Algorithms for the comparison of visual scan patterns

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Thomas Christian Alexander Kübler
aus Leutenbach

Tübingen
2016

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	27.01.2017
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Jun.-Prof. Dr. Enkelejda Kasneci
2. Berichterstatter:	Prof. Dr. med. Ulrich Schiefer

Acknowledgments

This work is the result of three and a half years of work at the Computer Engineering Department of the Eberhard-Karls-University Tübingen and the study course Ophthalmic Optics & Audiology at the University of Aalen. I am grateful for this unique opportunity and the many possibilities and challenges that have arisen during this work. I would like to express my gratitude to the supervisors and mentors of this work, namely Prof. Dr. Wolfgang Rosenstiel, Prof. Dr. med. Ulrich Schiefer and Jun.-Prof. Dr. Enkelejda Kasneci, for their continued support and scientific advice throughout the years. I have enjoyed freedom in my scientific development as well as guidance and opportunities that would not have been possible without your experience.

I have much appreciated working together with my colleagues Wolfgang Fuhl, Thiago Santini and David Geisler as well as the whole Perception Engineering, Vision Research and Neuro-teams. Thank you all for making this an unforgettable time.

Parts of the work at hand were designed and carried out together with talented students with whom it was a pleasure to work with. My thanks goes to Judith Ungewiss, Guilherme Schievelbein, Dennis Bukenberger, Tobias Rittig and Colleen Rothe for their excellent work.

Finally, I would like to mention my family and friends with their crucial contributions to who I am today. Thank you Melly, Evi, Herbert, Meike, Eva and Richard.

> *It is only with the heart that one can see rightly;*
> *what is essential is invisible to the eye.*
>
> Antoine de Saint-Exupéry

Abstract

Our actions and intentions characterize the movement patterns of our eyes. Our visual exploration is driven by a mixture of cognitive processes and a conflict between the inspection of detail and the maintenance of an up-to-date overview. As a consequence, the determination of the influence of separate behavioral factors is challenging.

The work at hand examines how eye movement sequences can be compared to each other. This process is at the core of almost every eye-tracking study as it answers questions such as: "*Does gaze behavior of a patient group differ from his/her control group?*," "*How does experts' visual exploration differ from novices'?*," "*How does the composition of a painting influence the observer's gaze?*"
Therefore, several eye movement processing steps are revised: The issue of data quality is discussed with focus on methods and benchmarks to assure good quality during pupil detection, gaze mapping and eye movement identification in dynamic scenarios. Furthermore, eye-tracking data is integrated with physiological parameters such as ECG, galvanic skin conductivity and pupil dilation. The fusion of these complementary physiological sensors helps to disambiguate gaze and attention allocation.
This thesis proposes a novel method for the comparison of visual scan patterns, which is based on the frequency of short snippets of the whole eye movement sequence. Combined with current techniques in machine learning, the method is adaptable to a multitude of applications. Visualization and aggregation procedures for frequently traversed gaze trails are demonstrated on the basis of the finding that these short patterns are highly characteristic to many applications.

The proposed comparison technique is evaluated against state-of-the-art approaches on a new collection of data from a broad spectrum of eye-tracking experiments, ranging from static viewing tasks to highly dynamic outdoor scenarios. It effectively predicts the observer's task in a conjunction search task and in the more complex Yarbus experiment significantly above chance level. Furthermore, it is possible to assess driving fitness and the driver's secondary task, as well as to classify the expertise of neurosurgeons.

This new approach is able to identify the influence of a single experimental factor upon eye movement sequences. In contrast to all competing methods, it generalizes well over a broad spectrum of experimental designs.

Zusammenfassung

Unsere Handlungen und Absichten spiegeln sich in den Bewegungen unserer Augen wieder. Die visuelle Exploration wird von einem Mix an kognitiven Prozessen und dem Konflikt sowohl Details erkennen zu können, als auch einen aktuellen Überblick zu bewahren, angetrieben. Deshalb ist es schwierig den Effekt einzelner Verhaltensfaktoren zu isolieren. Diese Arbeit untersucht, wie Augenbewegungssequenzen miteinander verglichen werden können. Dieser Vergleich ist Herzstück fast jeder Eye-Tracking Studie und beantwortet Fragen wie: *"Unterscheidet sich das Blickverhalten einer Patientengruppe von der Kontrollgruppe?"*, *"Wie unterscheidet sich das Explorationsverhalten von Experten und Anfängern?"*, *"Wie wirkt sich die Komposition eines Bildes auf den Blick des Betrachters aus?"*

Dafür sind mehrere Verarbeitungsprozesse der Augenbewegungen notwendig: Die Frage der Datenqualität wird mit Schwerpunkt auf Benchmarks zur Sicherstellung guter Pupillenerkennung, Blickrichtungsbestimmung und Augenbewegungsklassifikation in dynamischen Szenarien behandelt. Außerdem werden Blickdaten mit anderen physiologischen Signalen, wie EKG, Hautleitwert und Pupillendurchmesser kombiniert. Die Fusion dieser sich ergänzenden Sensoren ermöglicht es Blick- und Aufmerksamkeitszuwendung voneinander zu trennen.

In dieser Arbeit wird ein neuer Algorithmus für den Vergleich von visuellen Blicksequenzen vorgestellt, der auf Häufigkeitsverteilung kurzer Teilsequenzen in der Gesamtsequenz basiert. Kombiniert mit einem maschinellen Lernverfahren kann sich diese Methode selbstständig an eine Vielzahl von Applikationen anpassen. Visualisierungsformen und Aggregationsmethoden für häufig genutzte Blickpfade, die ebenfalls auf der Annahme basieren, dass diese kurzen Teilsequenzen charakteristisch für viele Anwendungen sind, werden demonstriert.

Die vorgestellte Methode wird auf einem neuen Datensatz, der ein breites Spektrum an typischen Eye-Tracking Experimenten enthält, gegen den Stand der Technik evaluiert. Diese Daten enthalten sowohl statische, als auch hochdynamische Realszenarien. So kann die Aufgabenstellung des Betrachters während einer seriellen Suchaufgabe aber auch während der komplexeren Aufgabenstellung des Yarbus Experiments signifikant über der Ratewahrscheinlichkeit klassifiziert werden. Außerdem ist es möglich Fahrtüchtigkeit und durchgeführte Nebenaufgabe eines Fahrzeugführers, sowie den Erfahrungslevel von Chirurgen zu bestimmen.

Dieser neuartige Ansatz ist in der Lage den Einfluss einzelner Faktoren auf eine Blicksequenz zu identifizieren. Im Gegensatz zu anderen Methoden generalisiert der Ansatz auf ein breites Spektrum an experimentellen Designs.

Contents

1 Introduction

1.1 How do we explore our surroundings?

The world that surrounds us contains decidedly too much visual information for our eyes and brain to process all at once. Of the 10^{10} bit/s deposited in the retina, only 10^4 make it to layer IV of the visual cortex. Therefore, humans have adapted a strategy of selective attention and "what we see" is supposedly in the range of 100 bit/s [1]. Perception is concentrated on objects that are considered currently relevant and withdrawn from less important ones.
This strategy is reflected in the anatomy of our visual system. The human eye is foveated, meaning that optimal visual perception is possible only within a small area of the retina: the fovea.
Our visual field, i.e. the area that can be seen without an eye movement, extends over 200° horizontally [2]. But as we move from the fovea towards the periphery, visual acuity drops rapidly. At an eccentricity of 5° from the fovea, only 50% of visual acuity is attained [3]. Figure 1.1 visualizes this rapid decline due to the amount of cortical issue devoted to processing input of a retina area and its distribution of receptors.

(a) Original image (b) Foveated version

Figure 1.1: (a) Original photograph taken from the Neckarbrücke in Tübingen and (b) with superimposed spatial resolution capacity of the human visual system. The effect is exaggerated for the purposes of exposition, when compared to the human visual system. If the blur would exactly correspond to the visual acuity decrease, the whole image would appear equally sharp to an observer looking at the center of the image

Consequently, our eyes are constantly in motion. Fast eye movements (saccades), which we are mostly unaware of, shift the fovea towards different parts of the world around us. This process of attention shift by eye movements is called *overt attention*. Complementary, *covered attention* shift is directed towards peripheral, low-resolution regions of the visual

field and is not necessarily linked to eye movement. However, it is usually performed before (and in preparation of) a saccade that targets the location of covered attention [4].
The small fovea and the relatively wide field of view with reduced perceptive capability are the outcomes of an evolutionary optimization as both are associated with high energy and brain capacity demands. Yet, our brain assembles a comprehensive impression of our surroundings, stitched together like a jigsaw from a sequence of attention foci.
The inspection of detail and the maintenance of an up-to-date overview of our large visual world are both essential, yet opposing goals. This internal struggle drives our visual exploration behavior to perform an average of three saccades per second [5].
The resulting spatiotemporal sequence of eye movements is known as the visual *scanpath*, a term coined by Noton and Stark [6]. In their scanpath theory a direct link between internal cognitive representation and eye motor control was suggested.

1.2 Differences in visual scanning behavior

Buswell [7] was probably the first to hypothesize that *"the mental set obtained by the directions given [...] obviously influences the characteristics of the perceptual process"*. The suggested connection between eye movements and cognitive processes was elaborated in a famous experiment by Yarbus. His subject, who was instructed to look at a painting (Figure 1.2) and was asked to perform seven different tasks [8], exhibited task-specific scanning patterns that differed clearly for each instruction.

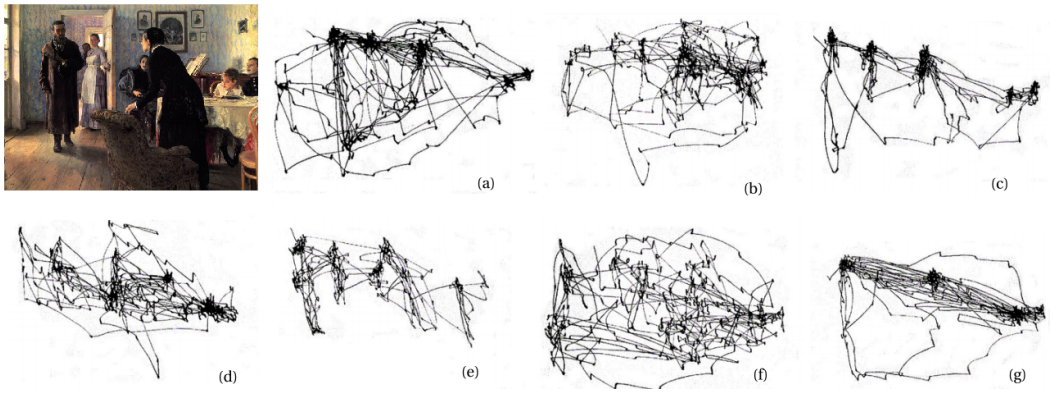


Figure 1.2: Eye movements of one subject examining Ilja Repin's *The Unexpected Visitor* for three minutes with different tasks assigned: (a) free examination, (b) estimate the material circumstances of the family, (c) estimate the age of the people, (d) surmise what the family had been doing before the arrival of the unexpected visitor, (e) remember the clothes worn by the people, (f) remember the position of the people and objects in the room and (g) estimate how long the unexpected visitor had been away from the family. Image source: [8], adapted from A.L. Yarbus: Eye movements and vision. 1967, New York Plenum Press.

However, in replications of this experiment the supposedly intuitive finding turned out to be a lot less distinctive than expected [9, 10]. Although behavioral patterns manifest in the scanpath, it is typically a challenging task to identify them.

A high natural variability adds up with inaccuracies of the measurement device, ocular properties [11] and a variety of other factors, such as damage of the visual system [12, 13], autism, eating disorders [14], or previous experience with a task [15]. All these factors exert a mixture of influences on the scanpath.
A common tool for the visual analysis of viewing behavior are heatmap visualizations (Figure 1.3). Areas hit by gaze gradually become *hot*, whereas areas that are not looked at remain *cold*.

(a) Original image (b) Heatmap

Figure 1.3: Heatmap of gaze samples recorded from nine subjects viewing Johannes Vermeer's *The Art of Painting* (detailed experiment description in Section 4.1.2). The color overlay represents the amount of gaze distributed towards the respective area. Often looked-at locations are represented by warm colors. The scale indicates the number of gaze samples required to aggregate at one location in order to produce the respective color.

A best practice for the construction of heatmaps is to average over at least 30 recordings; only then convergence of the heatmap is achieved [16]. This relatively high number of viewing passes stresses the large natural variability in the data.

1.3 Comparison of scan patterns: Challenges

The factors that influence eye scanning behavior are manifold. They make the analysis of eye-tracking data difficult - and rewarding at the same time. If we were able to infer a pathological state from eye movements, we could utilize an eye-tracker to measure eye movement biomarkers for medical diagnostics. They could be used to improve teaching of specific tasks to novices and would generally open a window to monitoring and understanding cognitive processes.

For this reason, automated methods for the objective comparison of eye movement sequences are required. We need to be able to identify and quantify features that both typify and distinguish between scanpaths.
To date, most approaches of scanpath analysis are based on a radical reduction of the whole scanpath to one characteristic measure, such as the average fixation duration or the number of fixations directed towards a specific region of interest (ROI). Thus, one of the most decisive attributes of a scanpath is ignored: its sequential nature, i.e., the order in which objects are looked at.
While for some experiments the analysis of fixation density is sufficient, in most cases the exact sequence of fixations is essential, e.g., does gaze follow composition lines when viewing fine art [17], during the exploration of driving scenarios for potential hazards [18], or in the context of activity recognition [19].
We are currently lacking the tools to harvest the information contained in the scanpath. The level of generalization of comparison metrics to other experimental contexts and applications (apart from the specific experiment they were designed for) remains mostly unexplored.

1.4 Contributions

An introduction to eye-tracking technology and the algorithms required therefor is provided in Section 2. I investigate data quality issues caused by technical restrictions of the eye-tracker. The first systematic study on the influence of eyeglasses on gaze mapping accuracy as well as novel methods for eye movement event detection are discussed in the first part of Section 3. The second part of this section is dedicated to the impact on data quality caused by attention constraints of the subject being tracked and the interaction of multiple physiological sensors.
Section 4 introduces novel data visualizations and gaze trail aggregations that focus on gaze trajectories instead of fixation allocation. These methods are applied to the viewing of fine art to provide insights about the correlation of composition principles and eye movements.
At the core of this thesis is a new method for the algorithmic comparison of general scanpaths. Subsequences, or short fragments of the original scanpath, are embedded in a string kernel support vector machine. The construction of the subsequences reduces dimensionality and compensates for common measurement inaccuracies, but conserves a notion of temporal sequence. The machine learning approach enables the method to adapt to different comparison problems. The method is introduced in Section 5 and extended with the machine learning step in Section 6. An extensive evaluation is performed in Section 6.3. Therefore, a broad spectrum of experimental designs, ranging from simple image viewing to complex real-world driving experiments, shows how the shortcomings of state-of-the-art scanpath comparison methods can be overcome. Section 8 summarizes and concludes this work.
This research has been presented in renowned conferences and published in scientific journals; the classification of eye movements was published in the *Proceedings of the Symposium on Eye Tracking Research and Applications* [20] (ETRA 2014), *Proceedings of the 23rd International Conference on Artificial Neural Networks* [21] (ICANN 2013) as

a follow-up book chapter *Artificial Neural Networks* [22], and in the *Proceedings of the Symposium on Eye Tracking Research and Applications* [23] (ETRA 2016).
Articels on data quality, pupil detection and recording software were published in the *Proceedings of the Symposium on Eye Tracking Research and Applications* [24, 25] (ETRA 2016), the *Proceedings of the 11th International Conference on Computer Vision Theory and Applications* [26] (VISAPP 2016) and *Computer Analysis of Images and Patterns* [27].
The automated creation of regions of interest was presented at the *5th European Workshop on Visual Information Processing* [28] (EUVIP 2014) and published in the *Proceedings of the 2nd International Workshop on Solutions for Automatic Gaze Data Analysis* [29] (SAGA 2015) and the *European Conference on Eye Movements* [30] (ECEM 2015).
A study on the fusion of physiological sensors with eye-tracking data was published in *Transportation Research Part F: Traffic Psychology and Behaviour* [18].
The visualization techniques (included in the Eyetrace software) were published in the *Proceedings of the International Conference on Health Informatics* [31] (HEALTHINF 2015), as a book chapter in *Communications in Computer and Information Science* [32], and were presented at the *ECCV Workshop* [33] (VISART 2016).
Methods for scanpath comparison were published in the *Proceedings of the Symposium on Eye Tracking Research and Applications* [34] (ETRA 2014) and Behavior Research Methods [35]. The application to driver monitoring was presented at *15. Internationales Stuttgarter Symposium Automobil- und Motorentechnik 2015* [19] and the application to microsurgery was presented in the *Workshop on Interventional Microscopy* [15] (MICCAI 2015). Studies on gaze and driving behavior of patients with visual field defects were published in *Optometry & Vision Science* [12] and in the Journal of Eye Movement Research [13].

2 Background

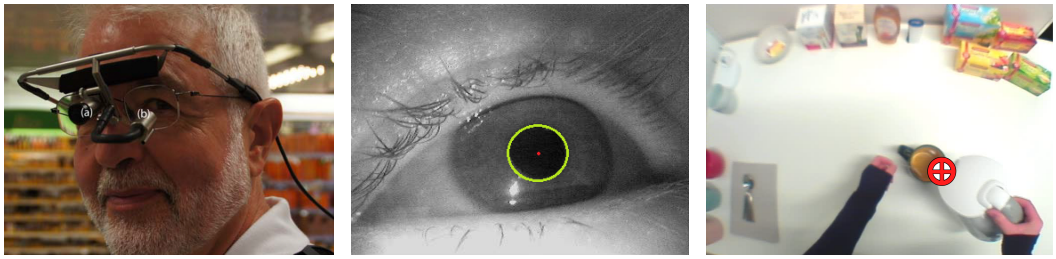
Eye-tracking has been a topic in academia and marketing research for quite a while [36]. Video based devices have reached a level where they can be handled by naïve users. Driven by this technological progress, eye-tracking technology it is now moving towards the mass market. Samsung and Microsoft include eye detection and interaction technologies in smartphones and tablets; the EyeTribe [37] and Tobii offer low-cost devices for video gaming; Oculus Rift and FOVE want to include eye-tracking in their virtual reality goggles. And as the hardware evolves, so need the algorithms. Whilst researchers enjoy the freedom to adjust and learn about parameters of the algorithms they employ, for everyday use we expect the devices to *just work*. Eye trackers need to become more robust and to adapt to different users and use-cases automatically.

2.1 Eye-tracking

Capturing gaze by placing a special contact lens onto the eye and measuring induced magnetic flux or by directly measuring the electromagnetic variation of the eyeball musculature dipole [36] has almost vanished. Nowadays video-based eye-tracking has superseded other technologies for most applications. Taking a picture of the eye is non-invasive and has minimal impact on normal viewing behavior.
This work focuses on video-based eye-tracking. Some methods and algorithms, such as pupil detection and calibration, are specific to this technique. Algorithms for event detection and scanpath comparison are relatively independent of the recording mode. They are likely to generalize well over other devices and techniques.

2.1.1 Video based eye-tracking

Head-mounted eye trackers such as the Dikablis device (Ergoneers GmbH, Manching Germany) shown in Figure 2.1 consist of at least two cameras. One to record a video of the infrared-illuminated eye from close-up, the other directed towards the region the subject is facing. The so-called scene camera is usually placed centered over the subject's nose like a cyclops eye.
There are two computational steps required to make an eye tracker work: First, the position of the pupil center in the image of the eye has to be detected. Figure 2.1(b) shows this pixel position as a small red dot in the midst of a green pupil ellipse.
Second, the transformation of the pupil center to a gaze point in the scene image. This pixel position in the scene image is shown as a red cross-hair in Figure 2.1(c). Gaze estimation usually requires a calibration process. During calibration, the subject needs to look at several points with a known pixel correspondence in the scene image (either by manual selection or



(a) Dikablis essential (b) Eye camera image (c) Scene camera image

Figure 2.1: The head-mounted Dikablis essential by Ergoneers GmbH, Manching/Germany, consists of two cameras, one directed towards the subject's eye, one facing straight forward, resembling the wearer's perspective. They are further called eye and scene camera. Pupil center coordinates in the eye camera image can be mapped to gaze positions in the scene camera image

by detection of a computer vision marker). These scene and eye image correspondences are then used to determine a mapping function. The process of transforming pupil center pixels to gaze location pixels is therefore often also called gaze mapping.

Most devices use infrared illumination as a reliable lighting, since it is invisible to the human eye and comparatively robust to ambient illumination. The infrared LEDs (IR-LED) are reflected at the inner and outer borders of the cornea and lens. These reflections are called Purkinje images. The first Purkinje image, i.e., the one at the outer cornea surface, is also called glint. These reflections can be used to simplify the calibration process, if the exact location of the LEDs relative to the camera is known. Given enough such reflections, the optical axis, i.e., the symmetry axis of the lens, can be calculated. The fovea is not located exactly on the optical axis, but at a temporal offset of roughly 5°. In order to determine the exact angular offset, a one-point calibration is sufficient. Each glint and camera allows us to determine a common plane on which all, the camera, pupil center, and the cornea center are located. By adding cameras or IR-LEDs, several planes can be intersected. For geometrical reasons, at least two such glints and for visibility reasons mostly the Purkinje image of the outer cornea surface are used in conjunction with a population average of the cornea radius in order to solve the resulting equation system.

A distinction can be made between the above head-mounted eye trackers and remote devices. Remote trackers do not require any equipment to get in touch with the subject. Their cameras capture an image with a wide field of view that contains the whole face (and usually some area around it). From an algorithmic point of view, a face detection step is added and the eye region is cropped from the large image. Some devices also add a processing step for the compensation of head movements.

One can further distinguish between monocular and binocular tracking, i.e., observing only one or both eyes. While for healthy sighted individuals both eyes should normally be directed towards the same object, more information than a mere redundancy can be gained from this signal. For example, the vergence angle (see Section 3.2.2.2) contains information about the distance towards a fixation target. Ocular misalignment may also be a pathological state (see Section 7).

In a nutshell, eye-tracking requires an image of the eye region in which the pupil can be detected. Through a calibration process the gaze orientation can be derived. Distinct eye movement events can be extracted from a temporal sequence of gaze orientations.

2.1.2 Pupil detection

As the first algorithmic step involved in eye-tracking, pupil detection has to be performed. Detecting the pupil in an image of the eye is an easy task for a human. But computer vision has a long way to go until the excellence of human pattern matching is reached.
The Visual Search Examination Tool Vishnoo [38] implements the popular Starburst algorithm [39]. Starburst sends out rays in multiple directions and collects all locations with a large difference of the intensity in consecutive pixels. The mean position is calculated and repeated until convergence.
Starburst has several severe issues. Tracking a pupil under laboratory conditions works properly. But once a tracking loss occurs, for example as consequence of a fast eye movement or a reflection at the pupil border, it re-initializes the pupil location with a strong center bias. This leads to frequent false detections. On real-world data that includes extreme viewing angles and reflections on the eyeball and eyeglasses, Starburst is unreliable [27].
Newer, more reliable algorithms do exist. Excuse [27] and Else [25] were created especially for coping with the difficult real-world conditions. They are implemented in the recording tool Eyerec [26].
Once the position of the pupil in an image of the eye is known, gaze orientation can be computed through a calibration process. This is covered in detail in Section 3.1.3. Therefore, we move directly onward to the topic of detecting eye movement events from a temporal sequence of gaze orientations.

2.1.3 Eye movement events

One of the early and most essential processing steps for eye-tracking data is the identification of fixations, saccades and smooth pursuit eye movements. The distinction of these movement types is necessary, since they are associated with different cognitive processes and implications for perception and attention allocation:

Fixation
During a fixation the eye position is relatively stable (with a spatial dispersion of $< 2^\circ$). Visual information is perceived during the duration of a fixation. An average of three fixations per second are performed when viewing a naturalistic scene [5]. But fixation rate and duration vary considerably between individuals and tasks [40]. Typical fixation durations range from 200-300 ms, but can last anywhere from some tens of milliseconds up to several seconds [36].
Small ocular movements, namely microtremor, drift and microsaccades, occur within a fixation. They are believed to counteract neural adaptation to a constant stimulus by repeatedly shifting the projected area on the retina [41]. Amplitude and frequency of these movements are below the measurement accuracy of many eye trackers. These

micro-movements are not directly relevant for the topics covered by this work.

Saccade
Saccades are very fast ballistic eye movements. They shift both eyes simultaneously towards the same direction (a version movement). The eyeball rotates at 30-500°/s. Peak velocity is so fast that visual perception is suppressed during most of the saccade to avoid large-field motion on the retina and to maintain perceptual stability [42]. Because saccadic suppression reduces visual sensitivity for a very short duration [43], saccades are often filtered from eye-tracking data as *not relevant* for perception and attention. Saccades are programmed before the beginning of the movement and the ballistic nature may result in an inaccuracy of the landing position and not hit the intended target. A corrective saccade may follow [44]. A glissade is a slow drift occasionally found at the end of a saccade. They might be produced by mismatches in the pulse involved in generating the saccade and the breaking pulse [45].

Smooth pursuit
Smooth pursuits are relatively slow (10-30°/s) ocular movements [36]. They are utilized to follow a moving target. Pursuits cannot be performed without a target to follow. They imply hard attention constraints on the moving target. If the eyes cannot keep up with a fast moving target, catch-up saccades are interspersed. In terms of velocity, movements caused by the vestibulo-ocular reflex - a compensation mechanism to maintain a fixation during head movement - may appear similar to a smooth pursuit. However, it is caused by an involuntary reflex and can reach much higher velocities.

2.1.3.1 Identification of eye movements

There are three general concepts for fixation and saccade filtering: velocity filtering, dispersion filtering, and probabilistic models. A multitude of different methods and implementations exist for each of these categories. A comprehensive review can be found in [36]. In the following a short, exemplary overview is provided:

Velocity based algorithms, such as the *Velocity-Threshold Identification* (I-VT) algorithm, separate eye movements by thresholding the velocity the eye is moving with. If the eye is moving *fast enough* a saccade is detected, otherwise a fixation. The challenging part is the *correct* definition of *fast enough*. Eye trackers are often calibrated to a pixel location on a screen or in a scene camera image. These pixel locations cannot always be directly converted to angular orientations of the eyeball. The distance between the image and the subject's eye that would be required for the conversion is not measured by the eye tracker and may vary during an experiment. Therefore, the choice of a threshold can be more challenging than simply using the literature values provided in the definition of fixations and saccades. Furthermore, the ideal threshold is highly individual and may even change during the experiment.
Dispersion filters distinguish eye movements by the distance between samples. For example, the *Dispersion Threshold Identification* (I-DT) algorithm considers the dispersion within a

temporal window [46]. Dispersion of samples within a fixation can be expected to be small; saccades will contribute to a large dispersion. The tricky part when using a dispersion filter is, just as for the velocity filter, the choice of the cutoff threshold between *small* and *large* dispersion.
The *artistic act* of parameter choice can be substituted by a data-driven procedure. The *Bayesian mixture of Gaussian model* (I-BDT) determines an optimal decision boundary by a probabilistic approach: two Gaussian distributions are trained, based on the data seen so far: one for the small velocities occurring during fixations, one for the large velocities of saccades. For each new sample the probability of belonging to either of the distributions can easily be determined by lookup in the probability density function of the distributions. Furthermore, new data can be used to update the distributions, allowing for a change in the decision boundary during execution [21, 20].

The above filtering algorithms distinguish only between fixations and saccades. But they all have straight-forward extensions for the detection of smooth pursuits: a second velocity threshold (I-VVT) or dispersion threshold (I-VDT) can be introduced [47]. The second threshold is located somewhere in-between the expected fixation and saccade velocities.
Berg et al. [48] proposed a method based on the assumption that samples recorded during a fixation should spread roughly circular around the fixation center. A saccade on the contrary results in a large variation of sample locations that is stretched in the direction of the saccade. This implies a large dispersion in the saccade direction and a very small spread in the orthogonal direction. By analyzing the ratio between the first and second principal component of samples within a moving time window, smooth pursuits can be identified (Principal Component Analysis Identification (I-PCA)). The principal components point to the directions of largest variance within the data. A more detailed description is given in Section 3.1.1.1.
For high-speed tracking data, an algorithm based on the Rayleigh test (I-VMPRay) was extended with four different spatial features (dispersion, consistent direction, positional displacement, and spatial range) [49]. There are also machine learning approaches on features such as velocity, slope and variance [50].

So we switch the focus of our visual attention by a quick sequence of fixations and saccades. During this exploration process, fixations are not randomly distributed but directed towards those regions that we consider currently relevant. At this point we make a transition from physiological eye movements to the semantic entity of *what* we are looking at.

2.2 Region of interest

A *region of interest* (ROI) is a distinct subregion of the stimulus that a researcher is particularly interested in. The concept of ROIs has its origin in the assumption that our brain's informational units are not spatial locations (as the eye tracker provides) but rather semantic object entities. For example the position and shape of a bouncing rubber ball changes whilst the entity of the ball does not.

Considering ROIs instead of positions adds a semantic layer and meaning to the data. The analysis is likely to be more robust as objects can be moved and interacted with. However, it comes at a cost: whilst humans can perceive, recognize, and track arbitrary objects with minimal effort, automated tracking via computer vision is computationally expensive and error-prone.

Emergent from the above definition of a ROI is that hypotheses of the researcher defining the ROIs are introduced into the data. Potential findings are thereby limited to the hypotheses and ROIs chosen, whilst other effects might be disregarded. A mask is put onto the data and only the areas visible through that mask are considered.

Even worse, researchers with identical hypotheses might come to different conclusions depending on the details of their ROI annotation. In [51] for example, the authors describe the problem of defining the exact boundaries and subregions of a human face and how slightly different annotations can lead to diverging results.

Generating ROIs for static stimuli

Besides all its downsides, the gold standard in generating ROIs is manual annotation. For static stimuli this procedure is feasible but highly time consuming. Every stimulus is annotated once. That annotation can then be used for every trial involving the stimulus.

A both trivial and practical way of computational ROI definition is to overlay a rectangular grid over the stimulus. Each segment of the grid is considered a separate ROI. Many researchers have criticized this method, since fixations close to each other and on the same semantic entity can be divided by the arbitrarily chosen ROI boundaries. A rectangular grid is only adequate for very few stimuli, such as a game of memory. Otherwise, the ROI division might not be semantically meaningful. It is often overlooked that such a semantic entity division might happen, but it is not the common case: fixations to the same object are usually close to each other ans therefore very likely to be assigned the same ROI label. Because of this the approach works for many practical applications as long as no motion is involved.

Alternatively, the stimulus can be processed to extract ROIs that correspond to more coherent regions. Computer vision algorithms can segment a stimulus image by low level features such as color, contrast, and edges [52]. These features are indicators of *objectiveness*. Sharp edges, color and contrast changes are likely to correspond to object boundaries. Object segmentation by computer vision algorithms is still imperfect, but even for improved algorithms the question of the desired level of detail remains. Are we interested to know where faces are located or do we want to distinguish between eyes and noses. Manual annotation is likely to result in an appropriate level of detail automatically, as the researcher utilizes his expectations about the results.

One solution to circumvent the computer vision step is clustering of the recorded eye movement data. Thereby, the semantic knowledge of the observers can be utilized to infer the location of relevant objects. In [53] an algorithm based on mean-shift clustering was introduced for this purpose, while in [54] k-means clustering is employed.

ROI annotation in dynamic settings

For dynamic stimuli, manual annotation is usually prohibitively time-consuming. For experiments where multiple subjects are watching the same video, it is still practicable: ROIs have to be defined for each frame. But as the scenario turns more dynamic, e.g., by allowing the subject to interact with the scene, the annotation effort increases dramatically: ROIs are then located differently for each subject (subjects can actively change their environment and the position of objects. Allowing head movements will totally change the stimulus material recorded by a head-mounted eye tracker and the relative location of all objects to the scene camera). A separate annotation not only for each frame, but for each frame of each subject/trial is necessary. Each of the (on average) three fixations per second has to be assigned to an object manually.

Manual annotation

Software manufacturers provide tools to help map fixations to ROIs or a reference image manually. By providing convenient graphical user interfaces (GUI) they promise speed-up factors of 10×-50× (SMI semantic gaze mapping technology). But a better GUI can simplify the task only marginally - it still boils down to aiming and clicking the mouse once per fixation (as shown in Figure 2.2). The phenomenal speed-up is almost completely due to the annotation of fixations instead of raw samples (a 20× speed-up for a 60 Hz tracker and 300 ms average fixation duration). Annotating fixations only results in similar dwell times as the annotation of each sample [55] (the cited study finds a high correlation even though they include samples recorded during saccades).

Figure 2.2: Manual ROI annotation of a tea cooking experiment (based on [56]): fixations (blue circle) in the scene camera image (right) are manually annotated by mouse clicks in a reference image (left)

Marker anchored ROIs

A partially automated approach is anchoring ROIs to computer vision markers (Figure 2.3). These markers are easy to detect by a computer. Based on the markers' appearance in the camera image, the 3D position and orientation of the camera relative to the marker can be

computed. Any other object with known position with respect to the marker can be virtually projected to the camera image. The intrinsic parameters of the camera are required (the parameters of the lens and sensor that are responsible for the camera's projection) in order to compute this projection.
Gaze can then be assigned to marker-anchored objects automatically. Essentially, a hardly traceable object was replaced by an easily traceable marker. Placing markers is not always possible. The major drawback is that these markers are by definition quite salient: in order to be easily detectable, they have to be distinguishable from their surrounding. Placing markers on very small or many objects can be difficult - for example a supermarket shopping experiment might require one marker per apple in the shelf.

Figure 2.3: Computer vision markers (highlighted by green squares) in the interior of a car. The gaze of the driver can be assigned to ROIs anchored to these markers. The red overlay visualizes the virtual ROI area of the dashboard

3 Data quality

The previous overview of algorithms involved in the recording and analysis of eye movement data leads inevitably to the question of data quality. How can we tell, whether the algorithms performed well and our data is of good quality? This section discusses data quality from two different points of view: Section 3.1 deals with signal quality, involving everything about the quality and performance of the involved algorithms and hardware. Was the pupil detected continuously, were eye movements identified correctly and is the calibration accurate?
Sections 3.2 and 3.3 are dedicated to the subject whose eyes are being tracked. Even a recording with perfect signal quality can be useless, if the subject is unwilling or unable to perform an experiment. Defects in the perceptual pathway or a gradual decrease in vigilance might result in useless but high signal quality data. Parameters of sustained attention that can easily be recorded and analyzed alongside the eye-tracking signal are introduced. Further, physiological parameters useful for the disambiguation of performance and perceptual failures are discussed.

3.1 Signal quality

3.1.1 Fixation filters and smooth pursuit movements

A lot of effort has been put into the *optimal* parameter choice of fixation identification algorithms and best practices have evolved [36]. Parameters were found to depend on technical characteristics of the eye-tracking device, such as the sampling rate [57, 58], but also to be task specific. Several studies discuss the influence of these parameters on key metrics [46]. Exact fixation identification is essential for the calculation of many characteristic values in eye-tracking research, such as the average fixation duration or saccade amplitudes. Consensus is that all methods are sensitive to their parameters and a comparison between studies is only applicable for identical algorithm and parameter choices.

3.1.1.1 Fixation filter validation

The question that arises from the multitude of fixation identification algorithms and their sensitivity to parameter choices is: are the fixations in a given data set identified *correctly*? As determining the optimal parameters for a fixation identification filter is non-trivial, Holmqvist et al. [36] suggest visual verification of the results by plotting the raw data next to the identified fixations.
The Eyetrace software developed during this work [31, 32] offers an extensive collection of data visualizations for eye-tracking recordings that can answer this question of data quality

on two layers:
I) Is a recording of good signal quality?
II) Is the fixation identification algorithm performing well?

Signal quality is determined by the tracking rate and tracking accuracy. Tracking rate states the proportion of frames in which the eye tracker successfully detected the eye. It depends on the performance of the device and the pupil detection algorithm. Tracking accuracy describes how close the measured gaze location is to the actual one. It depends on the quality of the gaze mapping algorithm and, therefore, on the goodness of the calibration. Tracking accuracy is hard to measure post-hoc, if no explicit calibration quality assessment was performed during the recording. Studies usually report tracking rate as a measure of data quality (if any at all).
Recordings of minor tracking rate are often discarded. Depending on the length of a trial and the cost of the experiment, this might be sufficient - for example one can simply exclude some trials from an analysis with hundreds of short recordings. But in expensive experiments or recordings pf long duration, e.g., a session of real-world driving, discarding a recording of poor tracking rate is an unfavorable option. Instead, one is interested in whether the whole recording is of degraded quality or there are consecutive parts with high and others with low tracking rate. A partial analysis of the high quality segments might then be an option. Figure 3.1 shows how Eyetrace can be used to visualize tracking quality and to get an impression of whether the data can be used, partially used, or has to be discarded.

