifacts.

When we compare the two proposed distortion-aware methods (bottom panel of Table 6.3) we see that the agreement between them stays roughly the same throughout

the equirectangular projection. As before, the Kappa scores do not reach a perfect agreement due to subtle changes in algorithm criteria, but they are very close to the results of the middle area (which has the least distortions) between the original algorithm with original data and the converted algorithm. The constant Kappa scores further prove the correspondence between our two proposed conversion methods.

Noise Kappa scores are more erratic between the three different comparison cases. For the results between the original and converted data the noise scores fall in line with the other eye movements and monotonically decrease as we move away from the middle line. For the other two cases, the noise Kappa score hovers around 0.4 in all four areas. The almost flat noise scores may be attributed to the blink detector that is included in our re-implementation of [Larsson et al., 2015]. The blink detection algorithm first detects saccades by using the algorithm of [Dorr et al., 2010] (explained in Section 6.1.2). It then searches on both sides of an area with eye-tracking loss and if it detects a saccade close in time it labels all the samples (including the saccade) as noise. Also before and after a blink the video-oculography based eye trackers tend to report a shift in gaze. In the case of the initial algorithm the probability of finding a saccade is higher because it does not compensate for the high distortions close to the top and bottom that result in erroneously detecting higher speeds and therefore more saccades.

To summarize, we have shown that the highest agreement between the original and converted versions occurs in the middle (least distorted) part of the equirectangular projection with high Kappa scores. As we move further away from the middle region the agreement for all three eye movement types drops with SP being the most influenced. Especially in the areas with the highest distortions at the top and bottom of the equirectangular projection the eye movement classification becomes completely unreliable with the initial algorithm. As expected the performance between the converted algorithms and converted data is not influenced by the distortions and stays constant throughout the whole area of the equirectangular projection. To conclude, the high Kappa scores for the middle area confirm experimentally the plausibility of both the algorithm conversion and the gaze reprojection methods.

6.2 I-S^{5}T algorithm

Apart from converting pre-existing algorithms we also developed a rule-based eye movement classifier for 360-degree stimuli that is almost a direct formalization of the eye movement definitions we consider in Section 2.2.1. It assigns primary and, po-

tentially, secondary labels to every gaze sample by analyzing the same gaze and head movement information that was available to the manual annotator (Section 3.3).

We first detected the saccades by analyzing the E+H speeds with the dual-threshold algorithm of [Dorr et al., 2010], which avoids false detections while maintaining high recall by requiring each saccade to have a peak gaze speed of at least $150°/s$, but all surrounding samples with speeds above $35°/s$ are also added to the detected episode (thresholds determined by a grid-search optimization on the training part of the annotated 360-degree data set of Section 4.3.3). We did not use the FOV speed of gaze as it is influenced by head motion and can easily reach speeds above $100°/s$ when the eyes compensate for fast large-amplitude head rotations.

Afterwards, blinks were detected by finding the periods of lost tracking and extending them to include saccades that were detected just prior to or just after these periods, as long as the saccades were not farther than 40 ms from the samples with lost tracking.

We then split the remaining intersaccadic intervals into non-overlapping windows of 100 ms and classified each such interval independently. For this, we calculated the speeds of the head and the eye (relative to the head and the world) as the distance covered from the beginning to the end of the window divided by its duration.

To formalize the concepts of "stationary" and "moving" head cases, we used a speed threshold of $7°/s$. For the gaze speeds, we applied the low and the high thresholds of $10°/s$ and $65°/s$, respectively (both for the eye-in-head and the eye-in-world speeds) in order to distinguish slow, medium, and fast movements. These were chosen via a grid-search optimization procedure as well. As gaze stability decreases with head motion [Ferman et al., 1987], we scaled the gaze speed thresholds according to the speed of the head: $thd_{\text{scaled}} = (1 + v_{\text{head}}/60) * thd$, where $60°/s$ is the "reference" speed of the head. This means that if the head was moving at e.g. $30°/s$, the gaze speed thresholds were increased by 50%.

A fixation was always labeled when the E+H speed was below the low gaze speed threshold. If the head speed was above the corresponding low threshold, a secondary VOR label was assigned.

Pursuit-type eye movement labels were assigned when the E+H speed was between the low and the high gaze speed thresholds, unless the eye-in-head speed was above the high threshold (in which case, a noise label was assigned). However, there are different label combinations possible here: (i) *head pursuit* in combination with the primary label of *fixation* was assigned when the FOV (eye-in-head) speed was below

Table 6.4: Threshold Values

Name	Used for	Threshold	Optimized
θ_{sacc}^{low}	saccades	35°/s	✓
θ_{sacc}^{high}	saccades	150°/s	✓
θ_{gaze}^{low}	fix., SP, VOR, head purs.	10°/s	✓
θ_{gaze}^{high}	fix., SP, VOR, head purs.	65°/s	✓
θ_{head}^{low}	VOR, head purs.	7°/s	-
θ_{head}^{high}	scaling $\theta_{gaze}^{\{low,high\}}$	60°/s	-

the low threshold and the head speed was above its own low threshold; otherwise, (ii) *smooth pursuit* in combination with *VOR* was detected when the head speed was above the low threshold, which implied that the head and the eyes were working in tandem (presumably, to follow a moving object); (iii) *smooth pursuit* without any secondary eye movement type was assigned when the head speed was below its low threshold, meaning that the eyes did not have to compensate for the head movement.

For the samples that did not fall into any of the previously listed categories it was then known that they had very high speed but were assumed not to be a part of any saccade (since saccades were detected already). Consequently, the noise label was assigned.

Overall, our approach uses five speed thresholds (plus a scaling parameter), and thus we refer to our algorithm as I-S^{5}T, *identification by five speed thresholds*. An overview of its parameters is given in Table 6.4: two thresholds for saccade detection, two to quantize eye speeds (scaled by head speed), and one to determine if the head was moving sufficiently to justify a potential VOR label. The first four of these thresholds were optimized via grid-search on half of the annotated data set, while the other half was used for testing.

We also implemented an algorithm for detecting OKN (or nystagmus), with its sawtooth pattern of gaze coordinates. This pattern is easier to detect in the FOV gaze data as it often occurred during high-amplitude head motion in our data. The idea behind our detector is similar to [Turuwhenua et al., 2014], but uses the already detected saccades for segmenting the recordings into slow and fast phases, instead of finding the maxima and minima in the speed signal. An OKN is detected when the overall direction of gaze movement during an intersaccadic interval is roughly opposite (angle $\geq 90°$) to the direction of the adjacent saccades, whereas the two

neighboring saccades are roughly collinear (angle $\leq 70°$). In case of an already assigned VOR label, OKN+VOR is labeled instead.

6.3 Overall evaluation

Evaluation of different frames of reference

To evaluate the performance of our algorithmic event detection as well as to explain the benefits of utilizing the data from both the eye and the head tracking, we compared the performance of our algorithmic detector I-S^{5}T against two versions of the same algorithm: one that only uses the speed of the eye within the head (e.g. directly applicable to mobile eye-tracking data), the other – E+H gaze data (e.g. in HMD recordings, if additional data were discarded) instead of a combination of all available movement readouts. For all algorithm versions, we selected the gaze speed thresholds (i.e. head speed threshold was not optimized) with a similar grid-search optimization procedure on the training set of the 360-degree data set – first, the two thresholds for saccade detection were jointly optimized, then the remaining two gaze speed thresholds.

We refer to the algorithm versions as (i) *combined* for the "main" proposed version – the I-S^{5}T algorithm – that uses both the eye-in-head and eye-in-world speeds, as well as head speed for threshold scaling, (ii) *FOV* for the version that uses the eye-in-head gaze speed only, and (iii) *E+H* for the one that only uses the eye-in-world speeds. Of course, the FOV and E+H versions do not detect the combinations of head and eye movements, so the secondary labels of VOR and head pursuit were not assigned. OKN detection is possible, however. Since there was much more OKN+VOR than pure OKN in our data, whenever OKN was detected based on the FOV or E+H algorithm versions, an OKN+VOR label was assigned.

We evaluated all three algorithm versions on the manually labeled test set of the 360-degree data set. Table 6.5 contains the sample- and event-level evaluation measures (in the form of F1 scores) for our approaches.

All three algorithms achieve relatively high F1 scores for fixation and saccade detection, with the FOV version yielding substantially lower scores, however. This indicates that saccades can be easily confused with the eyes compensating for the head movement. The difference is even more pronounced for SP detection, with the FOV version of the algorithm lagging far behind. The differences between the E+H version and the "combined" version are generally very small for the primary eye

Table 6.5: Classification Performance on the Test Set

		Sample F1			Event F1		
	EM type	Comb.	FOV	E+H	Comb.	FOV	E+H
Primary	Fixation	0.911	0.867	0.900	0.897	0.808	0.890
	Saccade	0.813	0.737	0.813	0.899	0.865	0.899
	SP	0.381	0.128	0.362	0.288	0.153	0.293
	Noise	0.758	0.743	0.758	0.744	0.729	0.742
	Average	0.716	0.619	0.708	0.707	0.639	0.706
Secondary	OKN	0.205	–	–	0.085	–	–
	VOR	0.600	–	–	0.636	–	–
	OKN+VOR	0.664	0.614	0.647	0.577	0.626	0.620
	Head Purs.	0.546	–	–	0.204	–	–

Where marked with "–", the respective eye movement is impossible to classify with the respective algorithm version.

movement classes (fixations, saccades, SP, and noise), with the combined variant achieving marginally higher scores. For the secondary labels, only the version that combined eye-in-head and eye-in-world speeds can detect the full spectrum of the defined eye movements, as most of the secondary labels require the knowledge of both the eye and the head movement information. OKN detection was comparable across the board.

These results demonstrate that eye movement classification algorithms could benefit from using all the available information about head and gaze in every frame of reference. This is especially important for distinguishing eye movements driven by the retinal input (e.g. smooth pursuit) and other sensory intakes (e.g. VOR), which is supported by the definitions of the eye movement types that we introduced in Section 2.2.1. However, when only one frame of reference has to be used choosing the E+H offers significant improvements in comparison to the FOV frame of reference.

Evaluation of algorithms

In Table 6.6 we provide a comparison between our proposed algorithm and the algorithms that have been converted to the 360-degree domain. The earlier uses both the E+H and FOV frames of reference while the latter are applied in the E+H frame of reference. Therefore, we can only compare the detection quality of the primary eye movements among the 6 algorithms.

The I-S^{5}T algorithm results are repeated from Table 6.5 and are presented here

Table 6.6: Evaluation results for the 6 algorithms that work in the equirectangular space. Results are presented as F1 scores for *sample-* and *event-level* detection.

	Sample-level F1			Event-level F1		
Model	Fixation	Saccade	SP	Fixation	Saccade	SP
Combined I-S^{5}T* (ours)	0.911	**0.813**	0.381	**0.897**	**0.899**	0.288
[Larsson et al., 2015]*	**0.922**	**0.813**	**0.429**	0.889	**0.899**	**0.395**
Saccade by [Dorr et al., 2010]*	–	0.757	–	–	0.847	–
Fixation by [Dorr et al., 2010]	0.791	–	–	0.765	–	–
I–VT	–	0.710	–	–	0.679	–
I–DT	0.893	–	–	0.815	–	–

Models marked with * have been at least partially optimized on the training set of the 360-degree data set. Cells marked with "–" identify eye movements that were not detected by the specific algorithm. In each column, the highest value is **boldified**.

as a reference. The algorithm of [Larsson et al., 2015] uses the algorithm of [Dorr et al., 2010] for saccade detection with the previously optimized speed thresholds and therefore achieves the best results together with our proposed algorithm. For fixation detection it reaches human-level performance and surpasses I-S^{5}T in the sample-level F1 score without any prior optimization. Even though its SP detection performance lags significantly behind the other two eye movement types it offers a substantial improvement over the I-S^{5}T algorithm, which is rather expected due to the utilization of more complex classification criteria instead of simple speed thresholds. Also the SP score in the current data set is between the scores of the two monitor-based data sets (Tables 5.1 and 5.2), which were computed with the original version of the algorithm before conversion to the 360-degree domain. But these improvements come with the shortcoming of assigning only primary eye movements. Secondary eye movements could be potentially detected by combining the outputs of the algorithm for the two different frames of reference in an approach similar to the transitions described in Section 4.3.2.

The saccade detector of [Dorr et al., 2010] uses the optimized thresholds of the previous section but it does not reach the performance levels of the I-S^{5}T and the [Larsson et al., 2015] algorithms because it does not utilize a blink detector. This results in the mislabeling of blinks as valid saccades with the effect of diminished overall performance. The performance of the I-VT algorithm [Salvucci and Goldberg, 2000] is markedly lower, especially regarding sample-level F1 score, due to the utilization of a single speed threshold.

Finally, for fixation detection the algorithm of [Dorr et al., 2010] returns the lowest scores with the I-DT algorithm returning higher scores, which are higher than its non-converted version scores (Tables 5.1 and 5.2).

Part IV

Applications in dynamic natural contexts

In the third part of the thesis we use the data sets and the algorithms that we developed in the previous two parts in new fields of application that help us to better understand different aspects of eye movements. In Chapter 7 we investigate the relationship between different brain areas and smooth pursuit while people were watching a Hollywood movie as an approximation to dynamic natural scene viewing. Eye movements were automatically detected with the *sp_tool* and they were then used in our fMRI analysis pipeline. The analysis and the results contained in this part have been published in [Agtzidis et al., 2020a].

Then in Chapter 8 we investigate how well eye movement characteristics observed through simple stimuli (i.e. dots moving on a screen) transfer to more complex scenarios that are a better approximation of natural gaze behavior. For this purpose we investigated the relationship between SP and the saccades that preceded it with regard to their relative angle and we measure some statistics (e.g. saccade position error) in three different scenarios of varying complexity or similarity to a natural viewing scenario. For the most complex scenario, we used the hand-labeled eye movements of the GazeCom data set. Then we created two new conditions with reduced levels of naturalness and new eye-tracking data were gathered by Alexander Goettker. The content of this chapter has been published in [Goettker et al., 2020].

Chapter 7

Understanding smooth pursuit brain activations in dynamic natural scenes

Humans along with other animals with foveal vision use eye movements to explore their surrounding space. The decisions about attending to an object together with the type of the performed eye movement are driven by a multitude of brain processes, which can be driven by either low-level or high-level features. Consequently, the neural implementation of gaze behavior is an active research topic. In particular, functional magnetic resonance imaging (fMRI) has been previously used along with eye tracking in order to identify brain areas (i.e. BOLD-responses) and networks related to specific eye movements such as fixations and saccades [Luna et al., 1998, Beauchamp et al., 2001, Sestieri et al., 2008, Ettinger et al., 2007, Lukasova et al., 2018]. However, brain areas subserving smooth pursuit (SP) eye movements have been studied to a lesser extent only [Petit and Haxby, 1999, Lencer et al., 2004, Kimmig et al., 2008], possibly due to technical challenges in the analysis of dynamic setups.

When segmented eye-tracking data are directly related to brain activation, the majority of experiments use specifically designed synthetic stimuli [Lencer et al., 2004, Kimmig et al., 2008]. Such stimuli can take the form of fixation crosses that change position when saccades are studied, or of linearly or sinusoidally moving dots when smooth pursuit is investigated. The biggest advantage of synthetic stimuli is that their properties are well defined and can explicitly represent specific features, which simplifies the analysis of both the eye-tracking and BOLD signals. But this simplicity comes at the cost of using a paradigm that is not representative of normal

human vision because ecologically valid visual input is much more complex and real-world SP does not occur in isolation but within sequences of saccades and fixations. Therefore, the use of synthetic stimuli moving on a uniform background ignores the possible influence of background information [Brenner and Smeets, 2015], crowding effects [Sanocki et al., 2015], and the overall eye movement planning process [Gold and Shadlen, 2007, Tatler et al., 2017]. Another important limitation is that following a uniform synthetic stimulus over a longer time interval can result in reduced maintenance of attention [Tagliazucchi and Laufs, 2014, Vanderwal et al., 2015].