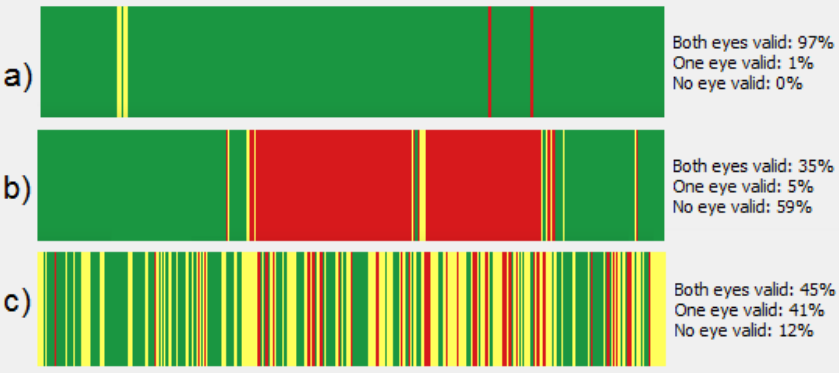


Figure 3.1: Quality plots of three recordings. Green areas mark samples with both eyes detected, yellow areas are samples with only one eye detected and red areas are tracking failures. a) a recording of overall good quality. b) consecutive tracking failures with good, usable data at the beginning and the end of the recording. c) dispersed tracking problems with one eye. Data from that eye should not be used for analysis

For the second step, the visual inspection of the performance of the fixation identification algorithm, a different visualization is available: instead of plotting all raw data of temporally and spatially overlapping fixations next to the detected fixations, a simplification of the raw data is performed first. An ellipse that represents the actual shape of the raw data assigned to a fixation is calculated and visualized (Figure 3.2). Depending on the sample rate of the eye

tracker, about 10-300 samples per fixation are summarized to such an ellipse. An optimal fixation identification will result in a roughly circular sample arrangement that resembles the Gaussian distribution of measurement errors and ocular micro-movements (for trackers with very high accuracy).
Non-optimal parameter choices will result in a more elliptical, less circular arrangement. The ballistic eye movements do not reach peak velocity immediately but require an acceleration phase. Therefore, samples from the beginning and end of a saccade are most likely to be wrongly assigned to a fixation. These samples of the beginning and the end of a subsequent or previous saccade are wrongly assigned to the fixation and contribute to a more elliptical shape.
This approach is the visualization pendant to the assumption of Berg et al. [48] that ellipse axes of highly different lengths can be used to identify saccades. Contrary to plotting all raw data, the amount of visual clutter produced by large data sets is minimized.

(a) circles of equal size (b) scaled by dwell time (c) ellipse fit

Figure 3.2: Fixation locations visualized by (a) location (b) location and dwell time as the relative scale of the circles. Larger circles mark longer dwell time. Dwell time is supposed to correlate with cognitive processes and the scaling of the circles can therefore be interpreted as attentional weight of the location. (c) ellipse fit to the samples acquired during the fixation

Additionally, the visualization provides an intuitive representation of measurement precision. Noise and ocular micro-movements add up to a confidence boundary for the actual fixation location. Small ellipses represent a high precision with all data points close to each other. Low precision results in a wide spread of samples and larger ellipses. This measure of precision must not be confused with the accuracy of the calibration. Calibration accuracy cannot simply be derived from the data, but high precision can be achieved with low accuracy.

For fitting the ellipse, eigenvectors of the data matrix are calculated. The data matrix contains two-dimensional gaze locations (Figure 3.3): $\mathbf{v}$ is an eigenvector of the data matrix A, if the following condition holds for the eigenvalue λ:
$A\mathbf{v} = \lambda\mathbf{v}$
As in a Principal Component Analysis (PCA), the first eigenvector represent the direction of

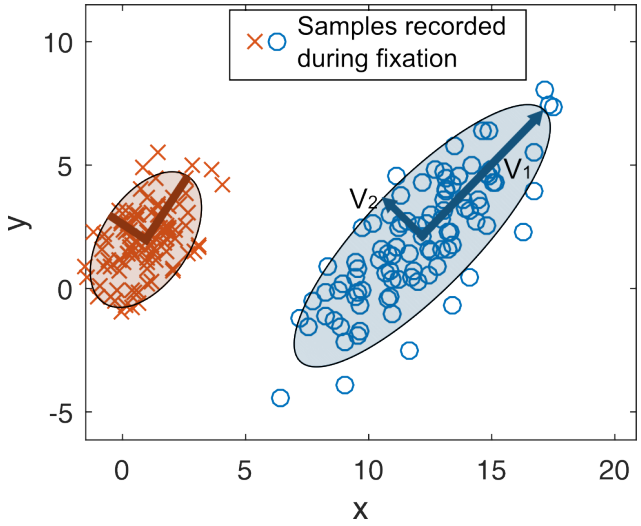


Figure 3.3: Ellipse fit of two Gaussian with different means and covariance matrices, corresponding to the distribution of raw eye-tracking data sampled during fixations. An ellipse fit can be performed by scaling the eigenvectors of the data by their respective eigenvalues and using them as major and minor axis of the ellipse. The ellipse center is determined as the data mean

the largest variance. The second vector is orthogonal to the first one. Therefore, they can be used as the ellipsis major and minor axis with their respective eigenvalues as scaling factors of axis length.

3.1.1.2 Smooth pursuit identification

Laboratory studies employ static stimuli whenever possible: they are easy to create and to analyze, e.g., by heatmaps and bounding boxes. For this type of stimulus, identifying eye movements reduces to a distinction between fixations and saccades.
The nature of encountered visual stimuli changes dramatically for experiments out of the laboratory: when we move with respect to our environment, a counter-movement of our eyes is necessary to maintain a stable fixation. Similarly, the vestibulo-ocular reflex causes an involuntary eye-movement during head rotation. Moreover, the environment is dynamic and objects in movement relative to the subject. Smooth pursuits are a lot more frequent in real-world settings. For example, in a driving experiment the ego-motion of the car causes all objects outside of the car to pass by. Maintaining an extended fixation requires a pursuit movement [22].
We studied data of a real-world and simulated driving experiment (details in Section 3.3) with the Bayesian mixture of Gaussian model (I-BDT) and found that the velocity threshold determined from the data varied across the recording. Variation is limited to a certain range with variation between subjects being much larger than variation within a trial. In fact, the degree of variation in the parameters of the Gaussian and their priors can be used as a feature for eye movement biometrics (see Section 3.1.2).
When studying the identified fixations more closely, we found that I-BDT learned to classify

the short smooth pursuits as fixations [12]. The frequent short pursuits bias the distribution of velocities within a fixation learned by I-BDT. It is shifted towards allowing higher velocities and includes the smooth pursuits caused by the ego-motion of the car and the subject's head.

When applying the ellipse visualization for the assessment of fixation identification quality (Section 3.1.1.1) to this data, we can observe an increased number of fixations with high differences between the two ellipse radii. This non-circular shape can be utilized to make a distinction between actual fixations and pursuit movements: in the driving scenario smooth pursuits are mainly driven by the ego-motion of the car. Therefore, they follow a primary direction (the direction of the optical flow). This mostly linear movement is resembled in the fitted ellipses as an increase in length of the primary axis. The proportion of axis lengths between major and minor axis can thus be used to separate the pursuits: movements with disproportional ellipse axes are likely smooth pursuits, circular ellipses represent fixations.

Results

A small quantitative analysis of the procedure was performed: data from a simulated driving experiment of one subject was chosen (more details on the experiment are given in Section 3.3). For a randomly chosen six-minute-long driving sequence, two persons manually annotated the data points as fixations, saccades, or smooth pursuits. This annotation task is very laborious, as the data has to be labeled frame-wise. 27 fixations, 8 smooth pursuits, and 11 saccades were labeled. Table 3.1 shows the per-class true-positive and false-positive counts. While the I-BDT algorithm detected all saccades correctly, the modification for the detection of smooth pursuits wrongly classified two fixations as smooth pursuits and one smooth pursuit as a fixation. It correctly identified 7 out of 8 smooth pursuits and 26 out of 27 fixations.

Table 3.1: True and false positive counts for the detection of fixations, smooth pursuits, and saccades in a simulated driving experiment

Eye movement	Annotation	TP	FP
Fixation	27	26	1
Smooth pursuit	8	7	2
Saccade	11	11	0

Obviously the amount of data available is not adequate for a thorough evaluation, but shows the principal functioning of the method and the strength of the adaptive approach followed by I-BDT. This method is focused on scenarios with a strong influence of ego-motion. This motion and the ocular movement response cause mainly linear pursuit movements (along the optical flow or the axis of head rotation). In general, smooth pursuits can appear in almost any shape and the proposed method will not be able to detect, e.g., spiral-shaped movements. The method was modified and extended to be applicable to general smooth pursuits in [23].

Implications for scanpath comparison

Scanpath comparison metrics are usually employed on sequences of fixations and saccades. Other movements, such as smooth pursuits, micro-saccades, drifts and micro-tremor are usually ignored: they are difficult to identify and their consideration in the scanpath would require far more complex representations.
Saccades are disregarded as visual perception is suppressed during these fast movements. Smooth pursuits are often represented as a quick succession of short fixations and low amplitude saccades.
Some scanpath comparison metrics come with their own fixation filter. But differentiating between fixations and saccades can be considered a separate algorithmic problem, even though it is an essential and non-trivial preprocessing step (e.g. [21, 59]).

3.1.2 Biometrics via eye movements

Biometrics is the process of identifying a person by biometric features. Usually a fingerprint or the iris structure are used. A relatively new approach is the incorporation of eye movements. The eye globe, its muscles and the brain's control are believed to be highly specific to a subject [60].
This section provides a deeper look into the parameters learned by the mixture of Gaussian model for fixation and saccade identification (I-BDT). The fact that a learning model can out-compete non-adaptive methods implies that there is a substantial variation in the data that can be learned. Indeed, we found a variation in the adjusted parameters of the Gaussian distributions, especially between recordings of different subjects. What remains unclear is the strength of this subject-specific effect.
Eye movements in the context of biometrics have a high counterfeit resistance potential. It is not possible to trick the system, e.g., by a high resolution image of the iris. The BioEye2015 challenge of biometrics via eye movements [61] was held to objectively compare the performance of different eye movement based methods. The challenge data contains two different sets of stimuli: random dots (RAN) and text (TEX). Both sets consist of two recordings per subject. The first recording is used as a training set. Characteristic, individual features are derived from this data. The task is to match the second recording, the test set, to the training recording of the same subject. Both sets contain data with a long and a short duration between the first and second recording session.

- RAN: random jumping-point-of-light. The total duration of an experimental trial was 1 minute 40 seconds, with the point changing its position every second.
- TEX: text reading stimulus. The total duration of an experimental trial (time given to subjects to read the text) was 1 minute.
- Short: 306 subjects. Sessions separated by approximately 20 minutes.
- Long: 74 subjects. Sessions were separated by approximately 1 year.

By applying an identification method based on features of the Gaussian mixture model, the subject specific influence on the Gaussian mixture model can be demonstrated.

Table 3.2: Identification rate and number of correctly identified subjects for the BioEye2015 challenge (excerpt from [62]). The baseline was calculated by the model introduced in [63]

	RAN Short	RAN Long	TEX Short	TEX Long
Baseline	34.0%	40.5%	58.2%	48.6%
Rank-1 Identification Rate	51.6%	40.5%	57.5%	40.5%
Baseline	52	15	89	18
Correctly identified subjects	79	15	88	15

The employed method is based on a feature vector of commonly used features in eye movement biometrics: saccade direction (frequencies binned by degrees as well as horizontal/vertical ratio), saccade amplitude, saccade velocity and acceleration profiles, fixation duration profiles, fixation and saccade mean velocity and velocity spread, blink rate and duration, and stimulus hit accuracy. Further, the Gaussian mixture model parameters (mean and standard deviation of the learned distributions) were added.
This feature vector constructed on the test data is then compared to each of the feature vectors constructed on the training data. Comparison can be done by simple absolute difference for scalar values and χ-square distance for differences in histograms (e.g., the saccadic velocity profiles). The resulting multi-dimensional distance vector is then merged to a single distance value by linear weighting. The subject with the smallest distance between his/her training feature vector and the test feature vector is the one that is most likely the creator of the eye movement. From the training data the weights of the features were adjusted by their discrimination power.
By reaching 4th place at the BioEye2015 challenge using (amongst other eye movement features commonly used for biometrics) the Gaussian mixture model parameters underlines that fixation identification needs to be subject specific.
In [64], these findings were used to classify the secondary task of a driver based on parameters derived from eye and head tracking.

3.1.3 Eye tracking and eyeglasses

The transformation of image features, such as the pupil center, into a gaze direction is called gaze mapping. It often involves a calibration step to calculate a correspondence between image features and gaze direction. One can distinguish the following functional principles: polynomial approximation, model-based and appearance-based mapping.
Simplest to implement and most often used is the mapping by a polynomial of second order. It requires a calibration procedure during which the subject is required to fixate a number of points (usually 9). The point correspondences between pupil center and scene camera coordinates are used to adjust the coefficients $(a-f)$ of the polynomial:

$$x = a + b \cdot p_x + c \cdot p_y + d \cdot p_x^2 + e \cdot p_y^2 + f \cdot p_x \cdot p_y$$

with x being the mapped horizontal gaze coordinate, p_x and p_y the pupil x and y-coordinates in the eye image and $a-f$ the polynomial coefficients. The equation for the vertical gaze

coordinate is identical. Only the value of the coefficients differs. One instance of this equation can be constructed per calibration point by substituting the variables for gaze and eye image points. This results in a linear equation system that is (over-)determined and can be solved for the coefficients.

Even though the above approach is commonly used and can reach high accuracy, there are applications where time-consuming and error-prone calibration processes are not applicable, e.g., when examining children or disabled persons who cannot follow the calibration instructions. Therefore, there is continuous effort to employ model-based gaze mapping techniques. Such models include a representation of the human eye, IR-LEDs and one or multiple cameras. The assumption of a model reduces the amount of parameters that require adjustment through the calibration process. Population averages can be used and adjusted over time in order to work completely calibration free.
Most model-based methods employ at least two IR-LEDs and often multiple cameras (e.g., [65]). But there are also approaches with only one camera that do not require IR reflections [66]. As a general rule, the more IR-LEDs and cameras are available, the more accurate and robust towards slippage and drifts is the system.
Appearance-based methods learn what different gaze directions *look like* in the image of the eye tracker from a huge amount of training samples, i.e., thousands or millions of images with annotated gaze direction. Acquiring such an annotated set of training data is not trivial. In [67], for example, a virtual 3D model of the human head is generated and images for a large variety of head and eye orientations are rendered. By comparing the actually captured image to the rendered images (for which gaze direction is known), the gaze can be estimated from the most similar images.
Appearance-based approaches rely on the availability of training data that is as similar as possible to the input. Model based approaches rely on the accuracy and completeness of the underlying model. To my knowledge, there is no such model nor good training data that includes a representation of eyeglasses. Consequently, gaze mapping techniques are usually evaluated on healthy-sighted subjects without correction.
This is quite surprising regarding the fact that 30% of young adults in the industrial nations need to wear eyeglasses or other optical correction [68, 69]. Eye-tracking technology should consider this fact, e.g., by including people with eyeglasses in their study populations - and their methodology.
This is especially important as the elderly, a subject population that is rarely considered in eye-tracking studies, have a high prevalence of eyeglasses and especially of more complicated varifocals. More specifically, most eye-tracking studies are conducted in the academic environment, where students are much easier to recruit than elderly subjects. Even students are already at a four times higher risk for myopia than persons with only primary schooling [68]. The need to be able to track through eyeglasses is obvious - but only few studies investigate how that can be achieved and how the error magnitude of gaze mapping methods is influenced by eyeglasses.
The hypothesis that the introduction of eyeglasses does not have a major impact on the accuracy of the polynomial fitting technique will be investigated. Due to this, researchers are able to track through glasses without worrying too much. Pupil as well as gaze positions

are measured through the glasses during both, calibration and measurement. Therefore, the coefficients of the polynomial are adjusted to incorporate the effect of the eyeglasses to some degree. While this is very useful for practical applications, little is known about the size of the remaining error. Geometrical models do not have such a built-in compensatory mechanism. They are likely to be more affected by the introduction of eyeglasses.

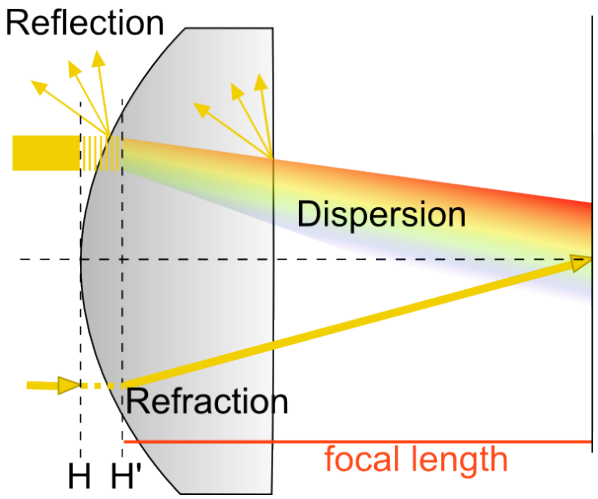


Figure 3.4: Most relevant effects of an eyeglass are refraction and reflection. Refraction is the effect of direction change of the light ray determined by the refractive index of the lens material as well as the thickness of the lens. Different wavelengths of light undergo refraction of different strength, an effect called dispersion. Reflective properties depend on the material as well as on the glass coating and its effectiveness for a certain wavelength of the light [24]

Geometrical models are based on calculating a ray from the pupil center (or other eye features) towards the camera. This ray gets refracted by the eyeglasses (Figure 3.4), introducing an angle that is not accounted for by the model and that depends on the strength of the optical medium.
To illustrate that this effect is indeed relevant, Figure 3.5 shows the change of the pupil edge solely due to the refraction caused by the cornea (simplified as one continuous optical medium): light rays towards the pupil are refracted by the cornea, altering the image of the pupil from the camera's perspective.
Some gaze mapping algorithms do compensate for the refractive effect of the cornea [65] when determining the pupil center. The otherwise introduced error has been quantified at 3°. A relevant deviation, similar in size to the overall error of most algorithms. The appearance change of the pupil caused by refraction at the cornea can be avoided by detecting and tracking the iris contour instead of the pupil [70]. The curvature and extend of the cornea results in only minor refractive influence on the image of the iris contour. However, the iris is harder to detect and track, since it is partially occluded by the eyelids and eyelashes most of the time.
This stability of the iris contour with regard to refraction does not hold for eyeglasses: their refraction changes the appearance of everything behind the glasses. Manufacturers

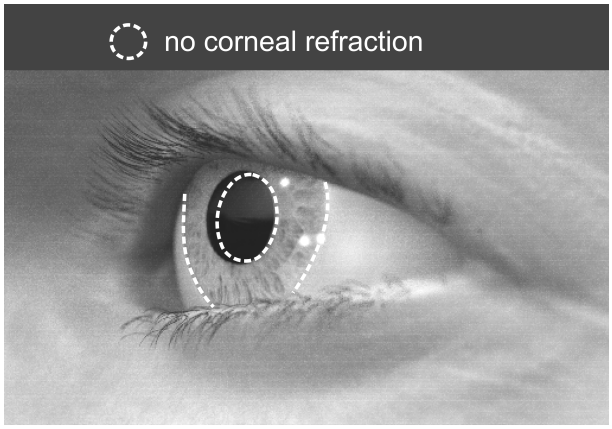


Figure 3.5: Effect of the corneal refraction on the pupil contour. Simulated image with normal refractive index of the cornea (1.336) and overlay of the pupil contour with cornea refractive index equal to the air so that no refraction occurs. Image after [24]

of calibration-free eye-tracking systems try to place the camera between the eye and the eyeglass in order to avoid this effect - resulting in the necessity to buy a special set of corrective glasses specifically designed for the eye tracker.
A first step towards the incorporation of eyeglasses is to model lenses and light rays. We generate synthetic eye images that enable appearance based gaze mapping to capture the effects of eyeglasses correctly. Beyond gaze mapping, synthetic ground truth data is essential for the development and targeted evaluation of pupil detection algorithms: we can now increase diopter and reflectance of the glasses and observe their effect on pupil detection.

3.1.3.1 Rendering artificial eyeglasses

In the following, an extension of a model by Świrski and Dodgson [71] for the creation of synthetic eye tracker images is described. The authors render a 3D model of the eye region to extend an eye tracker simulation model [72] by the image generation step. Our approach, published in [24], utilizes the same model and rendering pipeline and produces highly realistic images of the eye with different eyeglasses, as if recorded by a mobile eye tracker, including IR illumination. These images are then used to evaluate the influence of different lenses on the gaze mapping accuracy.

```
1  function calculateLens(dioptre, input_lens_thickness, F1, F2)
2
3  center_thickness = 7mm
4  while (not converged)
5    if(F1 given)
6      F2 = dioptre − F1 / (1−center_thickness/n_lens·F1)
7    else
8      F1 = (dioptre−F2) / (1+center_thickness/n_lens·(dioptre−F2))
9      [r1, r2] = recalculate_radii(F1, F2)
10     center_thickness = recalculate_center_thinkness(r1, r2, ...
```

```
11        input_lens_thickness)
12    return [r1, r2, center_thickness];
13
14
15  function recalculate_radii(F1, F2)
16    r1 = (n_air − n_lens) / F1
17    r2 = (n_lens − n_air) / F2
18    return [r1, r2];
19
20
21  function recalculate_center_thinkness(r1, r2, input_lens_thickness)
22    border_thickness = center_thickness - sagitation_depth(r1)
23        + sagitation_depth(r2)
24    center_thickness += input_lens_thickness - border_thickness
25    return center_thickness
26
27
28  function sagitation_depth(r)
29    return r·(1 − sqrt(1 − glass_height² / r²))
```

With n_{lens} being the refractive index of the lens material, F_1 and F_2 the front and back surface refraction, respectively. Diopter is the parameter that defines the desired lens refraction. Either F_1 or F_2 has to be given, the missing one is then calculated. This procedure is derived from [73]. The parameter *glass_height* is half of the diameter of the lens (in mm) and *input_lens_thickness* is the thickness of the lens (in mm).

The refractive index of the lens depends on the lens material and the wavelength of the light used. For the images shown here we chose indices of materials currently used in the manufacturing of eyeglasses (such as N-BK7 Schott, CR-39 polymer).

During calculation of the curvature radii (at the front and back) of the lens, the refractive index at 589 nm (Fraunhofer D line) is taken into account. Most eye trackers work with near infrared illumination. For generating the images, a refractive index at near infrared (900 nm) illumination was applied.

Modeling the reflective properties of a lens via ray tracing is complex for reflex reducing coated lenses. They consist of many different layers. The reflection process for the coating used in this model is only an approximation based on measurements of the overall reflectiveness of such lenses [74] at near-infrared illumination. It is notable that the efficiency of reflex reduction coating is wavelength dependent and not optimized for infrared illumination. Therefore, the eye tracker is likely to encounter more reflections than what humans are able to see.

Reflections on the eyeglasses will only occur if there is an environment to reflect. Therefore, an environment map was added to the simulation. Limited by the availability of infrared high dynamic range environment maps, we chose to simulate by a pseudo-infrared tone mapping technique taken from a preset of Adobe Photoshop.

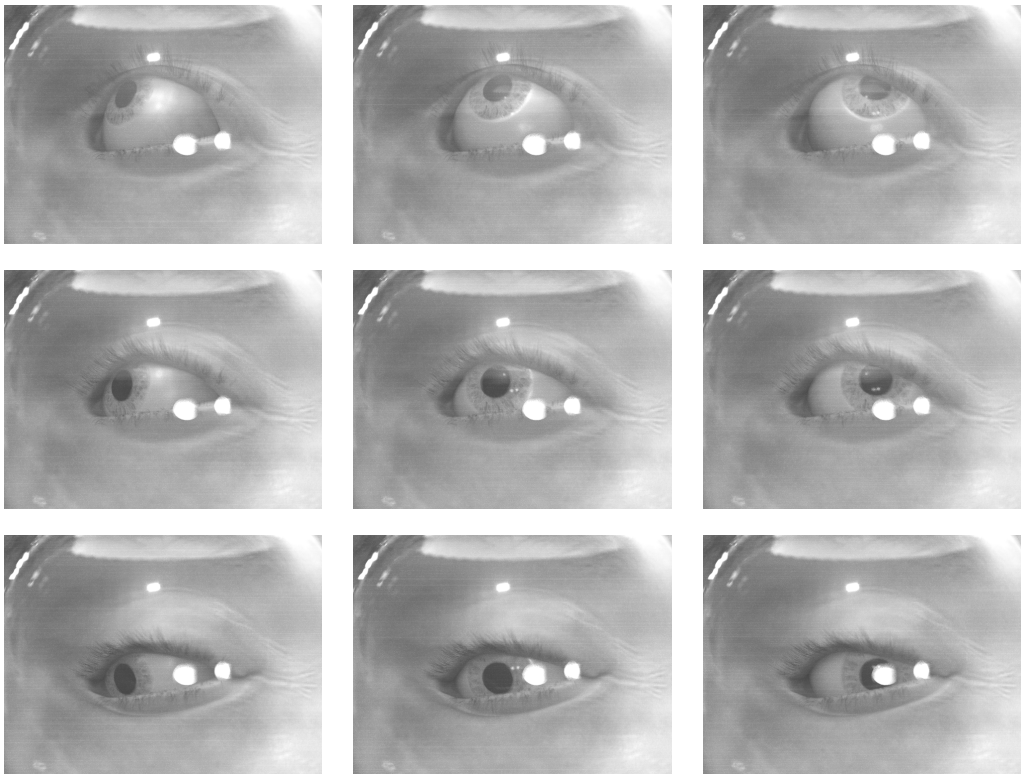


Figure 3.6: Generated calibration images with a -1 dpt uncoated lens. Gaze directions vary between -20° and 20° horizontally and vertically, as presented in [24]

Table 3.3: Mean accuracy (± standard deviation) of gaze prediction for two calibration algorithms and different eyeglasses: while the polynomial fit shows only a minor decrease in average accuracy and a minor increase in standard deviation, the geometrical model is strongly influenced by the refractive strength of the eyeglasses

Eyeglass	Polynomial fit	Geometrical model
+2 dpt	$1.36 \pm 0.55°$	$3.49 \pm 1.48°$
0 dpt	$1.33 \pm 0.46°$	$2.09 \pm 1.05°$
-1 dpt	$1.33 \pm 0.52°$	$1.95 \pm 1.10°$
-3 dpt	$1.37 \pm 0.57°$	$2.94 \pm 1.43°$
-5 dpt	$1.56 \pm 0.61°$	$5.38 \pm 2.16°$

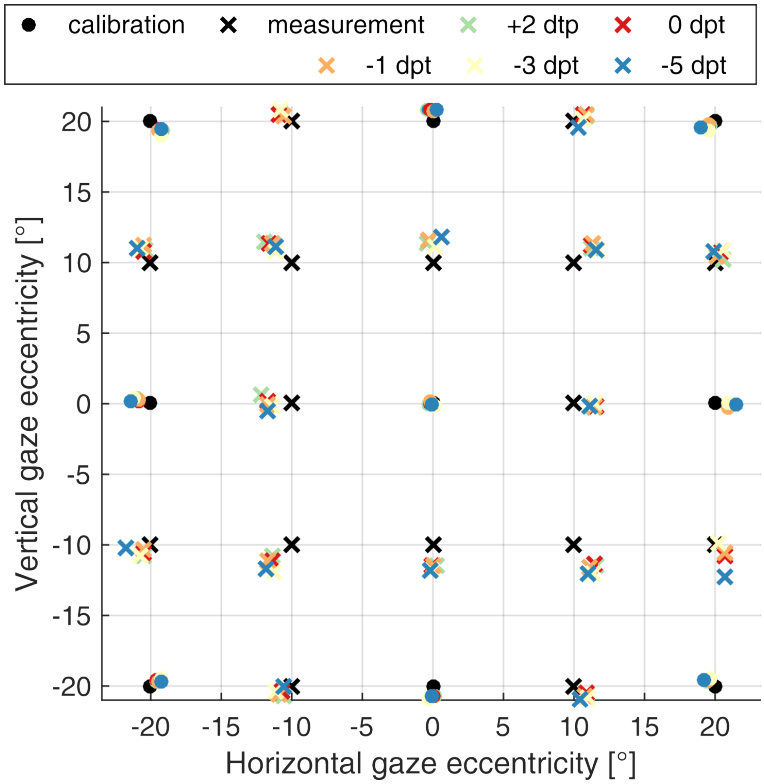


Figure 3.7: *Polynomial* gaze mapping for different eyeglasses (up to -5 dpt) and without glasses (0 dpt). Points at locations marked by black circles were used for calibration. Black crosses denote the true gaze orientation at the test points. No significant effect of eyeglasses on the calibration accuracy can be observed [24]

3.1.3.2 Evaluation of gaze mapping accuracy

We evaluate the effect of eyeglasses on two different gaze mapping techniques: polynomial fit and the geometrical model by Świrski and Dodgson [66]. Figure 3.6 shows the generated images for a 9-point calibration in primary (straight-ahead), secondary (up-down/left-right) and tertiary (diagonal) position at 20° eccentricity. These images were used for fitting the coefficients of the polynomial. Sixteen more images with gaze orientation as shown in Figure 3.7 were used to determine calibration error. Calibration errors usually increase with the distance from the calibration points. The calibration points were excluded from the calculation of the overall error. All test points are located within the calibrated area.

Pupil edges were determined by fitting an ellipse to ten manually annotated points on the pupil edge. Thus, we ensure that all inaccuracies are caused by the gaze prediction step, not by an insufficiently annotated pupil.

As described in [24], gaze mapping error was measured as the angular distance between the actual gaze target and the predicted one. This procedure was applied to images of the eye without glasses, with low power lenses (-1 dpt, +2 dpt), as well as high power lenses (-3 dpt

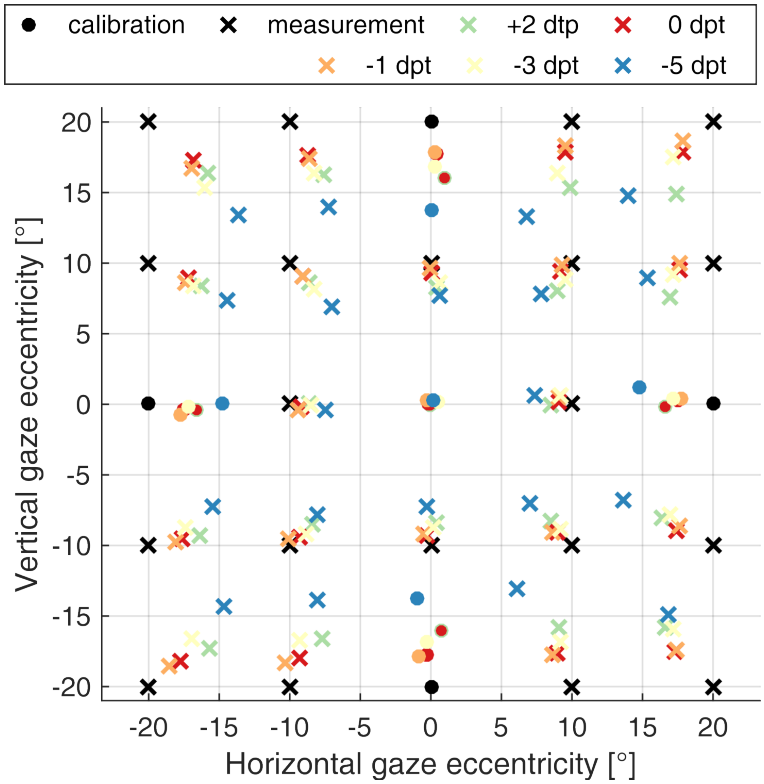


Figure 3.8: Gaze mapping with a *geometrical model* for different eyeglasses (up to -5 dpt) and without glasses (0 dpt). Points at locations marked by black circles were used for calibration, whereas black crosses denote the true gaze orientation at the test points. A relevant decrease of gaze prediction accuracy coupled to the refractive power of the eyeglasses can be observed [24]

and -5 dpt). Figure 3.7 shows the accuracy of the polynomial mapping, and Figure 3.8 the geometrical model. Numerical averages over all test positions are reported in Table 3.3.
The geometrical model calculates a 3D eyeball based on the ellipsoid shape of the pupil. It does not require additional calibration. However, we used five points in primary and secondary position to determine the gaze ray in relation to the eye coordinate system. This allows

- to easily determine the eye's primary position in the eyes's coordinate system (instead of the orientation relative to the camera).
- to correct for translation effects caused by the eyeglasses (this requires a 1-point calibration).

The original code of the geometrical model as presented in [66] was used.

Discussion

It is obvious that the polynomial method is indeed robust over a considerable range of lenses, while the error of the geometric method increases with optical power. It is worth

mentioning that this specific model was used as an example but any other geometrical eye model could be used as well. None of them includes eyeglasses and many require additional glints that cannot be easily determined when the subject is wearing eyeglasses and additional reflections are present. A very similar effect on gaze prediction accuracy can be expected for all of these models.
Appearance based models that are trained on a 3D model [67] are currently also unable to include eyeglasses. For them, eyeglasses are just a passive texture without optical effect placed on the 3D model. They would therefore likely show a similar effect.
It should be further noted that in this evaluation only a small subset of possible lens designs was tested. In particular, no prisms (except for the inherent prismatic effects of the lenses) or cylindrical optical effects were included. Therefore, this work can only give a glimpse into the effects of eyeglasses on gaze mapping algorithms.
Due to the more complex calculations and increased number of ray samples required for the refraction and reflection effects, rendering time has increased significantly when compared to the model without glasses. A ray has to hit the inner and outer border of the lens and the cornea twice, once on its way from the light source towards the eye, the second time on its way back to the camera. The cycles renderer of the 3D program (Blender) implements a path tracing algorithm and is thus unsuited for these complex reflections.
A different rendering engine such as Luxrender, which implements bidirectional pathtracing, is likely to improve run-time. This rendering algorithm connects both camera and light paths at each bounce and converges with fewer samples.
The simulated images are great for the evaluation of existing algorithms, however it should be noted that the artificial eye model has only a limited degree of variability. It does not reflect the whole spectrum of natural human eye variability. Creating an algorithm that is specially trained for the type of images generated by the model would be relatively easy. Thereby other methods could be outperformed on the evaluation data without providing a practically useful generalization. This could probably be remedied by using the proposed lens generation methods with a more sophisticated eye model that allows for a higher degree of variation (e.g., [75]).

3.1.3.3 Dirt and dust simulation

Świrski and Dodgson's rendering pipeline [71] includes a step to introduce an adjustable amount of camera noise to the image. The noise model is good for laboratory settings, where the camera's sampling characteristics are the major source of noise. However, if we want to transfer eye-tracking technology to everyday applications such as driver monitoring, public displays, or gaming, we face a much broader spectrum of noise sources. No regular cleaning plan for the eye tracker nor for the worn eyeglasses can be established. Eye-tracking algorithms have to be reasonably robust to dust and dirt on the eyeglasses as well as on the camera.
While it is possible to collect a large amount of noisy real-world data and to perform ground-truth annotation, i.e., labeling the pupil center in thousands of images, it is non-trivial to determine whether decreases in detection rate can be assigned to individual factors such as dirt, bad illumination or other characteristics. Simulating the data avoids the laborious

labeling process and allows introducing normalized, quantifiable amounts of noise. Intuition suggests that a larger amount of noise should increase the number of bad pupil detections. By varying the amount of virtual dust, the robustness of the algorithm can be quantified.
In this section a short demonstration of how the amount of dust can be objectively measured by rendering it into the eye image is given. Generally, the model allows to study a large variety of noise sources, e.g., problems with reflections on the eyeglasses can be continuously adjusted by modulating the reflection and transmission coefficient of the eyeglasses material in the simulation. Since the reflections of IR-LEDs is also rendered realistically, we are not limited to analyzing the quality of the pupil detection step, but the whole chain from pupil and glint detection over gaze mapping up to blink detection can be tested.
We analyzed some eye-tracking recordings by a head-mounted Dikablis essential eye tracker during real-world scenarios, such as shopping in a supermarket or driving a car and searched for the effect of dirt and dust on subjects' eyeglasses. It should be noted that in a remote tracking scenario, e.g., a driver monitoring system, there would be an additional dust layer at the lens of the camera that does not appear so clearly for the head-mounted device.

Figure 3.9: Images of the Dikablis eye-camera showing dirt and dust particles. The images were extracted from an eye-tracking experiment and represent a realistic level of dirtiness as it can be expected during a normal experiment

Figure 3.9 shows some images with a strong effect of dust on the eyeglasses. For illustration purposes images with the camera focus on the eyeglass depth were selected. During most recordings, however, the focus of the eye camera was not directly on the eyeglasses and the dust particles appeared more diffuse and blurry - not easy to spot in an individual image, but clearly visible in a video as spots of constant brightness. Since the dust particles are illuminated by the IR-LEDs, they appear bright in the camera image.
Based on the observed dust and dirt effects the parameters of a dirt simulation designed for testing the performance of optical sensors and image processing algorithms were adjusted [76].
A small selection of rendered, artificial dust particles that show a trend from in-focus to off-focus with regard to the dust layer is shown in Figure 3.10. Size, color, amount and focus of the dust can be adjusted. The produced data can be used for the assessment of robustness of the image processing methods.

Figure 3.10: Dust particles rendered upon a real eye tracker image. Intensity, size and focus are chosen to best resemble realistic conditions as found in other recordings

3.2 Concentration and sustained attention in low-resolution eye-tracking data

The signal quality of a gaze location measured by an eye tracker depends entirely on its technical specifications and the employed algorithms. But what we intend to measure is often not directly gaze location, but overt attention. Contrary to the mere gaze location, attention includes the associated cognitive processes. It is obvious that subjects do not perform an assigned task constantly with the same level of concentration and motivation for an unlimited duration. This section focuses on the subject's ability to maintain concentration and attention. The state of being able to maintain sustained attention is also called vigilance. It is often measured as the ability to react to a stimulus.
The longer the experiment, the more likely is a decrease in vigilance and concentration. As we have only minor control over the decay of vigilance, it is important to keep track of its current state. For some applications a high vigilance state is essential. During perimetry, i.e. the measurement of the visual field, for example, the information on whether a stimulus is detected is used to determine the extend of the visual field defect and detect visual field defects. The more stimuli can be tested with a good subject feedback, the more accurate the shape. However, with longer measurement time, vigilance decreases. In consequence, the number of false responses to the stimuli increases. Instead of refining the results only the level of noise is increased. It is therefore important to be able to filter data for cognitive states of the subject.
To approach this, the eye-tracking signal can be used to extract further parameters related to vigilance. Such parameters are for example pupil dilation and constriction. Besides this, binocular tracking and even the absence of a signal - a tracking loss (in the optimal case occurring only during blinks) - are commonly recorded alongside with gaze location.

In this section, we will explore signals derived from eye-tracking data that can be used as parameters to quantifying vigilance. This section is based on work published in [77] and focuses on parameters of attention and vigilance that are usually recorded alongside most eye-tracking experiments.
A lot of research has been conducted to detect attention, vigilance, and cognitive load from pupil dilation. Most of these studies require specialized conditions (such as low,

equiluminant light and no eye movements [78, 79, 80]). This work focuses on devices with low temporal resolution (up to 60 Hz) and no complicated eye model (meaning no pupil diameter in millimeters nor a valid calibration is required).

3.2.1 Background: attention and vigilance

William James defines attention in his Principles of Psychology [81, 82] in 1890 as:

> *"It is the taking possession by the mind [...] of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others [...]."*

Previously, we applied a theory of attention based upon looked-at locations. Due to the foveated built of the human eye there is solid support for this approach. Pomplun et al. investigated conflicting views of ambiguous figures and found that "not every fixation is filled with attention" [83]. An obvious example is staring into space while bored and the phenomenon of tunnel vision. Although a huge gaze duration is spent on the same target, it is clearly a low attentional state. On the other hand, the amount of attention associated with the fraction of a second when a car driver spots a hazardous situation is very high.

Vigilance is most commonly used as a term for sustained attention over a long time period. It is an indicator of the sleep-awake level and also of the level of cognitive performance [84]. Attention is a directed process whilst vigilance is a slow and global condition. Both of them are continuous rather than discrete variables that can be measured on a scale, e.g. with the pupillographic sleepiness test or the pupillary unrest index. These measures of pupillary oscillations primarily occur in scotopic conditions, but can also be recorded in photopic conditions, if a monotonous repetitive task is performed [85, 78].

In cognitive psychology, cognitive load refers to the total amount of mental effort being used in the working memory [86]. Cognitive workload can be described as the degree to which cognitive and perceptual capabilities are taxed with executing a task [87].

There are many applications that could benefit from an objective quantization of attention, vigilance and workload. Here we will focus on two: perimetry, the process of visual field testing, and image viewing.

3.2.2 Indicators for vigilance and attention in eye-tracking data

Pomplun et al. [83] constructed so-called *attentional landscapes*, i.e., heatmaps with fixation duration as a weighting factor of attention. But the mere information of where and how long a location was looked at was found to not completely reflect actual attention allocation. Mentioning this finding, Wooding phrases his definition of a *fixation map* as a three-dimensional overlay of Gaussian with a weighting factor d that is "deliberately left vague" [88]. The d factor stands as a variable for a quantification of actual attention and has not been investigated much further since. Wooding chose $d = 1$ to treat each fixation equally, well aware of the fact that this is not the case in reality. In this chapter we will consider different candidates for the d-parameter and begin with a review of commonly

employed measures of cognitive state.

Eye blink rates of healthy individuals range from 5 to 15 $\times$ per minute [89]. Blink rate, duration, lid cleft (the distance between the upper and lower eyelid) and standardized (with respect to the lid cleft) lid closure speed reflect the level of fatigue and sleepiness [90, 91]. Light fatigue is associated with an increase in blink frequency, sleepiness with an increase in blink duration [91].
Blink rate for a difficult mental arithmetic task is higher than for an easy one. However, the effect depends on the kind of task [92].

Pupil size is determined by two neurophysiological reflexes, the pupillary light reflex, which regulates the amount of light entering the eye and reaching the retina, and accommodation, resulting in changes in the curvature of the lens. Additionally, cognitive activity of the brain as well as emotional factors influence pupil size. For example, a mental multiplication task results in different levels of pupillary dilation for different difficulty levels of the task [80]. Nishiyama et al. found that in a driving simulator experiment a monotonic gradual miosis (i.e., constriction of the pupil) stage exists where subjects were not yet aware of their decreased vigilance but drifted towards sleepiness [85].

Henson and Emuh showed that it is possible to monitor vigilance during campimetry (i.e., similar to perimetry but with a flat screen instead of a cupola) under photopic conditions using pupillometry [78]. The equiluminant surrounding is in fact optimal to avoid any pupil diameter changes due to varying illumination. In darkness, pupil oscillations come in two flavors:
Slow waves of dilatation and constriction of 4 to 40 seconds duration and an amplitude of up to ± 0.5 mm. Superposed fast inextensive oscillations, of 0.5 to 1 second duration and with amplitudes of usually below 0.1 mm, but they reach up to 0.3 mm [93].
In the case of **fatigue waves**, the pupillary oscillations associated with decreased vigilance, the same oscillations occur but in an *exaggerated manner* [93] (up to an amplitude of 1 mm [78]). With decreasing vigilance during perimetry, the feedback loop that regulates pupil diameter, the central sympathetic inhibition, becomes unstable. This results in larger pupillary oscillations [79].

Saccadic velocity decreases with increasing sleepiness and it might be suited as an indicator of vigilance [85, 94]. However, saccades are very fast movements and determining their exact velocity is only possible using high temporal resolution (500-1000 Hz) cameras.
With the incidence of fatigue, a higher percentage of overlong (>900 ms) and express-fixations (<150 ms) occurs [91].

For the angle of **vergence** between both eyes (see Figure 3.11) the connection to vigilance and workload is not clear: heterophoria has been found to grow with fatigue and when performing an unfamiliar task [95]. Constant vergence errors of up to 2° occur. However, our vergence system is generally very accurate at maintaining its state with saccadic amplitude and fixation drift differences between both eyes at approximately 0.16° and

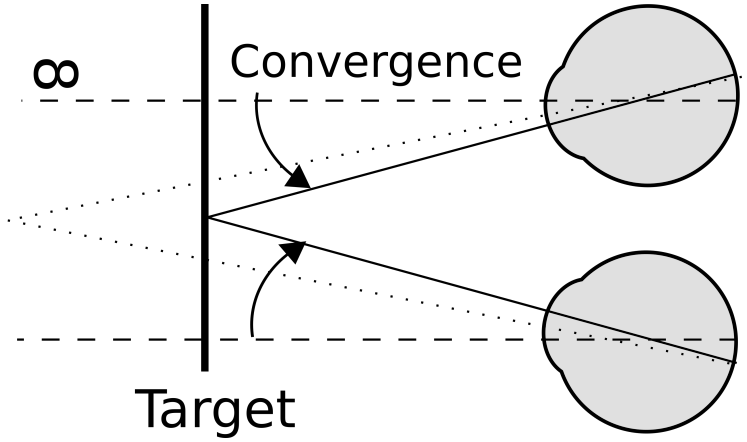


Figure 3.11: During vergence movements the eyes are guided contrarily and move in opposite directions. The dashed lines indicate parallel lines of sight. Focusing objects closer to the eye requires convergence, objects further away lead to divergence until the eyes are parallel to each other. A binocular recording is required to be able to calculate this measure

0.13°, respectively [96]. Some authors find our vergence system particularly fragile with fatigue [96], while it is not even considered a relevant factor in many fatigue detection studies.
Visual fatigue causes a large increase in the frequency of double-vergence events, i.e., two closely-spaced (160-400 ms), high-velocity initial components in response to a single vergence step [97].

3.2.2.1 Measuring vigilance during perimetry

During perimetry visual stimuli (light points) are presented at different positions in the subject's visual field. The subject has to press a button whenever a stimulus is perceived. This process is repeated for about 15 minutes at different positions in the visual field and different stimulus light intensities. Aim of the examination is to determine the threshold of local differential luminance sensitivity. Therefore, the testing involves mostly hard to detect stimuli, slightly above or below the perceivable threshold.
Visual field testing is a very exhausting task that requires a high level of vigilance. Once the subject is not fully focused on the task, the results of the test become unreliable: was the stimulus not detected because the subject did not perceive it or because the subject was not attentive and the related response time exceeded a critical limit? Due to the monotonous nature of the test it is hard to avoid vigilance decrease over time. Reducing the test duration is possible only within certain limits without a loss in resolution. Therefore, we focus on the detection of a vigilance decrease in order to intervene or stop testing once no reliable subject feedback can be expected. But not only vigilance is of interest during perimetry: cognitive workload has been shown to have an effect on visual field size and shape [98]. Therefore, new strategies for speeding up perimetry or making it more interesting must not include any additional cognitive workload on the subject otherwise they will alter the

results.

Figure 3.12: Octopus 900 perimeter by Haag-Streit AG [99]

Henson and Emuh showed that it is possible to detect fatigue waves during a campimetric examination [78], Müller et al. proved their existence during perimetry [100] with an Octopus 900 perimeter (Figure 3.12).
We aim at cross-correlating the occurrence of fatigue waves with other eye-tracking measures. Going for a multi-factor vigilance estimation might result in a more reliable, robust measure that allows a continuous estimation of the vigilance level.
Fatigue waves are usually associated with sleepiness. For perimetry we might also be interested in an earlier decrease of vigilance in order to be able to judge the reliability of the subject's feedback.

Methods

Determining the pupil diameter during perimetry is relatively easy as the subject should fixate the central fixation target at all times. Therefore, a change in the camera's image of the pupil can only be caused by a change of the actual pupil diameter. If the eye was allowed to move freely, changes caused by the perspective view would require a compensation.
For this study data recorded by Matthias Müller with the Octopus 900 perimeter's 320×240 pixel gray-scale camera at 20 fps was used. The pupil annotated data contained recordings of 8 subjects (age range 22 - 60 years, 4 male and 4 female). We analyzed the 7 recordings that Müller selected as good quality data and reanalyzed them.
Instead of static perimetry a *Method of Constant Stimuli* was used: it assesses luminance sensitivity in the central visual field and is especially fatigue inducing. Six stimulus intensities between 0.50 and 20.08 $\mathrm{cd/m^2}$ with a constant background luminance of 10 $\mathrm{cd/m^2}$ were repeated 30 times each (resulting in 540 questions asked per session). Stimulus size III (25.7') was used. Stimuli were displayed at (0°,0°), (-5°,-5°), (+5°,+5°). Subjects were

instructed to fixate the central fixation marker and to press a button once the stimulus was perceived.
Catch-trials were used to test whether the subject was attentive and giving reliable feedback: false-positive trials produced the sound of the perimeter shifting the stimulus location but did not show any luminance stimulus (this is not how the actual test was performed, see discussion). False-negative catch trials, i.e., stimuli clearly above the intensity threshold, should cause a response as long as the subject is attentive. Overall 10% of all trials were catch trials, equally distributed to false-positive and false-negative. Stimuli were presented for 200 ms and subjects could respond to the stimulus for 1,700 ms, starting with stimulus presentation. This response waiting time ended earlier in case a button was pressed. This leads to different trial durations from 10-14 minutes, depending on the average reaction time, the individual differential luminance threshold, and the number of detected stimuli.

Data processing

To measure a robust pupil diameter signal, the following processing steps are necessary: before and after each tracking loss (including blinks), the four adjacent frames before and after the tracking loss were labeled as invalid. They might contain a partially occluded pupil that cannot be detected and compensated by the employed pupil detection algorithm. Invalid samples were then replaced by a linear interpolation from the previous to the next valid sample.
Blink rate was calculated as the number of blinks in a 20 s sliding window and converted to a blink rate per minute. Pupil diameter variability was determined within a 10 s sliding window.
Fatigue waves are detected by transformation with the reverse biorthogonal wavelet (rbio3.7 in MATLAB) as proposed by Henson [78] but with an average level of 10 and detail level of 10 (Henson: 8) due to the different sampling rate, used in this experiment. By this adaptation we obtain fatigue waves of similar wavelength as in [78].
For the purpose of displaying the data in one combined graph, blink rate, pupil diameter variability and wavelet average component are normed to the range [0,1]. Wavelet detail coefficients are not normed as we expect an increase in amplitude for the fatigue waves and could loose visibility of this effect. Even continuous small oscillations would look like fatigue waves then.

Results

Figure 3.13 shows the fatigue measures available in the perimeter experiment. We can observe fatigue waves in subjects 2, 3 and some increases in oscillation that cannot be clearly assigned to fatigue waves or normal oscillation in subjects 4, 5 and 7. A decline in pupil diameter after the initial phase can be observed for subjects 1, 2, 3 and 4. Subjects 2 and 4 show an increase up to the initial pupil diameter after about 8 minutes in the experiment.
Subjects 1, 4, 5 and 7 exhibit an increase in blink rate, although overall blink rate is relatively low. Subjects were required to not miss any stimulus, and therefore, probably reduced

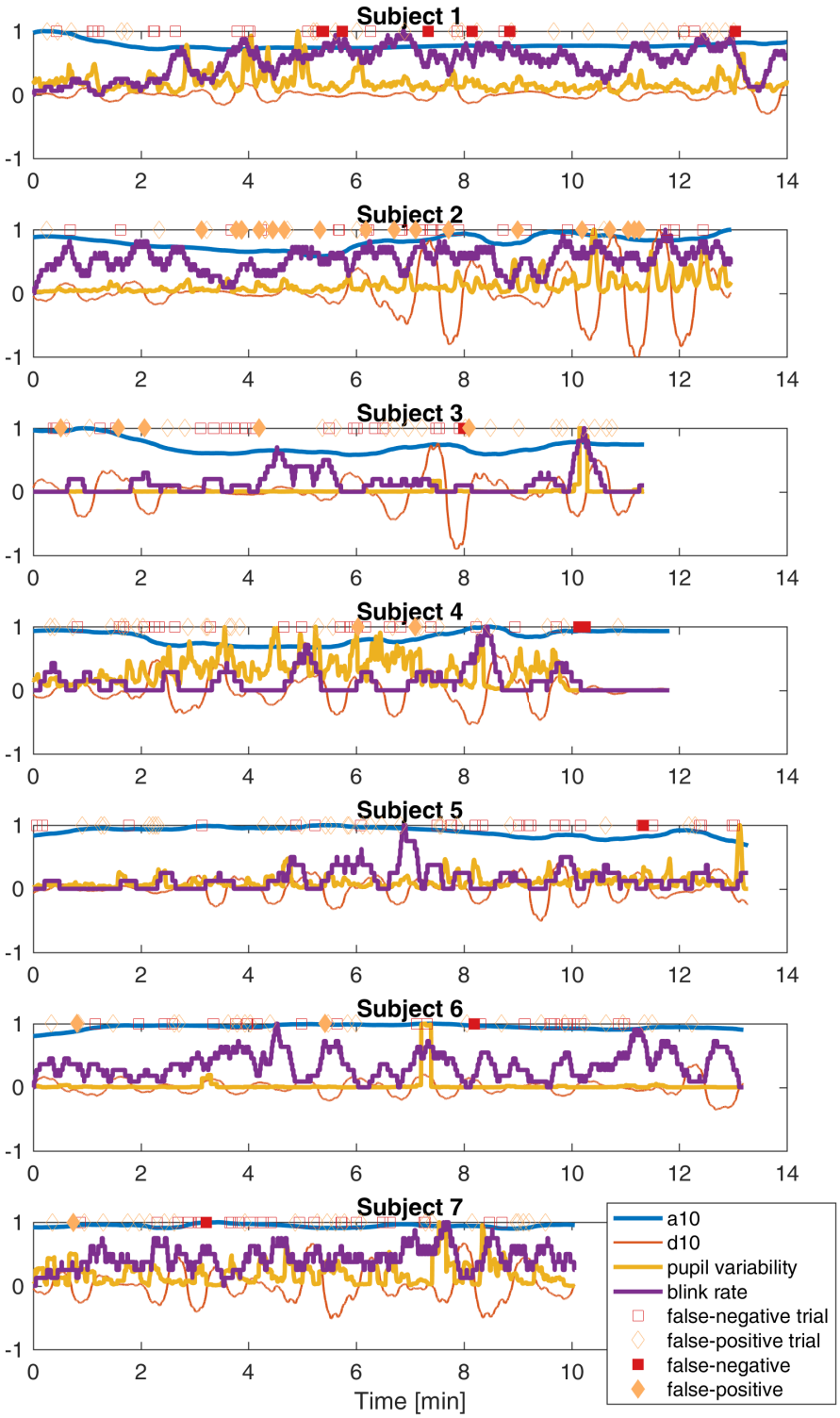


Figure 3.13: Physiological measures of fatigue during perimetry. On the top the subject's responses to the catch trials are depicted

their blinking. For some subjects a blink pattern after each stimulus could be observed. In summary, the examination procedure has a clear impact on the blink rate and makes it hard to interpret.

Discussion

With the recorded data it is difficult to correlate any signal to the error rate for the catch trials. There were only few incorrect responses to the catch trials. Instead, we can only assume a slow gradual decrease in vigilance over the test. But an examination of 15 minutes length is not exceptionally long and subjects might be able to maintain a high level of vigilance for the complete duration.
The infrared camera installed in the Octopus 900 perimeter used for this experiment records with approximately 20 fps. Finding the correct detail level for the wavelet decomposition turned out to be harder than expected. It is necessary to evaluate a fatigue wave based vigilance assessment for the exact sampling rate of the eye tracker used.
In a later code review that aimed at adapting the experiment to a design with an increased amount of catch trials some issues with the original test were found: false-positive catch trials did show a light stimulus that was within the perceivable range, although at the lower end. We can see this in the data of some subjects that were either very trigger-happy or simply able to perceive the "false-positive" stimuli. It was further possible (and not unlikely) that the randomized stimulus position resulted in testing the same position twice directly after one another. In this case the perimeter did not perform a movement of the stimulus presentation unit and no accompanying sound was produced by the perimeter.
To conclude, the experimental data has some flaws that need to be corrected in order to get more reliable results: since we are interested in the quality of the feedback from the subject, it is necessary to correct the problem with false-positive catch trials and to increase the number of total catch trials massively (currently there are only 1-2 catch trials per minute for some parts of the recording). Further a validation and quantification of fatigue waves, as determined by the wavelet transformation, has to be performed. The current comparison of co-occurrence with catch trial errors is insufficient as the catch trials are unreliable. A co-registration of additional vigilance measures, such as EEG or heart rate would be helpful.

3.2.2.2 Vigilance and image viewing

Image viewing results in relatively similar scanpaths amongst subjects. Bottom-up features of the image and interesting regions such as faces attract gaze. More subtle image composition techniques and lighting guide the observer's gaze through the image.
Both Buswell [7] and Yarbus [8] observed that the consistency of observers' fixation locations decreases over time. When free to choose, subjects view random pictures for an average of 9-50 s [9]. This huge variability in viewing time ranges from a relatively short time span that underlines the high capability of our visual system to capture the essential information of a picture in only few seconds up to a relatively long duration. Extending the viewing duration generally increases scanpath heterogeneity. A long duration may result in

either a more detailed look or induce eye movements that are not associated with attention, e.g., due to boredom and staring.
Not all pieces of fine art reveal all of their content and meaning on the first glance. There are several examples of image viewing tasks where high attentional states can be found episodically throughout a long viewing time. An obvious example is the *Wimmelbild*: an image that contains a large variety of details. It is impossible to find the most interesting parts within the first few fixations.
In this section we will search for a way of measuring cognitive workload and attention during image viewing. This measure could be used to assign different attention levels to individual image regions. Indicators of attention and vigilance in the eye-tracking signal that are commonly recorded during image viewing tasks are examined and evaluated in an image viewing experiment with extended viewing time.

Methods

Data from a replication and slight modification of Yarbus' image viewing experiment was analyzed. Four stimulus screens were presented to each subject. The stimulus order and duration is shown in Figure 3.14. Subjects were instructed to explore the images with two different tasks assigned: free-viewing and estimating the age of the people in the image.
20 subjects (age range 20 - 57 years, 6 male and 14 female), recruited from students and staff of Aalen University of Applied Sciences, participated in the study. Valid eye-tracking data were collected for 19 of the 20 subjects.

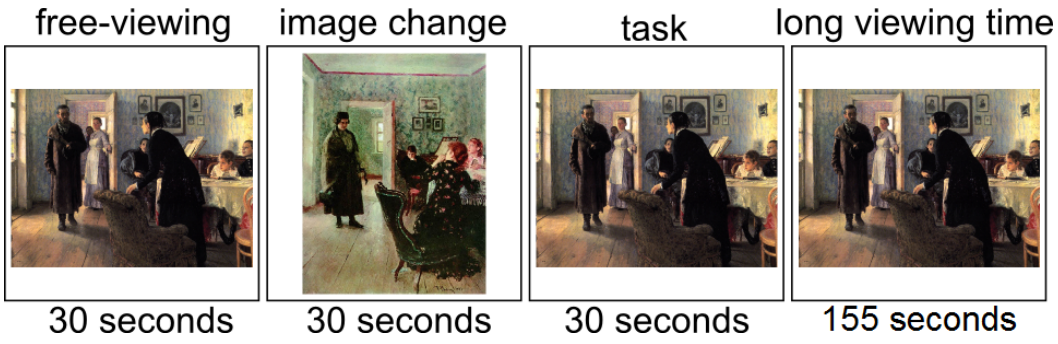


Figure 3.14: Stimulus image, assigned task and stimulus display time for the modification of Yarbus' experiment. Two similar images by the painter *Ilja Repin* were used as stimuli [101, 102]

Stimuli were presented on a computer screen (FUJITSU Display B24W-7 LED, screen size of 51.5×33 cm, resolution 1920×1200 px) at $1,196 \times 880$ px and a distance of 80 cm to the subject.
The Eye Tribe Tracker (The Eye Tribe Aps, Copenhagen, Denmark) recorded subjects' binocular eye movements at 30 Hz. It was placed in a distance of 70 cm to the subject. The system is able to estimate gaze in screen coordinates. According to the manufacturer an accuracy of 0.5-1° can be reached. We found the accuracy to be very sensitive to head motion. Therefore, a forehead and chin rest was used and subjects were instructed to sit still,

not to speak and not to perform chewing motions. A nine-point calibration was performed before the recording.
Instead of the originally intended 180 s viewing time only 155 s were recorded due to an error in the eye tracker recording software.

For the evaluation at hand we will focus on the last stimulus. At this point subjects had already seen the stimulus image twice and were assigned the free-viewing task combined with this image for the second time. The long viewing duration is likely to induce a considerable decrease in attention directed towards the painting. We can therefore assume a tendency of decreasing attention with viewing time.

Data processing

Blinks were identified as binocular tracking losses with a minimal duration of 100 ms. If both, the tracking of the eye and of the head was lost, no blink was registered. No upper limit for blink duration was set. Double blinks, i.e., multiple blinks in a short time interval, were counted as a separate blink each. Average blink rate was calculated as the number of blinks within a 20 s sliding time window and scaled (by a factor of 3) to derive the number of fixations per minute. The filter delay of the sliding window was corrected by shifting the final signal by half of the window width. A continuous signal for the blink duration was determined by linear interpolation between the discrete blink events.
Instead of calculating a vergence angle, the difference in horizontal screen coordinates between the left and right eye was used. As the head position was fixed during the experiment, we can expect vergence movements to manifest only in the horizontal plane (see Figure 3.11), the vertical direction is not relevant in this case.
The vergence angle was analyzed in two different ways:
I) averaged over all subjects. The angle can be expected to increase slightly over time. The resulting pixel difference between both eyes was smoothed by a moving average filter over 5 s. Since we are averaging over a lot of data, no further processing is necessary. Specifically, no compensation of the current gaze location as in II) is required as data of different gaze locations is averaged.
II) calculated separately for each subject. A correlation between pixel vergence and gaze location can be expected (caused mainly by the algorithmic pupil detection and calibration routines). The vergence angle was corrected for any effect that could have been caused by gaze orientation. This was done by fitting a polynomial of second order to the x- and y-coordinates of gaze position on the screen and the respective vergence measure. The component that can be predicted from gaze location was eliminated from the data. As an error of the measurement device it does not reflect a real change in binocular alignment.
For the calculation of pupil diameter variability and the smoothed pupil diameter, the eye-tracking signal was further preprocessed: all tracking losses and the three data samples (corresponding to 100 ms) before and after the tracking loss were removed. During the beginning and end of a blink, the eye tracker might record images where the pupil is partially occluded by the eyelid. The EyeTribe pupil detection algorithm cannot compensate for this and will therefore report a too small pupil diameter. By removing not only the tracking

losses but also some samples next to them, these partial pupil samples are eliminated from the data. Samples that were removed are interpolated linearly. The smoothing of pupil diameter is calculated via a wavelet transformation using the reverse biorthogonal wavelet (rbio3.7 in MATLAB) at a level of 10. Pupil diameter variability (PDV) was determined as the variance in pupil diameter within a 20 s sliding window.

Pupil size between subjects may vary, especially the used pupil size in pixels that the EyeTribe device produces. It depends on the distance between the tracker and the eye. To be able to compare recordings between subjects and to average over them, the range of occurring pupil diameters was normed to [0,1], for each subject separately.

For fixation duration and saccade length calculation the data was processed by a mixture of Gaussian fixation identification filter using a maximum likelihood fit for the two Gaussian. Saccade length was determined separately for both eyes as the Euclidean distance between start and end point of the neighboring fixations and averaged between both eyes. A continuous signal was derived from the discrete saccade/fixation events by linear interpolation.

Results

Figure 3.15 shows the time course of all indicators. The median over all subjects as well as the 75% confidence interval (corresponding to 12.5% and 87.5% percentiles as lower and upper bounds) are shown.

For the fixation duration we can observe that some relatively long fixations with up to 8 s dwell time occur after 60 s of viewing the image. They are even more frequent after 100 s viewing time. The median dwell time is quite constant over the whole trial duration.

We can observe relatively high saccadic amplitudes during the first 30 s, with a slow decrease over time.

Pupil diameter variability exhibits a peak at 120 s viewing time. A gradual miosis can be observed in the wavelet-smoothed pupil diameter, with the steepest decline over the first 30 seconds. Since no large luminance change occurred and subjects were looking at the same screen for the previous three trials (90 s) already, the initial pupil dilation and following slow constriction is likely to be associated with cognitive load: during the first seconds subjects perform the assigned task and explore the image - until they get gradually bored and attention decreases rapidly.

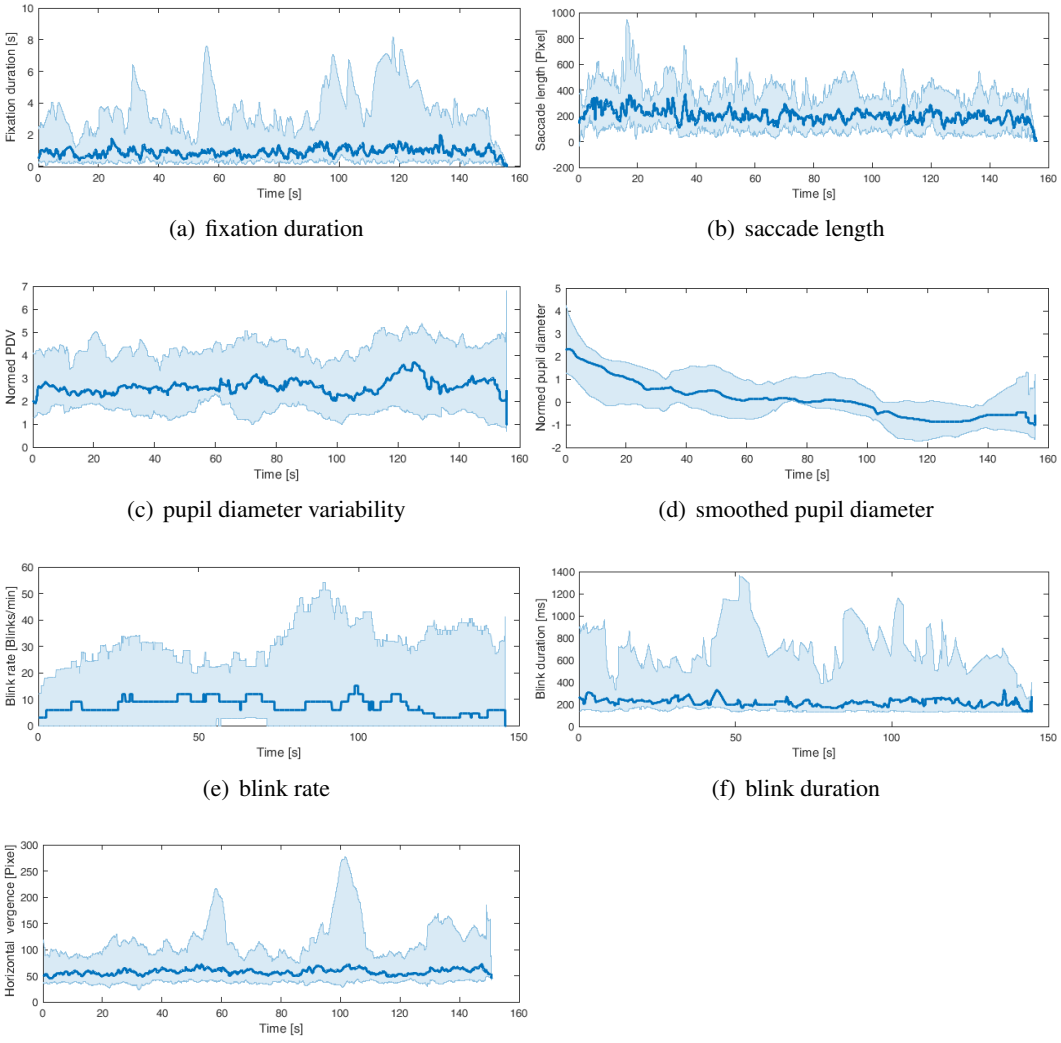
We can observe an increase in blink rate at 80 s trial duration. Blink rate roughly holds that level for the remaining trial. Blink duration on the other hand corresponds well to the 60 s and 100 s peaks that we found for the fixation duration signal.

This timespan is also highlighted by the pixel vergence, which shows a slight increase over time and two large peaks in the confidence interval at 60 s and 100 s.

Figure 3.16 reveals that most subjects exhibit an increase in the inaccuracy to adjust the eyes to the screen depth. Some evolve to an oscillatory pattern in vergence angle (probably to the plane of the computer screen and to a relaxed position further behind the screen). Most subjects exhibit short durations of vergence adjustment to the screen layer shortly after the presentation of a new stimulus.

Discussion

We can conclude that some subjects progressed into a phase where they simply stared at the image: long fixation durations, elongated blinks and an off-screen vergence angle are clear indicators for this behavior.
Generally, pixel vergence and the pupil diameter seem good parameters to detect a decrease in attention. The pupil diameter is well-known to reflect workload and attention. In this analysis, no clear conclusion can be drawn based on pupil diameter variability. This is probably due to the way the EyeTribe calculates and reports this measure. Since the algorithms are a black box we can only try to understand what happens internally. The EyeTribe reports a pixel size for the pupil diameter. This measure depends on gaze location as eyeball rotations will result in an ellipsoid shape of the pupil from the eye tracker's perspective. Thus the influence on the reported pupil diameter measure. Furthermore, we do not have any insight in how accurate the pupil detection is working or whether the iris contour is detected instead (leading to a similar center point as the pupil but very different diameter).
The strongly smoothed average pupil diameter is robust to this effect, but pupil variability is affected heavily. Extensive normalization of the pupil diameter was required to make it somehow comparable between subjects. This may have eliminated many interesting pupil diameter changes. Further analysis of the pupil diameter, such as the calculation of fatigue waves or the index of cognitive activity [103] was not done for these reasons.
The effects in pixel vergence and blink duration were only visible in the averaged signal but not that clear on a single-subject level. This renders the parameters an interesting choice for aggregated measures, e.g., as weighting factors for the calculation of attention heatmaps. Fixation locations with a pixel vergence that do not fit the image plane could be discarded or weighted less than fixation locations that hit the image plane.



(a) fixation duration

(b) saccade length

(c) pupil diameter variability

(d) smoothed pupil diameter

(e) blink rate

(f) blink duration

(g) pixel vergence

Figure 3.15: Candidate parameters for vigilance and attention indicators. The median over all subjects is depicted as a bold line, 75% confidence intervals are shown as area in a lighter color. Only the viewing of the fourth stimulus is depicted

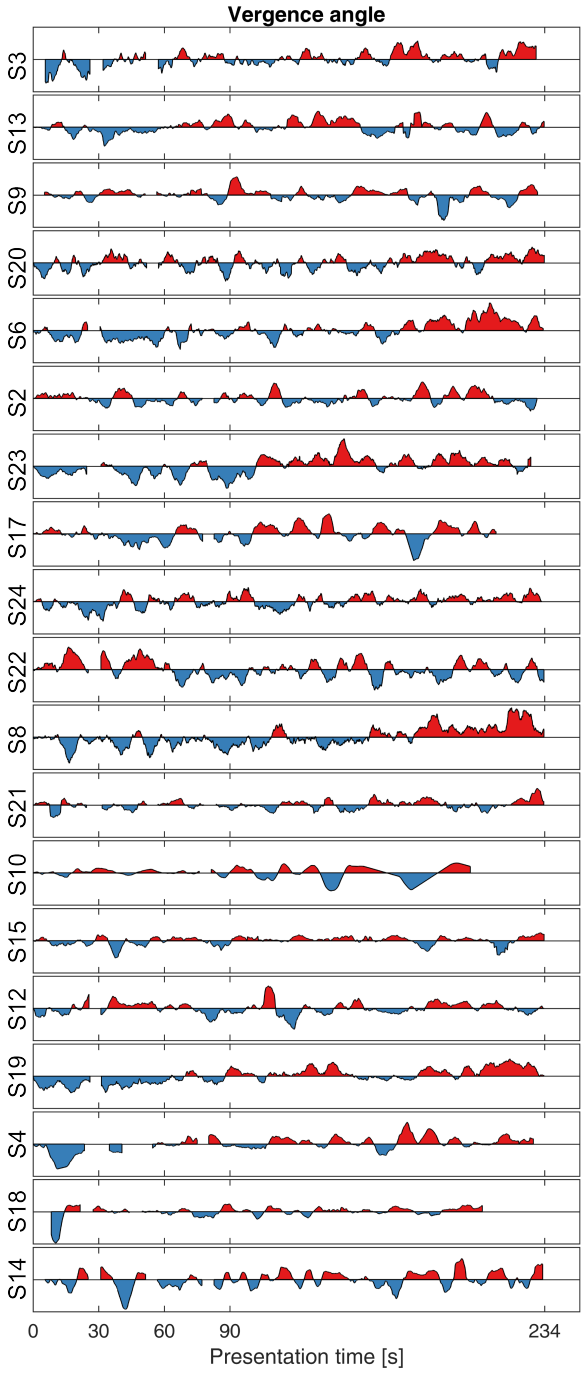


Figure 3.16: Pixel vergence over the total experiment duration (all stimuli). The timestamps on the abscissa mark stimulus changes, the ordinate displays vergence as pixel distance in the image plane. A smaller convergence angle is shown in red, larger convergence, i.e., a better match to the screen depth, in blue. Subjects are sorted by correlation of the signal, so that similar curves are closer to each other

3.3 Sensor fusion: stress parameters complement eye-tracking

Pomplun's statement that "not every fixation is filled with attention" [83] requires the extension that not every attention allocation requires a fixation towards its location. We are able to perceive peripheral stimuli and to distribute covert attention without an immediate and inevitable manifestation in eye movements [4].
This section investigates how physiological parameters such as ECG and galvanic skin conductance can be combined with eye-tracking to derive information on visual perception. The eye-tracking signal is used to identify the fixation of a hazardous object and a variety of biosignals is utilized to judge whether the object was perceived as a potential hazard.
In this work [18] gaze behavior of patients with binocular visual field defects is studied during a driving task in a driving simulator. The patient group involved in this study suffers from blindness in congruent parts of their visual fields. Especially for this group of subjects, a mere fixation of a target object may not be sufficient to reliably tell whether a target has been perceived.
A local overlap of the target and the visual field defect may prevent perception of the target. Furthermore, perception of a hazard does not automatically imply an adequate response (such as an emergency brake). Therefore, the intuitive assumption of non-perception derived from an inadequate driving response cannot be made. On the other hand, overlooking a hazard does not always result in an accident. Careful driving behavior such as reduced speed may resolve even non-perceived hazardous situations.
The occurrence and absence of three physiological measures, i.e., heart rate, pupil dilation and galvanic skin conductance, are correlated with hazard fixation and the driving response. This way the biosignals complement eye-tracking and may help to resolve situations, in which it is unclear whether the perception of a hazard failed or no adequate driving reaction was found in-time.
The above physiological measures have been considered as indicators of cognitive workload and stress during driving in related literature [104, 105, 106, 107]: safety-critical situations excite the sympathetic system, resulting in an increased heart rate and transpiration, measurable via ECG and a change in skin conductance.
Aim of this study was to
I) assess driving safety of subjects with binocular visual field defect.
II) correlate driving performance and compensatory eye and head movements.
III) disambiguate hazard fixation, perception and driving response by joint analysis of eye-tracking data and stress-associated biosignals.

3.3.1 Background: visual field defects

Homonymous visual field defects

Homonymous visual field defects (HVFD) occur as a consequence of stroke, brain tumors, or traumatic brain injury [108]. The prevalence has been estimated at 0.8% in a population aged 49 and above [109]. Driving safety has been studied in two ways mainly: first, police charts and self-reported accidents, which provide data of numerous subjects but a very

coarse characterization of individual defect areas and an overall low detail level [110].
Second, there are driving sessions in a simulator or on-road. These studies are expensive to conduct as subjects have to be recruited, examined and the driving session has to be performed. Thus, the number of subjects is usually low, but there is detailed information about the visual field defect and the driving situations affected. Contrary to large surveys than accumulate data over years, only very short time intervals (e.g., about 40 minutes of driving test) are considered. Whether these stressful test situations resemble actual everyday driving behavior is also questionable.
Most studies concur that there are difficulties with lane keeping, unstable steering, and inadequate hazard detection [111, 112, 113, 114, 115, 116]. Several studies found a subgroup of patients that showed no reduction in driving safety and driving performance at all [117, 118].
As a possible explanation for the unaffected driving performance, compensatory movements of both head and eye were suggested. These movements shift the intact visual field over the environment and thereby perform a successive scanning of the complete environment. An increased amount of scanning in the blind hemifield and towards collision-relevant objects can therefore be expected and was found in several studies [112, 118, 119, 117, 120].
But a robust correlation between compensatory movements and actual driving performance could not be proven [116, 121], probably because of severe limitations of the used driving simulations that restrict the generalization to more realistic driving scenarios [118]. Not all the studies mentioned in this section utilize eye- and head-tracking. Both are often based on manual analysis of video footage (i.e., manually counting *head movement events* instead of continuous measurement and objective quantification) [115, 117].