Because of the increased complexity of naturalistic stimuli, some studies have restricted themselves to the presentation of static naturalistic scenes, i.e. images [Kay et al., 2011, Mannion, 2015]. However, significant improvements in both vigilance and head motion by attaining the participant's attention were achieved by [Vanderwal et al., 2015], who used an abstract dynamic pattern together with fMRI resting-state analysis. An even better approximation to unconstrained human vision are fully naturalistic dynamic stimuli, and both the neuroimaging and eye-tracking communities have recently started to explore the possibilities of more immersive experiments [Hasson et al., 2004, Lahnakoski et al., 2012, Nardo et al., 2014, Andric et al., 2016, Marsman et al., 2016]. Some recorded data sets of naturalistic fMRI [Hanke et al., 2016] have even become publicly available.

In this chapter, we analyze the studyforrest fMRI data set [Hanke et al., 2016] with our smooth pursuit detection tool (sp_tool) of Section 5.1 and two different motion estimation algorithms [Barth, 2000, Revaud et al., 2015] with the aim of correlating brain activations with SP in dynamic natural scene viewing. The full analysis pipeline is made available online[1].

7.1 Methods

7.1.1 Data set

For our analysis we used the publicly available studyforrest data set as an approximation to a complex natural environment; for full experimental details, we refer to the paper presenting the original data set [Hanke et al., 2016]. Briefly, this data set includes 15 participants who watched the Hollywood movie "Forrest Gump" while their gaze was tracked in an fMRI scanner and another 15 participants with in-lab gaze only recordings (that we used here only to improve the automatic detection of

[1] `https://gin.g-node.org/ioannis.agtzidis/studyforrest_analysis`

smooth pursuit events, see below). The stimulus was presented to the in-scanner participants through an LCD projector in combination with a front-reflective mirror and to the in-lab participants through an LCD monitor. The gaze data were recorded with a high-frequency eye tracker (EyeLink 1000, set to 1000 Hz sampling rate with a telephoto lens attached for the fMRI recordings) and a 13-point calibration was performed at the beginning of each session. The fMRI recordings were acquired with a 3 T scanner (Philips Achieva dStream MRI scanner) with a repetition time (TR) of 2 seconds and $3 \times 3 \times 3\,mm^3$ voxel size.

7.1.2 Motion estimation in the stimulus

Because smooth pursuit behavior is tightly linked to moving targets, we estimated the overall motion per video frame with computer vision techniques. Despite all recent advances, such algorithms can still yield noisy outputs, so we used two different algorithms for additional robustness. The first algorithm computed motion based on the minors of the structure tensor as described by [Barth, 2000] with the aim to provide a sparse optic flow field by estimating motion only at points that are not susceptible to the aperture problem, i.e. corners. Initially, the input video was spatially subsampled by a factor of two, and then a spatio-temporal Gaussian pyramid with five spatial and two temporal levels was created. For each level of this multiscale representation, velocity per pixel was computed. These velocity estimates were normalized relative to the original video resolution and combined in a procedure similar to pyramid synthesis described by [Adelson and Burt, 1980]; higher speed values were clipped to the 90th percentile speed. The second algorithm uses edge-preserving interpolation of correspondences for optical flow (EpicFlow) computation as described by [Revaud et al., 2015]. The algorithm in the first step uses dense matching with edge-preserving interpolation followed by an energy minimization step. An example of content motion computation of the EpicFlow algorithm is provided in Figure 7.1b. For both algorithms, finally, the mean length of pixel displacements was computed per video frame.

7.1.3 Eye movement classification

From the provided data we created a quadruplet of values for each gaze sample that comprised time, x and y coordinates on the monitor coordinate system, and a confidence estimation of the eye tracking quality. Since the data set used monocular eye tracking, a confidence value of 1 meant good tracking of the eye and a value of 0

(a) Data set frame (b) Frame's flow

Figure 7.1: **(a)** Example frame from the studyforrest data set with superimposed gaze traces (over a 400 ms period; one color per subject) from the in-scanner participants. Smooth pursuit is evidenced by the elongated point clouds. **(b)** Optic flow computed by the EpicFlow algorithm. The estimated motion corresponds well with the actual motion in the video. Black lines indicate the sp_tool output, i.e. automatically detected smooth pursuit segments (in the 400 ms window).

meant tracking loss. After inspection of the data, lost tracking varied from 1.2 % to 16.7 % among subjects with the notable exception of subjects 05 and 20. For these two subjects the lost tracking was 86.7 % and 39.0 %, respectively, and they were excluded from all subsequent analyses. The remaining gaze traces were segmented into eye movements with the sp_tool.

Since the sp_tool was optimized for the GazeCom data set (Section 5.1), some parameters were adjusted (see parameters file on the online repository for full details). We further improved the SP detection by using both in-lab and in-scanner recordings together because the SP detection algorithm improves with an increasing number of gaze traces. Despite the different stimulus sizes, we used the same pixel-space for both sets of recordings by scaling the pixel-per-degree values for the (less noisy) in-lab recordings; the agreement in detected SP episodes for the two sets was high (r^2 of 0.84 for the share of SP in 2-second intervals).

7.1.4 fMRI analysis

The fMRI data analysis was performed with SPM12 using Matlab 9.2. We initially followed a standard preprocessing pipeline for each recording [Poldrack et al., 2011]. The process comprised realigning the functional data to the mean image of each session (without slice timing correction), coregistering them to the anatomical T1 scan, normalizing them to the MNI template and resampling them into $3 \times 3 \times 3\,mm^3$ voxels. Finally, we applied smoothing with a Gaussian kernel of $8\,mm$ at full width half maximum (FWHM).

During the recording of the studyforrest data set its authors split the movie stimulus into 8 different segments of approximately 15 minutes each with each one displayed separately in the scanner. In the first level analysis we combined all 8 recording sessions into one design matrix in order to model the full Forrest Gump movie. For each session in the design matrix we fitted an SP, a saccade, and movie motion regressor when needed. In order to account for variations in the onset and width of the hemodynamic response among subjects we used the canonical hemodynamic response function (HRF) along with its time and dispersion derivatives. Apart from the previous regressors we also used the six head movement components that were returned from the realignment step during preprocessing as nuisance regressors.

The eye movement and motion regressors were modeled as event time series with events placed 2 seconds apart, which by design coincides with the scanner's TR and therefore each event was representing the regressor variance between scans. The amplitude of each event was modulated by the prevalence of the corresponding eye movement or the amount of motion in the 2-second window and was having a value of 0 when it was the same as the overall mean and was linearly increasing up to a maximum value of 1. A detailed description of the regressor modeling procedure is given in the next section. As it becomes evident from how the regressors were modeled it would have been impossible to model both fixations and SPs with this process without creating strong (negative) correlations between the two. To make this interdependence more clear let's consider that a subject starts pursuing a target. Then consequently the amplitude of the SP regressor would increase with the fixation amplitude decreasing proportionally.

After fitting the GLM to the data of each subject independently we used the amplitude component of the HRF of each regressor that spanned 8 recording sessions in order to compute the contrasts of interest. These contrasts included the main effect of the eye movements and motion, the comparison between SP and saccades, and the comparison of the eye movements to motion. Finally, at the second level of the fMRI analysis we performed a one-sample t-test for each of the previous contrasts for the 13 valid subjects. The resulting clusters ($p < 0.05$ Family Wise Error [FWE] corrected with an initial threshold of $p < 0.001$) were overlaid on a three-dimensional brain and are presented in the results section.

Regressor modeling

As outlined above, our regressors were not modeling each eye movement event independently but were placed in 2-second intervals, which were modulated by the

amount of the respective eye movement in that window. For the experiments that were taking movie motion into account this was modeled through the mean movie motion and as before in consecutive 2-second windows.

More specifically, the computation of the magnitude of the eye movement modulation parameters was taking into account three main factors. (i) The first factor was capturing the changes in eye movements between different naturalistic stimuli and was represented by the mean percentage of each eye movement type of each subject and is equivalent to the mean viewing behavior. (ii) The second factor was capturing the differences in prevalence and variance between different eye movement types and it was a constant value with the modulation parameter being inversely proportional to it. The value of this factor was chosen from the data in order to bring approximately 95 % of the modulated values below 1 (for a visualization, see Figure 7.2). Therefore it was set to $modulationSacc = 1.5$ for saccades due to their small variance in relation to different input stimuli. For SP it was set to $modulationSP = 5$ in order to reflect the large variance of the eye movement, which cannot occur in the absence of a moving target but can be continuously performed for long periods of time when a salient moving object exists. (iii) The third factor was capturing the variance among subjects. This subject-specific factor was based on the observation that the prevalence of each eye movement type varies among subjects and it may directly or indirectly relate to the differences in brain connectivity [Mueller et al., 2013, Vanderwal et al., 2017]. In the case of the studyforrest data set saccades varied from 5.8 % to 12.4 % and SPs from 11.5 % to 19.3 % among the subjects and it becomes obvious that if the overall mean was used the relevant activations in some subjects would be suppressed and in some would be amplified.

As an illustrative example, consider a hypothetical subject which has an overall mean SP percentage of $overallSP = 15\,\%$ and performed SP $clipSP = 10\,\%$ of the time in a given clip. Now in a particular 2-second window $windowSP = 85\,\%$ of its duration was labeled as SP. The modulation magnitude will be $(windowSP - clipSP)/(modulationSP * overallSP) = (85 - 10)/(5 * 15) = 1$. After computing the modulation parameters across all 2-second intervals of the data set according to the previous formula we found that the SP and saccade regressors were uncorrelated (Pearson correlation $r = 0.02$), which is a good indication of no shared variability between the two.

For the magnitude of the motion estimation modulation parameters we followed a similar process as with the modeling of eye movement parameters. Again, here the steady state was captured through the mean content motion for each stimulus independently. The resulting value was normalized with the 90th percentile of the

motion values across all clips and was bound to a maximum value of 1 in order to limit the influence of the outliers.

7.1.5 Additional validation regressors

Apart from the eye movement regressors of SP and saccade we also used a motion regressor, which in the nominal case was modeling the global motion in the video. We further explored two variations of it. In the first variation of motion modeling we used a window around the gaze position to get a local estimation of the motion and in the second variation we subtracted the smooth pursuit velocity from the mean content velocity in the same window with the aim of approximating the retinal motion. Since the results with local motion were subpar in comparison to global motion we do not present them in the results section but we only discuss them later on.

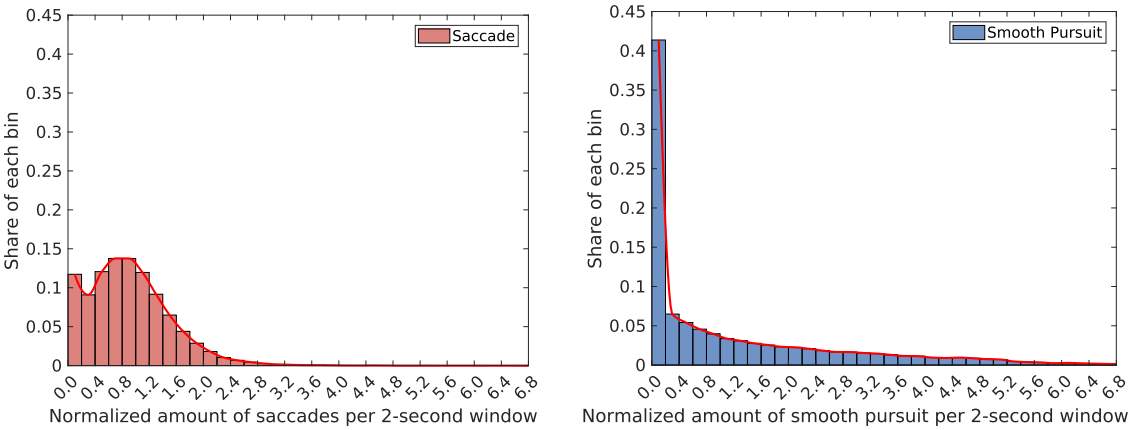
To further understand what drives eye movements we also ran models that included scene complexity and edge density estimation as additional regressors with their values being modeled identically to motion regressor as explained in the previous section. The scene complexity was computed as the entropy of the saliency of each frame using a standard saliency model [Itti et al., 1998]. Similar to the entropy of image saliency, we calculated edge density as the per-frame entropy of the absolute pixel values on the third level of a Laplacian pyramid (which represents edges in the spatial frequency range of approximately 3-6 cycles per degree, i.e. close to the peak of the human contrast sensitivity function). Again these results are discussed later on.

7.2 Results

The presented functional group results of this section were mapped to the three-dimensional cortical template of the "Population-Average, Landmark- and Surface-based" Atlas (PALS) [Van Essen, 2005] with the metric-enclosing-voxel algorithm in Caret (version 5.65) [Van Essen et al., 2001]. When needed the provided coordinates are reported in the Montreal Neurological Institute (MNI) coordinate system.

7.2.1 Eye movement statistics

Overall, in the valid in-scanner subjects the algorithm classified 53 % of gaze samples as fixations, 8.4 % as saccades, and 14.8 % as SP with the rest being labeled as noise (tracking loss, blinks, cluster noise, etc.). Because we here were interested in separating e.g. saccades and SP as cleanly as possible, this relatively high noise level was acceptable. Fixations showed the highest absolute variation among participants (std: 10.1 %), which is to be expected since the fixation detection is very sensitive to eye-tracking noise and our objective was not to model this type of eye movement. Saccades (std: 2.5 %) and SP (std: 3 %) had lower absolute variance but very high relative variations among participants. This relatively high between-subject variability was captured by the subject-specific modulation factor during the first level analysis.



(a) Saccade ratio distribution

(b) SP ratio distribution

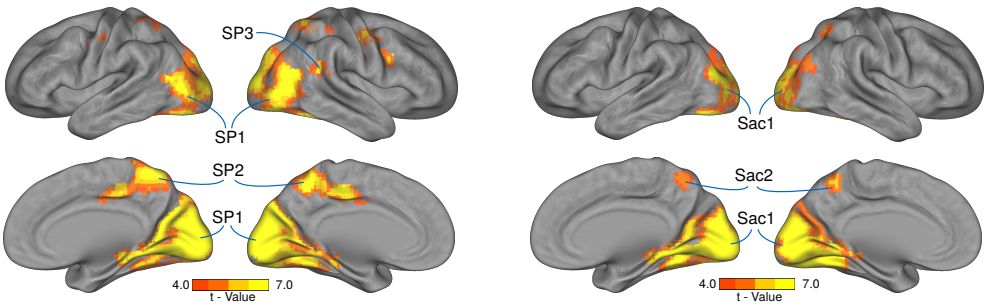
Figure 7.2: Probability distribution of saccade (a) and smooth pursuit (b) ratios as detected in the 2-second windows that were used during the event-related 1st level analysis, normalized so that 1 corresponds to each subject's mean. A wide distribution indicates high variability across subjects and time. Saccades (red) have lower variability and are centered around 1. SP ratios (blue) are more variable and the peak close to 0 represents the absence of SP (e.g. no SP target is moving in the scene).

Apart from the between-subject variability there exists within-subject variability, which varies for different eye movements. In Figure 7.2 we visualize the probability distributions of the ratios in 2-second windows of saccades and SP per subject in relation to the same subject's overall mean. Because the range of the distributions differed between eye movement types we chose the eye movement specific modulation factors of Section 7.1.4 with the aim of normalizing them into comparable ranges. Here, a value of 1 indicates that the share of each eye movement type in a given

interval is equal to the overall subject mean. A value of 0 denotes that the respective eye movement does not occur in that interval and values above 1 mean that we have above-average occurrence. As can be seen from Figure 2, saccades show lower within-subject variability and are centered around the mean ratio of 1. On the other hand, the occurrence of SP shows higher variability with a peak close to 0, which represents the absence of SP when no moving target is present in the stimulus, i.e. movie.

7.2.2 SP- and saccade-related activations

The mean effects of SP- and saccade-related BOLD-responses are given in Figure 3, where we present clusters at $p_{FWE} < 0.05$ using an initial threshold of $p < 0.001$. This procedure yielded three clusters related to SP (SP1-SP3) and two clusters related to saccades (Sac1-Sac2), see Table 7.1. The notable difference between the SP1 and Sac1 clusters is the strong activation within the middle temporal gyrus, presumably visual motion area MT+/V5 in SP1 but not Sac1, which is to be expected since this area is associated both with SP and motion processing. The second large cluster marked as SP2 mainly covers parts of the middle cingulate cortex and the precuneus. Also there exists a much smaller saccade-related cluster that covers part of the precuneus and is marked as Sac2. Finally, a small SP-specific cluster related to the right temporoparietal junction (rTPJ) is marked as SP3.