Glaucoma

66.8 million people worldwide are affected by Glaucoma, a progressive optic neuropathy. Prevalence increases with age and the numbers are expected to increase with demographic aging and the rise of emerging nations [122]. The shape of the defect area can range from nasal steps, temporal wedge defects, arcuate defects, paracentral defects, to generalized constriction and total vision loss.
For this study we focus on cases where corresponding areas of the visual field in both eyes are affected. This leads to a reduction of function in the binocular overlap area that cannot be compensated by the fellow eye. Such advanced stage glaucoma patients report restrictions in outdoor mobility and, therewith associated, in quality of life [123].
As for the homonymous visual field defects, accident reports show up to six times higher risk of involvement in a traffic accident for glaucoma patients ($n_{patients}$=48) [124] and they are also more likely to be at fault. However, self-regulation such as not driving at night or during rush hours, is common for glaucoma patients and leads to a lower likelihood to be involved in collisions ($n_{patients}$=576) [110] and no increased risk for injurious collisions ($n_{patients}$=234) [125].
The results of simulator driving sessions and real-world accident reports are ambiguous. Some studies report an increased risk of traffic accidents for a horizontal vision restriction to 100° ($n_{patients}$=40) [126], others find no significant change for mild and moderate

glaucomatous vision changes ($n_{patients}$=25) [127].
In [128] short video clips were presented to patients with glaucoma and their gaze behavior was analyzed ($n_{patients}$=9). The main finding was that a 16% higher number of saccades performed, but no difference in the looked-at points was found. In an on-road study, patients with binocular visual field loss (both HVFD and glaucoma, $n_{patients}$=20) who passed a driving test were found to perform more gaze, head and shoulder movements towards the defect area [117].
We can conclude that the ability to drive safely can not clearly be associated with presence or size of a visual field defect per se. There is evidence that a subgroup of patients retains the ability to drive safely whilst others are at an increased risk.

Driving regulations

Binocular visual field requirements for driving are regulated in the European Community Directive on Vision and Driving (2011) as greater than or equal to 120° on the horizontal meridian. There must be no significant visual field defect in the binocular field within 20° above or below the horizontal meridian.

3.3.2 Driving simulator experiment

Participants

Eight HVFD patients and six healthy-sighted, age- and gender-matched control subjects as well as eight glaucoma patients and eight healthy-sighted, age and gender matched control subjects were involved in this study. Detailed information on the visual field defects can be found in Figure 3.17. Age difference between the patients and their respective control group correspondence was at most 5 years.

Average time since first diagnosis of glaucoma was 15.0 ($\pm$5.8) years. Glaucoma diagnosis was confirmed based on optic nerve damage and visual field loss. Only advanced stage glaucoma patients with defects in the binocular visual field were included:

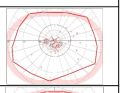
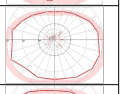
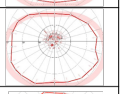
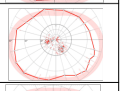
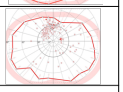
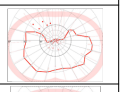
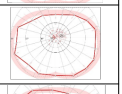
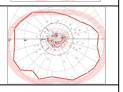
- defects in the upper visual field of both eyes (n=4)
- defects in the lower visual field of both eyes (n=1)
- defects in both the upper and lower visual fields of both eyes (n=3)

Three HVFD patients with right-sided, five left-sided defects were included. Two experienced neuro-ophthalmologists evaluated the visual field defects as

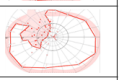
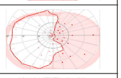
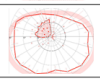
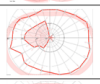
- complete homonymous hemianopia (n=3)
- incomplete homonymous hemianopia (n=2)
- incomplete homonymous quadrantanopia (n=3)

Four patients had macular sparing.

Visual fields were assessed by binocular semiautomated 90° kinetic perimetry (SKP, Octopus101, Fa. Haag-Streit, Koeniz, Switzerland), to provide representation of the patients'

Gender Age [yrs.]	Δt [yrs.]	Pathogenesis	Visual field defect (90° eccentricity, binocular, stimulus III4e)
m 66	18	Inferior arcuate defect and inferior paracentral scotoma within 20° eccentricity	
f 59	14	Circumscribed superior paracentral scotoma	
m 52	23	Superior arcuate defect and inferior paracentral scotoma within 10°-20° eccentricity	
m 74		Left hom. circular anopia in central visual field, Right hom. circular anopia in central visual field	
m 45		Hom. circular anopia in superior 30°visual field	
59 ± 10	*18±5*	*Passed subjects (mean ± standard deviation)*	
m 59	15	Superior altitudinal visual field loss respecting the horizontal meridian	
f 71	5	Circumscribed superior paracentral scotomas	
f 62	5	Inferior arcuate defect and superior circumscribed paracentral scotoma within 20° eccentricity	
64 ± 6	*8 ± 6*	*Failed subjects (mean ± standard deviation)*	

(a) Glaucoma

Gender Age [yrs.]	Δt [yrs.]	Pathogenesis	Visual field defect (90° eccentricity, binocular, stimulus III4e)
f 50	8	Ischaemic stroke (parieto-occipital area) Right incomplete superior quadrantanopia with macular sparing	
m 72	1	Ischaemic stroke posterior cerebral artery (PCA) Left incomplete homonymous hemianopia with macular sparing	
m 55	0,5	Ischaemic stroke (occipital area) Right complete homonymous hemianopia with macular sparing	
m 48	6	Ischaemic stroke (occipital area) Left incomplete superior quadrantanopia	
56 ± 11	*4 ± 4*	*Passed subjects (mean ± standard deviation)*	
m 38	7	Brain surgery Complete left homonymous hemianopia	
f 48	10	Traumatic brain injury Left incomplete homonymous hemianopia with sparing of the peripheral visual field	
f 41	16	Ischemia occipital Right homonymous hemianopia	
f 63	8	Ischaemic stroke (temporo-parietal area) Left incomplete superior quadrantanopia with macular sparing	
48 ± 11	*10 ± 4*	*Failed subjects (mean ± standard deviation)*	

(b) Hemianopia

Figure 3.17: Visual field defect and demographic information of all patients. Δt denotes the time since brain lesion for hemianopia [13] or the time since first diagnosis of glaucoma [12]

visual field in agreement with driving test requirements in Germany. The stimulus III/4e with background luminance $10\,\mathrm{cd/m^2}$ and 3° /s angular velocity was used. All patients had owned a driving license for years, but none of them met the legal criteria for driving because of their visual impairment at the time of inclusion.

Inclusion criteria were a best-corrected monocular visual acuity of at least 20/25, normal function and morphology of the anterior segment and visual pathways. Patients with significant cognitive decline, visual hemineglect or physical impairment that affected vehicle use were excluded. A mini-mental state examination [129] score of at least 24 was required (24-30 = no relevant cognitive impairment) [130].

Patients were recruited from the Department of Neuro-Ophthalmology at the University of Tübingen (Germany). The research study was approved by the Institutional Review Board of the University of Tübingen and was performed according to the Declaration of Helsinki. After verbal and written explanation of the experimental protocol, all subjects gave their written consent, with the option of withdrawing from the study at any time.

Driving simulation

The study was conducted in a moving-base driving simulator at the Mercedes-Benz Technology Center in Sindelfingen (Germany) [131]. A cabin with full 360° projection and a car body is moved via a hexapod and a 12 m long rail to enable the experience of full inertial characteristics of an actual motor vehicle.

Nine hazardous situations were placed along the route. It took an average of 37($\pm$2) min to drive the 37.5 km route that contained rural and urban areas with speed limits varying between 100 km/h and 50 km/h, respectively.

The nine hazard situations are described in Table 3.4 and their location on the route is shown in Figure 3.20. As a simulated crash would put additional emotional stress on the subjects, hazardous situations were resolved by the simulation shortly before a crash would happen. E.g., pedestrians leaped backwards onto the sidewalk and oncoming overtakers returned to their lane.

Table 3.4: Hazard description and their location during the driving course

Location (km)	Hazard
9.5	Pedestrian crossing from the left
15.7	Overtaker on left curve
24.9	Overtaker on right curve
30.7	Pedestrian crossing from the right
32.4	Parking car filters in from the right
33.9	Car crossing from left
35.1	Car crossing from right
36.8	Pedestrian crossing from the right
37.5	Pedestrian crossing from the left

The first kilometers were dedicated to simulator familiarization, beginning with a straight road segment and no oncoming traffic and slowly introducing more elements. The route was divided into two parts, the first 30.7 km resembling realistic driving with only four hazardous situations and the remaining 6.8 km with five densely packed hazards.

Since we expect reaction times to be increased on the side of the visual field defect (for the HVFD patients that have a primary defect side), available reaction time from hazard onset to crash has to be equal for hazards at both sides of the road. Under right-hand traffic conditions, hazardous objects approaching from the left have to cross the left lane first and would therefore be visible earlier on. This was avoided by hiding the objects, e.g., behind parked vehicles and advertising pillars.

A certified driving instructor who was masked to the participants medical status evaluated the driving responses to the hazardous situations as passed or failed according to the requirements of the official German driving test.

Biosignals

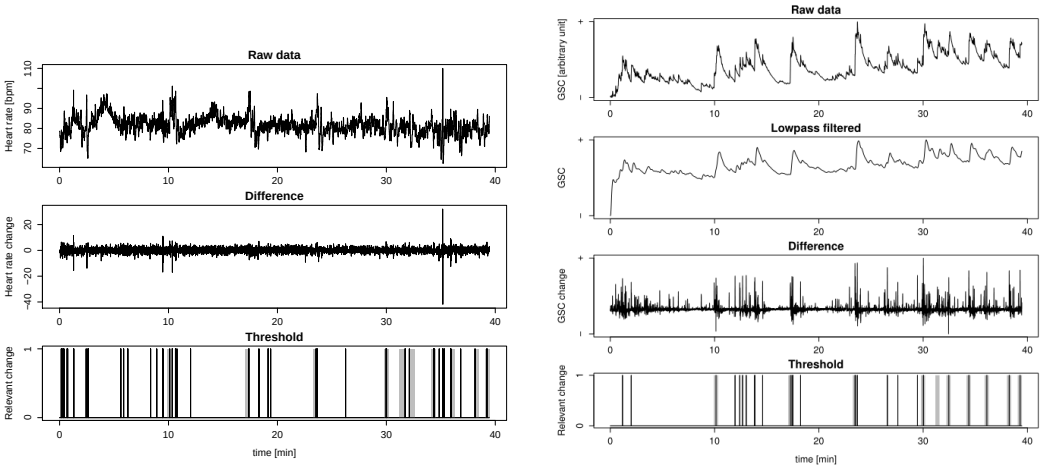
Galvanic skin conductance (GSC), heart rate (ECG) and eye movements were recorded. Figure 3.18 shows the experimental setup and devices attached to the subject.
Gaze behavior was captured by a monocular Dikablis (1.0.9) head-mounted eye tracker by Ergoneers GmbH, Manching/Germany. The device is able to measure through the participants' habitual glasses with a frame rate of 25 Hz. Nine-point calibration was performed with calibration targets projected to the driving scene.
The mobile 3-channel custo med ECG device was used for recording the heart rate. Skin conductance was measured by the Biotrace+ system using electrodes placed at the fingertips.
Head position and orientation was captured with theoretical millimeter accuracy by the optical laserBIRD system (Ascension Technology Corporation, Burlington, USA). It consists of a small sensor attached to a headband and a larger scanner device that was mounted above the passenger seat.

Figure 3.18: A subject equipped with the Dikablis eye tracker, a head-tracker headband and GSC sensors at the fingertips. The ECG electrodes are placed on the chest with only the cables visible [18].

Data processing

Biosignal data was available for 15 of 16 patients and 12 of 14 healthy sighted control subjects.
The ECG signal was preprocessed by the recording device to a continuous heart rate. For a stressful event we expect an increase in heart-rate. The exact amount of this rise and therefore, the threshold to classify it as caused by stress is highly subject specific. For subjects with a high overall variation in heart rate, a higher threshold is required. We chose

to set the threshold to three standard deviations σ of the individual's heart rate. Under the assumption of a normal distribution over 99% of the data is contained within this interval. GSC raw data was smoothed by a Butterworth low-pass filter in order to reduce noise. Identical to the ECG data, a threshold of $3 \times \sigma$ was applied to the GSC change, resulting in a signal similar to [132]: investigating changes between small time intervals results in a compensation for low frequency changes, such as a steady but slow increment in GSC. The slow progression is distributed equally over the whole recording, and therefore, becomes negligible for each single time step.



(a) ECG (b) GSC

Figure 3.19: (a) ECG data preprocessed by peak detection and calculation of the inter-beat interval as the inverse of the current heart rate. The change in heart rate at each time interval is calculated and a threshold of $3 \times \sigma$ applied. (b) Skin conductance is low-pass filtered, then the $3 \times \sigma$ threshold is applied to the differences between subsequent measurements.

Blinks and tracking losses were removed from the eye-tracking data. Fixations were identified using the mixture of Gaussian model [20] and ellipses were fit to the samples in order to estimate their spatial extent and to assess measurement accuracy. Eye-tracking calibration was accurate (defined as within 5°) for an average of 20 minutes. Afterwards, 2/3 of the recordings required offline calibration adjustment due to displacement of the eye tracker.

Hazardous objects in the driving simulation were annotated manually by bounding boxes. Each second frame of the eye tracker's scene image was annotated, the remaining frames interpolated. Hazardous situations were analyzed starting from the moment when the hazardous object first became visible to the driver up to the moment when it was resolved (either through a driver reaction or, in case of no response, by the simulation). The bounding boxes of the hazardous objects were intersected with drivers' fixations to determine whether the target was fixated.

ECG, GSC and eye-tracking were time-synchronized with the driving simulation: the simulator projection displayed a special frame at the beginning of each measurement that

was used to synchronize the simulator timestamps with the eye tracker. The same laptop that was used for recording the eye-tracking data also recorded ECG and GSC. ECG, GSC and eye-tracking timestamps are therefore synchronized with each other, except for the delay of the ECG and GSC devices. We did not compensate for this delay but assume that it is relatively small when compared to the low sampling frequency of the eye tracker. The synchronization of the simulator and the eye tracker requires a high accuracy as we intersect the spatial information of the eye tracker with the simulator image. But the biosignal devices do not require highly accurate synchronization as long as we do not want to determine the time delay between hazardous situations and a biosignal response within millisecond range. For determining whether a hazard was perceived as dangerous (not when exactly) synchronization within some fractions of a second is sufficient.
A stress response is manifested in a rapid pupil dilation, followed by a gradual return to its normal size. The challenge is the analysis of the pupil signal lies in filtering out all the large variations in pupil size caused by the ambient illumination. Only those peaks that correspond to stress, as the response to a hazardous situation, are of interest. The pupil diameter signal was processed by blink and invalid data removal. If less than 75% of the trial was valid data, the signal was marked as not available.
A least-squares quadratic fit of a parabola to peaks in pupil dilation is performed. Only peaks that exceed 1.5 standard deviations are considered as valid candidates. A time window of 15 s around the peak is extracted from the signal. On those snippets a transformation using Daubechies wavelets up to the level 4 is performed.
Amplitude, mean, area of the approximation coefficient A4, and the relative energy that corresponds to the detail coefficients D1-D4 is used as a feature descriptor for the peak.
A machine learning approach was employed to distinguish hazard responses from illumination changes based on these features. More specifically, a Support Vector Machine (SVM, see Section 6.1) with radial basis function kernel was trained to distinguish those peaks that correspond to hazardous situations from all others.
Leave one out cross-validation with random oversampling was performed. The SVM was trained on all available data except for data of the subject that we are currently classifying.
Furthermore, the driving performance parameters *average lane position* and *time to line crossing* (TLC) of second order was calculated (for a reference on TLC and the calculation see [133, 134]). TLC is defined as the average time without any steering or acceleration action until any wheel of the vehicle crosses either lane boundary. It is a measure of the driver's ability to keep a stable lane position.

3.3.3 Results

Driving Performance

Driving safety. Seven of the sixteen subjects with binocular visual field defect (4/8 hemianopia, 5/8 glaucoma patients) and one subject of the control group failed the driving test. Figure 3.20 shows the situations that led to a driving test failure for each subject.
Contrary to our initial assumptions, no increased collision rate for objects approaching from the visual field defect side was observed. Instead, most driving test failures occured at

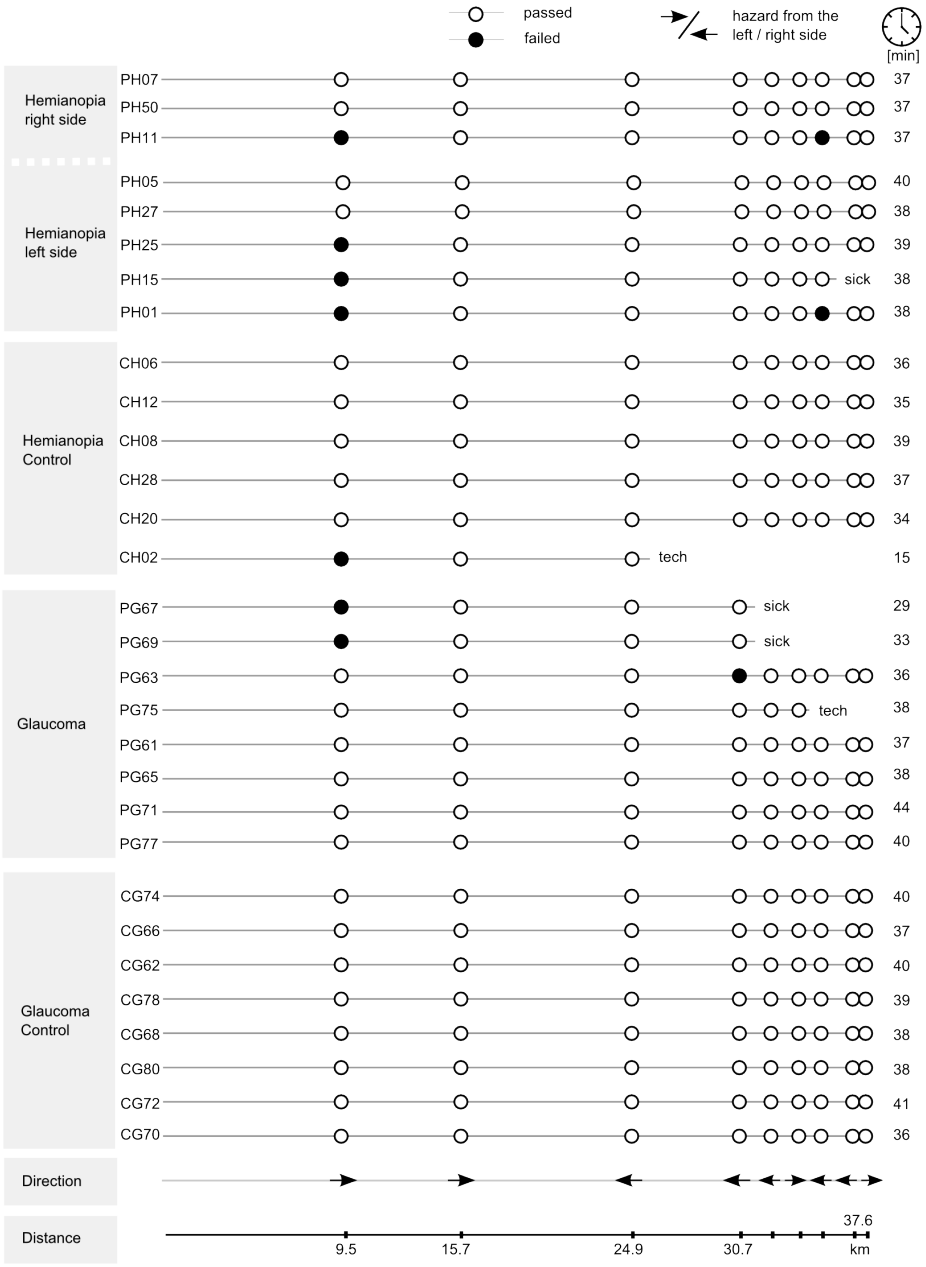
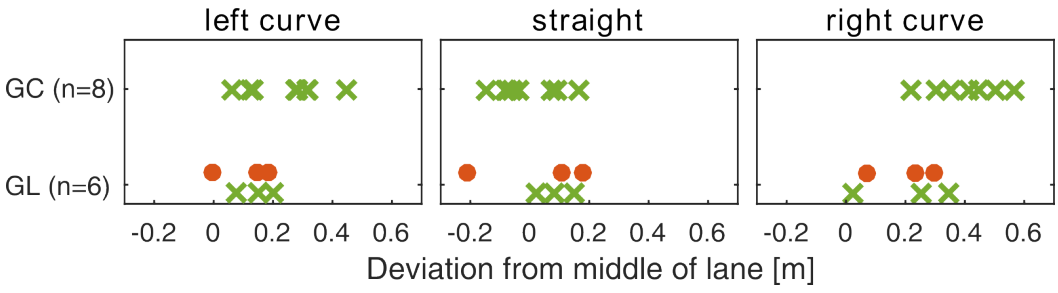


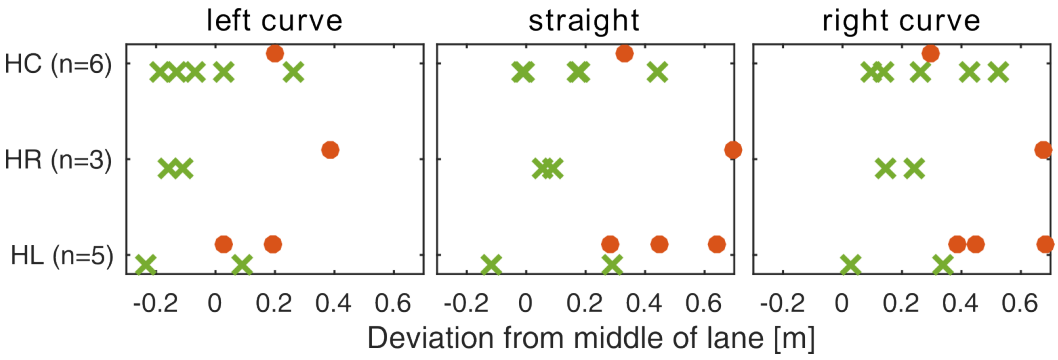
Figure 3.20: Outcome of the driving test for each participant (each row corresponds to one subject) and each hazardous situation (in the columns). Each circle represents a hazard encountered by one of the subjects. The last column shows the time required to complete the route. Arrows indicate the side the hazardous object was approaching from. Participants that aborted prematurely are marked with either *tech* if technical difficulties with the driving simulator facility or *sick* if simulator sickness led to the early stop. Figure adapted from [12, 13]

the first hazard situation. The situation required a very fast reaction. We can assume that subjects were unprepared for what was coming and drove more carefully afterwards. This may be the main reason for the lack of a side bias. Furthermore, no correlation was found between the mean time since brain lesion or first glaucoma diagnosis and driving fitness. Five subjects (4 patients, 1 control) had to abort the drive early due to simulator sickness or technical problems. Nevertheless, all participants completed more than half of the route.

Lane position. The lane position adopted by the control group is considered as a reference position. During curves and evasive driving the optimal lane position may vary significantly from the center of the lane. As all participants completed the exact same route, these effects can be canceled out when comparing to the control group (see Figure 3.21).



(a) Glaucoma



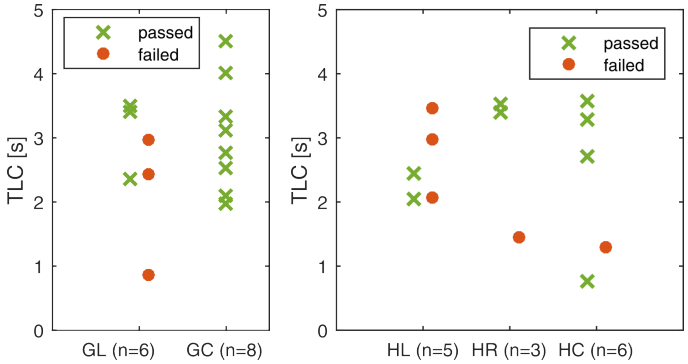
(b) Hemianopia

Figure 3.21: Average lane position for the HVFD groups compared to their respective control group on different road segments. Subjects passing the driving test are depicted as crosses, subjects failing the driving test as dots. GC = Glaucoma Control; GL = Glaucoma Patients; HC = Hemianopia Control; HR = Hemianopia Right side defect; HL = Hemianopia Left side defect. Figure adapted from [12, 13]

Glaucoma drivers showed no difference in lane position when compared to the control group. This is also reflected in the TLC measure (Figure 3.22) where no difference between patients and the control group was found.

HVFD patients who failed the driving test maintained a lane position to the right of the control group, both during curves and to a lesser extent on straight road segments. This is probably caused by a desire to maximize the distance towards oncoming traffic. However,

no significant difference in TLC was found. But a solid statistical subgroup analysis cannot be performed given the number of subjects per group in this experiment.



(a) Glaucoma (b) Hemianopia

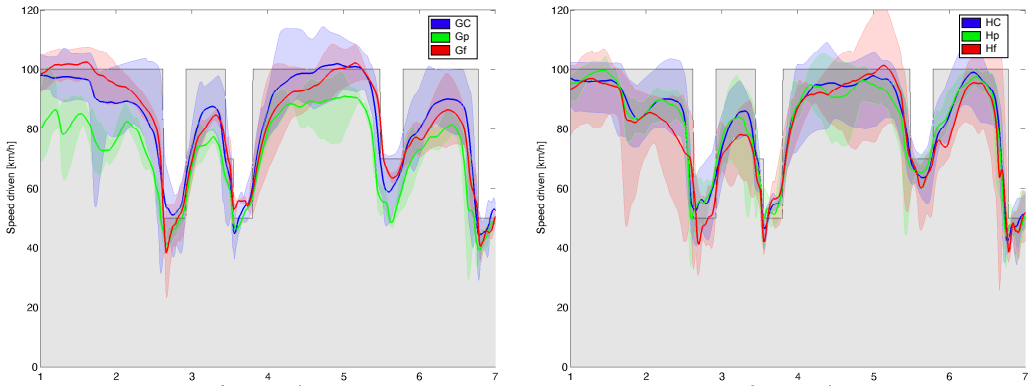
Figure 3.22: Time to line crossing, a measure of steering stability and driving performance. Figure adapted from [12, 13]

Speed. In support of the finding that glaucoma patients self-regulate their driving habits to avoid dangerous situations we found average driving speed was reduced compared to the control group in the group of glaucoma patients rated as safe drivers (55±25 km/h; control subjects 59±27 km/h) but not in the group of unsafe glaucoma drivers (63±24 km/h), see Figure 3.23.
All HVFD patients generally followed the speed limit, thus no problems with identifying speed limit signs or keeping a stable speed could be found. No significant speed differences between control and patient group was found (safe HVFD drivers 62 km/h, unsafe HVFD drivers 59 km/h, control group 60 km/h). HVFD patients did not show a speed reduction.

Head and eye movements

Head and eye movements were analyzed statistically. A linear mixed-effects model (using GnuR lme4 [135]) with the patient group as a fixed factor and the subject as a random factor was fitted. The model was evaluated against a baseline model without the group factor via Anova. An α-level of 5% was chosen for statistical significance. Tuckey's correction for multiple testing was applied where more than one test was performed. Although a statistical analysis is performed, one should be aware of the limited number of subjects and the thereby impaired meaningfulness of significance. While some effects show a clear tendency, others can only suggest possible directions for further research.

Head movements. Glaucoma patients judged fit to drive performed more head movements ($3.9 \pm 2.9^\circ/s$) than the age and gender matched control group ($2.5 \pm 0.4^\circ/s$). Unsafe drivers showed no such increase in head movements ($1.2 \pm 1.9^\circ/s$). Figure 3.24 shows the average amount of head rotations to the left/right for the subject groups. For hemianopia drivers the



(a) Glaucoma (b) Hemianopia

Figure 3.23: Speed driven during the first seven kilometers. The lines indicates the average speed in their respective groups, the boundaries of the colored area show the minimum and maximum speed in the group. The speed limit is shown as the gray area in the background. GC: Glaucoma control, HC: Hemianopia control, Gp: Glaucoma safe driver, Gf: Glaucoma unsafe drivers, Hp: Hemianopia safe drivers, Hf: Hemianopia unsafe drivers. Figure adapted from [12, 13]

difference in head movements between fit-to-drive patients and those who ailed the driving test is more evident.

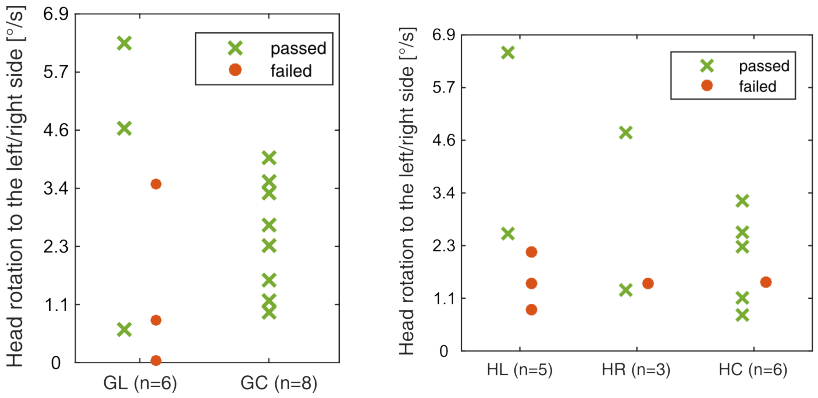
Analysis of head movements for the combined hemianopia and glaucoma group reveals that safe drivers performed significantly more (p=0.048) head movements than patients who failed the driving test. Unsafe drivers did not exhibit any increase in head movements when compared to the control group (p=0.601).

Eye movements. Glaucoma and hemianopia patients rated as unsafe drivers directed a higher proportion of gaze towards the right side of the scenario (Figure 3.25). Safe drivers with glaucoma showed the contrary effect of increased gaze direction towards the left side on cost of attention density to the central region. This can be interpreted as increased peripheral exploration activity in safe glaucoma drivers.

The hemianopia group of safe drivers did not exhibit such a behavior. Instead, their exploratory activity is indistinguishable from the control group.

Figure 3.26 presents various commonly used eye-tracking measures. Although the limited number of subjects does not allow for a robust statistical analysis, several tendencies can be discovered:

Comparison of saccadic amplitudes between glaucoma patients and the control group revealed a tendency for longer saccades in patients rated fit to drive, decreased saccadic amplitudes for unsafe drivers. Glaucoma patients who failed the test showed longer fixations when compared with glaucoma patients who passed the test. This behavior goes well together with the increased peripheral scanning activity that we found for glaucoma patients who were fit to drive. A quick switch of attention between peripheral and central gaze targets seems to be beneficial for driving safety. Horizontal-vertical saccade ratio suggests that safe



(a) Glaucoma (b) Hemianopia

Figure 3.24: Quantification of head movements (i.e., rotation to the left or right side) for glaucoma and hemianopia patients grouped by driving safety. Hemianopia patients are separated by the side of the visual field defect. Figure adapted from [12, 13]

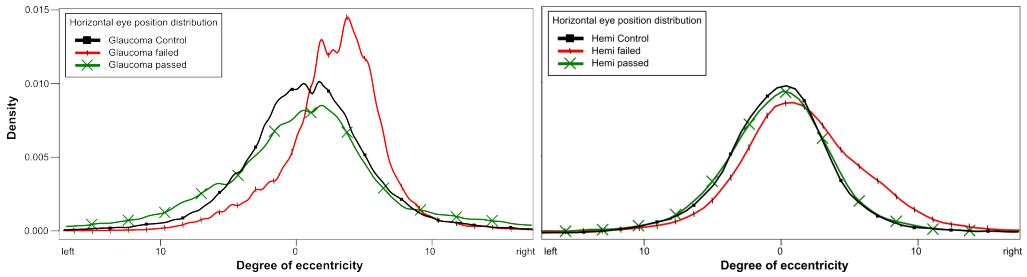
Table 3.5: Data availability for the biosignals in the patient and the control group, summed over all subjects

Measure	#Situations	Patient group	Control group
Target fixation	230	120	110
GSC	218	118	100
ECG	229	119	119
Pupil	159	78	81

drivers invest more scanning in the horizontal direction whilst unsafe drivers require more vertical scanning, perhaps as a consequence of their specific visual field defects.
Saccadic amplitude was also increased in fit-to-drive subjects of the hemianopia group when compared to the unsafe drivers (df=3,12 F(df)=3.52 p=0.049). Unsafe drivers can be characterized by more frequent (df=1,8 F(df)=8.56 p=0.020) and shorter fixations (df=1,8 F(df)=9.55 p=0.015). Surprisingly, the parameters of horizontal-vertical saccade ratio and saccadic orientation do not show any obvious tendency. For hemianopia patients such a change would have been a reasonable compensatory strategy of the half-sided visual field defect.

Biosignals

The number of hazardous situations available for the analysis of biosignals is presented in Table 3.5. As the measurement devices did not record good data for all subjects during the whole drive, the number of situations analyzed differs slightly for each sensor type.
In 93% of the hazardous situations with appropriate driving reaction a skin conductance peak was recorded. Heart rate changes were found in 69% of the hazardous situations associated with adequate driving response for the patient group (81% for the control group).



(a) Glaucoma (b) Hemianopia

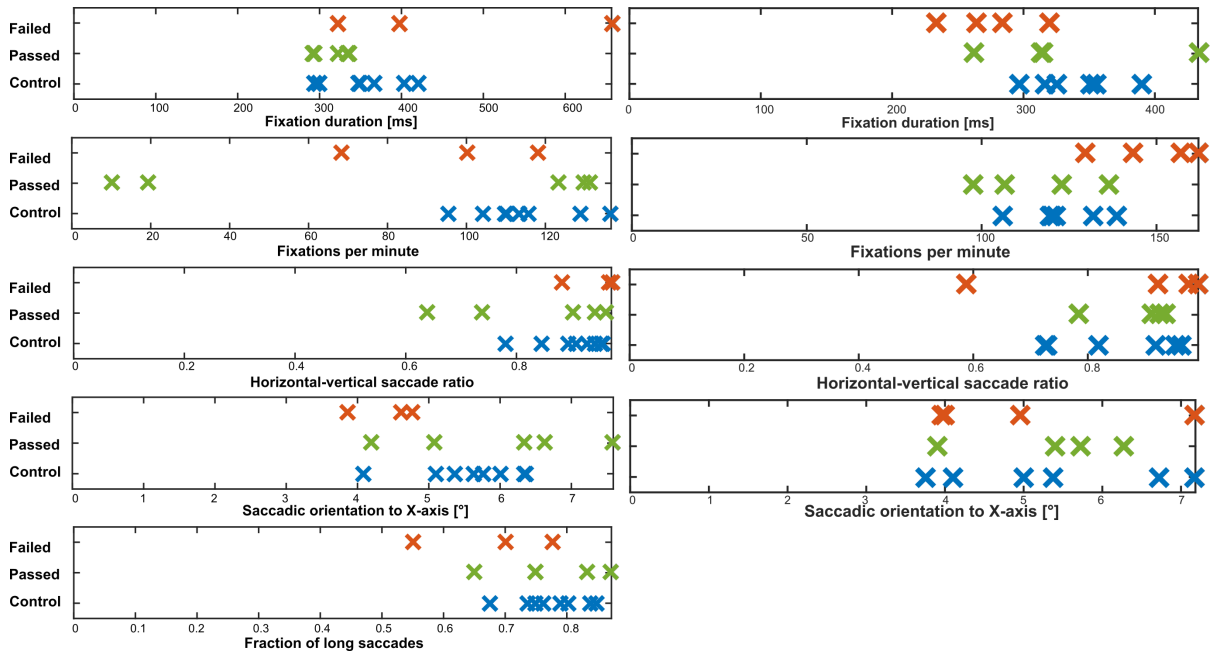
Figure 3.25: Horizontal eye position distributions within the central 20° of visual field. Different density estimation kernels were used for the two patient groups. A different level of smoothness results and the density scale was adjusted accordingly to allow for a comparison. Figure adapted from [12, 13]

Table 3.6: Combinations of target-fixation, behavioral driving response (D) and stress biosignal (galvanic skin conductivity / ECG) and their occurrence counts in the data (n). The 202 situations for which data of all sensors was available were analyzed

👁	D	✋ / 👂	n	Interpretation
x	x	x	140	Target perceived as threatening, appropriate reaction
x	x	–	0	Target perceived as non-threatening, appropriate reaction
x	–	x	6	Target perceived as threatening but no adequate reaction
x	–	–	0	Gaze movement towards target but not perceived
–	–	x	10	Target not perceived but driver realized threatening condition
–	x	x	20	Target perceived as threatening without looking directly at it
–	x	–	3	Target not perceived, no threat
–	–	–	23	Target was overlooked

The results in Table 3.6 allow us to conclude:

- all hazardous situations were reliably stress inducing, if perceived: target fixation always implied a stress response in at least one of GSC or ECG. This measure is valuable for the design of simulator test courses. Too easy situations might result in not relevant results (e.g. all visual field patients might pass the driving test, if the hazardous situations were very easy and detectable early-on).
- reaction to a non-fixated target is quite frequent (12% of adequate driving responses). Utilizing the biosignals, these situations can be subdivided into perceived and overlooked. The finding suggests that the mere lack of a fixation to a traffic hazard is not necessarily a sign that the hazard was overlooked.
- the utilized method does not allow for a prediction of driving response to a hazard as we did not test for in-time detection of the hazard. Therefore, there is a group of events (n=6) where fixation and biosignal data suggests correct hazard identification, but the driving response was inadequate.