(a) SP baseline activations (b) Saccade baseline activations

Figure 7.3: **(a)** SP-related activity with $p_{FWE} < 0.05$ with initial threshold of $p < 0.001$. Activations span bilaterally the visual areas of the brain (SP1: $k_E = 7647$, including the SP-related MT+/V5), bilaterally the middle cingulate expanding to the precuneus (SP2: $k_E = 2048$), and the right temporoparietal junction (SP3: $k_E = 109$). **(b)** Saccade-related activity with $p_{FWE} < 0.05$ with initial threshold of $p < 0.001$. Activations span bilaterally the visual areas of the brain (Sac1: $k_E = 6437$) and the precuneus (Sac2: $k_E = 245$). For a detailed list of the subareas refer to Table 7.2

Table 7.1: List of clusters with peak activation t-value and location along with cluster level FWE-corrected p-values that are related to SP and saccadic eye movements.

Cluster name	Peak activation			Cluster size	$p_{FWE-corr}$	Peak t-value
	x	y	z			
SP1	-3	-91	14	7647	< 0.001	15.11
SP2	6	-43	56	2048	< 0.001	8.48
SP3	57	-40	17	109	0.011	6.49
Sac1	-9	-82	17	6437	< 0.001	16.84
Sac2	-6	-52	56	245	< 0.001	6.25

Table 7.2 lists a more detailed description of anatomical and functional areas included in the identified clusters SP1-SP3 and Sac1-Sac2, respectively. Anatomical areas were parcellated with the automated anatomical atlas [Tzourio-Mazoyer et al., 2002, Rolls et al., 2015]. In order to avoid cluttering the table with anatomical areas that are represented by relatively few voxels, we applied a cutoff threshold as a percentage of the total voxel count in each cluster. For the two biggest clusters of Table 7.1 the threshold was set at $\sim 2\,\%$ and for the rest at $5\,\%$.

7.2.3 SP-saccade related activations

In the way that we structured our analysis, saccades were used as a proxy to represent the steady-state condition of our visual system, because of their lower variability, with smooth pursuits being the eye movement of interest. Hence we were interested in the specific differences between SP- and saccade-related brain activations during natural viewing. These contrasts with $p_{FWE} < 0.05$ and initial threshold of $p < 0.001$ are visualized in Figure 7.4.

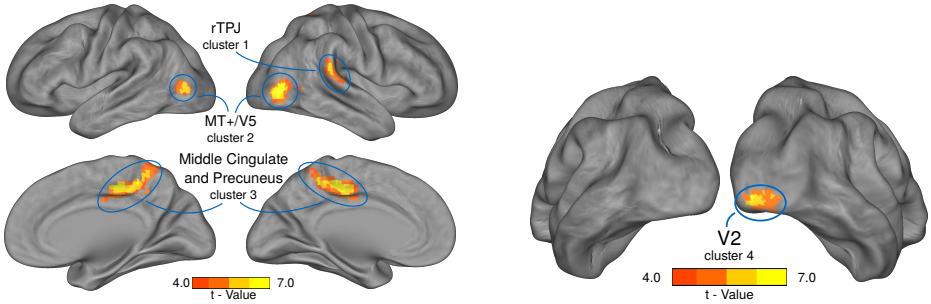
This procedure identified three areas with stronger activation during SP compared to saccades. In Figure 7.4a the first area has bilateral activations of the motion processing and SP-related area MT+/V5 (right: $k_E = 169$, left: $k_E = 89$), with the second area containing the middle cingulate and extending to precuneus ($k_E = 655$). Lastly, the third area comprises of an activation in the right temporo-parietal junction (rTPJ; $k_E = 158$). Figure 7.4b shows that the $saccade > SP$ contrast has significant activations in V2 (right: $k_E = 91$). The full list of anatomical areas that are part of these clusters is provided in Table 7.3.

Table 7.2: List of brain areas involved in both SP- and saccade-related clusters and areas that are unique to SP. The threshold for visualization was chosen at ∼2 % for the big clusters and 5 % for the smaller clusters of Table 7.1. Therefore the values do not sum up to the total number of voxels in each cluster.

Anatomical Area	Brodmann area (functional region)	SP activations						Saccade activations					
		Peak activation			Part of	Num. of voxels	Peak t-value	Peak activation			Part of	Num. of voxels	Peak t-value
		x	y	z				x	y	z			
L Lingual	17, 18	-15	-76	2	SP1	528	12.04	-18	-79	2	Sac1	522	14.00
R Lingual	17, 18	12	-85	13	SP1	578	8.59	9	-85	-13	Sac1	599	10.73
L Calcarine	17, 18, 30	-3	-91	14	SP1	505	15.11	-9	-79	14	Sac1	503	15.94
R Calcarine	17, 18, 30	12	-85	8	SP1	447	11.32	12	-79	14	Sac1	309	15.01
L Cuneus	18, 19	-3	-91	17	SP1	368	13.33	-9	-82	17	Sac1	329	16.84
R Cuneus	18, 19	12	-94	14	SP1	405	10.72	12	-79	17	Sac1	309	11.75
L Occipital Sup	18, 19	-15	-97	23	SP1	273	12.51	-15	-82	11	Sac1	255	12.49
R Occipital Sup	18, 19	15	-97	17	SP1	219	10.42	18	-94	5	Sac1	164	8.04
L Occipital Mid	19, 37 (V5)	-48	-73	5	SP1	532	10.47	-15	-10	8	Sac1	555	7.18
R Occipital Mid	19, 37	27	-88	14	SP1	303	8.09	36	-67	29	Sac1	291	6.45
L Occipital Inf	18	-27	-82	-10	SP1	142	8.30	-27	-70	-10	Sac1	131	7.36
L Fusiform	18, 19	-24	-79	-10	SP1	347	8.44	-33	-49	-10	Sac1	352	11.90
R Fusiform	18, 19	33	-79	-16	SP1	415	11.09	27	-82	-16	Sac1	365	12.03
L Cerebellum 6	–	-6	-73	-16	SP1	181	8.14	-18	-73	-16	Sac1	185	9.41
R Cerebellum 6	–	15	-85	-16	SP1	214	9.76	24	-82	-19	Sac1	187	11.86
L Precuneus	5, 7	-9	-49	47	SP2	325	8.20	-6	-52	56	Sac2	100	6.25
R Precuneus	5, 7	6	-43	56	SP2	278	8.48	6	-52	53	Sac2	69	5.30
L Temporal Mid	19, 39 (V5)	-48	-70	8	SP1	122	8.61	–	–	–	–	–	–
R Temporal Mid	19, 39 (V5)	41	-64	8	SP1	223	11.16	–	–	–	–	–	–
R Temporal Inf	37 (V5)	48	-46	-25	SP1	101	8.02	–	–	–	–	–	–
L Cingulate Mid	23, 24, 31	-9	-22	44	SP2	184	7.55	–	–	–	–	–	–
R Cingulate Mid	23, 24, 31	12	-25	44	SP2	177	6.32	–	–	–	–	–	–
R Paracentral Lobule	5	12	-40	56	SP2	111	6.63	–	–	–	–	–	–
R Temporal Sup	40	57	-40	13	SP3	55	6.94	–	–	–	–	–	–
R Supramarginal	40	51	-40	23	SP3	45	6.57	–	–	–	–	–	–

7.2.4 Accounting for movie motion

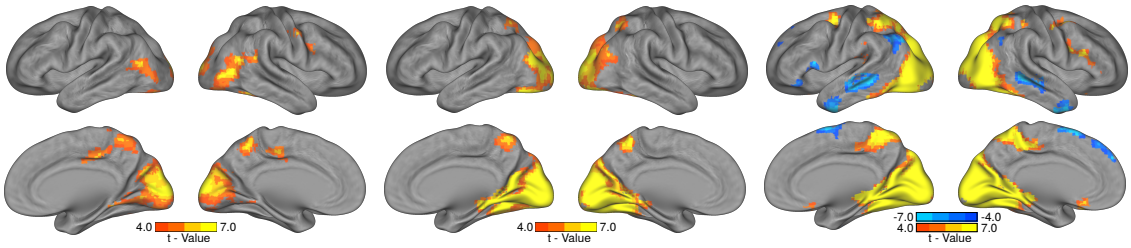
To differentiate SP from content motion-related brain activations we added an additional motion regressor during the first level analysis, which was again modeled as time-series with its values computed in a process similar to eye movement modulation of Section 7.1.4. Here, we present the results using the EpicFlow algorithm and whole frame mean motion modeling (results for the algorithm based on the minors of the structure tensor were qualitatively similar, data not shown). The resulting motion regressor was uncorrelated with the saccade regressor (Pearson $r = -0.11$) and the same held true for the SP regressor (Pearson $r = 0.18$). The mean effects of SP-, saccade-, and motion-related BOLD-responses are visualized in Figure 7.5. As can be seen from Figures 7.5a and 7.5b the activations for SP and saccades are qualitatively very close to the activations of Figure 7.3 but with reduced size and intensity for SP when motion was included in the model (Figure 7.5a). This reduction in SP-related activations followed by strong positive motion-related (Figure 7.5c) activations in roughly the same areas as the SP-related activations shown in Figure 7.3a. Moreover, the activity in the cortex lining the superior temporal sulcus



(a) $SP > saccade$ activations (b) $Saccade > SP$ activations

Figure 7.4: **(a)** Activations for SP > saccade at $p_{FWE} < 0.05$ with an initial threshold of 0.001. Activations in bilateral MT+/V5 (right: $k_E = 169$, left: $k_E = 89$), in the middle cingulate extending to precuneus ($k_E = 665$), and in the right temporoparietal junction (rTPJ) ($k_E = 158$). **(b)** Activations for $saccade > SP$ contrast with $p_{FWE} < 0.05$ with initial threshold of 0.001. Activation in V2 (right: $k_E = 91$).

(STS) and in the supplementary motor area including the supplementary eye field (SEF) was negatively correlated with our motion regressor.



(a) SP baseline activations (b) Saccade baseline activations (c) Motion baseline activations

Figure 7.5: **(a)** Activations for SP > saccade at $p_{FWE} < 0.05$ with an initial threshold of 0.001. Activations in bilateral MT+/V5 (right: $k_E = 169$, left: $k_E = 89$), in the middle cingulate extending to precuneus ($k_E = 665$), and in the right temporoparietal junction (rTPJ) ($k_E = 158$). **(b)** Activations for $saccade > SP$ contrast with $p_{FWE} < 0.05$ with initial threshold of 0.001. Activation in V2 (right: $k_E = 91$).

Following this model, the contrast $SP > saccade$ yielded significant activations only in the middle cingulate ($k_E = 106$) and rTPJ ($k_E = 104$) areas (not shown graphically). The MT+/V5 and precuneus activations of Figure 7.4a 4a did not reach the significance threshold of 0.05 FWE.

The $SP > motion$ contrast (Figure 7.6) revealed bilateral activations in the cortex lining the superior temporal sulcus (STS; right: $k_E = 536$, left: $k_E = 194$), the precuneus ($k_E = 102$), and the supplementary motor area including the supplementary eye field (SEF; $k_E = 177$). To the contrary, the $saccade > motion$ contrast did not reveal any significantly activated areas

Table 7.3: List of areas involved in *SP* > *saccade* and in the *saccade* > *SP* contrasts. The threshold for visualization was set to 15 voxels for all clusters and they do not sum up to the total voxel number for each cluster.

Anatomical area	Peak activation			Part of	Number of voxels	Peak t-value
	x	y	z			
R Temporal Sup	60	-31	20	cluster 1	73	5.95
R Supramarginal	60	-34	23	cluster 1	36	5.83
L Temporal Mid	-51	-73	8	cluster 2	23	5.79
R Temporal Mid	51	-70	-1	cluster 2	103	10.11
L Occipital Mid	-48	-76	5	cluster 2	89	7.46
L Cingulate Mid	-9	-31	44	cluster 3	157	10.18
R Cingulate Mid	9	-22	44	cluster 3	145	8.01
L Precuneus	-12	-44	44	cluster 3	30	5.63
R Precuneus	12	-52	58	cluster 3	92	6.72
R Paracentral Lobule	12	-37	47	cluster 3	32	6.36
R Postcentral	15	-49	68	cluster 3	32	6.77
R Parietal Sup	18	-49	68	cluster 3	26	6.06
R Occipital Inf	24	-97	-7	cluster 4	42	7.32
R Occipital Mid	39	-88	-1	cluster 4	18	7.11
R Lingual	24	-91	-4	cluster 4	15	5.64

7.3 Discussion

The aim of applying our algorithms and analyzing the studyforrest data set was to investigate brain activations related to SP and saccades in complex dynamic naturalistic scenes. To this end, we presented methods based on off-the-shelf algorithms and modeling techniques that can handle the noisy and unstructured nature of motion and eye-tracking data coming from scanner recordings when dynamic natural scenes are used as stimuli. Our main results are in line with previous studies showing activations in the MT+/V5 area during SP when SP and saccades were modeled separately. When an additional regressor representing motion content of the stimulus was included in the model, specific attention-related areas were identified while some other brain areas (including MT+/V5) fell below the significance threshold due to the similar SP and motion BOLD mean effects.

Validity of eye movement classification

To ensure that the sp_tool returned high quality output in the studyforrest data set, we manually tuned its parameters based on visual inspection of a small portion of the results. A full manual annotation of a data set as big as the studyforrest (ca. 30 hours) was not feasible given the fact that it takes approximately 15 to 75 s for one annotator to label one second of gaze and multiple annotators are needed for best results.

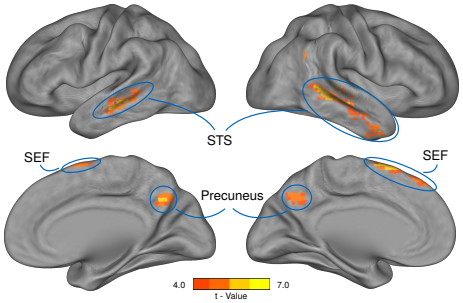


Figure 7.6: Activations for $SP > motion$ contrast with the motion regressor computed with EpicFlow and $p_{FWE} < 0.05$ with initial threshold of 0.001. Activations appear in the right superior and middle temporal gyri (posterior and anterior STS) ($k_E = 536$, peak $xyz : 60, -55, 26$) and in the left middle temporal gyrus (posterior STS) ($k_E = 194$, peak $xyz : -60, -28, 11$), bilaterally in the precuneus ($k_E = 102$, peak $xyz : 6, -58, 35$), and bilaterally in the supplementary eye field ($k_E = 177$, peak $xyz : -6, 11, 62$).

Validity of algorithms defining motion content

A potential weak point in using motion estimation algorithms to define the motion content of a stimulus is the fact, that they tend to give noisy results. For that reason, we validated the presented results by using two different motion estimation algorithms [Barth, 2000, Revaud et al., 2015]. In both cases the identified brain activations were comparable, underlining the validity of our approach. In the second analysis of our study, we were interested in specifically identifying what drives SP in humans in the presence of motion. To this end, we used the mean frame motion as an approximation to background motion. However, there exist many other ways of modeling motion and we investigated two of them in more detail. In the first approach we modeled motion in a five degree window around each gaze position. In the second approach we aimed at decorrelating the two regressors by modeling retinal motion. For this purpose, we subtracted the SP velocity (speed and direction) from the motion velocity in the same window and then used the magnitude of the resulting vector in our model. The resulting activations, while qualitatively similar, were weaker for both approaches in their extent and intensity. This can be partially attributed to the fact that in both of these approaches the correlation between the motion and SP regressors was higher than when only mean frame motion was used (window $r = 0.21$, window - SP velocity $r = 0.51$ vs. mean frame $r = 0.18$). The changes in the correlation values can be attributed to many factors. Generally the noisy results of motion estimation algorithms may become even noisier as we use the mean of a smaller window instead of the full frame. Also the reported gaze can be noisy and oftentimes has spatial offsets, which can result in missing completely or partially the moving target in the motion computation. As a result, SP velocity

disproportionately influences the result of its subtraction from the window motion and thus returns higher correlation values. A similar effect appears with targets of very small size. It should be noted that the reported gaze position from the eye tracker was much noisier in the scanner than in the lab: the median dispersion of 25 ms windows of gaze data was 31 pixels in the scanner vs. 10 pixels in the lab for the studyforrest data set.

Brain areas related to variance in smooth pursuit

The contrast of $SP > saccade$ with only the SP and saccade regressors included in the first level design matrix revealed activations in the middle cingulate and precuneus, which have been previously associated with SP eye movement control [Tanabe et al., 2002, Kimmig et al., 2008] and visuo-spatial processing [Berman et al., 1999, Cavanna and Trimble, 2006]. Additionally, this contrast yielded higher activation during SP related to the rTPJ, an area that is involved in guidance towards unattended areas [Corbetta et al., 2000, Wu et al., 2015, Marsman et al., 2016]. Most importantly, this contrast revealed bilateral activations related to area MT+/V5, which is regarded as a core motion processing area and has been associated with SP eye movements in previous studies [Petit and Haxby, 1999, Kimmig et al., 2008, Lencer and Trillenberg, 2008, Ohlendorf et al., 2010, Marsman et al., 2016]. Notably, the MT+/V5 area became non-significant in the same contrast when a third regressor modeling the overall stimulus motion was added. This may be best explained by the fact that the variance of the BOLD response in this area was now shared between two regressors (SP and motion) instead of one [Ohlendorf et al., 2010] as can be seen from the mean effect of SP and motion in Figures 7.5a and 7.5c. This demonstrates the difficulty in finding a single source of activation in natural scenes where many different factors may provoke activation of a specific area and a complete disentanglement of such confounds may prove elusive.

Benefits of considering motion content in the model

Adding motion as a regressor to the model allowed us to identify SP-related activations that were not per se driven by the overall motion of the stimulus (Figure 7.5a). Interestingly, motion itself additionally resulted in negative effects related to STS and SEF areas. Thus, when directly contrasting $SP > motion$, these two areas together with the precuneus occurred as being significantly stronger activated during SP than by motion content alone (Figure 7.6). STS is considered a hub for information processing including the processing of biological motion [Saygin, 2007, Jastorff

and Orban, 2009, Grossman et al., 2010] as well as the processing of faces in situations requiring social cognition [Allison et al., 2000, Lahnakoski et al., 2012]. In line with this model, inhibiting STS activity by transcranial magnetic stimulation (TMS) resulted in difficulties perceiving biological motion [Grossman et al., 2005]. Also, reduced activity in the STS has been associated with difficulties in understanding biological motion and emotional content in autism spectrum disorder patients [Alaerts et al., 2013, Nackaerts et al., 2012]. SEF activations have been associated with anticipatory eye movements, even in situations with invisible targets, reflecting cognitive input to smooth pursuit planning independent from visual input [Lencer et al., 2004, Ohlendorf et al., 2010].

When interpreting our finding related to motion content it should be considered that our motion regressor was based on a low-level account of pixel-wise motion energy, which might have failed to capture the semantic properties of natural scenes. Thus, high values of motion content from our analyses were related to background and camera motion (Figure 7.7), which are both extensively used in professionally shot cinematic videos [Cutting et al., 2011]. In contrast, moving mid-sized objects, i.e. socially meaningful targets, were linked to low motion content values. Thus, irrelevant motion modeled by our motion content regressor may have led to the observed negative activations bilaterally in the STS and the SEF unless SP to a meaningful target was performed.

Given the current rapid pace of progress in computer vision algorithms for high-level scene segmentation and understanding, more complex modeling of the semantics of different types of motion information might enable a more fine-grained analysis of such effects in the future.

Considering additional possible confounds

To at least partially alleviate the potential confounds of the motion energy analysis, we included additional regressors modeling basic video characteristics. In two control experiments, we modeled scene complexity based on saliency and edge density as attention-grabbing parameters in order to test whether these parameters interfere with the activations related to SP and motion content. In both cases the mean effect of the validation regressor showed significant activations in some very small clusters (approx. 150-300 voxels overall in the posterior part of the brain and mostly in the visual cortex) and did not influence the activations regarding the main contrasts of interest. From these observations we conclude that the eye movement planning process is predominantly driven by the underlying motion based on the way we

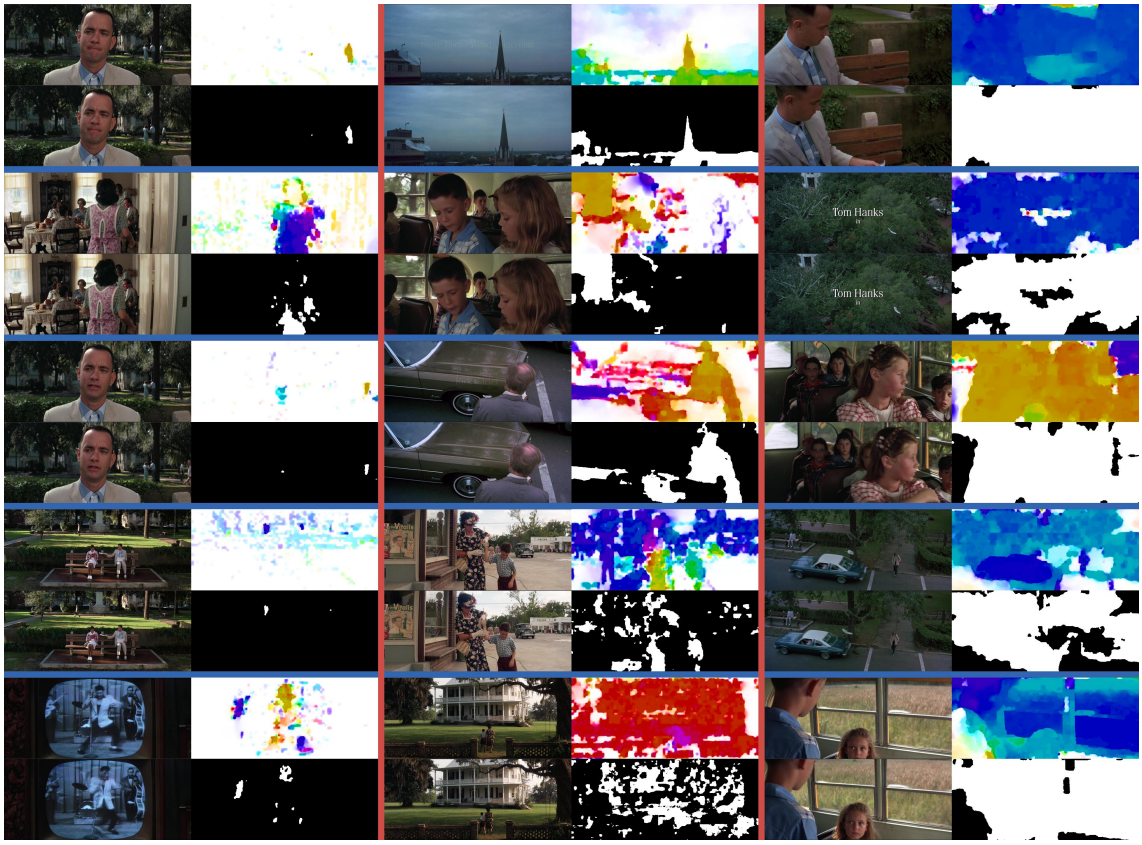


Figure 7.7: Visualization of 15 example frames and their motion content. Each of the three main columns (separated with red lines) represents different levels of frame motion (low, medium, high). Within each supercell we display the randomly chosen frame (top left), the progress of the movie 5 frames later (bottom left), the optic flow of the random frame (top right), and the 2 pixel absolute motion thresholded optic flow (bottom right). As we move towards higher motion content the amount of pixels above threshold increases, which indicates that our motion regressor returns higher values for background and camera motion.

modeled each characteristic. However, a more exhaustive search of all the potential parameters and modeling techniques may be required in future studies of SP in dynamic natural scenes.

Lack of associations with frontal eye fields under natural viewing conditions

We did not identify any activations related to the frontal eye fields (FEF), which have been described to be involved in the planning and execution of both SP and saccades [MacAvoy et al., 1991, Berman et al., 1999, Gagnon et al., 2006, Kimmig et al., 2008]. One possible explanation might be that in typical experiments participants switch between baseline periods of prolonged fixation and e.g. dot following

or scene viewing. Instead, in the data set used here, participants were likely to constantly engage in some form of eye movement planning during continuous movie viewing, which is more representative of real-world viewing behavior. Therefore, the variance of e.g. saccades in consecutive 2-second windows may not have been sufficient to identify all saccade-related activations, including FEF. Another limiting factor may be the small size of the FEF regions and the big variance in their reported location [Vernet et al., 2014] along with their activation being dependent on specific experimental conditions and instructions [Lencer et al., 2004].

7.4 Chapter conclusion

In this study, we demonstrated brain networks specifically related to the often-overlooked smooth pursuit eye movements in complex dynamic naturalistic scenes. Our findings underline the notion that special care has to be taken to model variance across subjects, within subjects, and for different eye movement types. We also identified some of the confounds which arise from the semantic variation in movie content and which cannot be captured by a low-level image-based analysis alone. Nevertheless, our results show that findings from previous research with impoverished synthetic scenes can be qualitatively confirmed for highly complex, ecologically valid naturalistic stimuli.

Chapter 8

Saccade and smooth pursuit initiation interactions

Generally, eye-tracking experiments can be characterized by many attributes, but the most important distinctions are the stimulus dynamics: static versus dynamic, and the elaborateness of the stimulus: highly controlled, simple stimuli versus ecologically relevant natural stimuli in everyday scenes of different complexity. While static stimuli mainly evoke saccades and fixations, dynamic stimuli can also evoke sequences of slow smooth pursuit eye movements (for reviews, see [Schütz et al., 2011, Kowler, 2011, Lisberger, 2015]). Synthetic motion stimuli such as dynamic random dot patterns allow the control of the different attributes under test, i.e. strength of motion signal versus noise, size, and direction of dot displacement, but they come at the cost of potentially diminished ecological validity [Heinen and Watamaniuk, 1998, Schütz et al., 2010]. Naturalistic stimuli such as photographs or videos of everyday scenes provide more ecologically valid targets in their natural environment, but they come at the cost of little control because of the complexity of natural images and the many relevant factors that might influence and modify the response behavior of observers.

Synthetic stimuli have been used to measure, under controlled conditions and with high precision, eye movement characteristics such as latency and accuracy of saccades [Saslow, 1967, Munoz et al., 1998]. They can also be used to study more abstract concepts such as neural response times [Eagle et al., 2007], and cognitive states [Heuer et al., 2013]. Static naturalistic stimuli have been used to investigate how instructions change saccadic eye movements [Mills et al., 2011], as Yarbus studied this already in his classical experiments [Yarbus, 1967, p. 174], working memory [Wolfe et al., 2011], reading comprehension [Jacobson and Dodwell, 1979], and the

prediction of future gaze locations through different saliency models [Itti and Koch, 2000, Kümmerer et al., 2016].

A significantly smaller number of studies have used dynamic stimuli. Dynamic synthetic stimuli have been used to investigate smooth pursuit eye movement characteristics [Tychsen and Lisberger, 1986] (for a review see [Lisberger, 2015]) and their neural correlates [Gellman and Carl, 1991, Pack and Born, 2001, Nagel et al., 2006]. Dynamic natural stimuli have been used to understand eye movement behavior in everyday scenarios and tasks [Land and Hayhoe, 2001, Dorr et al., 2010, Tatler et al., 2011]; for a review see[Hayhoe, 2017], but these studies are rare.

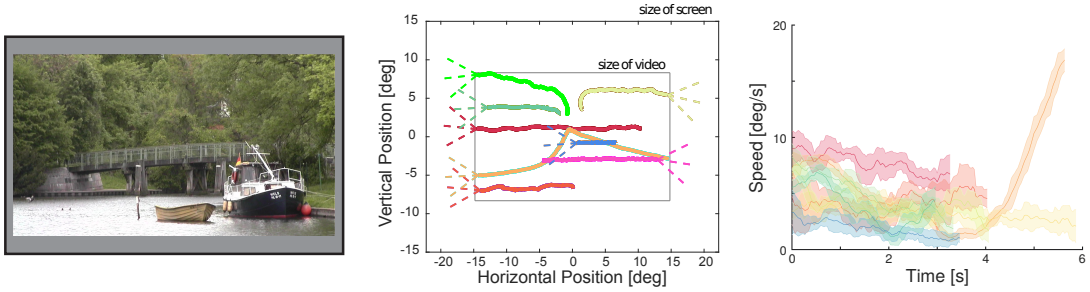
The important question remains whether and how the results of studies with synthetic stimuli transfer to the relatively underexplored dynamic natural environments. Some studies have investigated the transferability of such results by using paradigms that presented stimuli at different levels of complexity. [Foulsham and Kingstone, 2010] used static images and investigated how saccade directional asymmetries changed between computer-generated images and natural scenes. [Martens and Fox, 2007] investigated how fixation patterns changed with the familiarity of a scene across two levels of naturalness. For this purpose they taped videos and eye movements during repetitive driving under real conditions and compared them with repetitive viewing of a prerecorded driving clip.

In the present study we wanted to compare voluntary tracking behavior (in terms of saccadic and pursuit latency and errors of both movements) in response to synthetic and naturalistic dynamic targets and to test for generalizability between them. To bridge the large gap between free-viewing of natural scenes and controlled lab conditions with a single synthetic stimulus, we created two experiments. The starting point of our experiments was the GazeCom data set. Based on this, we determined a set of targets (the baseline) that were followed by a large number of observers with smooth pursuit eye movements (Figure 8.4). We then used this baseline and the corresponding video parts for our new experiments to measure eye movements to targets moving along the same trajectories as in the baseline. In both experiments, eye movements always started from the same position, but the trajectory was either represented by a video clip containing the target of interest (naturalistic experiment), or by a Gaussian blob that moved in the same pattern as the original target (synthetic experiment) (Figure 8.2).

8.1 Methods

8.1.1 Selection of baseline trajectories

The two new experiments together with the free-viewing analysis relied on the GazeCom data set. For the design of our experiments (as summarized in Figure 8.2) and the measured statistics (e.g. interaction between smooth pursuits and targets) eye movements alone were not sufficient and we needed the accurate trajectory of the targets in the natural videos of the data set's videos. Contrary to experiments with synthetic stimuli, it is very challenging to obtain accurate target trajectories in dynamic natural contexts. Even though many automated algorithms exist for motion estimation and optical flow extraction they would have been too noisy for our use case. For this reason we manually labeled target trajectories: We selected 45 targets[1] based on the longest-duration smooth pursuit clusters (from 0.8 to 5.9 sec) as detected by an initial version of the sp_tool and manually labeled a representative point of each target (e.g. nose of a walking person) in each video frame.



(a) Free viewing example (b) Target trajectories (c) Target speeds

Figure 8.1: Free-viewing condition and selected trajectories. **(a)** A Still shot from one video clip of the GazeCom data set. Over the course of the video, a duck is flying by (orange trajectory in panel B) that many observers follow with smooth pursuit. **(b)** Hand-labeled trajectories of 8 different targets from 8 different video clips (different colors) that were used in the two new experiments. The horizontal, up-, and downward tilted line segments represent the three possible positions of the fixation point at the beginning of each trial as well as their distance (5 deg) from the initial position of the subsequent target. The different frames depict the size of the experimental monitor and the size of the presented videos. **(c)** Speed profiles of the 8 selected targets as presented in the synthetic and naturalistic experiments. Shown is the average speed in a 100 ms time window, with the error bars showing the standard deviation of the labeling induced by the labeling of the target. The colors correspond to the trajectories shown in panel b.