(a) Glaucoma (b) Hemianopia

Figure 3.26: Different eye movement measures for both patient groups. Tendencies for the patients to show altered horizontal scanning with longer saccades and shorter fixations can be derived. Figure adapted from [12, 13]

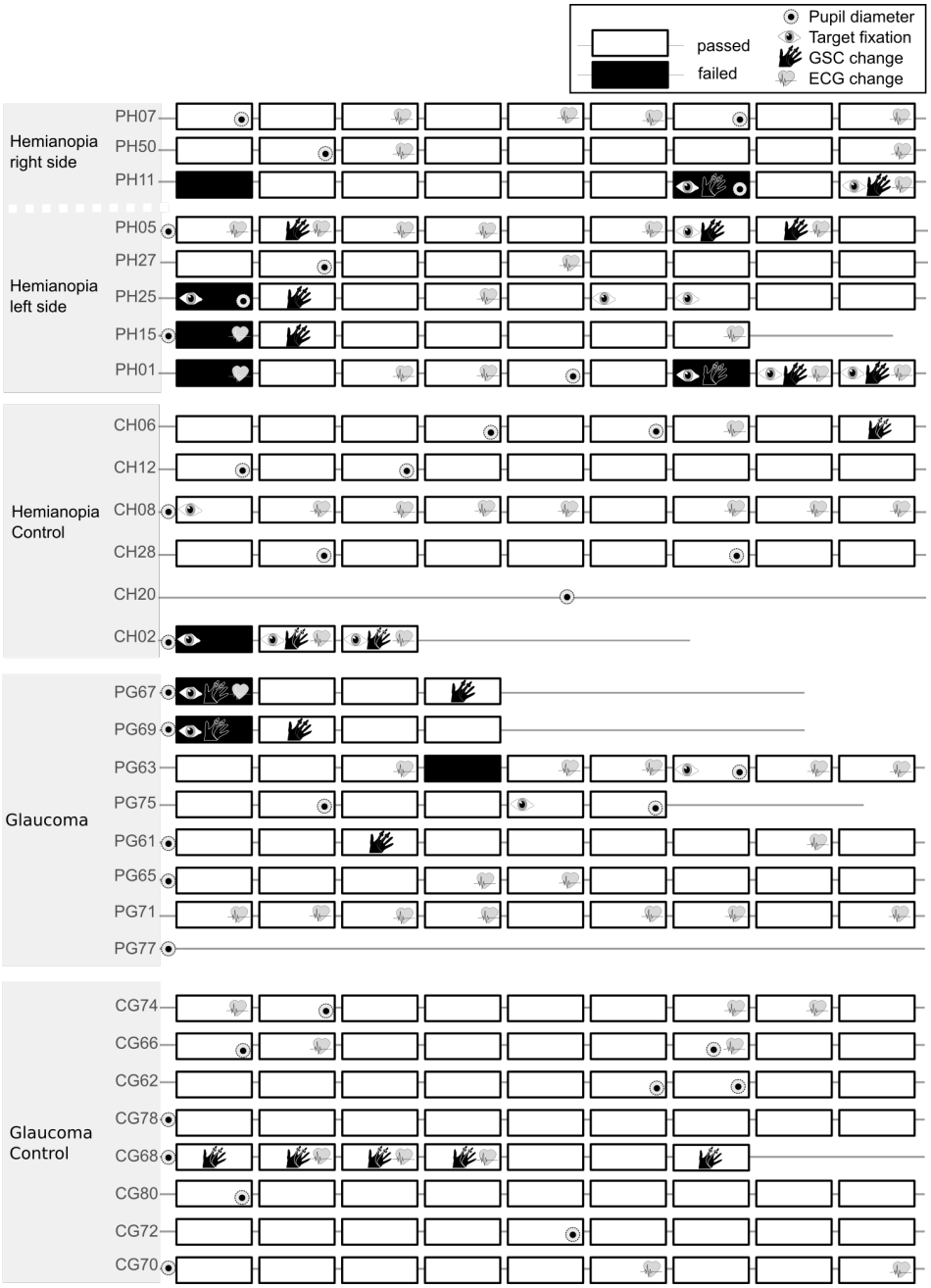


Figure 3.27: Comparison of the biosignal indicators for perception of traffic hazards. Signals that differ from the expected signal are marked. For example, one would expect a GSC change to be present when a hazard and an adequate driving reaction occur but no GSC change for hazards with no adequate driving reaction. Therefore, for passed situations (white boxes) a biosignal symbol indicates the absence of that signal, while for failed situations (black boxes) a biosignal symbol marks the presence of a signal. The pupil diameter marker in front of a row (before the first situation) marks that this parameter is not available for that recording. A blank line means that no biosignal is available for the subject. Figure adapted from [18]

The pupil diameter classifier returned an average of 8 (standard deviation ± 8) false-positives for each drive, i.e., peaks that were classified as hazard response but were not associated with one of the hazardous situations. Without the classification step there were an average of about 50 peaks per drive.

Discussion

HVFD patients as a group show more inadequate driving responses to hazardous situations when compared to the control group. However, when driving performance is analyzed individually, half of the patients are rated fit to drive. As reported in a recent on-road driving study [117], there is evidence for differences in both visual exploration strategy and driving behavior between patients who are fit to drive and those who exhibit unsafe driving behavior. Hence, we aimed at identifying the exploratory patterns and driving performance measures that are associated with safe driving.

In this driving scenario the very first hazardous situation lead to the highest number of driving test failures. The first situation occurred unexpectedly for the driver and likely resulted in a higher alertness. Thus, less driving errors occurred at the following situations.

Prior simulator studies reported varying results with regard to driving safety [136, 137, 138, 113]. The discrepancies to the present study are mainly due to the level of realism of the simulator. Especially the 360° projection in this study enabled a realistic viewing behavior. Patient inclusion criteria (for example the exclusion of hemineglect) also play an important role.

Some studies reported much higher success rates for HVFDs patients, possibly due to lower visual and cognitive demands of a simpler simulation (e.g., a restricted field of view of only $16^\circ \times 21^\circ$ [136]). In a collision avoidance experiment, Papageorgiou et al. [137] reported higher avoidance rates for HVFD patients (up to 77%, $n_{patients}$=26). Hence, a high complexity and difficulty of the driving scenario that exceeds a certain level of cognitive load seems to ne necessary to discriminate between patients who compensate and those who are unable to do so. Our findings are therefore not consistent with [136] where no driving-related disabilities in patients with HVFDs were found ($n_{patients}$=9).

On the other hand, some studies reported very low success rates for HVFD patients. Lövsund et al. [138] found that only 3 out of 26 patients with HVFDs were able to compensate during a detection task. However, the above study used static stimuli that were not part of the driving scene. Similarly, Szlyk et al. [113] reported that performance of patients with HVFDs was worse than or similar to an older control group. However, they included three patients with hemineglect, two months post-stroke and the test was of short duration (5 minutes).

Similar considerations apply to some on-road studies. Tant et al. [114] and Kooijman et al. [139] report a pass rate of HVFD patients less than 15%. These studies, however, included only patients whose driving was suspected to be unsafe.

This study is consistent with Coeckelbergh et al. [140], where 43% of patients with peripheral VFDs passed an on-road test ($n_{patients}$=87) and Kasneci et al. [117] with 40% of patients with binocular glaucomatous visual field loss passed an on-road test ($n_{patients}$=20).

Compensatory gaze patterns

In accordance with previous studies, our results confirm increased head movements and longer saccades in patients who were judged as fit-to-drive [137, 117, 139]. Compensation by increased saccadic amplitudes in patients with HVFDs has been reported by various authors in less advanced driving simulators [137, 141]. Furthermore, in a simulated collision avoidance task, the authors reported larger mean gaze eccentricity in the group of adequately compensating patients (who had few collisions) compared to the control group [137].
Recently, Bowers et al. [115] also quantified head scanning and found that hemianopia drivers had impaired detection of blind side stationary pedestrians at simulated intersections, either due to not scanning or an insufficient scan magnitude. The same authors found that successful detection of a pedestrian moving on a collision course in the blind field was associated with a saccadic eye movement towards the target [116]. Pedestrians were presented at a relatively small eccentricity of 14°, which is within the range of a typical saccade (rarely greater than 15° in normal individuals) [142].
Similarly, more frequent compensatory saccades to the defect side but no head movements were found for one patient who had no collisions in a simulator study [120]. However, the horizontal field of view was only 58° and the contribution of head movements to gaze compensation might have been underestimated. Possibly for the same reason Zangemeister et al. [143] reported smaller head movement proportions in combined head-eye saccades, and suggested that head movement programming is more time-consuming for HVFD patients ($n_{patients}$=6).
We found that glaucoma patients safe to drive showed increased exploration activity in terms of more eccentric head movements when compared to glaucoma subjects who failed the driving test. Thus, the present simulator study replicated the findings of a recent on-road study [117] and the study of Coeckelbergh et al. [140] by means of sophisticated eye and head tracking and suggests that active scanning by means of head movements is indeed an efficient compensatory strategy.
However, other studies investigating visual search behavior of patients with HVFDs in everyday tasks have confirmed the use of compensatory eye and head movements [117].
In accordance with previous literature, our findings indicate that patients with homonymous visual field defects perform more and shorter fixations than control subjects [140]. Additionally, we can link this mainly to the subgroup of drivers who failed the driving test. Unsafe glaucoma drivers displayed a tendency for shorter saccadic amplitudes, a gaze bias to the right, and a more straight-ahead eye position. Increased gaze concentration toward the road center with increasing cognitive load, a phenomenon commonly coined as *tunnel vision* was reported by Engström et al. in a driving simulator study [144].
Our results regarding longer saccadic amplitudes and more lateral eye position in safe drivers are consistent with recent studies using video-based hazard perception tasks [128]. The authors also reported that patients with binocular glaucomatous visual field defects performed more eye movements than control subjects.
Our results are consistent within a study regarding eye movement behavior when viewing a dynamic driving scene. In the latter study, glaucoma patients produced more and, thus, shorter fixations than the control group when searching for hazards in the hazard perception

test [145]. Hence, viewing behavior appears to be related not only to compensatory potential but also to the task complexity and quantity of visual information.
Interestingly, unsafe drivers in our study showed a gaze bias to the right. This was probably an attempt to maintain a stable lane position, because no differences in lane position were found between safe and unsafe drivers. This is in line with Vega et al. [146], who attributed this finding to the optimal control theory of manned-vehicle systems. A possible explanation is that safe glaucoma drivers paid more attention to avoiding traffic hazards, whereas unsafe glaucoma drivers attempted to maintain a stable lane position but failed to recognize traffic hazards because of limited gaze compensatory reserves.

Driving performance measures

Problems with steering stability and lane keeping in patients with hemianopia have been reported for simulator and on-road studies [111, 113, 117]. A previous study [147] showed that safe hemianopic drivers maintained a central lane position. We found that hemianopia patients who failed the driving test maintained a lane position towards the right side of the road (both with left and right sided visual field defect). One explanation is that drivers try to increase their safety margin towards the oncoming traffic; however, for patients who cannot compensate, this behavior leads to an unacceptable rightward lateral deviation and test failure.
In [148] drivers with HVFDs adopted a lane position toward their seeing field. Contrarily, in [114] patients with right sided defect drove too close to the right side of the road. Our study and [139] find no relationship between the side of the lateral deviation and the side of the visual field defect. Oncoming traffic seems to be a source of feedback and correction for the deviant lateral position.
Consistent with [127], TLC and lane position were similar between glaucoma patients and control participants, probably because of behavioral hypervigilance in the patient group.
Speed management skills have been reported to be problematic in patients with HVFDs, as safe drivers drove at higher speeds in the Wood et al. study [147]. Other authors reported that patients with HVFDs dramatically reduce speed in an attempt to compensate [114]. No difficulties were found in our study population in accordance with previous simulator experiments and a recent on-road study [113, 136, 117, 149].
During the more dense second part of the drive, an increased rate of driving test failures can be observed for the hemianopia group. Additionally, the biosignals indicate an increased number of non-perceived hazardous objects that did not result in an immediate threat due to careful and anticipatory driving. Interestingly, this finding was observed only in the hemianopia group and may be associated with the visual impairment: under the assumption that gaze patterns change with regard to the extent of the visual field, central visual processing capacities and working memory availability [150], an increasing memory involvement during the driving task and effective gaze adaptation must be implemented to track vehicle movements continuously [137].
Previous research suggested an impaired working memory of HVFD patients that influences adequate compensation in visual search tasks [151, 152].

Therefore, problematic visual scanning or reduced working memory availability of HVFD patients may lead to an increased rate of failed situations.

Biosignals

Our hazard situations were created to be easily overlooked and resembled a looming emergency. They are therefore very attention arousing and stress inducing. For less challenging scenarios where the hazardous objects can be detected earlier and sufficient reaction time is available, no stress signals would be expected. Eye-tracking measures would then be sufficient. Indeed, GSC and ECG signals were more often absent in situations that allowed for an extended reaction time (e.g., situations 2 and 3).

In conclusion, the GSC signal was superior to ECG in terms of confirmatory power for a driver reaction. Congruous with other studies [153] we found the heart rate signal not specific enough for traffic hazard perception. Many other factors influence the heart rate, such as respiration, emotional status and fitness level [132]. Heart rate is probably sensitive to traffic hazards; however, it is not specific. Our approach to consider only the largest changes in heart rate in order to achieve a similar false positive rate to GSC may have eliminated these peaks.

We are able to disambiguate several hazardous situations by the biosignals: for example, patient PH11 did not perceive the first situation at all and showed no reaction to the hazard. Consequently, the patient failed the driving test at this point. The seventh situation of that patients looks exactly the same in terms of an inadequate driving response. But we can proof that the patient did perceive and realize the hazard. On the other hand, situation nine was passed although obviously not perceived at all. Careful driving behavior may have prevented an accident in this situation.

We found differences in several key measures of driving performance as well as eye- and head movements between subjects who passed and subjects who failed the driving test. These findings support that averaging over the whole population of patients, as it is done when analyzing accident rates and police reports, does not reflect the whole nature of the defect. Instead of a generally higher risk of accidents, subjects can be assigned to subgroups of safe and unsafe drivers.

Limitations

The total number of 30 subjects (16 patients and 14 controls) in this study is considerable. However, the heterogeneity of visual field defects (glaucoma, hemianopia, quadrantanopia) and driving performances resulted in the need to subdivide the subject pool to 3-8 subjects per subgroup. At this level statistical methods are only of limited use.

The standardized traffic conditions in a driving simulator lead to comparable conditions of the driving test. Traffic density can be standardized and hazardous situations provoked without any actual danger to the subject. However, the validity and relevance for real-world driving can be questioned. Furthermore, subjects were aware that they were performing a driving test, which may have influenced their typical driving behavior and alertness.

Although real-time knowledge of driver attention would be extremely useful, e.g., for interaction with driving assistance systems, the methods employed here are not suited for this purpose: skin conductance changes occur with a relatively long and irregular delay (approximately 1–4 s) and the pupil dilation processing chain requires the whole dilation wave for processing. We are further not able to synchronize the biosignals with the occurrence of the hazard on a millisecond level due to technical limitations. The exact moment when a hazardous object becomes visible is ambiguous for some situations.
Thus, in their current implementation, the methods cannot be used as input and predictors for triggering driver assistance systems. Further studies and a different experimental setup would be required to be able to study the exact moment of hazard perception in-depth.

Conclusion

In conclusion, this study supports the hypothesis that a considerable subgroup of subjects with binocular visual field loss attributed to glaucoma shows a safe driving behavior in a virtual reality environment, because they adapt their viewing behavior by increasing scanning. By means of a driving simulator and sophisticated eye and head tracking, individual performance differences in terms of driving safety were related to visual exploratory behavior. This type of compensation improves traffic safety and may have practical implications in planning individualized driving fitness tests and driver rehabilitation programs.
We have seen how gaze behavior can be roughly described by some characteristic key values (such as the average fixation duration, the fixations per minute or saccadic orientation). In the following this thesis will focus on more involved methods for the comparison and visualization of scanning patterns.

4 Scanpath visualization and visual analytics

Recording good quality eye-tracking data is key to any eye-tracking study. But it is only the starting point in the process of data analysis. Making sense of the data is often facilitated by a good visualization. This section introduces two novel scanpath visualization techniques that focus not only on the attended areas, but mainly on the transitory patterns between them.

4.1 Saccade trajectories

Eye-tracking data is mostly visualized based on the allocation and local density of fixations. Attention maps, fixation clustering [54], and aggregated values (e.g., dwell time on a region of interest) are frequently used tools for eye-tracking data analysis. In fact, fixation-based analysis is motivated by the way our visual perception works; perception is only possible during fixations and suppressed during saccades, i.e., high velocity movements of the eyeball [36]. Therefore, saccadic patterns have mostly been studied indirectly, e.g., as transitions between ROIs [154]. Obviously, a large proportion of saccades will occur as ROI transitions, e.g., between the faces of people in a painting. However, ROIs might be ambiguous in an art work. For example, when viewing abstract art, gaze is supposed to follow artistic composition principles [155] or in medieval art, by inserting reflective gold leafs in the painting [17]. Yarbus defined composition as *"[...] the means whereby the artist to some extent may compel the viewer to perceive what is portrayed in the picture"* [8]. Especially for abstract paintings, the definition of meaningful ROIs is questionable and an analysis of saccades and gaze transitions would be restrained to these ROIs.
This work focuses on the analysis of saccade trajectories. Thus, instead of asking *what* is looked at, we aim at proving techniques to tackle the *how* and *why* our gaze is driven and guided over a stimulus in a particular way.
First, it is shown how saccade trajectories form patterns that are characteristic for the stimulus material. Then two methods to analyze saccade trajectories are introduced:
(I) a novel visualization method, a saccadic heatmap
(II) a clustering technique to cluster saccades for eye-tracking data of low temporal resolution.
Both methods are compared to ROI transition diagrams and a trajectory clustering approach (attribute-driven edge bundling [156]) in an art-viewing experiment.
Both methods are implemented in the Eyetrace [32] software and have been accepted for publication to the ECCV 2016 workshop VISART. Eyetrace is a visualization and analysis tool for static stimulus experiments, such as the viewing of fine art. It provides a variety of

state-of-the-art algorithms for each processing step: Identification of fixations and saccades (e.g., [157, 20]), clustering of fixation locations, automatic ROI annotation, and scanpath comparison.

4.1.1 Related work

In eye-tracking recordings, data samples recorded during fixations outweigh by far the saccade samples. A first version of Eyetrace already tried to implement a feature for sampling saccades [158, 159].
Dong et al. were among the first to work on simple heatmaps of saccades for the evaluation of enhanced imagery in cartography [160]. However, most studies are based on fixation heatmaps; sometimes even heatmaps containing both fixation and saccade data are employed, especially when no event filter is applied [161, 162]. In [163] a space-time cube visualization is proposed, where saccades make up a large part of the visualization: as all samples are connected by a line, saccades result in the longest line segments.
Probably the best metaphor for a saccade heatmap is a grassland, where the grass is trampled down in paths that are frequently walked over. Trails emerge and enlarge as they are used more frequently. Similarly, the saccade heatmap visualizes frequently traversed gaze trails derived from the saccade point in eye-tracking data. Corresponding to the ROIs emerging from hot spots in the fixation heatmap, we will explore so-called saccade bundles, i.e., clusters of saccades.
Other methods, such as attribute-driven edge bundling techniques, to cluster general trails have successfully been applied to eye-tracking data [156]. A visually appealing and fast implementation can be found in the CUBu software [164] that can be accessed via Eyetrace [32]. In another approach, so-called saccade plots [165], saccades can be visualized in a more abstract way. Similar to a ROI transition diagram, saccades are split into x- and y-components and visualized, e.g., by arcs that connect different stimulus regions.

4.1.2 Eye-tracking during art viewing

The proposed methods were applied to eye-tracking data collected during the viewing of paintings. Two paintings were chosen for this experiment that are at the center of a controversial methodological discussion in art history for several decades. In 1961 Kurt Badt argues that in order to interpret a painting one has to describe the path taken by the eye to go through it. His foremost examples are Jan Vermeer's *Art of Painining* at the Kunsthistorisches Museum, Vienna and Jacopo Tintoretto's *The Last Supper* in S. Giorgio Maggiore, Venice [166]. Badt's argument was often discussed. But it could not yet be confirmed or falsified with empirical evidence.

Experiment 1: The Art of Painting

In the first experiment, nine subjects viewed Johannes Vermeer's *The art of painting* on a screen (iiyama ProLite E2607WS, 1920×1200) for one minute. Eye movements were recorded by means of an EyeTribe (The Eye Tribe Aps, Copenhagen/Denmark) eye tracker

at 30 Hz sampling rate at a distance of 60 cm from subject to screen. Fixations and saccades were determined via a Gaussian mixture model [157, 21].

Experiment 2: The Last Supper (Tintoretto)

This data set was recorded at the University of Vienna and contains eye-tracking data of 40 subjects viewing Tintoretto's *Last Supper* for two minutes each. An IViewX RED 120 tracker was used and the painting shown on a 30" display (2560×1600 px) with a distance of 90 cm to the observer. The 20 art historians and 20 novices were instructed to judge whether they liked the picture to induce a sense for aesthetics.

4.1.3 Saccade heatmaps

Characterizing saccades requires at least two points, the origin and the target of the saccade. In addition, the representation of a saccade may contain its direction, amplitude, velocity and a whole trail of samples to show its ballistic nature. Clustering saccades is in contrast to clustering fixations a challenging task. Instead of comparing 2D fixation locations to each other, we need to assess the similarity of whole saccade trajectories. In this context, saccade direction, amplitude and the position of intra-saccadic measured points might also be relevant. The visualization of saccades without further post-processing might not be very informative. For example, Despite the relatively short viewing time of one minute and the small number of subjects, we extracted overall 959 saccades from the eye-tracking data collected during Experiment 1, Figure 4.1. Each saccade is visualized by an arrow, resulting thus in a visual clutter and overlapping shapes. Given this visualization, it is pretty hard to derive any pattern; this is probably an additional reason why saccades are usually excluded from further analysis in many studies.

Construction of a saccade heatmap

To process saccadic data, we introduce a novel computational method for saccade heatmaps. The aim is to visualize the density of saccades, where frequently traversed areas gradually become *hot* while other areas stay *cold*. To achieve this, we have to (1) define a density function for a saccade, (2) integrate the density functions over all saccades, and (3) apply some post-processing, such as weighting. Each of these processing steps is described in detail in the following paragraph.

Density functions

In the computation of fixation heatmaps, the density around a fixation location is usually modeled by a Gaussian. The mean of the Gaussian distribution is placed at the center of the fixation location and the standard deviation adjusted to represent 2-5° of the visual angle. Thus, it is supposed to represent the area of the fovea, the accuracy of the eye-tracker, or the area of *sharp, high-resolution* vision.
To compute saccade heatmaps, such a Gaussian with equal spread towards each direction obviously does not represent saccades very well. But the approach can be adapted by

(a) Original artwork (b) Raw saccades

Figure 4.1: All saccades contained in a recording of nine subjects viewing *The Art of Painting* for one minute. Each saccade corresponds to one arrow in the visualization

stretching the Gaussian along the saccade to cover its origin and target. More specifically, we apply the following Equations 4.1 and 4.2 to calculate the standard deviations of the Gaussian, where *dist* is the length of the saccade.
In our implementation we used the pixel distance. To guarantee scale invariance, the pixel distance is calculated based on mm distances in the real world.

$$std_{dir}(dist) = \sqrt{dist} \cdot (1 + \ln(dist)) \tag{4.1}$$

$$std_{orto}(dist) = \sqrt{dist} \tag{4.2}$$

Note that in this case we have a covariance matrix that can be split into the contribution in the direction of the saccade std_{dir} and its orthogonal vector std_{orto}. The orthogonal contribution is chosen much smaller than the contribution along the saccade's major direction. This way a slim, ellipsoid shape is produced. We used the natural logarithm of the distance as stretching factor with e as base of the Gaussian and the idea that $e^{\ln dist} = dist$. For the Gaussian this is not completely correct (the standard deviation is the denominator), but the effect is as expected. The density function is then rotated and translated to align with the position and direction of the saccade vector.

$$g(x,y) = \frac{1}{2 * std_{dir} * std_{orto} * \pi} * \mathrm{e}^{-\frac{1}{2} * (\frac{x^2}{std_{dir}^2} * \frac{y^2}{std_{orto}^2})} \tag{4.3}$$

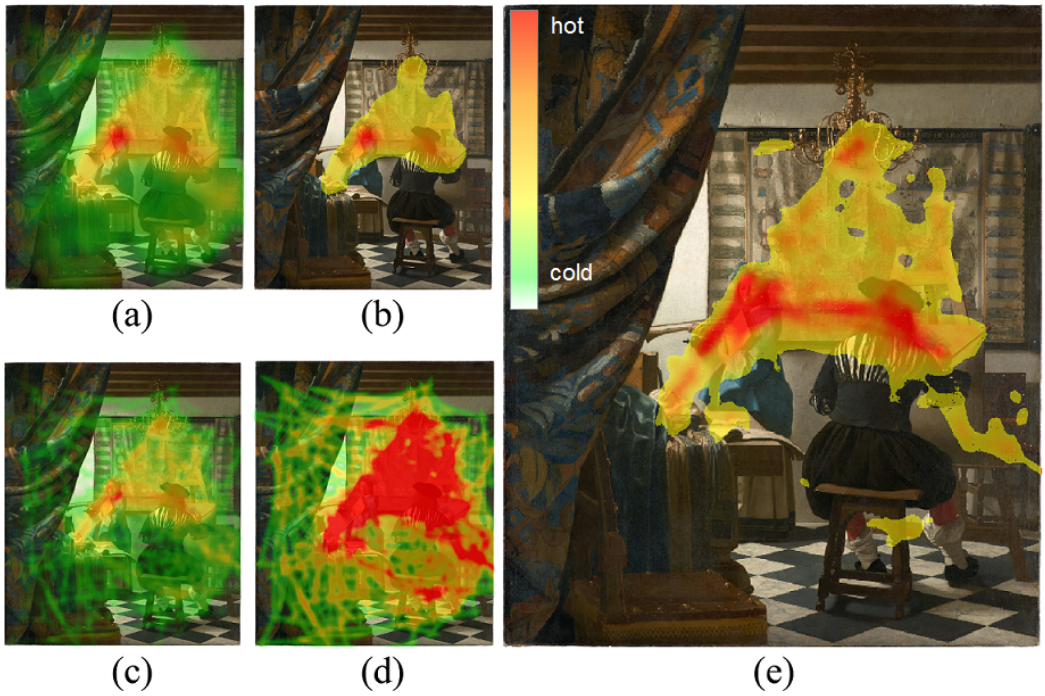


Figure 4.2: (a) raw saccade heatmap with Gaussian density functions stretched to cover the saccade trajectories. (b) thresholded at a minimum density to pronounce the most important paths. (c) raw heatmap with a small standard deviation for the saccade density function. As a result the heatmap is less smooth and more precise. (d) the heatmap was capped at a maximal density. The color resolution available for the remaining areas is therefore enlarged, but the resolution of the most frequently traveled paths is decreased. (e) saccade heatmap with a low standard deviation, capped at a maximum density and with a minimum density threshold applied

Equation 4.3 shows the complete Gaussian function, where x and y are the positions shifted from the saccade center.

Figure 4.3 visualizes the Gaussian distributions stretched along saccades. We can observe that the peak density is reached in the middle between the origin and target of the saccade and that positions along the saccade are not weighted equally. Furthermore, the start and end point are not contained within the high-density area.

Figure 4.2(a) and (c) show the saccadic heatmap for two Gaussians with different standard deviations. In (a) the lines are smoothed and blurred by the high standard deviation. In (c) individual saccades are still visible and crisp. There is no obvious real-world equivalent to the spread of the Gaussian, like with the fovea for the fixation data. Instead, the parameter depends mainly on the eye-tracker's accuracy and the homogeneity of eye movements that the stimulus material invokes. If we want to study fine-grained details, a small standard deviation needs to be chosen. When general saccadic patterns are of interest, a larger standard deviation contributes to a faster convergence of the heatmap. Saccade trajectories are more likely to overlap when the spread is larger.

To achieve an equal weight of the whole saccade trajectory, the central cross-section of the

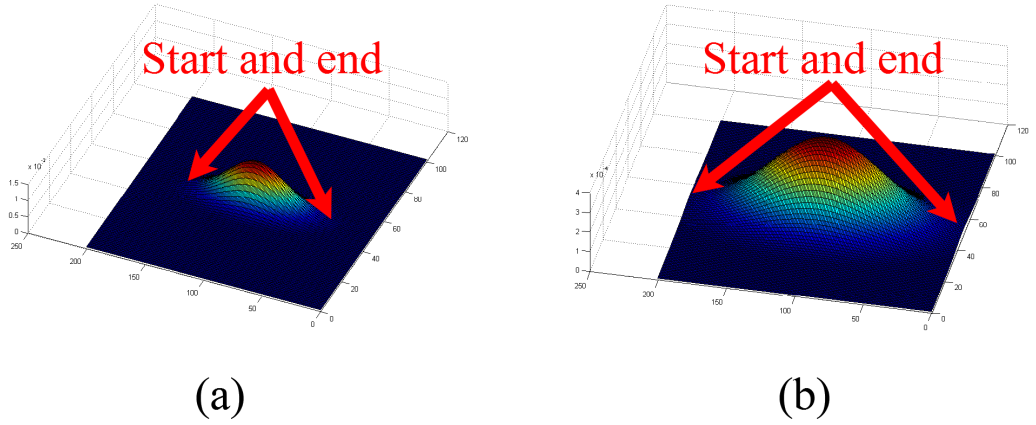


Figure 4.3: Normal distribution density functions for two saccades. The height and color of the surface represent the density assigned to the respective position: (a) shows a short, (b) a long saccade

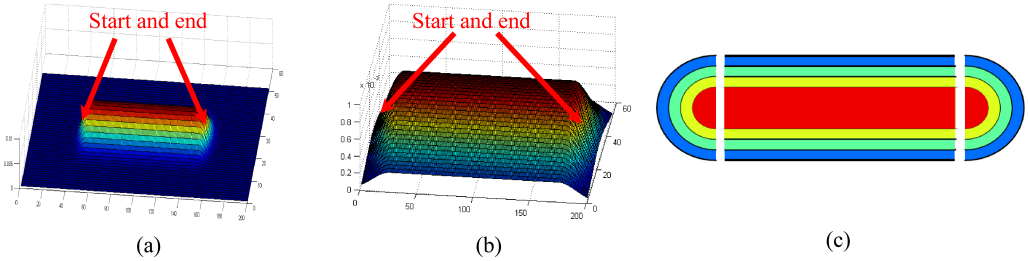


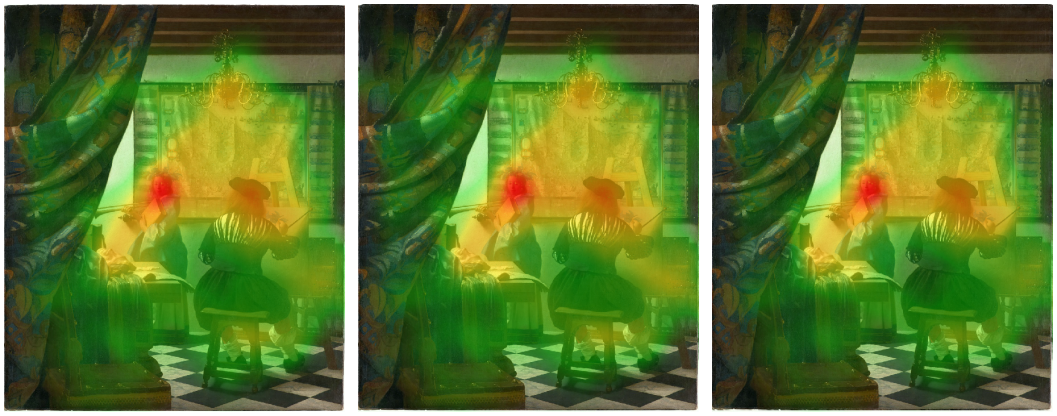
Figure 4.4: Modified normal distribution density function that assigns the same gradient to each position along the saccade trajectory. (a) low standard deviation, leading to a crisp saccade representation. (b) larger standard deviation, resulting in a smooth but blurry heatmap. The width of the Gaussian is based on the standard deviation chosen by the user. (c) contour plot of the three components: two caps for the saccade start and end point, and a length-variable adapter piece between them

2D Gaussian density function is *copied* along the trajectory (see Figure 4.4). The start and end of the saccade are then modeled as a dissection of the Gaussian with one half applied to the start, the other to the end of the saccade (Figure 4.4(c)).

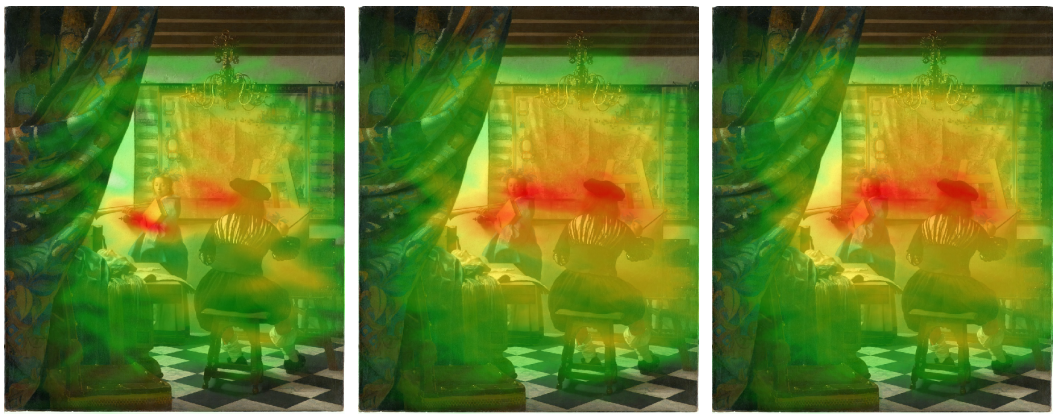
Integrating density functions

Integration over all saccades in a recording is simple, as the density functions can be added. Using the modified density function without further modification would result in an increased number of saccade overlaps within the smaller ROIs (just as it happens in the example shown in Figure 4.2(a) with the face of the woman): transitions to multiple other locations originate here, overlap each other, and cause the saccade heatmap to highlight the overlap region instead of the saccadic trajectory. A simple approach to avoid this effect is to apply a small negative offset to the endpoints of the density functions; endpoints will not accumulate anymore as the small shift assigns a lower weight to the periphery of the

trajectory. Note that this effect is already built-in for the non-modified Gaussian density approach described above.
Endpoints will not accumulate anymore as the small shift assigns a lower weight to the periphery of the trajectory. Note that this effect is already built-in for the non-modified Gaussian density approach, since there exists only one maximum at the center of the saccade as described above.



(a) Repeated (b) Length (c) Duration



(d) Stretched (e) Length (f) Duration

Figure 4.5: (a,b,c) in the top row, the modified density function is applied. (a) unweighted heatmap. (b) heatmap weighted by the length of the saccade; highlighting longer trajectories. (c) heatmap weighted by the duration of the enclosing fixations. (d,e,f) the bottom row employs the non-modified, stretched Gaussian density

Figure 4.5 shows the practical consequences of using either the modified density function (top row) or the non-modified density function (bottom row). When the modified density distribution is applied, frequent traversals between the painter, the woman and the mask are highlighted (first column). But there are also unwanted effects of saccadic overlay within the ROI regions of the faces (those are in fact the overall *hottest* areas). The stretched normal

distribution compensates for this effect: hottest regions in this map are located in-between the face regions. However, the trail of gaze is not clearly visible. Especially the triangle between the two faces and the chandelier is not visible anymore.
Figure 4.5 shows the practical consequences of using either the modified density function (top row) or the non-modified density function (bottom row). When the modified density distribution is applied, frequent traversals between the painter, the woman and the mask are highlighted (first column). But there are also unwanted effects of saccadic overlay within the ROI regions of the faces (those are in fact the overall *hottest* areas). The stretched normal distribution compensates for this effect: hottest regions in this map are located in-between the face regions. However, the trail of gaze is not clearly visible. Especially the triangle between the two faces and the chandelier is not visible anymore.
A relevant drawback of the current implementation is that saccades sum up with each other independently of their direction. Theoretically, it would be possible to calculate separate heatmaps for saccades towards different directions. These heatmaps could then be merged by adding up only those heatmaps that stem from saccades with a similar direction. Heatmaps from different directions can be merged non-additively by taking the maximum of both maps.

Weighting and post-processing

Just as the contribution of a fixation to a heatmap can be weighted by the fixation duration, the contribution of a saccade towards the saccade heatmap can be scaled. Figure 4.5(b) and (e) are weighted using the length of the saccade. Longer saccades contribute more towards the final heatmap, emphasizing the long-distance gaze transitions. For (e) and (f) saccades were weighted using the duration of both adjacent fixations. In these weighted heatmaps we can observe that the scaled normal distributions highlight the relevant gaze trails with only a minor overlap effect in the ROI regions when compared to the modified distribution.
Heatmaps of both, fixations and saccades, often suffer from the effect of one location that is so frequently looked at (or traversed), that all other areas are covered by the large effect. For heatmaps this means that most of the color space is required to represent one spot and the remainder of the image has to be visualized with only a limited diversity of available colors. To cope with this effect, a parameter to cap the heatmap at some user defined maximum density is implemented. On the cost of resolution at the high density areas, low density effects can be studied in more detail.
Figure 4.2 shows the saccadic heatmap with capped maximum and a minimum density threshold that cuts off non-relevant areas.