[1] `https://gin.g-node.org/ioannis.agtzidis/gazecom_annotations/src/master/targets_arff`

Baseline trajectories for data collection

For the synthetic and naturalistic experiments we chose 8 of the hand-labeled trajectories from the GazeCom data set. We based our choice on the duration of the target movements (> 3 sec) and on a pilot study, where we tested whether participants were making a saccade to the relevant targets from the initially presented fixation dot. The trajectories of the 8 selected targets and their speed profiles are shown in Figure 8.1b and 8.1c.

Baseline trajectories for validation

To investigate how well our results generalize to the fully unconstrained free-viewing condition, we analyzed SP responses to all 45 targets comprising about 3000 SP intervals (about 7 % of the total GazeCom viewing time) for the baseline targets.

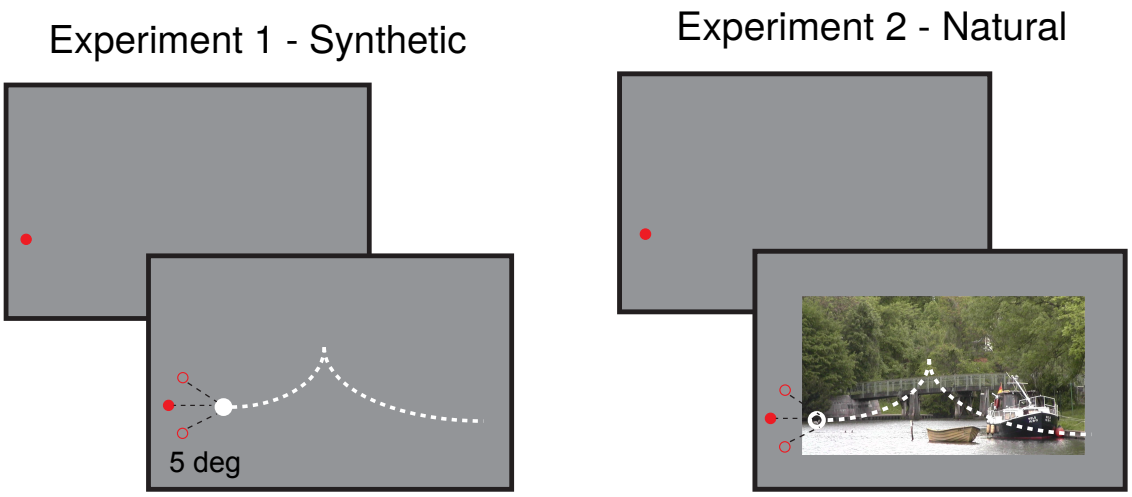
In the synthetic and naturalistic experiments, we varied the relative angle between saccade and smooth pursuit to collinear and $+/-30$ deg, and our analyses were based on at least 150 ms of post-saccadic smooth pursuit. From the GazeCom hand-labeled ground truth [Startsev et al., 2019b], we selected only SP intervals that were within the previous criteria and obtained 238 saccade-SP pairs (down/collinear/up in 30 deg wide bins: 69/114/55).

8.1.2 Experimental design

Every single trial in both experiments started with a fixation cross, which was placed based on the initial direction and position of the upcoming target trajectory. The initial direction of the natural trajectories was defined based on the slope of the line between the first position of the target and the position of the target after 250 ms. The fixation cross could appear collinear with the motion direction at 5 deg from the initial target position or at 5 deg distance but rotated by 30 deg up- or downwards. Participants looked at the fixation cross and pressed a button to start the trial. We used the button press to perform a drift correction at the fixation location. After the button press a red dot replaced the fixation cross and stayed there for a random duration between 1 and 1.5 sec (see Figure 8.2). After the dot disappeared, the two experiments presented different types of stimuli. In the experiment with synthetic stimuli a target appeared and directly started to move. The target was a white Gaussian blob (SD = 0.5 deg, max contrast = 0.5) on a uniform gray background and its movement was following one of the eight hand-labeled target trajectories.

In the naturalistic experiment the part of the respective scene that contained the actual hand-labeled target motion was shown.

Participants randomly started with either of the two experiments and each experiment contained 4 blocks of 72 trials each (8 scenes $*$ 3 orientations of the starting position $*$ 3 repetitions). One block lasted approximately 15 minutes with participants taking breaks between blocks and typically performing 3 blocks per session. In our analysis, we included only trials where the saccade started from within 2.5 deg of the initial fixation position and the position error measured between the saccade end position and target position was below 3 deg. The second criterion made sure that participants were tracking the correct target, which was especially relevant in the naturalistic experiment, where multiple potential targets may have been present. Overall, 6077 out of 7488 trials (81 %) were included in the analysis. As the target was more clearly defined in the synthetic condition we only had to exclude 390 trials in comparison to 1021 trials in the naturalistic experiment. There were some small differences in the exclusion rate for the individual scenes, as in some scenes there was more additional information that attracted the gaze as well.



(a) (b)

Figure 8.2: Depiction of the two new experiments. After an initial fixation, in experiment 1 (synthetic) participants saw a white Gaussian blob moving on a uniform gray background along the same trajectory as the duck (or other hand-labeled objects in the natural videos). In experiment 2 (natural) participants initially fixated at the same position, but then saw the part of the natural video that included the moving target. The trajectory of the flying duck corresponds to the orange line in Figure 8.1b.

8.1.3 Experimental setup

Participants sat at a table facing a 32 in monitor (Display ++, LDC; Cambridge Research Systems Ltd) in a dimly illuminated room. We used a chin and forehead rest to stabilize the participant's head and minimize its movement. In this setup the eyes of the participants were approximately at the height of the screen center and at a distance of 90 cm from it. We recorded eye movements from the right eye with a desk-mounted eye tracker (EyeLink 1000 Plus, SR Research) with a sampling frequency of 1000 Hz. The experiments were programmed in MATLAB using the Psychtoolbox [Kleiner et al., 2007]. Before each block we used a 9-point calibration to align the gaze data with the screen.

As the temporal and spatial resolution of the stimulus as well as the viewing distance used to collect the GazeCom data (30 Hz, 1280 $*$ 720 pixels, 45 cm) were different from the ones used for the semi-controlled stimuli (120 Hz, 1920 $*$ 1280 pixels, 90 cm) we had to perform some transformations. To account for the different temporal resolution between the original monitor and that of experiment 1 (synthetic presentation), we resampled the x and y pixel values from the hand-labeled trajectories to 120 Hz with linear interpolation. In experiment 2 (natural scene presentation) the videos were still presented at 30 Hz by just updating the monitor's content every fourth frame. Regarding the difference in spatial resolution, we decided not to resize the presented scenes or the position of the labeled targets to the new monitor size but we instead decided to present them in the central 1280 $*$ 720 pixels of the larger monitor. This choice led to differences in the visual field between our experiments (30 $*$ 17 deg) and the GazeCom recordings (48 $*$ 28 deg) but this decision was driven by two factors: First, rescaling the scenes may have led to blur. Second, and more importantly, for our two new experiments we wanted to control the initial fixation position of the participants. By presenting the smaller target trajectories or scenes in the center of the screen, we were able to present a fixation dot 5 deg from the target starting position (Figure 8.2). This would not have been possible if we had used the full size of the monitor to present the stimuli.

8.1.4 Participants

For our experiments we recorded the eye movements of 13 volunteers (mean = 23.5 years old, SD = 3.5; 11 females). All of the participants were naïve to the purpose of the study and had not seen any of the videos before. They were mainly students of Giessen University and had normal or corrected-to-normal vision. Before the

start of the experiments they gave informed consent (Declaration of Helsinki) and all experiments were approved by the local ethics committee (LEK FB06 2017-08). Participants received 8 Euro per hour as monetary compensation.

8.1.5 Data analysis

For each subject we quantified the characteristics of initial saccades and smooth pursuit and tested for differences with respect to the stimulus complexity and the relative angle between both consecutive eye movements. We examined these interactions for saccade latencies and saccade position errors as well as for pursuit gain and pursuit directional accuracy. Below, we describe how each measurement and statistic was calculated and mention potential differences or limitations for each type of experiment.

For the synthetic and naturalistic experiments we calculated the saccade latencies as the difference between the start of the target movement and the onset of the first saccade. For the (GazeCom) free-viewing validation data such latencies are not defined, because of the continuous presentation. The saccade position error was defined as the Euclidean distance between the saccade landing position and the labeled target position. For pursuit parameters, we analyzed the interval between 50 ms and 150 ms after the saccade completion. This choice excluded post-saccadic oscillations from our calculations and thus returned more robust results while limiting the influence of additional new retinal information after the end of the saccade. Specifically, pursuit gain was defined as the mean of the ratio between pursuit and target speeds. Since the targets were not moving linearly we projected the sample-to-sample gaze direction onto the linearly interpolated target direction at each moment in time. The gain was computed as the average ratio of the projected gaze speed and the target speed during the relevant interval. Pursuit accuracy was defined as the pursuit angular error, which is calculated as the absolute difference between the pursuit direction and the target direction, calculated between the first and the last point during the pursuit interval. We also computed the pursuit precision for each scene, which was defined as the width of a Gaussian distribution fitted to all available segments of the signed pursuit direction error across participants. These included the direction errors measured in the 50 to 150 ms interval, but we additionally included segments after this interval if the eye stayed closer than 3 deg to the target and had less than 45 deg of direction error, as here the participants were presumably still tracking the target. We used a sliding window of 100 ms in 10 ms intervals to find these new segments.

To test for systematic influences on these statistics we used repeated measures ANOVA with factors relative to angle (down, collinear, up) and stimulus complexity (synthetic vs. natural scene).

8.2 Results

Here we present the results of different eye movement parameters for the synthetic and naturalistic experiments that represent different levels of stimulus complexity. Where possible, and in order to better bridge the gap to complete free viewing of naturalistic scenes we also present the equivalent statistics from the original GazeCom data set.

8.2.1 Saccadic eye movements

The latency of initiating saccades towards a target can function as an indicator for the processing time that is required for the programming and execution of target-directed saccades. Our experiments (Figure 8.3a) contain two stimulus conditions and three different saccade-target angles (down, collinear, up), which were used as factors in a repeated measures ANOVA that tested for significant influences on the saccade latencies. We observed significant main effects both for the stimulus complexity ($F(1, 12) = 89.745, p < .001$) and relative angle ($F(2, 24) = 6.137, p = .006$) but no significant interaction between the two. The simpler stimuli that comprised of a single Gaussian blob moving on a uniform background had saccade latencies of 183 ms on average, which was significantly lower than the 235 ms for the natural targets. Moreover, saccades had lower latency when they were collinear with the subsequent pursuit target across both stimulus complexities (synthetic: 177 ms vs. 186 ms, $t(12) = 4.21, p = .001$, natural: 231 ms vs. 237 ms, $t(12) = 2.06, p = .06$).

Figure 8.3b illustrates the saccade accuracy with regard to the different stimulus complexities and different directions. The saccade accuracy was defined as the saccade position error between the endpoint of the saccade and the starting position of the target and was computed per subject as the average error across the 8 different scenes. In a similar procedure as before we ran a repeated measures ANOVA, which revealed a significant effect of relative angle ($F(2, 24) = 5.446, p = .011$) and stimulus complexity ($F(1, 12) = 109.262, p < .001$) and no significant interaction between the two ($F(2, 24) = 2.489, p = .104$). Saccade position error was lower in the synthetic condition and also displayed lower variance. Also in Figure 8.3b we

display the saccade position error for the GazeCom data set, which is on average closer to the natural experiment. But overall the errors in the free-viewing condition follow a similar pattern as with the two new experiments with saccades been more accurate in the collinear condition.

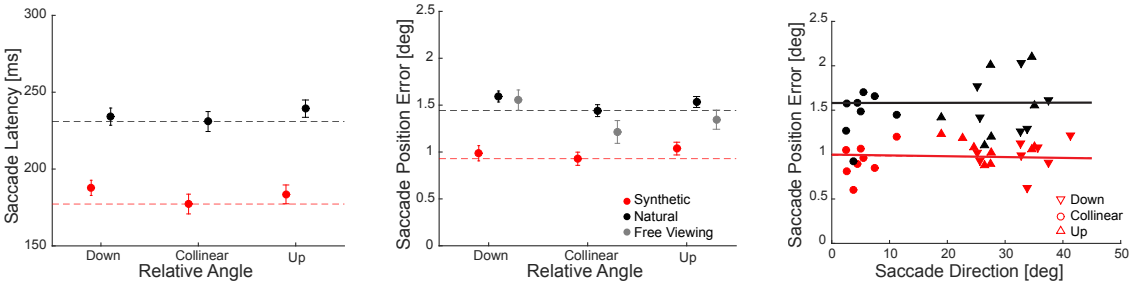
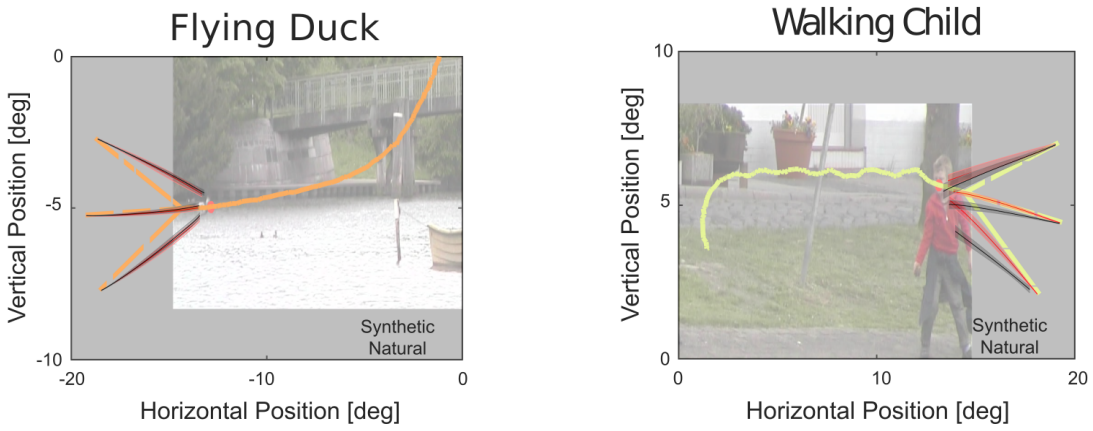


Figure 8.3: Comparison of initial saccades across experiments. **(a)** Stimulus complexity strongly influenced saccadic latency. The relative angle between the initial saccade and the upcoming pursuit had a much smaller but significant effect. The two dashed horizontal lines represent saccadic latency in the collinear case for synthetic (red) or natural (black) targets. Error bars depict the standard error of the mean (panels a and b). **(b)** Saccade position errors, defined as average Euclidean distance between saccade endpoints and target positions, are shown for the synthetic (red) and for the natural (black) condition. For comparison, saccade position errors from GazeCom free viewing validation data are plotted (light gray). The graph shows that the saccade position errors are larger by about 0.5 deg for targets in natural scenes. **(c)** Saccade error as a function of the deviation of the saccade direction from the horizontal axis. The three symbols represent the different relative angles between initial saccades and pursuit for the synthetic (red) and the natural (black) experiments. The solid lines represent a linear regression fitted to the data.

We found that the stimulus complexity has a strong effect on initial saccades; in natural scenes initial saccades to moving objects had significantly longer latencies and larger position errors. However, there is a confound we need to clarify because the saccade error is closely related to object size. For a small object, such as a duck flying in the distance, a single point may be sufficient to describe it; for larger targets, such as a moving child or a car close to the camera, however, a single dot is not enough to represent the target. The problem of object size is evident in the end positions of saccades starting from different locations as shown in Figure 8.4 for two target positions chosen for two objects of different sizes: the beak of the flying duck on the left and the nose of the moving child on the right. For the duck the saccades from all three positions land on it while for the child the landing position is sometimes dependent on the starting position. Saccades starting at the top position almost always land on the child's face (the marked representative point) but as we move downwards the saccade landing position start to deviate more often

towards the center of the child's body. As a result, when only one target location is used for the representation of a larger object saccade position errors will most probably increase because observers will attend different parts of the same object. In Section 8.3.3 and Figure 8.6a we discuss this observation in more detail.