4.1.4 Saccade bundles

This section introduces a new method for hierarchical clustering of saccades. Parts of the work were carried out by Guilherme Schievelbein as his Bachelor thesis [167]. Aim of the method is to summarize the most frequent gaze trails (similar to the warm regions in the saccade heatmap).

As for fixation clustering, this would enable us to combine data from multiple recordings and subjects in order to reach a convergent gaze trail. The extraction of the most repetitive elements can be described as a denoising process that deletes individual variation from the data and highlights only the most common sequences.
When compared to visualization methods, clustering has several advantages: The method is working with the actual data, not any indirect (simplified) representation such as a density distribution. Each saccade can uniquely be assigned to one saccade bundle. These bundles can be quantified and compared to each other. Filtering and bundle selection can be applied (for example a visualization can select and display only the three most important gaze trails).

Clustering algorithm

A hierarchical clustering [168] of a set of objects can be displayed as a binary tree (Figure 4.6). Each node represents one object of the set and the distance between two nodes represents the dissimilarity between the two objects.

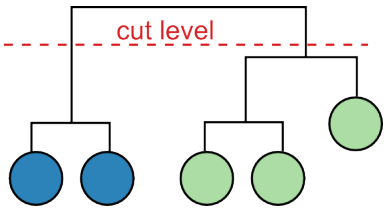
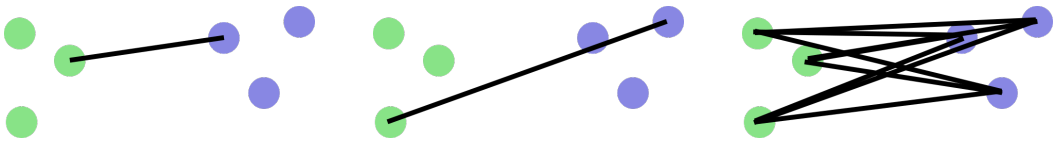


Figure 4.6: Binary tree with an indicated cut level that separates the tree into two clusters (blue and green)

This tree implies a hierarchy on its nodes, with nodes that are joined further upwards in the tree being more dissimilar to each other.
Cutting the tree at any height will result in groups of nodes ending up in the same subtree. Each such subtree can be considered a separate cluster. Depending on where the tree is cut, very small and highly similar clusters or more dissimilar, general but larger clusters can be extracted.
Constructing such a hierarchy tree consists of two steps. First calculating the dissimilarity between two nodes and second a linkage method for the dissimilarity between groups of nodes.
Popular linkage criteria are maximum (or complete) linkage, minimum (or single) linkage and average linkage. Figure 4.7 visualizes the between-cluster distances that are used for the different linkage criteria. Basically, maximum linkage returns the largest distance between any two elements of the clusters, minimum linkage the smallest distance and average linkage the mean of all distances between all nodes.
The linkage step is repeated until all elements of the binary tree are linked to each other.
For the definition of a distance metric between saccades we will consider both the orientation and the Euclidean distance between start and end points of the saccades. Data obtained from many or long recordings can easily contain some thousands saccades. As the dissimilarity



(a) minimum (b) maximum (c) average

Figure 4.7: Distances between clusters that contribute to the between-cluster distance for different linkage criteria

calculation needs to be performed pairwise (resulting in a runtime of $O(n^2)$), computational efficiency is an important issue.
That said, the clustering of saccades is done in two subsequent steps, as depicted in Figure 4.8: in a first clustering step the orientation between saccades is used as dissimilarity measure, afterwards a second clustering step by the Euclidean distance between the saccades is performed within the previously found clusters.
Runtime can be reduced by filtering very short saccades that are unlikely to contribute much to driving gaze over the picture and by scanpath simplification, i.e., merging of temporally sequential saccades into the same direction.

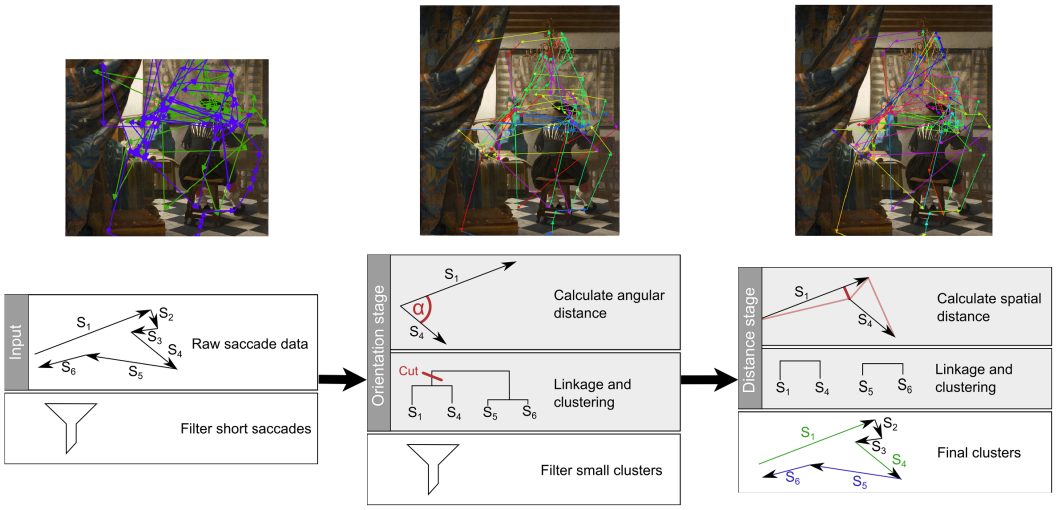


Figure 4.8: Saccade clustering workflow. On the right side the raw data and clusters after the angular clustering as well as the subsequent distance clustering step are shown. Saccades of the same cluster are colored the same

For the definition of a distance metric between saccades we will consider both the orientation and the Euclidean distance between start and end points of the saccades. Data obtained from many or long recordings can easily contain some thousand saccades. As the dissimilarity calculation needs to be performed pairwise (resulting in a runtime of $O(n^2)$), computational efficiency is an issue.
Saccade clustering is computed in a two-step approach as depicted in Figure 4.8: in a first

clustering step the orientation between saccades is used as similarity measure. Afterwards clustering based on the Euclidean distance between the saccades is performed, but only within the previously found clusters.
Runtime can be reduced by filtering short saccades that are unlikely to contribute much to driving gaze over the picture and by scanpath simplification, i.e., merging of temporally sequential saccades into the same direction.
Given a saccade with start point $A = (x_a, y_a)$ and end point $B = (x_b, y_b)$, its angle towards the horizontal axis in $(-\pi;\pi]$ is calculated as:

$$\measuredangle(A,B) = atan2(y_b - y_a, x_b - x_a) \tag{4.4}$$

With the angle towards the horizontal axis precalculated, the angular difference between two saccades S_1 and S_2 is computed efficiently as:

$$d(S_1,S_2) = \begin{cases} |\measuredangle(S_1) - \measuredangle(S_2)|, & \text{if } |angle(S_1) - angle(S_2)| \leq \pi \\ 2\pi - |\measuredangle(S_1) - \measuredangle(S_2)|, & \text{otherwise} \end{cases} \tag{4.5}$$

The above equation can easily be adjusted for direction independence such that the saccades $S_1 = (A,B)$ and $S_2 = (B,A)$ are considered identical.
For the second clustering step the spatial distance between two saccades $S_1 = (A,B)$ and $S_2 = (C,D)$ is calculated as the minimal distance of the start and end points towards the line from start to end of the other saccade:

$$d(S_1,S_2) = \min(d(A,(\overline{CD})), d(B,\overline{CD}), d(C,\overline{AB}), d(D,\overline{AB})) \tag{4.6}$$

The distance between a point A and a line segment $\overline{BC}$ is defined as the Euclidean distance between A and the closest point on the line segment $\overline{BC}$.
The construction of the clustering tree with the currently implemented method requires $O(n^3)$, cutting the hierarchy tree can be done in $O(n)$. The computational bottleneck is therefore the computation of the first clustering tree that includes many saccades.
The hierarchical clustering parameters can be adjusted easily and fast, allowing choosing between average cluster size and within-cluster similarity.
This ability to choose the detail-level (i.e., cluster size ans within-cluster similarity) after the computationally expensive part of the algorithm, allows for an easy and fast adjustment of the relevant parameters.

4.1.5 Application of the proposed techniques to art viewing

The above approaches were applied to eye-tracking data of both free-viewing experiments introduced previously. For both eye-tracking data sets, the orientation clustering step was performed with maximum linkage criterion and direction dependency, whereas the distance-based clustering step with minimum linkage. Cutoff values were determined by successively easing the restrictiveness of the cutoff (i.e., increasing the cutoff threshold), until relatively many saccades were contained in the clusters. This parameter is necessarily subjective, as the homogeneity of saccades depends on the stimulus material. While increasing the

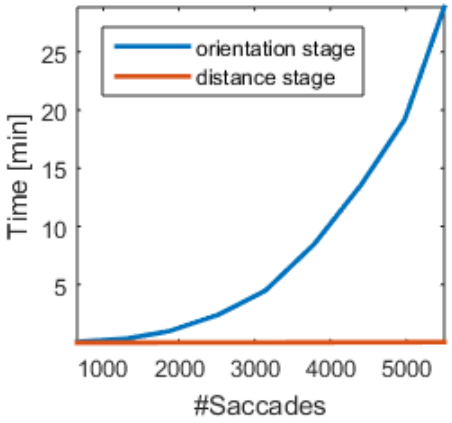


Figure 4.9: Runtime of the saccade clustering on different data sizes. The two processing stages were run on a i5-4200M 2.5 GHz processor

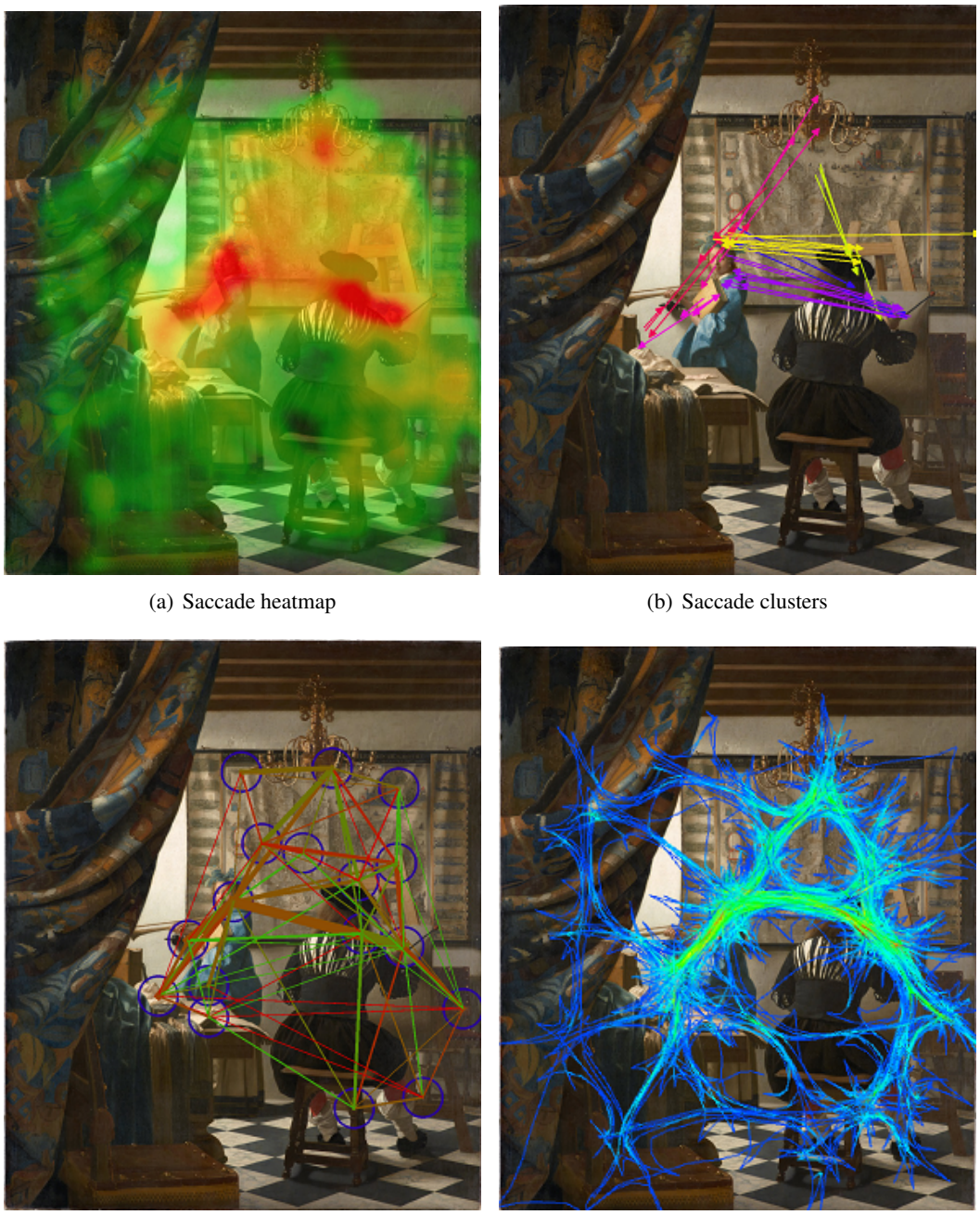
threshold, a transition from very detailed saccade bundles (often with only one saccade contained in each bundle) towards a more general, coarser summary takes place.

Results on Experiment 1

Figure 4.10(b) visualizes the clustering result for Experiment 1. Remarkably, saccade clustering reveals that the eye movements of the observers were driven by the social cue in the painting. The painter and the woman in this painting are displayed in a way that their gaze target can be estimated by the viewer. We can observe that the most frequent gaze trails computed by our clustering approach follow these social cues between the painter, the woman, and the plaster mask. We can further observe that the composition line of the painting that connects the mask, woman, and chandelier has a strong effect on gaze behavior.

In addition, Figure 4.10 displays the spectrum of visualizations for saccades that is currently available in the Eyetrace software, where (a) visualizes the result of the saccade heatmap computation and (b) the result of our saccade clustering technique. Besides the different look, the main distinction between these methods and ROI transitions lies in the amount of simplification that is performed. More specifically, the ROI transition graph as visualized in (c) builds upon the identification of ROIs (e.g., via mean-shift clustering of fixations). Its major advantage is that scanpath transitions instead of direct saccades between ROIs can be considered. Scanpath transitions may contain saccades to non-ROI areas in-between two ROIs and do not require a saccade directly from one ROI to another ROI.

The edge bundling approach shown in (d) consists of various different steps (clustering of fixations, clustering of the trajectories, relaxation, color choice,...). Each step is associated with a set of parameters that require adjustment. Parameters were adjusted to emphasize the same effect that was also found by the other methods and we can clearly observe the primary gaze trajectories along the faces and towards the chandelier.



(a) Saccade heatmap

(b) Saccade clusters

(c) ROI transitions

(d) Cubu Edge bundling

Figure 4.10: Comparison of the different approaches proposed in this work (a,b) and other state-of-the-art visualization techniques (c,d)

The most recent and impressive example of such a clustering is attribute-driven edge bundling [156, 164]. Edge bundling performs the mean-shift algorithm on both, saccadic start and end points as well as samples distributed equally along the saccadic trajectory. Therefore, clustered trajectories get an organic look, as it the exact ballistic eye movement was measured with an extremely high sampling rate and accuracy. When looking at the results it is important to keep in mind that the displayed data represents a considerable simplification and that the suggested level of detail is in fact not contained in the data. The samples along the saccade trajectory are interpolated.
Contrary to the approach suggested here, the whole saccadic trail can be clustered - if a recording at a high enough frame rate is available. The proposed clustering approach uses only the start and end point of each saccade and can therefore also be run on the CPU while edge bundling requires the massive parallelization of a GPU.

Results on Experiment 2

Tintoretto was the first painter who represented the table of The Last Supper from the side, hence foreshortening it in the depth of the space. The main composition lines, as they have been described by Badt on the left (Apostes) and right (cat, servant, sideboard) lead into the depth of the space. Most saccade trajectories in Experiment 2 are along those composition lines with almost no transitions across the table. Also gaze escapes towards the light source in the top left corner mainly via the woman in-between the central image area and the light. The empirical experiment confirms the assumption of a correlation between composition lines (as generally analyzed by art historians) and eye movements of beholders. In one crucial point the experiment falsifies Kurt Badt's analysis: his central assumption is that the viewer starting on the lower left corner will be refrained from following with his eyes the apostles along the table, and will instead follow the high-lighted leg of the left apostle, the dog and cat up to the servant and the right foreground. In the experiment, this connection was extremely rare [169].

(a) saccade heatmap (b) saccade bundles

Figure 4.11: *The Last Supper* by Tintoretto (a) the saccade heatmap (b) saccades clustered together are shown in the same color

Discussion

Two novel computational techniques to process saccades were introduced: (1) the saccade heatmap and (2) a completely data-driven method for saccade clustering.
Both methods were applied to two art viewing experiments alongside ROI transition diagrams and edge bundling. As they work without a definition of regions of interest, our methods are relatively easy to apply. The methods are powerful tools when static stimuli are examined. But for dynamic scenarios, saccades won't sum up in a heatmap nor are they easily clustered. For these scenarios ROI annotation is almost inevitable.
We were able to demonstrate that there are indeed characteristic patterns by which gaze traverses through a painting. These patterns correlate with composition elements of the painting. The proposed approaches facilitate the study of saccadic patterns and might contribute to a better understanding of the influence of image composition on visual scanning.

4.2 Semi-automated annotation of ROIs in dynamic scenarios

Assigning ROI labels to a specific object is more challenging in dynamic scenarios than in static ones. In dynamic scenarios, labeling has to be performed continuously instead of once per stimulus, since the object of interest may move. In interactive scenarios that involve ego-motion of the subject, objects will change their position relative to the subject all the time.
The intuitive approach to track all objects in a video [170] or to use video segmentation [171] suffers from the restrictions of current computer vision algorithms. The requirement to track all relevant objects across all frames is computationally expensive. Tracking losses do occur and judging the identity of an object before and after a tracking loss is not trivial. Therefore, it is not sufficient to track only the currently fixated object. Instead, all ROIs have to be tracked through all frames in order to maintain object identity over the whole recording.
Currently no professional eye-tracking software offers automated, video-based ROI labeling. Computer vision algorithms are not yet able to solve this problem reliably.
In this section, a semi-automated approach is presented. The method was published in [28]. Even a non-perfect labeling that is able to assign the correct label to a relevant fraction of fixations, can save a lot of time. Therefore, an image segmentation and image feature based approach performs an initial labeling, followed by a manual correction step in a fast drag & drop GUI.

Algorithm

First, a preprocessing step is required to filter fixations from the raw eye-tracking data, utilizing the same speed-up effect as SMI's semantic gaze mapping (see Section 2.2).
The next step in semi-automated ROI labeling is the detection of the currently fixated ROI, i.e., finding the boundaries of the currently looked-at object. Then, features are extracted from within the segmented area. These features are then compared against a wordbook of already seen ROIs in order to assign a label based on feature similarity.
Two processes are triggered on the fixation location in the scene camera image:

- image segmentation on an area around the fixation.
- a region growing approach to find scale invariant features (SIFT) starting at the segmented area and moving outwards.

Image segmentation is not a trivial problem, but much easier than video segmentation: there is no need to track the identity of an object between frames. This decreases complexity and run-time dramatically. Different segmentation methods can be used and the choice of an optimal method depends on the image material and amount. For a large quantity of images one might choose a segmentation method that requires a start seed. Starting at the fixation location, the area is enlarged outwards until a border is hit. Since not the whole image has to be processed, run-time is very fast. Segmenting the whole image requires more time, but it offers the possibility to compensate for measurement errors of the eye tracker. Based on the hypothesis that individual objects in the image are rather small, one can correct fixations that hit a large (background) surface, if a smaller area can be found nearby (see Figure 4.12). In this experiment the mean shift algorithm was employed [172].
The *scale invariant feature transformation* searches for distinct feature points in an image, which are invariant to translation, rotation and scaling. This invariance allows finding the same feature points of an object at different positions and orientations in an image, regardless of the perspective transformation. For finding distinctive features, the image is convoluted with Gaussian filters at different scales. Large Gaussian result in a smooth image that contains mainly the background. Smaller filters contain more details. If the different detail levels are subtracted from each other, low frequency changes that are caused by illumination or color gradients are discarded and high contrast, salient points stand out. Local extrema are calculated for different detail level contrasts. If a feature point stands out from its neighbors at several scales, it is considered a good feature candidate. Filtering steps follow to discard low contrast and edge points. Depending on how much the features stand out from their neighbors, they can be ranked and assigned a score.
A set of SIFT features can therefore be used to describe an object and to find it in an image. In our ROI annotation method, SIFT features are collected from the segmented image area around the fixation location until a number m of features is found. m is a parameter that increases the number of features used for the description of an object when chosen to be large, but also decreases feature specificity (features with a smaller ranking score are used). We chose m dependent on the resolution of the eye tracker image as follows:

$$m = \lfloor \frac{\sqrt{width} + \sqrt{height}}{2} \rfloor$$

For the Dikablis device[1] this results in 25 feature points for a 576×768 px scene video.
If not enough good feature points satisfy the quality score within the segmented image region, the region is enlarged, and at the same time the minimal quality requirement for the features eased. There are situations, where the image segmentation step returned a bad segmentation or where relevant feature points are mainly located at the boundary of the object. In these cases, the region growing can compensate for an unsatisfactory image

[1] Dikablis essential monocular eye-tracker (Ergoneers GmbH, Manching) with 25 fps.

segmentation step. The simultaneous quality decrease helps to keep the features within a certain range around the segmented area, so that the region growing is limited.
SIFT features are used as descriptors of the currently looked-at object. A comparison to all previously looked at ROIs is performed by a FLANN-Matcher (Fast Approximate Nearest Neighbor Search [173]). If the ratio test [174] is successful (meaning that the detected features and features of the already looked at ROI are similar to each other), we have found the matching ROI of the current fixation and can assign a label accordingly. The newly collected features can then be used to update the descriptor of the ROI: with each fixation, new SIFT features are added to the ROI descriptor. Due to this, the singularity of the descriptor grows and the ratio test dynamically adjusts to that growing reliability.
If no or multiple matches are found, the fixation is considered a new ROI. It is assigned a new, unique label and its features are stored alongside. The reason to establish a new ROI for a new fixation with several matching candidates is that the found SIFT features may not be unique and, consequently, adding them to the (at the moment) best fitting ROI might degenerate the ROI's reliability.
After each fixation has initially been assigned to a ROI, similar regions are merged: an ROI that was isolated during the initial process because of none or several unreliable matches has a good chance to find a match during this merging process, as most regions are likely to have grown and their descriptors became better. Computationally, the merging process is a simple repetition of the initial assignment but applied to the ROIs instead of individual fixations.
After this merging, the automated part of the algorithm has finished and user interaction is required. ROIs are shown in a graphical user interface (GUI) and the user can correct false assignments via drag & drop, merge or split ROIs (Figure 4.12).

Evaluation on a tea cooking experiment

In order to evaluate the manual annotation demand with the proposed method, we established a new scanpath data set of 10 subjects performing a tea-cooking task (similar to the one introduced by Land et al. [56]). Subjects (age 20-56) were recruited from students and staff of Aalen University (Germany).
Subjects were instructed to make tea for themselves and repeated with the more specific instructions to prepare herbal tea with honey in a blue cup for the experimenter. The standardized setting of the experiment (Figure 4.13) contains a sink, 5 different cups, a spoon, 5 kinds of tea, two kinds of sugar, honey, artificial sweetener, a bowl to dispose of waste, and a water boiler. All objects are directly visible to the subjects so that no searching in shelves is required.
A monocular Dikablis eye tracker by Ergoneers GmbH (Manching/Germany) was used to track the movements of the left eye at 25 fps. We performed the 4-point calibration provided by the manufacturer using a printed poster with colored and numbered circles. The calibration grid was placed on the table so that the distance resembles the working distance for cooking tea.
Mean recording duration was $56(\pm 13)$ s, excluding the calibration phase. For each subject, calibration accuracy was measured after the experiment by repeating the calibration routine.

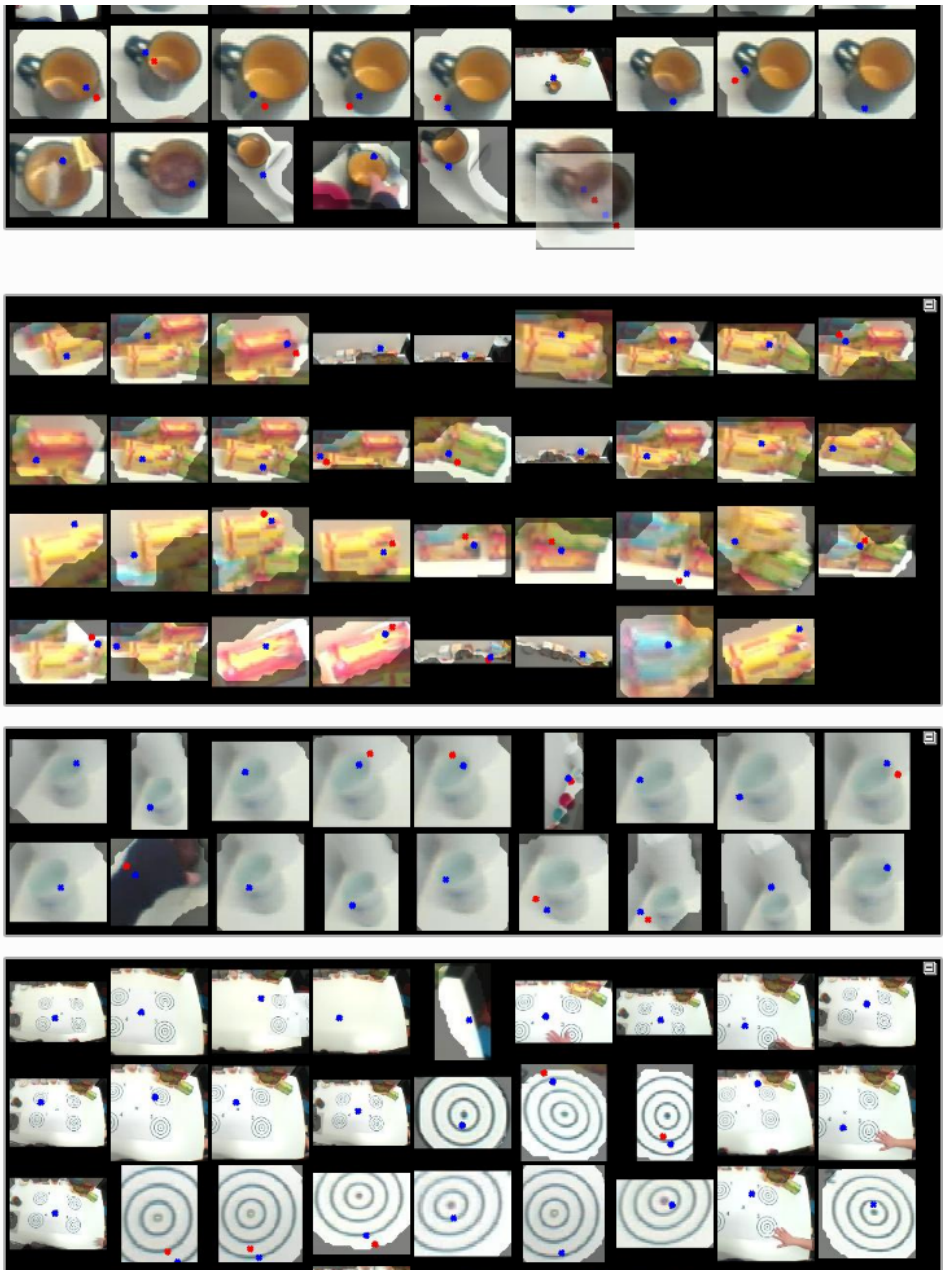


Figure 4.12: ROI annotation tool with four different ROIs and thumbnails of the fixation locations assigned to them. Blue dots mark the measured fixation location, red dots are fixation locations that were relocated to optimize the segmentation, i.e., to produce small ROIs that are likely to correspond to objects instead of the background, such as the table surface [28]

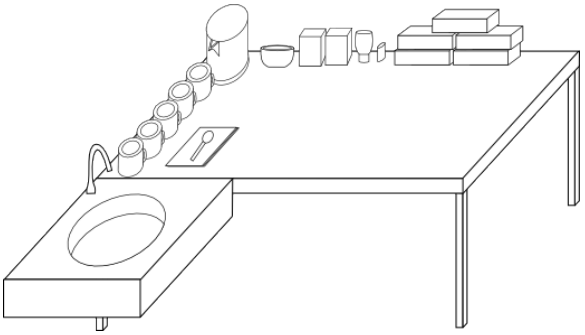


Figure 4.13: Sketch of the experimental setup as used for the tea cooking experiment [28]

Table 4.1: Eye-tracking accuracy achieved after the experiment for the first and second set of instructions (1/2) and each subject (given by the ID). The third and fourth column show the number of manual operations required to refine the automatically created ROIs and the number of unique ROIs that were created by the algorithm, respectively. The number of fixations in each data set corresponds to the number of required labeling operations, if each fixation was annotated manually

ID	Accuracy (1/2) [°]	Manual (1/2)	Auto (1/2)	number of fixations (1/2)
1	4.5 / 3.5	162 / 96	173 / 103	574 / 260
2	3.5 / 5.0	137 / 119	127 / 134	338 / 317
3	3.0 / 2.0	243 / 108	124 / 100	430 / 190
4	3.0 / 2.0	133 / 136	115 / 137	287 / 243
5	5.0 / 4.0	167 / 111	102 / 85	283 / 283
6	5.0 / 4.5	102 / 109	121 / 121	343 / 274
7	7.0 / 7.0	83 / -	77 / -	200 / -
8	6.0 / 5.5	109 / 101	113 / 95	319/ 248
9	>7 / 4.0	53 / 63	88 / 79	357 / 226
10	7.0 / 2.0	89 / 97	131 / 125	284 / 203

Table 4.1 shows the number of ROIs created by the algorithm. On average, the automation step reduces the number of objects for manual labeling to 40% of the fixations, resulting in an average speed-up factor (in terms of required labeling operations) of 2.8×.

4.2.1 Application in scanpath comparison

In the following we will take a look into how insights about the performed eye movements can be gathered from the ROI-labeled data. For this purpose, the commonly used Needleman-Wunsch algorithm (more details on this algorithm are given in Section 5.1) is applied to the ROI labeled scanpaths of the tea cooking experiment. The algorithm pairwisely compares to strings to each other and calculates a string similarity score (parameter choice for this experiment: match score 1, mismatch score -1, gap penalty 3). In concordance with related work [56], the comparison shown in Figure 4.14 reveals that the task of tea cooking is accomplished in a specific order in which the different required steps are traversed - by

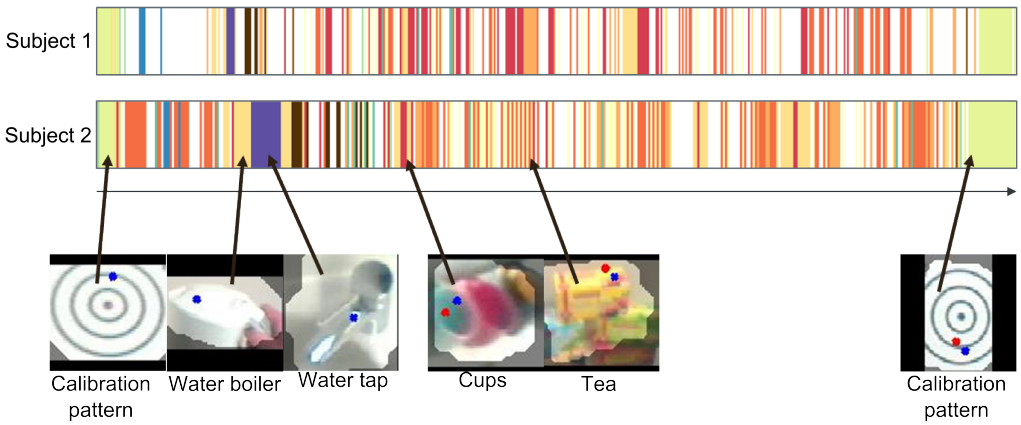


Figure 4.14: Sequence visualization of fixated ROIs by two subjects performing a tea-cooking experiment. Task specific patterns are visible for some ROIs (calibration pattern at the beginning and the end, water tap occurs only once per sequence) [28]

actions required for cooking, preceded by gaze towards the relevant objects.
The high level of similarity between scanpaths means that the global string alignment approach of the Needleman-Wunsch algorithm is applicable to this experiment. For the comparison of scanpaths of unequal length, normalization by the length of the longer scanpath was applied.
Resulting pairwise distances can be compared to each other statistically. Of special interest are the distances between groups of scanpaths when compared to the distances within these groups - just as an Anova tests variance between groups against variances within the groups. In this experiment we tested the effects of:

- cooking tea for the first time (and for oneself) versus cooking tea for the second time (for the experimenter).
- whether subjects repeated the task more similar than the between-subject similarity.

A Wilcoxon rank sum test was applied. We found significant differences ($p \leq 0.01$) for distances between the first and the second run, but none for the inter-individual differences. For visualization purposes, the dissimilarity matrix was squeezed into two dimensions by multidimensional scaling [175]. Figure 4.15 visualizes three samples and the three distances between them as radii of circles. Obviously, there is no way to place the three samples in 2D space such that all the distance requirements are fulfilled - otherwise the blue and red circle would intersect in at least one point. This simple example illustrates the problem of stress minimization during the dimension reduction step. The similarity matrix is high dimensional and we need to break it down to two or three dimensions for visualization purposes. Thereby, we necessarily introduce a small error.
Figure 4.16 shows that some subjects (1,2,3,6,9) perform quite similar scanpaths for both trials (i.e., their points are relatively close together), whilst others vary a lot. Subject 8 for example spoiled water on the table during the first trial and started cleaning the table.

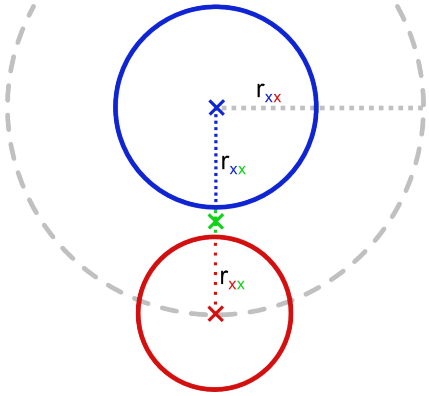
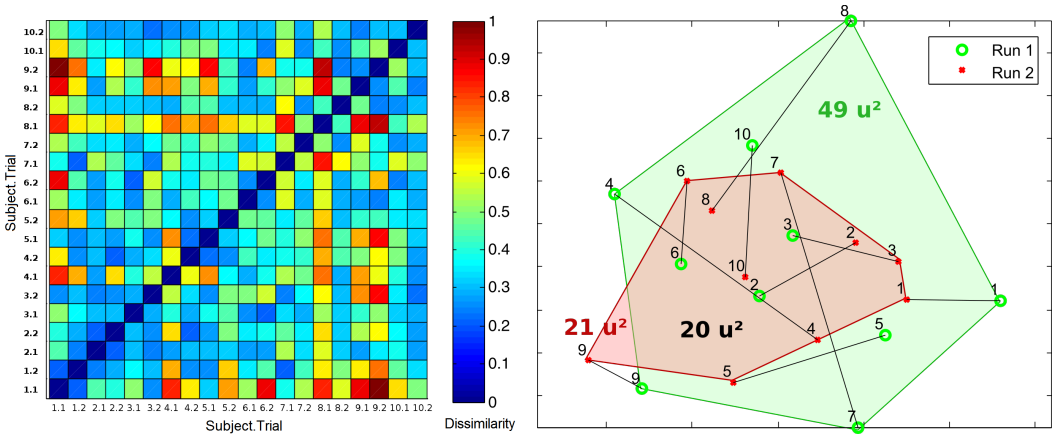


Figure 4.15: Multidimensional scaling of three samples (marked as X) with relative distances towards each other. When inserting the green sample point we must choose a position on both the red and the blue circle. Since this is not possible, stress is induced in the final positioning (green dotted line). The sample is placed at a location where stress is minimized

That resulted in a quite unique scanpath. It can be found quite far off from the centroid of the MDS plot, the *average tea-cooking scanpath* area. We can hypothesize that scanpaths of the first trial tend to be further off from the centroid than scanpaths of the second trial. This finding is also represented in the much larger area of the convex hull of the first trial when compared to the second one. That indicates a more regular, task-oriented behavior for the second trial. This could be due to the subject knowing exactly where the required objects are placed or due to the more standardized instructions to prepare a specific tea in a specifically colored cup.
In this section a string comparison was employed in order to compare scanpaths to each other. There is a variety of other algorithms for the same purpose. Some of them are mere extensions of this approach, others differ fundamentally. The following section provides a review of the state of the art and proposes a new method.



(a) Dissimilarity matrix (b) MDS

Figure 4.16: (a) Dissimilarity matrix and (b) multidimensional scaling to two dimensions of the scanpaths derived from eye-tracking data during the tea cooking task. Each point corresponds to one scanpath. Scanpaths of the same subject are connected by a line, scanpaths from the same trial are summarized by their convex hull. The total areas of both convex hulls as well as the area of their intersection is provided in arbitrary units [u]

5 Scanpath comparison based on subsequence frequencies

Technological progress has continuously facilitated eye movement recording. The analysis of eye movement data, however, is still a laborious, highly manual task. The variety of experimental designs and research hypotheses is infinite. But the central question behind many eye-tracking experiments is the same: is there a difference in gaze behavior between two groups or experimental conditions?
Some more specific examples are: does gaze behavior differ between a patient and a control group? Are there differences in how a painting is perceived by expert art historians and by novices? Does the task given to a subject alter gaze behavior?
There are several approaches to tackle this question of scanpath similarity. They differ in their data representation as well as specific assumptions about the conditions of the experiment. A good metric of scanpath similarity will produce a high similarity score between scanpaths with a common general shape and order, and a low similarity score between heterogeneous, spatially and temporally different scanpaths.
This definition is equivalent to defining a *distance* between scanpaths, where dissimilar scanpaths are characterized by a large distance to each other.
It is important to understand that scanpath similarity can occur on very different levels and scales: laboratory experiments with static stimuli allow for low noise levels and are likely to produce highly similar scanpaths. Quite subtle differences between groups can still be detected. In contrast to that, real-world experiments are generally characterized by a higher baseline heterogeneity due to increased measurement noise and a larger freedom in movement of the subjects. Differences between groups need to be more pronounced in order to be detectable in the midst of a high noise level.
Section 5.1 provides a review of the current state-of-the-art in scanpath comparison. With the knowledge about functioning, abilities and shortcomings of these algorithm, a novel method based on subsequence frequencies is proposed in Section 5.2. The method is then evaluated and compared to its competitors in Section 5.3.

5.1 State of the art in scanpath comparison

One can distinguish between similarity metrics that compare scanpaths as a whole, by their overall shape and sequence, and those that reduce the scanpath to just one specific parameter (such as the average fixation duration or saccade length) prior to the comparison. A comprehensive summary of the latter can be found in [36]. The GazeAlyze [59] software and several commercial products calculate those measures.

Only sequence-sensitive measures are in the scope of the work at hand. They might reduce the scanpath representation prior to the comparison, but not to one single value.
How exactly scanpath similarity and heterogeneity is defined depends mainly on the model of scanpath representation employed by the algorithm. A recent review can be found in [176]. The following review will introduce and discuss different data representations and comparison algorithms that use these representations.

5.1.1 String alignment

The string representation is a letter encoding of fixation locations. The spatial sequence of fixations is transcribed to a sequence of letters. Each fixation is encoded by one or multiple letters. They are assigned based on the ROI that contains the fixation location. An example of such an encoding is shown in Figure 5.1: the presented scanpath is encoded by the first letter of the ROI label, e.g., in the order of fixations as `PWMPW` for the left scanpath, and `PWMWM` for the right scanpath. The string representation conserves spatial as well as sequence information.

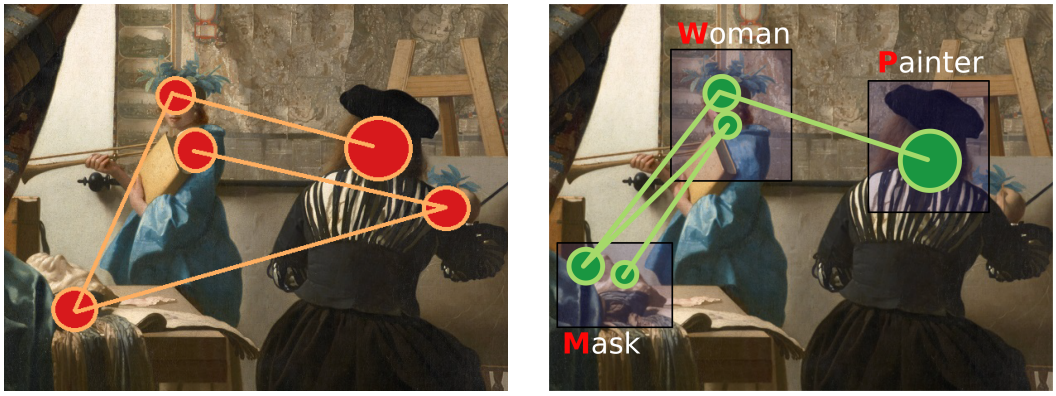


Figure 5.1: Two exemplary, artificial scanpaths in red (left) and green (right). Fixations are shown as circles, scaled by their dwell time. Saccades are shown as connecting lines. ROIs are annotated by bounding boxes. Painting: *The Art of Painting* by Johannes Vermeer

Through this encoding, the problem of defining a scanpath similarity has been mapped to the definition of a string similarity - a well known problem in bioinformatics and spell-checking software. Various measures do exist to approach such a comparison. They are instruments to calculate the similarity of genome sequences, or to suggest corrections for a misspelled word.
A basic and often employed string similarity measure is the Levensthein distance [177], i.e., the minimal number of insertion, deletion, and substitution operations that are required to convert one string into the other. For the above example, two edit operations are required (a substitution of the second `P` by `W` and the last `W` by `M`).
The more sophisticated Needleman Wunsch algorithm is employed by ScanMatch [178]. It can model relationships between ROIs (e.g., spatially close or semantically similar ROIs

can be treated as such by the algorithm). Furthermore, *gaps* can be added to the alignment. In the case of one scanpath containing a continuous subsequence of fixations that does not match any segment of the second scanpath, it can be treated as one continuous gap in the second scanpath and scored differently from a series of individual insertion operations.
Up to now we discarded the information of fixation duration in the string encoding. It is however possible to represent the fixation duration by repetition of a letter. In SubsMatch, Cristino et al. suggest an interval of 50 ms per letter to encode the temporal information [178]. A fixation of the painter in our example for a total duration of 150 ms would result in the representation *PPP*.
For the letter encoding Cristino et al. recommend either a ROI-based approach or a regular grid (in case no ROIs are available). As discussed in Section 2.2, we know that there is no objectively *correct* way of labeling ROIs. In the context of scanpath similarity, too large ROIs will result in scanpaths appearing more similar to each other than they in fact are (having just one ROI would make any scanpaths appear almost identical, up to their difference in length). Very fine-grained ROIs will overestimate heterogeneity. An implementation of the Levensthein distance and the Needleman-Wunsch algorithm for the analysis of eye-tracking data can be found in the eyePatterns software [179], a modified edit distance is implemented in ProtoMatch [180].
The most important caveat of the proposed string representation is the comparison of sequences of different lengths: one has to be aware of the necessity for normalization. The longer the sequences, the more edit operations are required to align them. Therefore, the number of edit operations per letter (that is the probability for an edit operation per letter) seems a good normalized measure as the first glance. But, non-intuitively, it is only a good approximation for sequences of roughly the same length.
For different sequence lengths, normalization is usually performed by division of the alignment score by the length of the longer sequence. But this procedure will return higher alignment scores for unequal sequence lengths than for equal sequence lengths: when comparing a very short sequence to a significantly longer one, there is a high chance of achieving a very good alignment score just by placing the short sequence at the best-matching segment of the longer one. The likelihood of a good match solely by chance increases with the difference in sequence lengths. Against this effect plays the fact that each unmatched position due to the different sequence lengths has to be counted as a gap or insertion/deletion. Depending on the overall sequence similarity (if scores are more likely to be positive than negative), similar length comparisons are preferred because there are more opportunities to get a positive score. The interplay of these two effects is hard to estimate by statistics.
Consequently, the score normalization is implemented incorrectly in all string-alignment algorithms used for scanpath comparison. They exhibit insufficient normalization when comparing sequences of different lengths.
The iComp method [181, 182] circumvents the subjective ROI labeling step: fixations are clustered using the mean-shift algorithm (see Figure 5.2 for an explanation). This leads to a data-driven string conversion, where a unique letter is assigned to each fixation cluster without any human interaction.
As iComp applies the clustering separately to each scanpath, the produced clusters differ.

For the string alignment, a coherent label is essential. Therefore, it is necessary to match corresponding clusters between scanpaths. Those are identified by intersection of the clusters. A similar procedure was employed by Privitera and Stark to determine the overlap between automatically generated and human annotated ROIs [54].

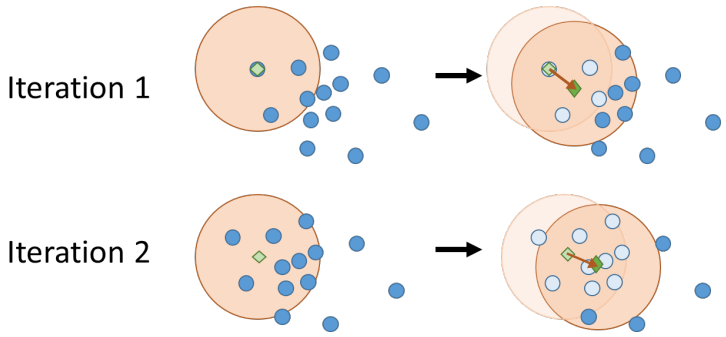


Figure 5.2: Mean-shift clustering algorithm. The algorithm assumes that samples (blue) stem from a Gaussian distribution. In each iteration a window around each sample (red circle around green mid point) is considered. The mean of all samples within the window (light blue) is calculated and the window shifted towards that midpoint (green diamond). With each step the center of the window converges towards the area of highest density. Depending on the size of the window a local or the global area of highest density can be found

Alternatively, Voronoi cells can be constructed around fixation locations [183], leading to small cell sizes in densely fixated regions and larger cells in homogeneously viewed regions. Zangemeister et al. [184] encode not fixation location but saccade directions by letters. Those indicate the compass directions (N, NE, E,...). The method was applied to identify differences in visual scanning behavior of hemianopic patients [184]. The authors utilize different weights between ROIs and a stronger weighting of the first fixation.
Some experimental factors may have a huge influence on this measure, others only a small one. For evaluating the goodness of such a measure, it is possible to test for significant differences in the similarity within and between scanpath groups. However, this measure does not imply anything about actual group separability. Zangemeister et al. applied a post-hoc statistical test (Kolmogorov-Smirnov test) to the scanpath distances to reveal significant scanning differences between the patient and the control group.
Feusner and Lukoff [185] propose a permutation test to determine the significance of differences between scanpath distance distributions. While the permutation test is easy to apply (since it is parameter free and does not require any prior conditions), its sensitivity is limited.

5.1.2 Fixation map comparison

iMap [51] is based on the comparison of 3D fixation maps (the three dimensions are x and y position of gaze, and $fixation\ density$). Random field theory is applied to each pixel of the fixation maps to statistically test for group differences. During this process massive multiple testing (one comparison per pixel) is performed. A correction for multiple testing

is essential to derive any significant effects. The random field theory applied therefore is lent from the analysis of MRI images: fixation maps are usually smoothed by a Gaussian filter so that neighboring pixels are not independent of each other but likely very similar. This has implications for the statistical analysis. Later versions of iMap [186] use different statistics, such as pixel-wise linear mixed effect models and a bootstrapping approach. With a better compensation for the multiple testing, statistical power increases and more subtle scanpath alterations can be detected.
The main advantage of fixation maps over string-based methods is that no semantic a-priori expectations of the experimenter are introduced to the data analysis, since no ROI annotation is required. The huge simplification in fixation location required by the string conversion is not necessary for the fixation map. Attention constraints can be integrated into the fixation map, e.g., by weighting a fixation's contribution to the density function by the fixation duration or by modifying the spread of the contribution to the extent of the fovea region or the eye tracker accuracy.
However, a fixation map is unable to represent the temporal dimension and order of a scanpath. Different fixations to the same location will simply sum up to the same effect as one long fixation. There are fixation map approaches that try to incorporate a time dimension to a limited degree (e.g., by considering several fixation maps from different time slices). But only few provide a robust statistical test [17].
The straight-forward extension of fixation maps by a fourth dimension that represents time is prohibitively time-consuming and makes statistical testing - and thereby the objective comparison of fixation maps - even more difficult. The number of subjects required for the map to converge is much higher than for normal 3D attention maps. No working implementation of such a 4D fixation map comparison is known to the author, although several visualizations (without the comparison step) do exist [163, 187, 88].

5.1.3 Geometric representation

Given the numerical output of eye trackers, the probably most straight-forward approach to scanpath comparison is a geometrical representation. The problem of scanpath comparison can then be formulated as finding an optimal mapping between fixations in both scanpaths. That is done by matching the most similar fixations of both scanpaths to each other. A solution to this is given by the Mannan distance [188]. This approach uses the vector coordinates of fixation location and is easy to implement. Basically, the pairwise Euclidean distance between all fixations is calculated and those fixations with a small distance towards each other are matched. For a three second long image viewing experiment Mannan found that respecting the sequence of fixations is not necessary, as only very few sequences are conserved [189]. However, the devil is in the detail: should we allow the matching of several fixations from one scanpath to just a single fixation in the other scanpath? Eyenalysis [190], for example, performs a double mapping (see Figure 5.3) to circumvent this problem of the Mannan distance. How do we proceed with scanpaths of different lengths, where some fixations do not have a matching partner in the other scanpath? How to include the time dimension? How to define a distance between fixations?
Mathôt et al. leave all of these questions open with their Eyenalysis algorithm [190]. Which

features of the scanpath to use, how to compute a distance between them and how to weight them (e.g., are 2 cm distance between fixation locations a larger distance than 2 s time delay?). Thereby, they formulate an extremely general measure.
As an extension to the Mannan distance, Mathôt et al. normalize by the length of the longer sequence and suggest including a time stamp in the feature vector in order to achieve sequence sensitivity [190].

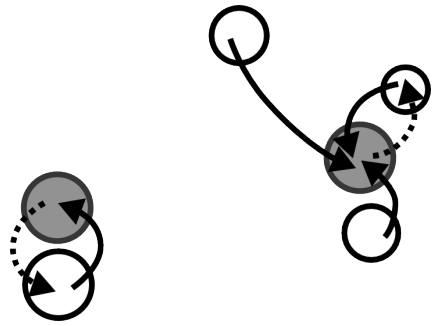


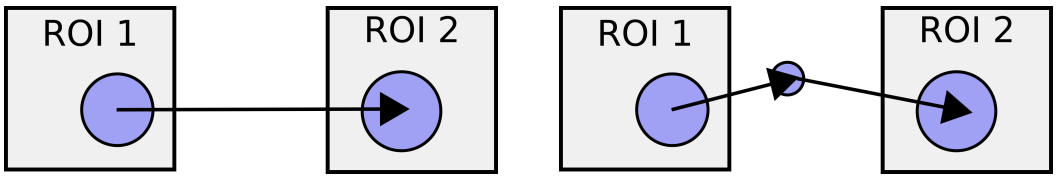
Figure 5.3: Double mapping between two scanpaths S_1 (light) and S_2 (dark) represented by their fixations. In the first step, the spatially closest fixation of S_2 to every fixation in S_1 is determined (solid arrows). The step is repeated the other way round (every fixation in S_2 is assigned to its closest neighbor in S_1, the dotted arrows). Graphic adapted from [190]

The MultiMatch algorithm [191, 192] is a more sophisticated vector-based method. It produces independent scanpath distances in various dimensions, such as location and duration. Therefore, fixations are converted to a vector representation after simplifying the scanpath shape (i.e., deleting *small* saccades and merging subsequent saccades towards the *same* direction). Representative values such as the location and the fixation duration are then chosen as vector dimensions. An optimal mapping of fixations is determined by the Dijkstra algorithm. It finds the shortest path through a fixation vector distance matrix. The method conserves fixation sequence.
FuncSim [193] splits a task into different functional units (subtasks) and performs comparisons only within a functional unit. This way, the normalization problem for different scanpath lengths does not occur. A plausible way to segment a task into subtasks and a way to label the data (automatically or manually) is required. Saccade length, direction, fixation duration and spatial characteristics of the scanpath can be modeled. The split into functional units represents a sequence conservation, but the algorithm can also include fixation duration in its alignment step. Additionally, FuncSim provides a similarity score baseline by comparing the similarity of a scanpath to its own permuted derivative. Statistical significance can be tested that way.

5.1.4 Probabilistic models

Natural variability in scanpaths is high and may mask the subtle but existing differences between scanpaths. Probabilistic representations can handle this by learning the level of variability in scanpaths and comparing against the level of variability found between two

groups of scanpaths.
An early mentioning of the transition matrix approach based on fixation location can be found in [154]. A transition matrix contains the number of transitions from one ROI to any other ROI. One advantage of this approach is that transition frequencies can be normalized for each scanpath separately, thus resulting in a good normalization for different sequence lengths. It is possible to calculate either only direct transitions from one ROI to another ROI, or to include indirect transitions. Indirect in this context means that a saccade starts within one ROI and ends outside of any ROI, then an arbitrary amount of saccades that do not hit any ROI follows, until another ROI is entered (Figure 5.4).



(a) direct transition (b) indirect transition

Figure 5.4: Direct and indirect ROI transitions. The indirect case may contain an arbitrary amount of saccades outside of any ROI region in-between the saccade that leaves and the saccade that enters a ROI

To better cover general scanning patterns, changes in saccadic directions (encoded as compass directions N, S, E, W, NE,...) instead of location features were proposed [194]. The authors mention the necessity of dimension reduction (i.e., deleting seldom occurring directions from the transition matrix), especially when employing statistical testing (e.g., χ^2 test) on the transition matrices.
Another common representative of this group is the Hidden Markov Model (HMM) or the simpler form of a Markov chain (e.g., [195, 196]). States of the model are typically represented by ROIs and state transitions are calculated as transition probability between ROIs [197]. Sophisticated implementations model emission probabilities by 2D Gaussian that represent fixation location probabilities. Markov models are restricted by the Markov assumption, i.e., a specific state depends only on the previous state, not on the whole history of prior states.
Whether this assumption holds for eye movements is questionable. Holmqvist et al. [36] mention that "the usefulness of transition matrices longer than 2 can be disputed" and at that time there was "no study reporting longer sequences than 4" - just to follow it with the note "even though, in theory, there are situations where longer sequences would be interesting". In Section 4.1.4 on saccade trajectory bundling we have already observed clearly longer patterns. In fact, we have observed that at least for the image viewing task the history of looked-at positions in the image has a major influence on gaze behavior, leading to saccade pathways that guide the observer's gaze over a painting. Yet, the finding by Holmqvist is not surprising: the number of possible unique transition patterns increases dramatically with the number of transitions considered. The transition count matrix gets sparse and a comparison is more difficult. In the following sections, we will have an in-detail look at

exactly these longer sequences and the information they hold.
Mast and Burmeister [198] went into this direction by employing a t-pattern detection [199]. They were able to find repetitive scanning patterns. Based on a statistical critical interval test, repetitive sequences are identified. It is even possible to find patterns with a contained constant time-delay (similar to a string alignment with a sequence of gaps). Short t-patterns can be combined to longer, more complex ones.

5.2 SubsMatch - Comparison of subsequence frequencies

The new method SubsMatch for scanpath comparison was published in [34]. SubsMatch is based on the idea that typical behavioral patterns manifest in repetition. Therefore, an elongation of transition matrices [154] should be able to capture those patterns. The same assumption was followed by the t-pattern approach by Mast and Burmeister, who assemble larger patterns from shorter ones based on the frequency of their co-occurrence.
Whether it is the shoulder check during driving or reading a line of text, the behavioral pattern is employed repeatedly throughout the task. Either by the same subject or between subjects and trials.
Transition patterns much longer than two do occur and are in fact usable in data analysis. The difficulties of sparse occurrences and the high uniqueness of long patterns is only a hindrance when we aim at investigating one individual, special pattern. But for the comparison of scanpaths as a whole, integration over all patterns in the scanpath is possible. The averaging is an efficient denoising and smoothing operation. Small, individually hardly detectable differences in the transition patterns add up to a measurable effect.
The algorithmic steps in SubsMatch are visualized in Figure 5.5 and will be described in detail in the following.

5.2.1 String conversion

Region of interest

Generally, semantic ROI labeling is advantageous for any scanpath comparison algorithm and SubsMatch will also profit from a semantically meaningful ROIs [28]. However, as discussed in Section 4.2, this step is not always feasible and usually very time consuming. The string representation used by SubsMatch differs from string alignment techniques in several essential ways: the string conversion is only used as a data simplification step. The individual letters are not required to reflect semantic ROIs in the scene, but are more similar to Zangemeister's approach [184] of encoding saccade direction - a high level abstraction of the actual scene content.

Binning

If no ROI annotations are available, the most trivial alternative is the use of a regular grid. While this is very easy to implement, there are several disadvantages associated with it: grids do not correspond to semantic entities. They might even be placed in a way that they

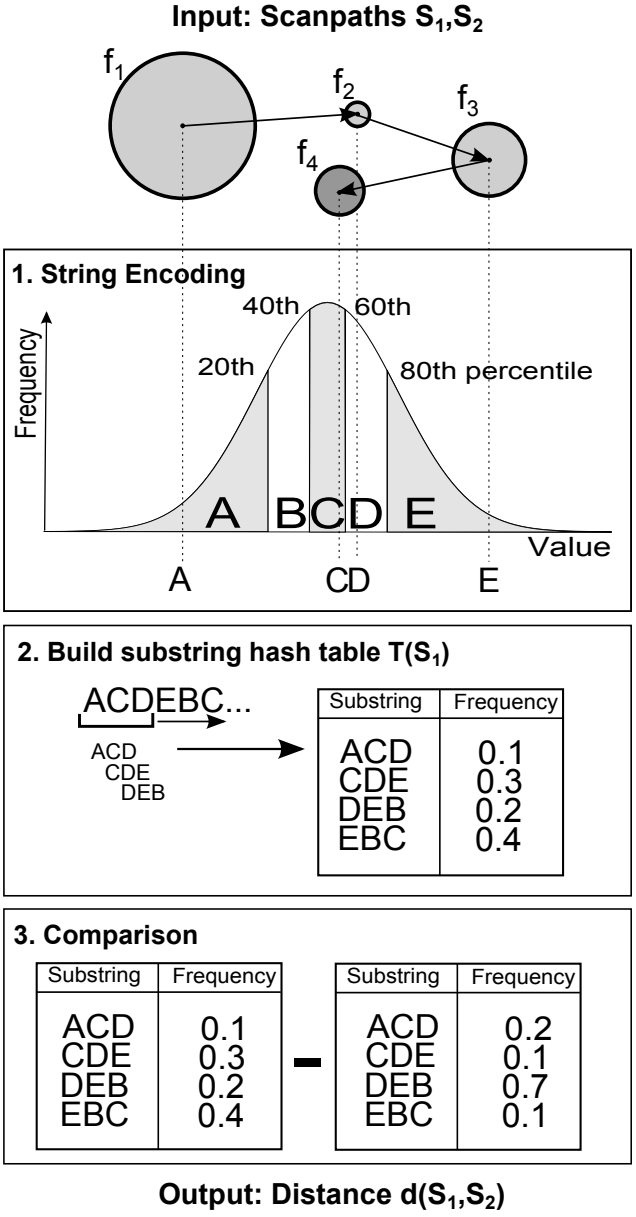


Figure 5.5: SubsMatch processing flow from the input scanpaths (top) to a final distance measure (bottom). First, the scanpath is converted to a string sequence representation. Small substrings are cut from the overall sequence and their occurrence frequencies are counted in a hash table. Two such hash tables can be compared to each other to produce a distance measure

split semantic entities in parts and therefore assign different labels to fixations directed at the same object.
Fixations with a different label might be very far apart or very close to each other. Information on where exactly the grid border separates the fixations is lost.

Percentiles

SubsMatch constructs the scanpath string by binning the data by percentiles. This results in a data-to-letter mapping in which the number of occurrences of each letter is the same. Thereby, an efficient use of the available spatial resolution is guaranteed.

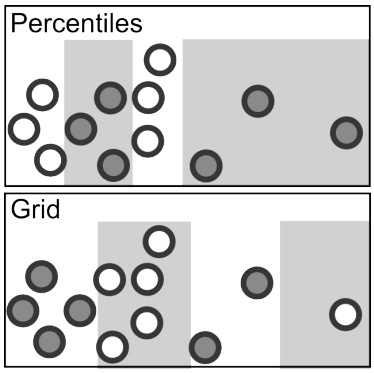


Figure 5.6: Grid binning versus percentiles approach for assigning letters to data

Binning by percentiles (Figure 5.6) is a very useful data-driven approach for the analysis of real-world eye-tracking data: depending on the distance between an observer and the target object, the target can appear at a different scale. Calibration drifts and slight positional changes of the observer will result in an offset between recordings of different individuals and trials. The use of percentiles basically centers the data set to its mean and scales locally by continuous gaze density. We can observe that for specific tasks a gaze towards a certain data percentile is also associated with a certain action. An example of a driving task is shown in Figure 5.7. Different semantic entities also correspond to different areas in the heatmap. Since there is a significant offset and scaling difference between subjects, the simple grid approach would require prior normalization in order to label coherently.
This representation is similar to SAX [200], a technique frequently used in data mining applications. There are also similarities with shapelets [201], a way to encode 2D shapes. Dimension reduction (piece-wise aggregate approximation) could be performed, similar to the scanpath simplification done by other algorithms. However, this step is out of our scope.

5.2.2 n-gram feature embedding

The next step is to split the scanpath into small chunks, so-called n-grams or sub-sequences. n denotes the length, i.e., the number of letters, of the chunk. At the first glance, this step might appear non-intuitive. Contrary to other methods, the scanpath strings are not directly

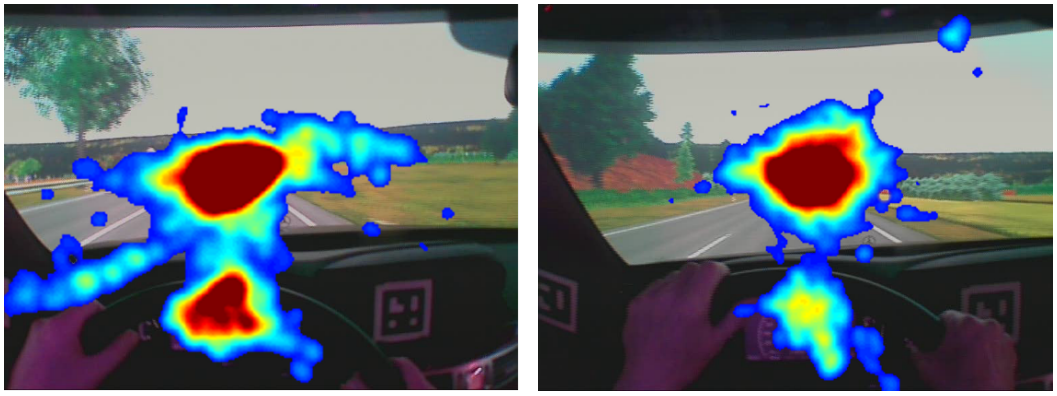


Figure 5.7: Gaze density heatmaps of two different subjects recorded during a driving task. Although they are generally similar, locations do not match exactly.

aligned to each other. Important sequential information of the scanpath is discarded by splitting it into smaller sub-sequences.
In static image viewing tasks this kind of information is very important. For example, saliency maps are only valid for the few very first fixations. If we consider a dynamic scenario however, there is a continuous time-line and it is not possible to clearly define when a certain scene was viewed first. If we think of a driving scenario, the subject can already perceive most of a scene's information at a time shortly before and needs to update his knowledge only in the parts that are subject to relevant change - if there was a large house to the driver's right a second ago, it will likely still be there by now.
On the other hand there are strongly sequential patterns during the driving task, for example a glance to the rear or side mirror. Such a pattern is usually associated with a driving action, like overtaking, taking a turn, or checking speed. We can assume that during dynamic tasks the exact temporal order of the whole scanpath is not as important as it is for static tasks. Instead, focus should be on relatively short, repetitive behavioral patterns. Only within these patterns, the sequence is highly conserved. The nature and frequency of these patterns tells us something about the subject's behavior. For example, whether a driver checks traffic and speed frequently is an indicator of driving performance.
For the case of $n = 1$ this means basically examining glances towards individual ROIs. n=2 corresponds to the transition matrix / Markov chain approach.

5.2.3 Normalization

Normalization is already inherent to the n-gram frequencies: the sum of all n-gram frequencies sums up to 1 for each scanpath, independent of its length. Nevertheless, it is worth some words:
In theory, an alphabet of size a can yield a^n different n-grams. The scanpath length l forms another upper limit of $l - n + 1$. For large alphabet sizes in the string encoding stage, a large n in the feature embedding stage or short scanpaths, the subsequence frequencies

tend to become sparse. n-grams become more and more unique and their frequencies stop adding up. Both, alphabet size and n-gram length, regulate the number of unique n-grams that are possible. The scanpath length has an influence on how many n-grams are actually present. The calculation of n-gram frequencies, implying that most n-grams occur at least more than once (remember the n-grams are representatives of typical, repetitive behavioral patterns), limits the parameter range in practice. The number of actually appearing n-grams is therefore often smaller than the theoretically possible a^n.
Once these boundaries are left, overlaps in the sparse feature vectors become a gambling game that depends on the sparsity of both scanpaths. In these cases the algorithm does not report an objective scanpath distance, but a value that is basically up to chance. As we will see, this fact has to be kept in mind when choosing adequate parameters for the algorithm.

5.2.4 Histogram comparison

Once n-gram frequencies of both scanpaths are known, the calculation of a distance measure is nothing but a histogram comparison. The simplest way is to calculate the absolute difference of all frequencies. This results in a distance $\in [0,2]$. SubsMatch norms this distance $d(S_1,S_2)$ between scanpaths S_1 and S_2 to $[0,1]$.
More elegant histogram comparison measures such as the Earth Mover's distance [202] or the Wasserstein distance [203] cannot be applied, since they are based on the assumption that neighboring bins in the histogram share some kind of similarity. This does not hold for our n-grams and it is not possible to sort them in a way to guarantee this.
A simple example of why n-grams cannot be ordered linearly by similarity is given by the following n-gram set:
AA, AB, BA, BB; The best order in terms of the edit distance is:
BA, AA, AB, BB; but the edit distance between BA and BB is only 1, the distance in the ordered set is 3.
Another pitfall of the primitive histogram comparison is that the measure does not discriminate between groups. It can only capture an overall sense of similarity. In Section 6 we will look into how the approach can be improved in order to achieve group separability.

5.3 Evaluation

There is a huge variety in the application areas of eye-tracking technology, ranging from controlled laboratory experiments to real-world scenarios. A good scanpath comparison measure is required to perform well on the whole spectrum. In this section the SubsMatch measure is applied to both, a laboratory task as well as a highly dynamic driving scenario. Additionally, a practical use case is demonstrated by comparing the scanning behavior of expert and novice surgeons.

5.3.1 Conjunction search task

For the evaluation of SubsMatch under laboratory conditions a conjunction search task was chosen. A number of colored geometric shape stimuli is presented on a computer screen at

three different tasks. Subjects were instructed to count

- all objects of a specific color.
- all objects of a specific shape.
- all objects of a specific shape *and* color.

Stimuli that fulfill the criteria (i.e., have the correct color and/or shape) are called targets, all stimuli with the remaining combinations of color and shape are called distractors. The third case, where both color and shape property of the stimulus have to be combined, is called the conjunction case, after which the test is named. Only the conjunction of two stimulus properties together marks a target. The total amount of stimuli, i.e., targets that match the instruction properties plus distractors that do not match, regulates how fast the task can be solved.
This experiment design was inspired by Machner et al. [151], who tested the hypothesis that visual search patterns in patients with homonymous visual field defects are caused by visual-sensory deficits. Machner et al. demonstrated that different viewing strategies can be employed and described them as circular, line-wise, column-wise, 8-shaped or chaotic [151].
Subjects' search duration (and the number of fixations performed) correlates with the task and its difficulty, i.e., fewer fixations are required for color search than for shape search. The authors concluded that objects with a defined color were significantly faster to find than objects of a specific shape. For this experiment the pop-out effect of color is stronger than that of shape.
For this standardized test, all parameters that modulate scanpath shape and complexity are controllable. It provides different tasks and difficulties by variation of distractor and target counts. The experiment was conducted by Colleen Rothe as part of her Master thesis on scanpath comparison algorithms [204].

Methods

Target objects (1, 4 or 8 per stimulus screen) were displayed amongst distractors, resulting in a total stimulus count (targets + distractors) of 40, 60 or 80. The trials were balanced between the different settings and presented in a randomized order.
Subjects were seated in front of a 24" monitor (Fujitsu Display B24T-7 LED) with 1920×1080 px resolution. A chin-rest was used to minimize head motion. Distance to the screen was 60 cm.
Stimulus shapes were triangles, squares and circles colored in the CIE-coordinates red [0.601/0.322], blue [0.218/0.5790] and green [0.417/0.070]. Minimal distance between stimuli was 1°. The screen background was set to black and a black paper was placed around the screen to avoid distraction by the frame or status LEDs. An EyeTribe eye tracker was placed in front of the subject within a distance of 45-75 cm. Distances vary between subjects depending on where optimal tracking could be performed. The EyeTribe Server 0.9.36 software was used for the recording.

21 subjects (10 female, 11 male) recruited from students and staff at the university of Aalen participated in the study (age range 22-43, average 26.5±4). Nine subjects wore glasses during the tracking, four contact lenses.
A 9-point calibration was performed prior to the experiment.

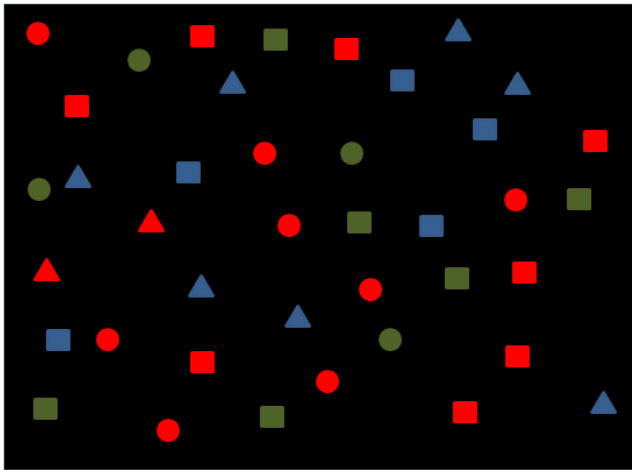


Figure 5.8: Conjunction search task screen with 40 stimuli

Each ten stimulus screens a recalibration was performed to compensate for eventual drifts. Further, a calibration check was done after the last trial. Each subject completed 35 different search tasks. Counting of target objects was performed by clicking the left mouse button, progressing to the next stimulus display was done by clicking the right mouse button.
Seven different scanpath metrics were included in the evaluation, namely MultiMatch, ScanMatch, FuncSim, iComp, HMM, Eyenalysis and SubsMatch. I used implementations provided by their respective authors where possible. For some scripts minor changes were required in order to enable the algorithms to work with pre-identified fixations and to load custom data. The core components of the comparison were not changed, though. For the HMM, a mean-shift clustering was applied to the data. The resulting clusters were used as HMM states and the cluster transitions as transition probabilities between the states. Gaussian emissions were assigned with covariances set to cover the spread of the cluster.
For those algorithms that offer a parameter choice, grid search was employed to find the optimal settings. The results of the grid search showed that all algorithms are quite robust to parameter choices within a reasonable range, so that the influence of the optimization should be limited.
All algorithms perform pairwise scanpath comparisons. For each such comparison of two scanpaths, one similarity value is calculated. A pairwise comparison of n scanpaths towards each others results in a n×n distance matrix. If the experimental design implies several groups of scanpaths, as in the case of the conjunction test the three different tasks, we can subdivide the distance matrix into

- Comparisons between two scanpaths of the same group (e.g., shape vs shape)

Table 5.1: The objective performance parameters for the different settings of the conjunction search task (namely the type of task, the number of targets and the total number of stimuli). Reported values are aggregated over all trials with the respective factor and all combinations of the unrelated factors (e.g. the factor color contains trials with all different amounts of target objects and stimuli)

	Conj.	Col.	Shape	1	4	8	40	60	80
Number of errors	10	8	30	12	10	26	16	13	19
Number of fixations	18	10	24	16	18	20	15	18	21
Task perf. time [s]	6.6	3.1	8.8	5.9	6.7	8.7	5.1	6.7	7.9

- Comparisons between scanpaths of different groups (e.g., shape vs color or shape versus conjunction)

The assumption that there are differences between scanpath groups can then be tested post-hoc by comparing the distance distributions of the within-group comparisons to those of the between-group comparisons. The Kolmogorov-Smirnov test can be used in order to compare the distance distributions statistically.
In the following evaluation, I performed a multitude of tests for the different experimental settings and for each algorithm. The reported p-values are false discovery rate (FDR) corrected with the method of Benjamini & Hochberg [205].

A total of 735 trials were recorded (21 subjects $\times$ 35 tasks). 51 trials were excluded from the analysis due to insufficient tracking quality, as defined by 2.7σ of the median tracking rate of all trials.
In the following, three groups of scanpaths were compared to each other. We can distinguish scanpaths by

- the *task*
- the total *number of stimuli* on the screen
- the *number of search targets*

Thereby, the scanpath measures were evaluated against the objective parameters of task completion time and counting error rate. We would expect scanpaths of settings with a similar task completion time and error rate to be reported as more similar to each other than scanpaths of settings with large differences in completion time and error rate.
In Section 5.1, the problem of length normalization for state-of-the-art algorithms was mentioned. Especially for string comparison algorithms this is a relevant issue. In order to demonstrate the practical impact, we will take a look at the outcome of the distance measures with respect to scanpath length, i.e., the number of fixations performed.