(a) (b)

Figure 8.4: Saccades to moving targets of different sizes. The images in the background are cutouts of the two natural videos: Flying duck on the left and Walking child on the right. The colors of the trajectories correspond to the colors of the target trajectories in Figure 8.1b. Overlaid are the averages of saccade trajectories starting from three different fixation locations to two targets of different sizes. Note that for the flying duck scene the black and red curves obscure each other. The shaded areas depict the standard area of the mean of the trajectories. Note how the saccade landing positions depend on the starting positions and the target size. When moving objects are small like the duck, deviations of the trajectories and landing locations are very small and comparable to the synthetic experiment; when target objects are large like the child, saccades sometimes aim at different locations of the same object, either at the center of the body when starting from the lower fixation dot or towards the center of the face when fixating the central or upper fixation dot.

8.2.2 Pursuit eye movements

For pursuit gain as well as the pursuit direction error we only analyzed the pursuit in the interval close to the end of the saccade (50 to 150 ms after saccade end) to minimize the influence of post-saccadic oscillations and of any new retinal information after the saccade. We again computed a repeated measurement ANOVA with the factors relative angle (down, collinear, up) and stimulus complexity (synthetic vs. natural). For pursuit gain (Figure 8.5a) we observed a significant main effect of the relative angle ($F(2, 24) = 11.365, p < .001$), while there was no influence of

stimulus complexity ($F(1, 12) = 0.028, p = .870$).

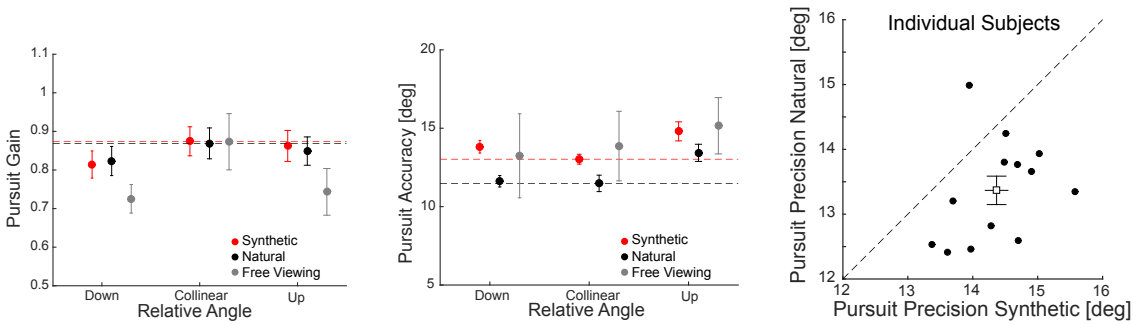


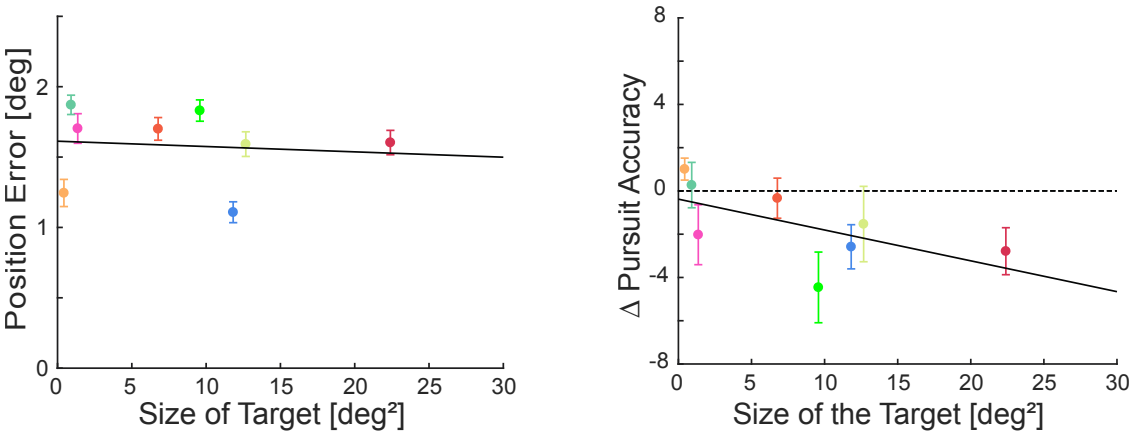
Figure 8.5: Pursuit behavior. **(a)** The average pursuit gain depended on the relative angle between the direction of the initial saccade and the upcoming pursuit and on the stimulus complexity. The dashed horizontal lines depict the value of the collinear condition. **(b)** The absolute pursuit direction errors depended on the relative angle between the direction of initial saccade and the upcoming pursuit and on the stimulus complexity. As in panel a the dashed horizontal lines depict the value of the collinear condition. **(c)** Comparison between the standard deviation of the pursuit direction errors for the synthetic and naturalistic experiments. Each black dot represents a single subject, the open black square the average. All error bars depict the standard error of the mean.

Despite the qualitatively similar results we still observed one significant benefit of the natural and richer information when watching the video. When we analyzed and compared pursuit accuracy (Figure 8.5b), we found a significant main effect of stimulus complexity ($F(1, 12) = 20.715 p < .001$), with lower error when tracking a natural moving target in the video. This benefit was highly consistent across our observers and the different targets/scenes (Figure 8.5c; $t(12) = 4.311, p = .001$), indicating that the additional information led to an improved tracking performance in natural scenes.

8.2.3 Effect of object size

In order to better understand the observed effects in the naturalistic experiment, we compared them against the different object sizes. The object size for each target was estimated by manually fitting a bounding box on a representative frame and was kept constant under the assumption that each target's shape did not change substantially during its presentation (50 to 150i ms). Even though none of the measured statistics was significantly correlated with the size of the pursued object, below we present some results of interest. Initially we correlated the saccade position error against

the object size because the definition of the earlier becomes more ambiguous with larger objects. Here we did not observe a relationship between the two (Figure 8.6a: $r(8) = -0.11, p = .8$) suggesting that higher position errors in the naturalistic experiment were not purely driven by the larger object sizes. Additionally, we present the correlation between pursuit accuracy and object size in Figure 8.6b. With our limited sample size of only 8 different scenes, which correspond to 8 targets, we found a correlation of $r(8) = -0.6(p = .12)$.



(a) (b)

Figure 8.6: Role of object size. **(a)** Saccade position error as a function of target size. The eight video scenes are depicted with different colors as in Figure 8.1b. The black solid line depicts a linear regression fitted to the data. **(b)** Comparison of pursuit direction error between the synthetic and naturalistic experiments. The relative pursuit error for the eight video scenes is depicted in the same colors as in panel **a** and negative values indicate an improved performance for the natural condition. The black solid line depicts a linear regression fitted to the data. The black dashed line depicts zero. All data were first averaged across the different relative angles and then across participants. The error bars depict the standard error of the mean across participants.

8.3 Discussion

These experiments tried to tackle whether eye movement behavior differs for natural moving objects, such as a duck flying across a river in a park, and a simple Gaussian blob moving across a gray monitor screen.

To answer this question we performed two experiments to investigate the generalizability of results gained in eye movement studies with highly controlled, simple laboratory stimuli and more complex animate objects in natural scenes. To enable

a direct comparison, we used the trajectory of a real object from a natural video (such as a flying duck) for simple Gaussian blob targets so that they moved exactly along the same path and at the same velocity. Overall, we observed qualitatively comparable results, suggesting that one can generalize between eye movement performance under highly controlled conditions with simple stimuli to eye movement behavior when watching natural video scenes. For pursuit, we observed that the gain was comparable for synthetic or natural targets and both conditions showed a significant benefit of collinear saccadic and pursuit eye movements. However, there were also interesting differences: latencies of initial saccades to moving Gaussian blob targets were significantly lower compared to saccades to moving objects in video scenes, and pursuit accuracy in video scenes was significantly higher.

8.3.1 Effect of scene complexity on saccadic eye movements

Comparative studies of eye movement properties in response to the same target shown in different contexts, i.e. uniform backgrounds versus naturalistic scenes are rare. However, most results revealed that the characteristic properties of eye movements are quite similar [Foulsham et al., 2011, Henderson et al., 2013, Walshe and Nuthmann, 2015]. We contrasted eye tracking responses across two conditions with moving stimuli, either a Gaussian blob or objects in video scenes, and also found similar eye movement behavior. Initial saccades rapidly aligned the gaze with the moving peripheral target, which in our study was either a simple blob or a complex natural object. While we kept the starting position and the movement trajectory the same we were able to compare the effects of backgrounds, which consisted of a uniform monitor screen or the natural scene. Doing so we made two observations: (i) latencies were shorter for saccades to synthetic stimuli, (ii) the position error was higher for the natural stimuli, despite having a qualitatively similar pattern.

The latency difference seems to be based on the fact that in real-world scenes the selection of a target object is potentially hard if one thinks of the target selection in some kind of a race model [Gold and Shadlen, 2007, Tatler et al., 2017]. Computational saliency models used to predict gaze behavior [Itti and Koch, 2000, Einhäuser et al., 2008, Kümmerer et al., 2016] produce more variance for naturalistic complex scenes in comparison to the clearly defined simple targets in synthetic scenes. An indication for this higher demand with respect to selection and decision is the overall increase in saccade latency by 52 ms for the natural background compared to the uniform blank screen (see Figure 8.3a). Higher saccadic latencies were also reported by [Walshe and Nuthmann, 2015] for initial targets in their uniform con-

dition (203 ms) versus their scene condition (214 ms). Thus, while it is possible to use less complex scenes as a proxy for naturalistic images, there seem to be some additional caveats that one needs to take into account [Foulsham and Kingstone, 2010]: saccade parameters like the position error are qualitatively comparable between varying stimulus complexities, but the conditions are less clearly defined with respect to the target and saccades have different latencies.

Even though saccadic position error was significantly lower in the synthetic experiment, the position error for the complex videos was still within roughly 1.5 deg from the target and far from being inaccurate. The way we defined the position error was more suitable for the blob condition, as here the target was symmetrical around the labeled target position. For the video conditions, as only one point on the object was labeled, trials in which a participant looked at the relevant object, but not on the labeled position (see for example Figure 8.4), were assigned a high position error. This probably led to an overestimation of the position error in the video conditions, as we did not ask participants to look at the labeled part of the object, but they could freely choose their gaze position. The preferred landing position of saccades in the natural condition seemed also to depend on the former fixation position since observers directed their saccades to different parts of larger objects after fixations on higher or lower positions as found for the child shown in Figure 8.4b. This seems to be in line with the analysis of saccade landing positions when animals were shown in natural scenes, where the landing position also revealed a preference toward the head of the animal as well as the center of gravity [Drewes et al., 2011]. Also other studies comparing fixation patterns across different stimulus complexities found changing gaze patterns [Martens and Fox, 2007, Foulsham et al., 2011]. Foulsham and colleagues compared the gaze behavior in people in the real world walking outside to buy coffee with that of people watching in the lab videos that were taped during the walk. They found that during actions in the real world the allocation of gaze depended much more on the current task requirements compared to free viewing conditions of the same video sequences in the laboratory.

Interestingly, based on these reasons one could assume that this should lead to higher position errors in the video conditions, especially for larger objects. However, if we compare the size of the target object in the video conditions, there seems to be no relationship with the magnitude of error (Figure 8.6a), indicating that there might be idiosyncratic differences in where people look for certain objects [de Haas et al., 2019] and for which objects these differences happen.

8.3.2 Saccade pursuit interaction

We have observed that when the saccade and subsequent pursuit are collinear the saccade position error is reduced and pursuit gain is closer to 1 (optimal). Because we measured the pursuit gain immediately after the saccade completion (50 to 150 ms) the increase in pursuit gain is unlikely to be the result of new retinal input. Up to 150 ms is often considered as the open-loop interval [Rasche and Gegenfurtner, 2009, Buonocore et al., 2019], where due to processing delays no new incoming retinal information is affecting the pursuit response. The benefit in pursuit gain could potentially be explained by muscle synergies, as for collinear eye movements the eye simply can keep moving in the same direction, whereas for the other two conditions the eye needs to decelerate more in order to change direction.

On the other hand, the increased saccade accuracy in the collinear case cannot be explained by muscle synergies or the saccade orientation (see Figure 8.3c). A possible explanation can be provided by early interactions between the saccadic and pursuit systems [Goettker et al., 2019], in which the saccade landing position is influenced by the subsequent pursuit direction. This interaction points to a shared network between the two [Deravet et al., 2018, Goettker et al., 2019] that increases the tracking performance during the transition phase from saccade to pursuit.

8.3.3 Effect of scene complexity on pursuit eye movements

Across the different levels of scene complexity we observed no significant differences with regard to pursuit gain, which suggests that results obtained with lab stimuli can be extrapolated to natural videos. A similar effect was observed across stimuli complexities between the relative angle of the initial and the subsequent pursuit. Also this similarity in pursuit gain suggests that the reported differences in the saccade position errors were probably affected by the larger object sizes in the natural experiments, which allowed the participants to target different areas of the same object.

Interestingly, we observed a clear benefit in terms of the pursuit direction error and its variability for the more complex scenes. This suggests that participants actually could make use of the additional information and the embedded context to improve their tracking performance. This is in line with recent evidence, showing that during tracking of a flying ball oculomotor control system is integrating the physical properties of the scene into its planning and it is adversely affected when some of these properties are artificially changed [Delle Monache et al., 2019]. Certain

expectations can also drive anticipatory pursuit [Kowler, 1989] and there is also evidence that prior knowledge is incorporated in the planning process [Darlington et al., 2017, Deravet et al., 2018]. Thus, our results suggest that prior knowledge of the constraints and behavior of physical objects allows for a better pursuit tracking in comparison to the tracking of an artificial Gaussian blob, which lacks any meaning and could move at seemingly random patterns.

One additional interesting suggestion is that larger objects lead to an increased pursuit performance and fewer catch-up saccades [Heinen et al., 2015]. In Figure 8.6b we visualize how the size of our 8 different targets is benefiting the pursuit accuracy. The negative error values indicate that the target tracking is more accurate in the naturalistic experiment than in the synthetic experiment and a positive value the opposite. Although not significant with our 8 scenes there seems to be a trend towards a benefit for the larger stimuli in the natural condition. Thus, the observed benefit to pursuit accuracy when tracking targets in natural videos seems to based on better motion integration due to larger object size and the use of prior knowledge [Watamaniuk and Heinen, 2015], which results in better prediction of the target motion.

8.4 Chapter conclusion

Thus taken it all together, we found that it is possible to compare and generalize oculomotor behavior across different stimulus complexities. However, some intricacies have to be noted: (i) Different levels of complexity lead to latency differences depending on how easily identifiable the targets are. (ii) It is difficult to measure positional accuracy for ill-defined and asymmetrical natural targets. (iii) Pursuit eye movements become more accurate for larger targets or if context information allows a better prediction of the target movement.

Chapter 9

Conclusion

Eye movements in dynamic natural contexts have been relatively underexplored in comparison to static scenes. Even when dynamic stimuli are used they are predominantly synthetic targets that usually translate in space and not videos of natural scenes presented on a monitor or a head-mounted display. Also oftentimes the eye movements that account for the dynamic nature of the content are disregarded and binned together with other eye movements. For example, researchers have defined SP as a fixation on a moving target by potentially assuming that SP comprises a negligible part of the overall gaze signal. Also a similar approach of detecting fixations and saccades only was followed by some of the first algorithms because they were either assuming static stimuli or they were intentionally merging SP with the other two. Only in recent years, algorithms that detect dynamic eye movements have been developed with varying levels of success.

To tackle the previous shortcomings, in this thesis we provide new data sets, algorithms, and new innovative applications of these. Our eye movement data sets are the largest to date hand-labeled data sets that span from videos of everyday scenes and Hollywood movies to immersive 360-degree content. We find that SP can be performed up to a quarter of the time on average across subjects. This observation shows the importance of separately labeling SP from fixations and saccades but also our results demonstrate the challenge of doing this automatically with simple thresholds due to the overlapping basic characteristics among the three. However, the large size of our data sets enabled us to develop and optimize more elaborate algorithms that achieved state-of-the-art performance across all eye movement types that were used in this thesis including SP, which is much more challenging to robustly detect in comparison to fixations and saccades. These algorithms reached human-level performance for fixation and saccade detection and overall high-quality

results for SP detection. Taken together our algorithms allow for robust annotation of eye movements without the need for tedious and time-consuming hand labeling.

More specifically, with our algorithms we were able to analyze the large studyforrest data set and to correlate SP with brain areas in a naturalistic free viewing experiment, which would have been very difficult and almost unattainable otherwise. Moreover, our labelings of the GazeCom data set together with new controlled recordings formed the basis for understanding the interactions between saccades and SP during the initiation phase of the latter across different modalities. To conclude, the infrastructure presented in this thesis (data sets, tools, algorithms) provides the foundation for a better understanding of human eye movements in naturalistic conditions since they enable the automatic analysis of large amounts of data recorded in more complex and ecologically valid environments.

Bibliography

[Adelson and Burt, 1980] Adelson, E. H. and Burt, P. J. (1980). *Image data compression with the Laplacian pyramid*. University of Maryland. Computer Science.

[Agtzidis and Dorr, 2019] Agtzidis, I. and Dorr, M. (2019). Getting (more) real: Bringing eye movement classification to HMD experiments with equirectangular stimuli. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, pages 18:1–18:8.

[Agtzidis et al., 2020a] Agtzidis, I., Meyhöfer, I., Dorr, M., and Lencer, R. (2020a). Following Forrest Gump: Smooth pursuit related brain activation during free movie viewing. *NeuroImage*, 216:116491.

[Agtzidis et al., 2016a] Agtzidis, I., Startsev, M., and Dorr, M. (2016a). In the pursuit of (ground) truth: A hand-labelling tool for eye movements recorded during dynamic scene viewing. In *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)*, pages 65–68.

[Agtzidis et al., 2016b] Agtzidis, I., Startsev, M., and Dorr, M. (2016b). Smooth pursuit detection based on multiple observers. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ETRA '16, pages 303–306.

[Agtzidis et al., 2019] Agtzidis, I., Startsev, M., and Dorr, M. (2019). 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 1007–1015.

[Agtzidis et al., 2020b] Agtzidis, I., Startsev, M., and Dorr, M. (2020b). Two hours in Hollywood: A manually annotated ground truth data set of eye movements during movie clip watching. *Journal of Eye Movement Research*, 13(4).

[Alaerts et al., 2013] Alaerts, K., Woolley, D. G., Steyaert, J., Di Martino, A., Swinnen, S. P., and Wenderoth, N. (2013). Underconnectivity of the superior temporal

sulcus predicts emotion recognition deficits in autism. *Social Cognitive and Affective Neuroscience*, 9(10):1589–1600.

[Allison et al., 2000] Allison, T., Puce, A., and McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, 4(7):267–278.

[Andersson et al., 2017] Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., and Nyström, M. (2017). One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*, 49(2):616–637.

[Andric et al., 2016] Andric, M., Goldin-Meadow, S., Small, S. L., and Hasson, U. (2016). Repeated movie viewings produce similar local activity patterns but different network configurations. *NeuroImage*, 142:613–627.

[Bahill et al., 1975] Bahill, A., Clark, M. R., and Stark, L. (1975). The main sequence, a tool for studying human eye movements. *Mathematical Biosciences*, 24(3):191 – 204.

[Barth, 2000] Barth, E. (2000). The minors of the structure tensor. In *Mustererkennung 2000*, pages 221–228. Springer.

[Barz, 2015] Barz, M. (2015). Pupil fixation detection. `https://github.com/pupil-labs/pupil/blob/master/pupil_src/shared_modules/fixation_detector.py`.

[Beauchamp et al., 2001] Beauchamp, M. S., Petit, L., Ellmore, T. M., Ingeholm, J., and Haxby, J. V. (2001). A parametric fMRI study of overt and covert shifts of visuospatial attention. *NeuroImage*, 14(2):310 – 321.

[Behrens et al., 2010] Behrens, F., MacKeben, M., and Schröder-Preikschat, W. (2010). An improved algorithm for automatic detection of saccades in eye movement data and for calculating saccade parameters. *Behavior Research Methods*, 42(3):701–708.

[Berg et al., 2009] Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., and Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision*, 9(5):1–15.

[Berman et al., 1999] Berman, R. A., Colby, C., Genovese, C., Voyvodic, J., Luna, B., Thulborn, K., and Sweeney, J. (1999). Cortical networks subserving pursuit and saccadic eye movements in humans: an fMRI study. *Human Brain Mapping*, 8(4):209–225.

[Brenner and Smeets, 2015] Brenner, E. and Smeets, J. B. J. (2015). How moving backgrounds influence interception. *PLOS ONE*, 10(3):1–21.

[Buonocore et al., 2019] Buonocore, A., Skinner, J., and Hafed, Z. M. (2019). Eye position error influence over open-loop smooth pursuit initiation. *Journal of Neuroscience*, 39(14):2709–2721.

[Bylinskii et al., 2016] Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2016). What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*.

[Cavanna and Trimble, 2006] Cavanna, A. E. and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583.

[Cohen, 1960] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

[Collewijn and Tamminga, 1984] Collewijn, H. and Tamminga, E. P. (1984). Human smooth and saccadic eye movements during voluntary pursuit of different target motions on different backgrounds. *The Journal of Physiology*, 351(1):217–250.

[Corbetta et al., 2000] Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., and Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience*, 3(3):292–297.

[Coutrot et al., 2012] Coutrot, A., Guyader, N., Ionescu, G., and Caplier, A. (2012). Influence of soundtrack on eye movements during video exploration. *Journal of Eye Movement Research*, 5(4):2. 10 pages.

[Curcio et al., 1990] Curcio, C. A., Sloan, K. R., Kalina, R. E., and Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, 292(4):497–523.

[Cutting et al., 2011] Cutting, J. E., Brunick, K. L., DeLong, J. E., Iricinschi, C., and Candan, A. (2011). Quicker, faster, darker: Changes in hollywood film over 75 years. *i-Perception*, 2(6):569–576.

[Dar et al., 2019] Dar, A. H., Wagner, A. S., and Hanke, M. (2019). REMoDNaV: Robust eye movement detection for natural viewing. *bioRxiv*. doi:10.1101/619254.

[Darlington et al., 2017] Darlington, T. R., Tokiyama, S., and Lisberger, S. G. (2017). Control of the strength of visual-motor transmission as the mechanism of

rapid adaptation of priors for bayesian inference in smooth pursuit eye movements. *Journal of Neurophysiology*, 118(2):1173–1189.

[David et al., 2018] David, E. J., Gutiérrez, J., Coutrot, A., Da Silva, M. P., and Callet, P. L. (2018). A dataset of head and eye movements for 360° videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 432–437.

[de Haas et al., 2019] de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., and Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, 116(24):11687–11692.

[Delle Monache et al., 2019] Delle Monache, S., Lacquaniti, F., and Bosco, G. (2019). Ocular tracking of occluded ballistic trajectories: Effects of visual context and of target law of motion. *Journal of Vision*, 19(4):13–13.

[Deravet et al., 2018] Deravet, N., Blohm, G., De Xivry, J.-J. O., and Lefèvre, P. (2018). Weighted integration of short-term memory and sensory signals in the oculomotor system. *Journal of Vision*, 18(5):16–16.

[Diaz et al., 2013] Diaz, G., Cooper, J., Kit, D., and Hayhoe, M. (2013). Real-time recording and classification of eye movements in an immersive virtual environment. *Journal of Vision*, 13(12):5–5.

[Dodge, 1904] Dodge, R. (1904). The participation of the eye movements in the visual perception of motion. *Psychological Review*, 11(1):1.

[Dodge et al., 1930] Dodge, R., Travis, R. C., and Fox, J. C. (1930). Optic nystagmus: III. characteristics of the slow phase. *Archives of Neurology & Psychiatry*, 24(1):21–34.

[Dorr et al., 2010] Dorr, M., Martinetz, T., Gegenfurtner, K. R., and Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, 10(10):28–28.

[Drewes et al., 2011] Drewes, J., Trommershäuser, J., and Gegenfurtner, K. R. (2011). Parallel visual search and rapid animal detection in natural scenes. *Journal of Vision*, 11(2):20–20.

[Drewes et al., 2014] Drewes, J., Zhu, W., Hu, Y., and Hu, X. (2014). Smaller is better: Drift in gaze measurements due to pupil dynamics. *PLOS ONE*, 9(10):e111197.

[Duchowski et al., 2002] Duchowski, A. T., Medlin, E., Cournia, N., Gramopadhye, A., Melloy, B., and Nair, S. (2002). 3D eye movement analysis for VR visual inspection training. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*, ETRA '02, pages 103–110.

[Eagle et al., 2007] Eagle, D. M., Baunez, C., Hutcheson, D. M., Lehmann, O., Shah, A. P., and Robbins, T. W. (2007). Stop-signal reaction-time task performance: Role of prefrontal cortex and subthalamic nucleus. *Cerebral Cortex*, 18(1):178–188.

[Einhäuser et al., 2008] Einhäuser, W., Rutishauser, U., and Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, 8(2):2–2.

[Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD Proceedings*, volume 96, pages 226–231.

[Ettinger et al., 2007] Ettinger, U., Ffytche, D. H., Kumari, V., Kathmann, N., Reuter, B., Zelaya, F., and Williams, S. C. R. (2007). Decomposing the neural correlates of antisaccade eye movements using event-related fMRI. *Cerebral Cortex*, 18(5):1148–1159.

[Everingham et al., 2015] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338.

[Ferman et al., 1987] Ferman, L., Collewijn, H., Jansen, T. C., and den Berg, A. V. V. (1987). Human gaze stability in the horizontal, vertical and torsional direction during voluntary head movements, evaluated with a three-dimensional scleral induction coil technique. *Vision Research*, 27(5):811 – 828.

[Findlay and Gilchrist, 2003] Findlay, J. M. and Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Number 37. Oxford University Press.

[Foulsham and Kingstone, 2010] Foulsham, T. and Kingstone, A. (2010). Asymmetries in the direction of saccades during perception of scenes and fractals: Effects of image type and image features. *Vision Research*, 50(8):779–795.

[Foulsham et al., 2011] Foulsham, T., Walker, E., and Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, 51(17):1920 – 1931.

[Fox and Dodge, 1929] Fox, J. C. and Dodge, R. (1929). Optic nystagmus: II. variations in nystagmographic records of eye movement. *Archives of Neurology & Psychiatry*, 22(1):55–74.

[Gagnon et al., 2006] Gagnon, D., Paus, T., Grosbras, M.-H., Pike, G. B., and O'Driscoll, G. A. (2006). Transcranial magnetic stimulation of frontal oculomotor regions during smooth pursuit. *Journal of Neuroscience*, 26(2):458–466.

[Gellman and Carl, 1991] Gellman, R. and Carl, J. (1991). Motion processing for saccadic eye movements in humans. *Experimental Brain Research*, 84(3):660–667.

[Giannopoulos et al., 2015] Giannopoulos, I., Kiefer, P., and Raubal, M. (2015). Gazenav: Gaze-based pedestrian navigation. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15, pages 337–346.

[Goettker et al., 2020] Goettker, A., Agtzidis, I., Braun, D. I., Dorr, M., and Karl, G. R. (2020). From gaussian blobs to natural videos: Comparison of oculomotor behavior across different stimulus complexities. *Journal of Vision*, 20(8):26–26.

[Goettker et al., 2019] Goettker, A., Braun, D. I., and Gegenfurtner, K. R. (2019). Dynamic combination of position and motion information when tracking moving targets. *Journal of Vision*, 19(7):2–2.

[Gold and Shadlen, 2007] Gold, J. I. and Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience*, 30(1):535–574.

[Goldberg and Schryver, 1995] Goldberg, J. H. and Schryver, J. C. (1995). Eye-gaze-contingent control of the computer interface: Methodology and example for zoom detection. *Behavior Research Methods, Instruments, & Computers*, 27(3):338–350.

[Grossman et al., 2005] Grossman, E. D., Battelli, L., and Pascual-Leone, A. (2005). Repetitive TMS over posterior STS disrupts perception of biological motion. *Vision Research*, 45(22):2847 – 2853.

[Grossman et al., 2010] Grossman, E. D., Jardine, N. L., and Pyles, J. A. (2010). fMR-adaptation reveals invariant coding of biological motion on human STS. *Frontiers in Human Neuroscience*, 4:15.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

[Hanke et al., 2016] Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., Nigbur, R., Waite, A. Q., Baumgartner, F., and Stadler, J. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Scientific Data*, 3:160092.

[Hasson et al., 2004] Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640.

[Hayhoe, 2017] Hayhoe, M. M. (2017). Vision and action. *Annual Review of Vision Science*, 3:389–413.

[Healey and Enns, 2011] Healey, C. and Enns, J. (2011). Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188.

[Heinen et al., 2015] Heinen, S. J., Potapchuk, E., and Watamaniuk, S. N. (2015). A foveal target increases catch-up saccade frequency during smooth pursuit. *Journal of Neurophysiology*, 115(3):1220–1227.

[Heinen and Watamaniuk, 1998] Heinen, S. J. and Watamaniuk, S. N. (1998). Spatial integration in human smooth pursuit. *Vision Research*, 38(23):3785 – 3794.

[Henderson et al., 2013] Henderson, J. M., Nuthmann, A., and Luke, S. G. (2013). Eye movement control during scene viewing: Immediate effects of scene luminance on fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, 39(2):318.

[Hessels et al., 2018] Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., and Hooge, I. (2018). Is the eye-movement field confused about fixations and saccades? a survey among 124 researchers. *Royal Society Open Science*, 5(8):180502.

[Heuer et al., 2013] Heuer, H. W., Mirsky, J. B., Kong, E. L., Dickerson, B. C., Miller, B. L., Kramer, J. H., and Boxer, A. L. (2013). Antisaccade task reflects cortical involvement in mild cognitive impairment. *Neurology*, 81(14):1235–1243.

[Holmqvist et al., 2011] Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., and Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

[Hooge et al., 2015] Hooge, I., Nyström, M., Cornelissen, T., and Holmqvist, K. (2015). The art of braking: Post saccadic oscillations in the eye tracker signal decrease with increasing saccade size. *Vision Research*, 112:55–67.

[Hooge et al., 2017] Hooge, I. T. C., Niehorster, D. C., Nyström, M., Andersson, R., and Hessels, R. S. (2017). Is human classification by experienced untrained observers a gold standard in fixation detection? *Behavior Research Methods*, 50:1864–1881.

[Hoppe and Bulling, 2016] Hoppe, S. and Bulling, A. (2016). End-to-end eye movement detection using convolutional neural networks. *arXiv preprint arXiv:1609.02452*.

[Huey, 1908] Huey, E. B. (1908). *The psychology and pedagogy of reading*. The Macmillan Company.

[Itti and Koch, 2000] Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489 – 1506.

[Itti et al., 1998] Itti, L., Koch, C., and 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.

[Jacobson and Dodwell, 1979] Jacobson, J. Z. and Dodwell, P. C. (1979). Saccadic eye movements during reading. *Brain and Language*, 8(3):303–314.

[Jastorff and Orban, 2009] Jastorff, J. and Orban, G. A. (2009). Human functional magnetic resonance imaging reveals separation and integration of shape and motion cues in biological motion processing. *Journal of Neuroscience*, 29(22):7315–7329.

[Javal, 1878] Javal, E. (1878). Essai sur la physiologie de la lecture. *Annales d'Ocilistique*, 80:61–73.

[Katsanis and Iacono, 1991] Katsanis, J. and Iacono, W. G. (1991). Clinical, neuropsychological, and brain structural correlates of smooth-pursuit eye tracking performance in chronic schizophrenia. *Journal of Abnormal Psychology*, 100(4):526.

[Kay et al., 2011] Kay, K. N., Naselaris, T., and Gallant, J. L. (2011). fMRI of human visual areas in response to natural images. `http://dx.doi.org/10.6080/K0QN64NG`.

[Kienzle et al., 2009] Kienzle, W., Franz, M. O., Schölkopf, B., and Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision*, 9(5):7–7.

[Kimmig et al., 2008] Kimmig, H., Ohlendorf, S., Speck, O., Sprenger, A., Rutschmann, R., Haller, S., and Greenlee, M. (2008). fMRI evidence for sensorimotor transformations in human cortex during smooth pursuit eye movements. *Neuropsychologia*, 46(8):2203 – 2213.

[Kleiner et al., 2007] Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, 36(14):1–16.

[Ko et al., 2016] Ko, H.-K., Snodderly, D. M., and Poletti, M. (2016). Eye movements between saccades: Measuring ocular drift and tremor. *Vision Research*, 122:93 – 104.

[Komogortsev, 2014] Komogortsev, O. V. (2014). Eye movement classification software. `http://cs.txstate.edu/~ok11/emd_offline.html`.

[Komogortsev et al., 2010] Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., and Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, 57(11):2635–2645.

[Komogortsev and Karpov, 2013] Komogortsev, O. V. and Karpov, A. (2013). Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior Research Methods*, 45(1):203–215.

[Kowler, 1989] Kowler, E. (1989). Cognitive expectations, not habits, control anticipatory smooth oculomotor pursuit. *Vision Research*, 29(9):1049–1057.

[Kowler, 2011] Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research*, 51(13):1457–1483.

[Kümmerer et al., 2016] Kümmerer, M., Wallis, T. S., and Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*.

[Lahnakoski et al., 2012] Lahnakoski, J. M., Glerean, E., Salmi, J., Jääskeläinen, I. P., Sams, M., Hari, R., and Nummenmaa, L. (2012). Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Frontiers in Human Neuroscience*, 6:233.

[Land and Hayhoe, 2001] Land, M. F. and Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, 41(25-26):3559–3565.

[Larsson et al., 2015] Larsson, L., Nyström, M., Andersson, R., and Stridh, M. (2015). Detection of fixations and smooth pursuit movements in high-speed eye-tracking data. *Biomedical Signal Processing and Control*, 18:145 – 152.

[Larsson et al., 2016] Larsson, L., Nyström, M., Ardö, H., Åström, K., and Stridh, M. (2016). Smooth pursuit detection in binocular eye-tracking data with automatic video-based performance evaluation. *Journal of Vision*, 16(15):20–20.

[Larsson et al., 2013] Larsson, L., Nyström, M., and Stridh, M. (2013). Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. *IEEE Transactions on Biomedical Engineering*, 60(9):2484–2493.

[Lencer et al., 2004] Lencer, R., Nagel, M., Sprenger, A., Zapf, S., Erdmann, C., Heide, W., and Binkofski, F. (2004). Cortical mechanisms of smooth pursuit eye movements with target blanking. An fMRI study. *European Journal of Neuroscience*, 19(5):1430–1436.

[Lencer and Trillenberg, 2008] Lencer, R. and Trillenberg, P. (2008). Neurophysiology and neuroanatomy of smooth pursuit in humans. *Brain and Cognition*, 68(3):219 – 228.

[Lisberger, 2015] Lisberger, S. G. (2015). Visual guidance of smooth pursuit eye movements. *Annual Review of Vision Science*, 1:447–468.

[Lord and Wright, 1949] Lord, M. P. and Wright, W. (1949). Small voluntary flicking and following eye movements. *Nature*, 163(4151):803–804.

[Lukasova et al., 2018] Lukasova, K., Nucci, M. P., Neto, R. M. d. A., Vieira, G., Sato, J. R., and Amaro, Jr, E. (2018). Predictive saccades in children and adults: A combined fMRI and eye tracking study. *PLOS ONE*, 13(5):1–17.

[Luna et al., 1998] Luna, B., Thulborn, K. R., Strojwas, M. H., McCurtain, B. J., Berman, R. A., Genovese, C. R., and Sweeney, J. A. (1998). Dorsal cortical regions subserving visually guided saccades in humans: An fMRI study. *Cerebral Cortex*, 8(1):40–47.

[MacAvoy et al., 1991] MacAvoy, M. G., Gottlieb, J. P., and Bruce, C. J. (1991). Smooth-pursuit eye movement representation in the primate frontal eye field. *Cerebral Cortex*, 1(1):95–102.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297.

[Mannion, 2015] Mannion, D. (2015). fMRI responses of human visual cortex (V1, V2, V3) to natural image patches obtained from above and below the centre of gaze of an observer freely-navigating an outdoor environment. `http://dx.doi.org/10.6080/K0JS9NC2`.

[Marsman et al., 2016] Marsman, J.-B. C., Cornelissen, F. W., Dorr, M., Vig, E., Barth, E., and Renken, R. J. (2016). A novel measure to determine viewing priority and its neural correlates in the human brain. *Journal of Vision*, 16(6):3–3.

[Martens and Fox, 2007] Martens, M. H. and Fox, M. (2007). Does road familiarity change eye fixations? A comparison between watching a video and real driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 10(1):33 – 47.

[Mathe and Sminchisescu, 2012] Mathe, S. and Sminchisescu, C. (2012). Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *European Conference on Computer Vision*, pages 842–856.

[Mills et al., 2011] Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., and Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, 11(8):17–17.

[Mital et al., 2011] Mital, P. K., Smith, T. J., Hill, R. L., and Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation*, 3(1):5–24.

[Mueller et al., 2013] Mueller, S., Wang, D., Fox, M. D., Yeo, B. T., Sepulcre, J., Sabuncu, M. R., Shafee, R., Lu, J., and Liu, H. (2013). Individual variability in functional connectivity architecture of the human brain. *Neuron*, 77(3):586–595.

[Munoz et al., 1998] Munoz, D., Broughton, J., Goldring, J., and Armstrong, I. (1998). Age-related performance of human subjects on saccadic eye movement tasks. *Experimental Brain Research*, 121(4):391–400.

[Nackaerts et al., 2012] Nackaerts, E., Wagemans, J., Helsen, W., Swinnen, S. P., Wenderoth, N., and Alaerts, K. (2012). Recognizing biological motion and emotions from point-light displays in autism spectrum disorders. *PLOS ONE*, 7(9):1–12.

[Nagel et al., 2006] Nagel, M., Sprenger, A., Zapf, S., Erdmann, C., Kömpf, D., Heide, W., Binkofski, F., and Lencer, R. (2006). Parametric modulation of cortical activation during smooth pursuit with and without target blanking. An fMRI study. *NeuroImage*, 29(4):1319–1325.

[Nardo et al., 2014] Nardo, D., Santangelo, V., and Macaluso, E. (2014). Spatial orienting in complex audiovisual environments. *Human Brain Mapping*, 35(4):1597–1614.

[Nyström and Holmqvist, 2010] Nyström, M. and Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, 42(1):188–204.

[Nyström et al., 2013] Nyström, M., Hooge, I., and Holmqvist, K. (2013). Post-saccadic oscillations in eye movement data recorded with pupil-based eye trackers reflect motion of the pupil inside the iris. *Vision Research*, 92:59–66.

[Ohlendorf et al., 2010] Ohlendorf, S., Sprenger, A., Speck, O., Glauche, V., Haller, S., and Kimmig, H. (2010). Visual motion, eye motion, and relative motion: A parametric fMRI study of functional specializations of smooth pursuit eye movement network areas. *Journal of Vision*, 10(14):21–21.

[Pack and Born, 2001] Pack, C. C. and Born, R. T. (2001). Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain. *Nature*, 409(6823):1040–1042.

[Petit and Haxby, 1999] Petit, L. and Haxby, J. V. (1999). Functional anatomy of pursuit eye movements in humans as revealed by fMRI. *Journal of Neurophysiology*, 82(1):463–471.

[Poldrack et al., 2011] Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge University Press.

[Rasche and Gegenfurtner, 2009] Rasche, C. and Gegenfurtner, K. R. (2009). Precision of speed discrimination and smooth pursuit eye movements. *Vision Research*, 49(5):514–523.

[Rashbass, 1961] Rashbass, C. (1961). The relationship between saccadic and smooth tracking eye movements. *The Journal of Physiology*, 159(2):326–338.

[Revaud et al., 2015] Revaud, J., Weinzaepfel, P., Harchaoui, Z., and Schmid, C. (2015). Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1164–1172.

[Robinson, 1965] Robinson, D. A. (1965). The mechanics of human smooth pursuit eye movement. *The Journal of Physiology*, 180(3):569–591.

[Rolls et al., 2015] Rolls, E. T., Joliot, M., and Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, 122:1 – 5.

[Rothkopf et al., 2007] Rothkopf, C. A., Ballard, D. H., and Hayhoe, M. M. (2007). Task and context determine where you look. *Journal of Vision*, 7(14):16–16.

[Salvucci and Anderson, 1998] Salvucci, D. D. and Anderson, J. R. (1998). Tracing eye movement protocols with cognitive process models. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pages 923–928.

[Salvucci and Goldberg, 2000] Salvucci, D. D. and Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, pages 71–78.

[San Agustin, 2010] San Agustin, J. (2010). *Off-the-shelf gaze interaction*. PhD thesis, IT-Universitetet i København.

[Sanocki et al., 2015] Sanocki, T., Islam, M., Doyon, J. K., and Lee, C. (2015). Rapid scene perception with tragic consequences: Observers miss perceiving vulnerable road users, especially in crowded traffic scenes. *Attention, Perception, & Psychophysics*, 77(4):1252–1262.

[Santini et al., 2016] Santini, T., Fuhl, W., Kübler, T., and Kasneci, E. (2016). Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, ETRA '16, pages 163–170.

[Saslow, 1967] Saslow, M. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *Josa*, 57(8):1024–1029.

[Sauter et al., 1991] Sauter, D., Martin, B. J., Di Renzo, N., and Vomscheid, C. (1991). Analysis of eye tracking movements using innovations generated by a Kalman filter. *Medical and Biological Engineering and Computing*, 29(1):63–69.

[Saygin, 2007] Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. *Brain*, 130(9):2452–2461.

[Schütz et al., 2011] Schütz, A. C., Braun, D. I., and Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, 11(5):9–9.

[Schütz et al., 2010] Schütz, A. C., Braun, D. I., Movshon, J. A., and Gegenfurtner, K. R. (2010). Does the noise matter? Effects of different kinematogram types on smooth pursuit eye movements and perception. *Journal of Vision*, 10(13):26–26.

[Sestieri et al., 2008] Sestieri, C., Pizzella, V., Cianflone, F., Romani, G. L., and Corbetta, M. (2008). Sequential activation of human oculomotor centers during planning of visually-guided eye movements: a combined fMRI-MEG study. *Frontiers in Human Neuroscience*, 2:1.

[Sharpe et al., 1979] Sharpe, J. A., Lo, A. W., and Rabinovitch, H. E. (1979). Control of the saccadic and smooth pursuit systems after cerebral hemidecortication. *Brain: A Journal of Neurology*, 102(2):387–403.

[Sitzmann et al., 2018] Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., and Wetzstein, G. (2018). Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1633–1642.

[Startsev et al., 2019a] Startsev, M., Agtzidis, I., and Dorr, M. (2019a). 1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits. *Behavior Research Methods*, 51(2):556–572.

[Startsev et al., 2019b] Startsev, M., Agtzidis, I., and Dorr, M. (2019b). Characterizing and automatically detecting smooth pursuit in a large-scale ground-truth data set of dynamic natural scenes. *Journal of Vision*, 19(14):10–10.

[Startsev and Dorr, 2018] Startsev, M. and Dorr, M. (2018). Supersaliency: Predicting smooth pursuit-based attention with slicing CNNs improves fixation prediction for naturalistic videos. *arXiv preprint arXiv:1801.08925*.

[Startsev et al., 2019c] Startsev, M., Göb, S., and Dorr, M. (2019c). A novel gaze event detection metric that is not fooled by gaze-independent baselines. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19, pages 22:1–22:9.

[Steil et al., 2018] Steil, J., Huang, M. X., and Bulling, A. (2018). Fixation detection for head-mounted eye tracking based on visual similarity of gaze targets. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18, pages 23:1–23:9.

[Tagliazucchi and Laufs, 2014] Tagliazucchi, E. and Laufs, H. (2014). Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. *Neuron*, 82(3):695–708.

[Tanabe et al., 2002] Tanabe, J., Tregellas, J., Miller, D., Ross, R. G., and Freedman, R. (2002). Brain activation during smooth-pursuit eye movements. *NeuroImage*, 17(3):1315 – 1324.

[Tatler et al., 2017] Tatler, B. W., Brockmole, J. R., and Carpenter, R. H. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review*, 124(3):267–300.

[Tatler et al., 2011] Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, 11(5):5–5.

[Thaler et al., 2013] Thaler, L., Schütz, A. C., Goodale, M. A., and Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, 76:31–42.

[Turuwhenua et al., 2014] Turuwhenua, J., Yu, T.-Y., Mazharullah, Z., and Thompson, B. (2014). A method for detecting optokinetic nystagmus based on the optic flow of the limbus. *Vision Research*, 103:75–82.

[Tychsen and Lisberger, 1986] Tychsen, L. and Lisberger, S. G. (1986). Visual motion processing for the initiation of smooth-pursuit eye movements in humans. *Journal of Neurophysiology*, 56(4):953–968.

[Tzourio-Mazoyer et al., 2002] Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1):273 – 289.

[Van Essen, 2005] Van Essen, D. C. (2005). A population-average, landmark-and surface-based (PALS) atlas of human cerebral cortex. *NeuroImage*, 28(3):635–662.

[Van Essen et al., 2001] Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., and Anderson, C. H. (2001). An integrated software suite for surface-based analyses of cerebral cortex. *Journal of the American Medical Informatics Association*, 8(5):443–459.

[Vanderwal et al., 2017] Vanderwal, T., Eilbott, J., Finn, E. S., Craddock, R. C., Turnbull, A., and Castellanos, F. X. (2017). Individual differences in functional connectivity during naturalistic viewing conditions. *NeuroImage*, 157:521 – 530.

[Vanderwal et al., 2015] Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C., and Castellanos, F. X. (2015). Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage*, 122:222 – 232.

[Vernet et al., 2014] Vernet, M., Quentin, R., Chanes, L., Mitsumasu, A., and Valero-Cabré, A. (2014). Frontal eye field, where are thou? Anatomy, function, and non-invasive manipulation of frontal regions involved in eye movements and associated cognitive operations. *Frontiers in Integrative Neuroscience*, 8:88.

[Võ et al., 2019] Võ, M. L.-H., Boettcher, S. E., and Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, 29:205 – 210. Attention & Perception.

[Walshe and Nuthmann, 2015] Walshe, R. C. and Nuthmann, A. (2015). Mechanisms of saccadic decision making while encoding naturalistic scenes. *Journal of Vision*, 15(5):21–21.

[Walther and Koch, 2006] Walther, D. and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407.

[Wang et al., 2018] Wang, W., Shen, J., Guo, F., Cheng, M.-M., and Borji, A. (2018). Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4903.

[Watamaniuk and Heinen, 2015] Watamaniuk, S. N. and Heinen, S. J. (2015). Allocation of attention during pursuit of large objects is no different than during fixation. *Journal of Vision*, 15(9):9–9.

[Westheimer, 1954] Westheimer, G. (1954). Eye movement responses to a horizontally moving visual stimulus. *AMA Archives of Ophthalmology*, 52(6):932–941.

[Wolfe et al., 2011] Wolfe, J. M., Võ, M. L.-H., Evans, K. K., and Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, 15(2):77–84.

[Wu et al., 2015] Wu, Q., Chang, C.-F., Xi, S., Huang, I.-W., Liu, Z., Juan, C.-H., Wu, Y., and Fan, J. (2015). A critical role of temporoparietal junction in the integration of top-down and bottom-up attentional control. *Human Brain Mapping*, 36(11):4317–4333.

[Yarbus, 1967] Yarbus, A. L. (1967). Eye movements during perception of complex objects. In *Eye Movements and Vision*, pages 171–211. Springer.

[Zemblys et al., 2018a] Zemblys, R., Niehorster, D. C., and Holmqvist, K. (2018a). gazeNet: End-to-end eye-movement event detection with deep neural networks. *Behavior Research Methods*, 51(2):840–863.

[Zemblys et al., 2018b] Zemblys, R., Niehorster, D. C., Komogortsev, O., and Holmqvist, K. (2018b). Using machine learning to detect events in eye-tracking data. *Behavior Research Methods*, 50(1):160–181.

[Zhong et al., 2013] Zhong, S.-h., Liu, Y., Ren, F., Zhang, J., and Ren, T. (2013). Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *Twenty-seventh AAAI Conference on Artificial Intelligence*, pages 1063–1069.