Results

The number of counting errors shown in Table 5.1 are largest for the *shape* factor, followed by the factors *8 targets* and *80 stimuli*. This proves our initial assumption that the task gets more challenging with an increasing number of distractors and target objects. Consistently,

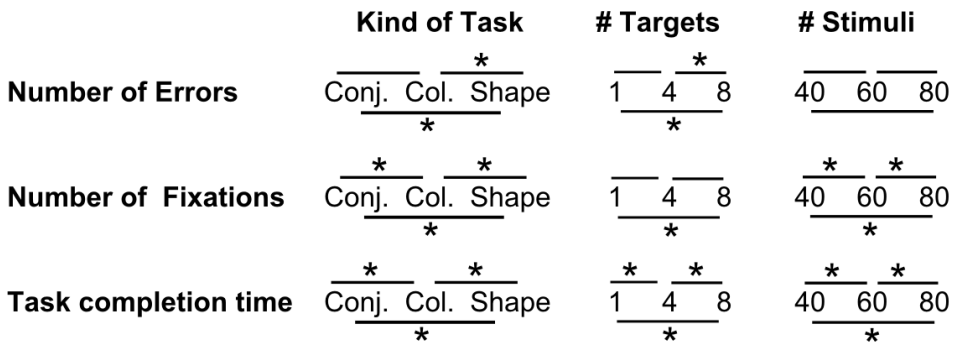


Figure 5.9: Overview of experiment design factors and their influence on task performance measures. Significant values of a Wilcoxon rank sum test ($p \leq 0.05$) are marked by a *. The within group distances of the respective group are tested versus the between group distances to the two other groups

more fixations are required to solve the more difficult stimulus screens and task completion time is enlarged.

Figure 5.10 summarizes the result of a Wilcoxon rank sum test:

The performance measures indicate that it is very easy to perform the pop-out color task, but difficult to distinguish targets by shape. The distribution of task completion times (median conj. 6.6 s, color 3.1 s, shape 8.8 s) as well as the number of counting errors (conj. 3.7%, color 3.8%, shape 11.9%) suggest that the conjunction task is more similar to the color task than to the shape task. We can assume that there is a strong color pop-out effect that allows for efficient filtering by color, leading to a large number of distractors being negligible for the more demanding shape feature comparison. This finding is consistent with the results of the healthy control group in the study of Machner et al. [151].

Figure 5.10 shows the results of a statistical test of scanpath distances within versus between the different factors. Some algorithms offer multiple measures (such as the five by MultiMatch). In this case each measure is evaluated separately.

Many algorithms are able to detect differences between the tasks. But the results are not quite as expected: given the above assumptions about scanpath similarity, we would have expected the shape task to clearly stand out and the color and conjunction task to be more similar to each other. This evaluation should result in an easily distinguishable shape task and harder to distinguish color and conjunction tasks (since between-group distances between color and conjunction task should be relatively small). The MultiMatch direction and length measures are not sensitive to this. They fail to separate the shape task. The result of the HMM shows what we would expect for an algorithm that exceeds its sensitivity range: the conjunction task, a mixture somewhere in-between the other two tasks, cannot be separated correctly anymore. iComp, ScanComp and SubsMatch, as well as some MultiMatch measures, are sensitive to all three tasks. It should however, be noted that the detection of a difference between groups does not imply that the algorithms are able to separate the groups from each other in the way a classifier would. Depending on the number of samples, a large overlap between groups would still be detected as significant, as

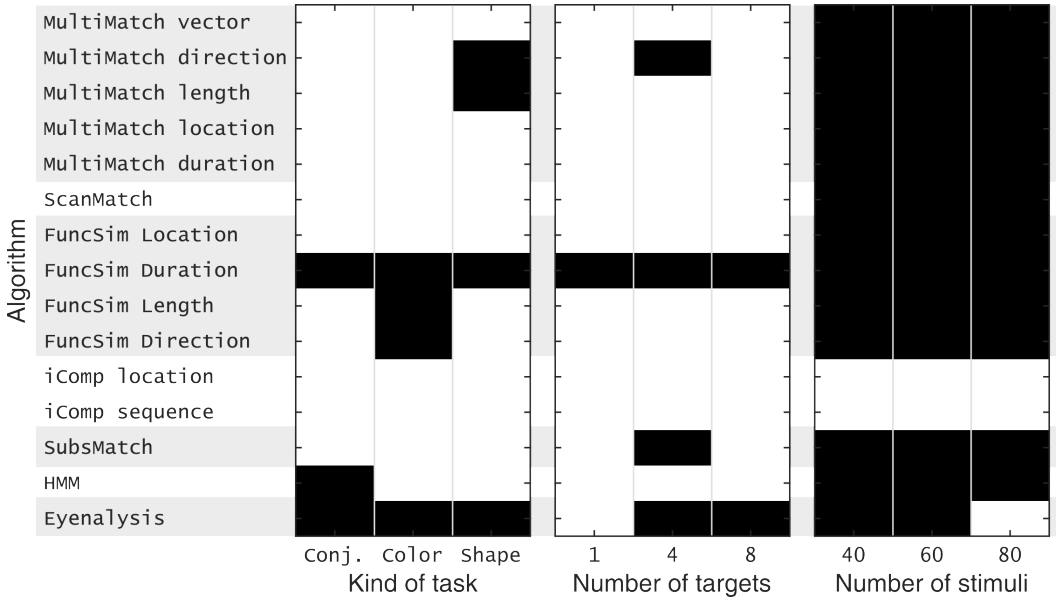


Figure 5.10: Wilcoxon rank sum test results for the three experimental factors. Significant differences between scanpath distance distributions ($p < 0.05$) are shown in white

long as a difference in the mean is present.

Figure 5.10 shows that all algorithms but FuncSim duration and Eyenalysis are able to distinguish between the different target counts. The sensitivity of MultiMatch direction and SubsMatch was exceeded: error rate and the number of fixations performed are more different between the one and eight target trials than towards the four-target trials. Therefore, the group with four targets should be the first one that cannot be distinguished as a distinct group anymore. For Eyenalysis we can observe an untypical behavior that will be discussed later on.

The total number of stimuli displayed did not have an impact on the error rate. Only the number of fixations performed and the task completion time were increased. During this extended time period the normal, unaltered visual search behavior continued. No change in search strategy, but only an extension in search time is reflected in the fact that almost no algorithm is sensitive to the change in stimuli count.

Only iComp and Eyenalysis are able to distinguish these conditions. The creators of Eyenalysis compared their algorithm to ScanMatch by the use of artificial eye movement data and performed much better than ScanMatch. This is in contrast to the results gained in this experiment, where Eyenalysis was not sensitive to most factors.

The algorithm includes a step where the produced similarity score is corrected for the length of the compared scanpaths. From our findings we can assume that the incorrect scanpath length normalization performed by Eyenalysis did not make a difference in their evaluation with synthetic (probably almost same-length) scanpaths, but has a huge impact for the scanpaths recorded during this experiment. iComp and Eyenalysis favor differences in scanpath length over scanpath shape changes.

5.3.2 Driving with visual field defects

Driving is a highly dynamic and interactive task: a looked-at object changes its relative location to the subject with vehicle speed, steering angle and head movement. In this experiment subjects with visual field defects and a control group performed a driving test in a simulator. The acquisition of the data used for the evaluation in this section is described in detail in Section 3.3.

Studies indicate that no general decrease in patients' fitness to drive can be observed and that some individuals show safe, others highly unsafe driving behavior. Many studies agree that so-called compensatory gaze movements, i.e., eye movements that enlarge the perceivable scene, may help to compensate for the visual impairment (see Section 3.3). However, characterization and identification of these movements is still vague.

Thus, we are interested in the comparison of viewing behavior of fit-to-drive patients, unsafe drivers, and the control group. For this analysis, drivers were assigned to one of three groups: (I) Glaucoma or hemianopsia patients passing the driving test (II) Glaucoma or hemianopsia patients failing the driving test, and (III) fourteen healthy-sighted control subjects.

The gaze compensation theory suggests that either patients who pass the driving test exhibit a different gaze pattern than the control group or that exploratory behavior of patients who fail the driving test is altered.

In terms of scanpath similarity comparisons this means we have to test whether within-group distances differ from between-group distances for those three groups. More specifically, we hypothesize that the scanpath distances within the groups are smaller than those between groups.

In contrast to the conjunction search task the data contains not hundreds of short trials, but few and long trials. The extracted scanpaths contain about 40 minutes of eye-tracking data per driving sequence, resulting in an average of 5,500 fixations per drive. Scanpath distances were computed using the SubsMatch algorithm and its competitors ScanMatch [178], MultiMatch [191] and transition matrices [194]. As a common preprocessing step fixations were extracted using a mixture of Gaussian methods [22].

Grid search resulted in an optimal parameter choice for SubsMatch n-grams with $n = 6$ and alphabet size $a = 6$. MultiMatch offers 5 different similarity measures. The best results (which are reported here) were achieved for position similarity. The optimal grid size for ScanMatch was determined at 5 bins on the horizontal, 3 on the vertical axis. Transition matrices do not produce a single similarity measure, but a whole matrix. Therefore, the absolute difference in transition frequencies was computed, just as for the SubsMatch frequencies. This implies that SubsMatch will always perform at least as good as the transition matrix approach. Transition matrices with such a post-processing step applied are just a specific parameter choice for SubsMatch with $n = 2$.

A permutation test (100,000 permutations) was performed on the scanpath distances computed by the above algorithms. The actually found within and between group distances were compared to this random baseline. Reported p-values in Table 5.2 are FDR adjusted for multiple testing.

The null hypothesis that scanpath distance distributions within and between the groups are equal, can be rejected ($\alpha = 0.05$) for two of the comparisons (Table 5.2): (I) the group

Table 5.2: Comparison of the distances within and between scanpath groups (G for glaucoma patients, p/f for passed or failed the driving test, C for control subjects). The FDR adjusted p-values of a permutation test are shown for the different algorithms. Significant ($p \leq 0.05$) values are presented in bold

	$\{G_f; G_p\}$	$\{C; G_f\}$	$\{C; G_p\}$
SubsMatch	**0.03**	**0.03**	0.67
ScanMatch	0.05	0.05	0.65
MultiMatch	0.89	0.89	0.89
Transition matrix	0.80	0.81	0.81

of glaucoma patients who passed the test (G_p) and the group of glaucoma patients who failed (G_f), and (II) glaucoma patients who failed (G_f) and control subjects (C). The null hypothesis cannot be rejected for the G_p and Glaucoma control (C), nor for any of the hemianopia group comparisons (not shown in the table).

None of the competitor algorithms was able to identify any differences between any of the groups, although ScanMatch came quite close ($p = 0.05$).

SubsMatch showed robust significant differences for the parameter range $n = [5;7]$ and $a = [4;9]$. We can conclude that the increased pattern length compared to the transition matrices also increases the sensitivity of the algorithm for the effect of compensatory gaze movements. The range in which parameters can be chosen without altering the significance of the results is relatively large.

The occurrences of those patterns that exhibit the highest difference in their frequency between the groups are shown in Figure 5.11. There are patterns that often occur in the control group, but not in the group of patients who failed the driving test. These deficits may be interpreted as indicators of decreased visual exploration performance. On the other hand, specific patterns occur in the patient group only, possibly manifestations of the visual field defects in exploratory behavior. Examining these patterns in more detail reveals that they consist mainly of large, alternating saccades. Peripheral vision may be required in order to target these saccades. They may be important to perform a fast scanning of the environment.

SubsMatch can successfully find patterns in this dynamic scenario as the key concept is not the question of *what* is being looked at, but a higher level process of visual exploration.

This important difference has implications for the scenarios when SubsMatch is favorable over its competitors. SubsMatch is well applicable to dynamic scenarios and scanpaths of sufficient length. It is likely to be sensitive to repeated patterns and to generalize well even with a high noise level.

For other applications, such as still image viewing, where ROIs are easy to define and viewing times are short, the scaling and translation invariance of SubsMatch might erase important patterns that are likely to be detected by the competitor algorithms.

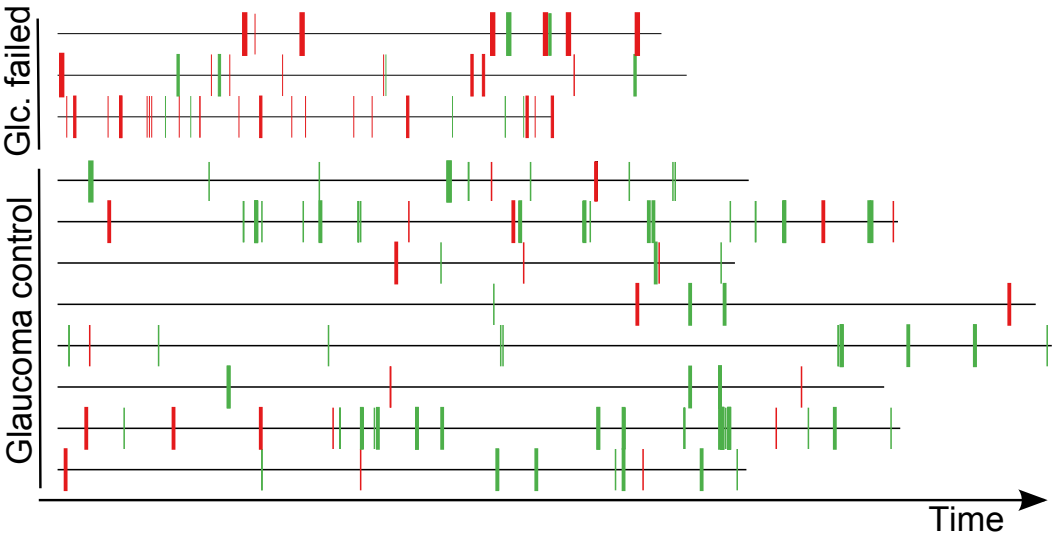


Figure 5.11: Occurrences of the ten patterns with the highest frequency difference between Glaucoma patients who failed the driving test and the control group. Each row indicates the complete scanpath of one subject, with the length of the black line corresponding to the length of the respective scanpath. Red vertical lines mark patterns that occur more frequently in patients that failed the driving test, green lines are frequent patterns in the control group

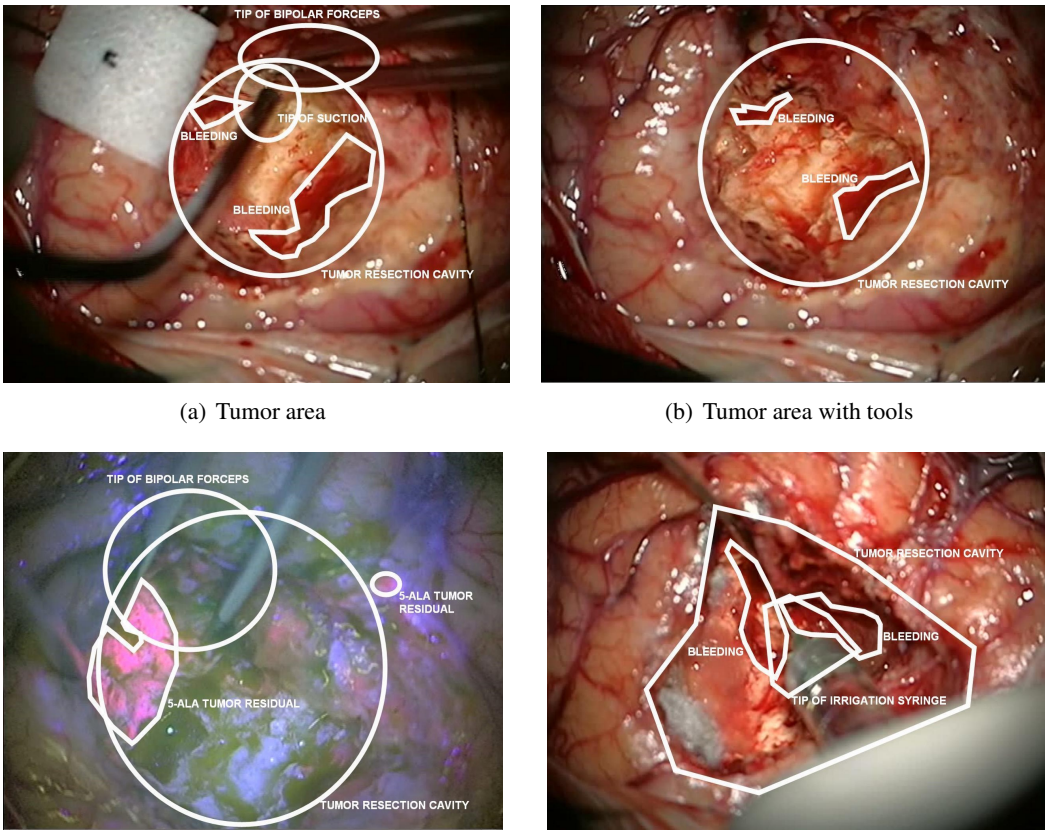
5.3.3 Neurosurgery under the operating microscope

Performing neurosurgery is undoubtedly demanding with regard to both, medical expertise and visual performance. For example, during a tumor removal surgery the visual challenge is the identification of all tumor among normal tissue. Such surgeries are often performed with an operating microscope that is itself complex to handle correctly. E.g., one has to adjust the focus on the correct depth to produce a sharp image.
Scanpath analysis has been applied for the assessment of expertise in a variety of challenging visual domains, such as medicine [206, 207, 208, 209], arts [7] or chess [210]. The aim of studying the influence of expertise on eye movements and visual search behavior is to get a better understanding of the cognitive processes of the experts. This could be beneficial to improve and speed up the training of novices.
In this section, we examine expert neurosurgeons' gaze behavior while viewing images of a surgery. Their gaze patterns are compared to those of novices.

Methods

Data used in this section was recorded by Eivazi et al. [206]. The authors thankfully provided data of one more participant than published in their study. Seven expert surgeons and seven novices viewed four images of a tumor removal surgery (Figure 5.12) for 10 s each. Subjects' gaze was recorded by a Tobii T120 eye tracker. ROIs were annotated for the tumor cavity, the instruments, and the bleeding areas.
Eivazi et al. found differences in the viewing behavior with regard to the amount of gaze

directed towards the instruments. In the third image (Figure 5.12(c)), a fluorescence marker was applied and experts deployed more gaze towards the highlighted areas. Furthermore, longer fixation durations as well as shorter saccadic amplitudes were found for the expert group. The overall viewing behavior of experts was characterized as *more compact*, especially for the third and fourth stimulus (Figure 5.12(c,d)). We will investigate whether this finding can be supported by automated scanpath analysis.



(a) Tumor area

(b) Tumor area with tools

(c) Flourescence marker applied

(d) Tumor area with tools

Figure 5.12: The sequence of stimuli images in the order in which they were presented. ROI annotations were not shown to the subjects. Image source: [206]

Scanpath similarity was calculated for the above data (14 subjects viewing 4 stimuli) with different measures, namely SubsMatch [34], ScanMatch [178], MultiMatch [191], FuncSim [193], iComp [182], Eyenalysis [190], and a HMM trained on mean-shift clusters. Each of these measures produces a similarity matrix of dimension $14 \times 14 \times 4$: for each stimulus image the scanpath of each participant was compared to all other scanpaths from that stimulus.
Some measures require a choice of one or more parameters. Since those cannot easily be derived from data nor are there meaningful defaults, grid-search was applied in order to find the optimal parameter settings. For algorithms that provide multiple distance measures

(such as MultiMatch), all dimensions are treated as a separate, independent measure.
Aim of this analysis is to find differences in the viewing behavior between expert and novice surgeons. These should be reflected in the distributions of scanpath similarity within and between the two groups of scanpaths. Generally, several results would be possible:

- the distributions could be identical, speaking for no difference in viewing behavior.
- the within-group similarity could be higher within the surgeon and within the novice group when compared to the inter-group distances. An indicator for distinct expert and novice viewing pattern.
- the within-group similarity of only one of the groups could be small when compared to the intra-group distances and the distances of the other group. One of the groups employs a more standardized viewing pattern, for example a learned strategy or a strong saliency guidance, while the other group shows a higher variability.

In order to detect these changes in similarity distributions, the Kolmogorov-Smirnov test was applied. The parameter free test imposes no prior assumptions on the shape of the distributions and has a high robustness at the cost of a relatively low statistical power. Correction for multiple testing was applied (Benjamini-Hochberg false discovery rate for 14 measures$\times$4 stimuli).
A p-value < 0.05 denotes a significant difference in the distributions between groups for an α-level of 0.05.

Results

Figure 5.13 shows the results of the statistical tests for all similarity measures and stimuli. We can observe that seven measures were able to determine differences in the viewing behavior between the expertise groups for at least one of the images. Only SubsMatch identified differences between the expertise groups for the first as well as for the second stimulus (Figure 5.12(a,b)).
Algorithms that rely on fixation location information (ScanMatch, SubsMatch) and respecting the temporal order of fixations (ScanMatch, FuncSim, SubsMatch) show good performance. For the static image stimuli local information also corresponds to ROIs (such as the location of tools). Results will probably change in favor of other methods (such as FuncSim direction) once a video or real operation would be used as stimulus.
Each of the boxes in Figure 5.13 hides a full similarity matrix - and a lot more information than the mere significance of group distances is available. To visualize this, we chose the similarity matrix produced by the SubsMatch algorithm for the fourth stimulus as an example (it is significant for most algorithms and likely to contain clearly differing scanpaths of both groups).
Any measure that produced a significant result is likely to exhibit some kind of group separability in its similarity matrix, but the degree to which this is the case varies a lot.
Figure 5.14(b) was created by a multidimensional scaling (see Section 4.2.1) of the distance matrix of Figure 5.14(a). We can observe a subgroup of three novices showing very similar gaze behavior, while experts' scanpaths spread around them. This finding can be interpreted

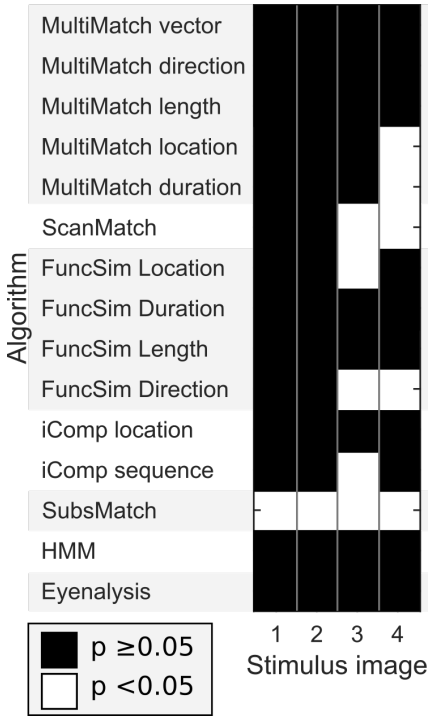


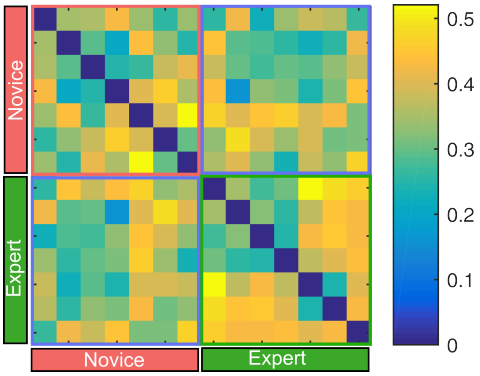
Figure 5.13: Result of automated scan pattern comparisons on four stimulus images. Tests were performed per image on the similarity distributions within and between the expertise groups. Reported are the false discovery rate corrected p-values of a Kolmogorov-Smirnov test applied to the distances within and between expertise groups

as a more homogeneous viewing behavior of novices while experts exhibit heterogeneous strategies. The novice group can probably be split into two subgroups, one with very homogeneous gaze behavior and one more similar to the experts' gaze behavior.
Figure 5.14(c) suggests that experts focus on certain image regions with the first few fixations and continue then with a broader exploration phase. Novices show a more repetitive viewing behavior, similar to the first few seconds of the experts. Possible explanations would be that experts understood all important aspects of the image during the initial viewing phase and were confident to have realized everything important. The second phase might be a broader search in order not to miss anything.
A geometrical separation of the MDS plot into the two expertise groups would be possible as a function of distance towards the centroid. This indicates that the level of expertise is a major cause of systematic variance within the eye-tracking data and that, consequently, differences in viewing behavior between experts and novices do exist.

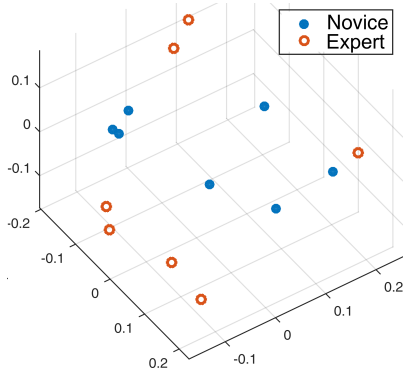
Discussion

Among the state-of-the-art algorithms SubsMatch revealed scanpath differences for all four stimuli. The above findings of the automated analysis correspond to the findings of the

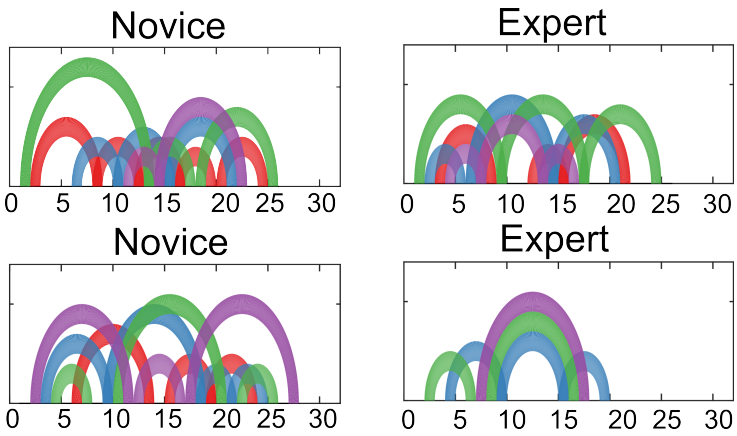
manual data analysis.
Eivazi et al. [206] found a significant effect of expertise on the number of fixations performed. Furthermore, the authors reported that the average fixation duration was found to be significantly longer (for all stimuli) and average saccade length larger (for stimuli 3 and 4) for the expert group.
In the light of the automated scanpath analysis we can now conclude that longer saccade lengths probably resulted in or from the more heterogeneous scanpaths found for the expert group. Most algorithms performed better on the third and fourth stimulus image, for which Eivazi et al. were also able to find more and stronger effects than for the other images. It should be noted that significant differences derived by the automated analysis does not necessarily imply that groups are clearly separable. Although higher distances between than within groups indicate this fact, a large overlap between the groups is possible. Therefore, the visualization step is useful to reveal the actual extend and nature of the differences.
As of now we can only speculate about what happens during the second, heterogeneous exploration phase in the expert's gaze behavior. Whether mimicking these viewing patterns increases the learning rate of novice surgeons or contributes to a better surgery remains completely unexplored.
In the MDS plot of Figure 5.14(b) we can see a separation of novice and expert gaze behavior, which suggests that the two groups could be separated based on their scanpaths. For the previous driving experiment such a separation would be extremely useful, as it would be an objective measure of driving safety. But SubsMatch is not a classification method. The measure does not aim at optimal group separability but only quantifies the level of similarity between scanpaths. In the next section SubsMatch is extended to be useful for classification based on the scanpath.



(a) Distance matrix



(b) MDS plot showing the magnitude along three virtual dimensions



(c) Repetitive patterns along the first 30 fixations

Figure 5.14: (a) Distance matrix of the pairwise scan pattern comparisons with SubsMatch for the fourth stimulus image. Each row and each column corresponds to one individual. Individuals are grouped by their expertise label. Entries along the diagonal represent comparisons of a subject towards itself (which results in zero distance, as we compare a single scanpath per subject towards itself) (b) Scatterplot of the multidimensional scaling of the distance matrix in (a). The dimensions resulting from the scaling have no unit, but correspond to a distance between the scanpaths. Novices show an overall smaller distance towards each other, resulting in a denser MDS plot while expert scanpaths exhibit larger distances towards each other. (c) Repetitive patterns as found by SubsMatch. Each arc connects two occurrences of a pattern and the colors code the identity of a pattern. A repetitive viewing behavior can be found for the novices (i.e., more arcs). Experts focus their repetitive scanning to the first few fixations, then transition to broader exploration. The height of an arc corresponds only to the distance covered between the repetition of the pattern and was introduced to reduce the overlap between arcs for purposes of illustration.

6 Scanpath Classification

The SubsMatch algorithm utilizes scanpath features that are sensitive to a large variety of different experimental factors. On the one hand, this high sensitivity is beneficial, since it enables us to identify a large diversity of even subtle factors and slight alterations in scanpath shape. The main drawback, however, is that the similarity score calculated by SubsMatch integrates over all those factors and thereby mixes the effects on the final measure, i.e., strong effects superimpose weaker ones and can easily dominate the final score. Weak effects are detectable only when a large amount of balanced data is available. More specifically, highly frequent subsequences have a large influence on the similarity score, no matter whether they are discriminative.
Clustering scanpaths by the SubsMatch similarity score gives insight only on the most prominent effects. In the previous section, post-processing techniques of the similarity matrix (such as MDS) were investigated. They can compensate for this effect, if enough data is available.

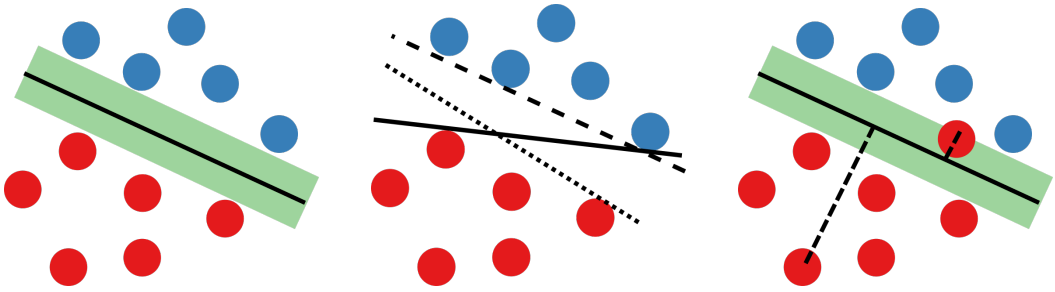
In scope of this section is a technique that enables us to identify the influence of one targeted experimental factor on scanpath similarity. Such methodology is required to answer questions such as "Is this the scanpath of a currently attentive or a distracted driver?" or ideally "Is this patient fit to drive a car?". This problem is called scanpath classification, as we want to assign a label (attentive or distracted) to a scanpath.
One simple extension to transform SubsMatch into a classification method is to perform a k-nearest-neighbor label assignment: the k scanpaths with the least distance to the scanpath that is to be classified cast a vote for its label and the majority of votes decides, e.g., if the scanpath is similar to many scanpaths of attentive drivers and few of distracted drivers, it is likely to belong to an attentive driver.
A possible extension would be the introduction of archetype scanpaths for each class: a scanpath that is highly representative for its group would be assigned a high weight, scanpaths that are less characteristic vote only with low weights. The voting is then performed by the weighted majority. Hembrooke et al. [211] follow a similar approach by performing multiple sequence alignment - an extension of the string alignment approach that results in an *average scanpath*.
However, at this stage we are already working with a noisy measure that integrates over all experimental factors and effects. It is favorable to tackle the problem at an earlier stage, before SubsMatch calculates the similarity score.

This section discusses the calculation of a discriminating similarity score, meaning some sort of weighting of subsequence frequencies' contribution to the final similarity score, based on whether the subsequence discriminates between two groups.

This way we can emphasize different aspects of the similarity measure and are able to calculate a score that is sensitive to one specific factor, in the presence of other factors or noise.
This challenge is tackled by means of machine learning. Section 6.1 demonstrates how SubsMatch features can be embedded in a machine learner. In Section 6.2 possible expansions are discussed.

6.1 Support vector machines for scanpath classification

There is a broad spectrum of powerful machine learning methods available (neural networks, HMMs, Bayesian approaches and many more). Theoretically, most of them could be used for scanpath classification. Often the choice of good input features is much more important for the final performance than the learning method employed.
In practice, Support Vector Machines (SVMs) are a commonly used tool. Figure 6.1 illustrates the basic principle of a linear SVM at the example of a two-dimensional feature space. SVMs offer several benefits. SVMs with a linear kernel allow for the extraction of learned feature weights. These can either be used for feature selection [212] (e.g., to minimize computational load) or to get an impression of which scanpath features are changing the most between two groups. It is in part possible to take a look into the abstract machine learner and to understand what the decision is based on.
Further, using a trained SVM has low run-time demands. In fact, Braunagel et al. [213] use a classifier, based on the method described here, for secondary task detection during conditionally automated driving in an online setting.



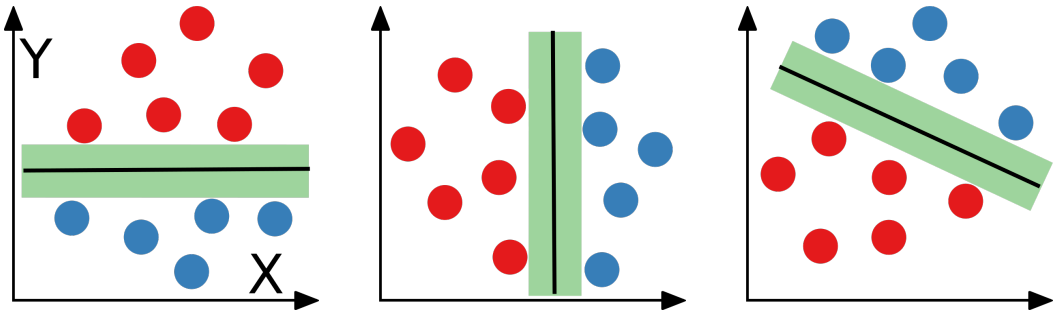
(a) Hyperplane with maximal margin

(b) Other suboptimal hyperplanes

(c) non-perfect class separation

Figure 6.1: (a) Samples of two classes (red and blue) characterized by the two-dimensional feature of their location that can be linearly separated by a hyperplane. The hyperplane with maximal margin (the green area) towards the support vectors is used by the SVM and determined during the learning phase. Only samples close to the hyperplane are considered as support vectors and influence the course of the hyperplane. (b) Other hyperplanes that separate both classes perfectly but with inferior margin. For classifying new samples, the maximal margin approach is most likely to lead to the correct result even for samples close to the hyperplane. (c) In some cases the classes cannot be separated perfectly. Generally, the larger the distance from a sample towards the separating hyperplane (dashed lines), the more certain the classification

The same scanpath features as in the SubsMatch algorithm will be used, frequencies of short scanpath subsequences. The *n*-gram feature embedding is described in the SubsMatch section.
As feature matrices easily become large and sparse, the tool Sally [214] can be used to construct them. It has an impressive computational performance with such data. As observed for the SubsMatch algorithm, the number of potentially possible *n*-grams increases quickly but the actually occurring ones are much less in numbers. That results in a highly sparse feature matrix. Sally was designed for mining large amounts of text documents such as computer logs or DNA sequences. It implements a generalized form of the bag-of-words model (that can contain *n*-grams as a special sort of word) to map a set of strings to a vector space spanned by strings. The string frequency is the amplitude in the direction of the string in the vector space.
Normalization of the feature vectors can be done in two ways: feature-wise, in order to avoid the masking of weak features by stronger ones, or vector-wise, which corresponds to the frequency normalization applied by SubsMatch and compensates well for different sequence lengths.
In this work, the libSVM [215] implementation and its variant for linear SVMs, libLinear [216], were used. During training, the SVM simultaneously minimizes classification error and maximizes margin. The trade-off factor (how easily false classifications are tolerated in favor of a larger margin), also called the soft-margin, is one of the parameters that needs to be set before training of a SVM.
LibLinear offers a dual-based L2 regularized solver that is specially suited for processing large sparse data where the number of instances (in this case scanpaths) is distinctly smaller than the number of features (in this case *n*-grams). This solver is known to be quite insensitive to the SVM C-Parameter, which influences mainly the time consumption for training. Elegant solvers are available for linear kernels and feature weights can be retrieved. On high-dimensional data, as at hand (each *n*-gram spans its own dimension), linear kernels often perform similar to more complex kernels (such as the radial basis function) [216]. Figure 6.2 visualizes how feature weights can be extracted from the SVM.



(a) separable by Y-coordinate (b) separable by X-coordinate (c) separable by combination of X and Y coordinates

Figure 6.2: Linear SVM feature weight visualization. The two feature dimensions are named X and Y. (a) in this special case, only the Y-coordinate holds any information about class identity and therefore has weight 1. The non-informative X-coordinate has weight 0. (b) the other way around. (c) for most practical applications the combination of several features contributes to a good classification. Both feature dimensions are assigned (positive or negative) weights unequal to 0

6.2 Alternative approaches and possible extensions

6.2.1 Mismatch kernel

The SVM approach cannot utilize feature similarity for the comparison. The patterns *AAA* and *AAB* are considered just as different from each other as *AAA* and *BBB*. Since SubsMatch cannot utilize histogram comparison methods such as earth-movers distance, the string kernel has no notion of n-gram similarity. This problem can be tackled by *smoothing* frequency scores between similar n-grams. Depending on the number of mismatches between two n-grams (i.e., the number of edit operations required to turn one into the other), they are considered more or less similar to each other. Mismatch kernels [217] implement a notion of string similarity by adding pattern occurrence counts not only to the exactly matching n-gram, but also to all n-grams within its mismatch equivalence class, i.e., those sequences with a certain edit distance to it. This way the histograms get smoothed.

The major drawback of this approach is the immense computational overhead [218]: n-gram similarities have to be calculated and counts increased for a large amount of n-grams instead of just one. Relatively good performances can be achieved by a clever choice of an indexing structure over the n-grams. But the major reason, why feature embedding without mismatches can be done really fast is that the feature matrix is sparse, i.e., many of the possible n-grams never occur. For the mismatch kernel all possible n-grams have to be considered (since they might be in the equivalence class of an occurring one).

In prototype tests, allowing mismatches did not have a huge impact on the results: if there is enough data, i.e., many and long scanpaths, reducing the alphabet size slightly has a similar effect on the results as allowing for mismatches.

The Sally tool provides a computationally more efficient way of handling special kinds of mismatches, i.e., sorting the n-grams alphabetically. The n-grams *AAB*, *ABA* and *BAA*

would be sorted into the same bag *AAB*. While the run-time decrease is dramatic when compared to other mismatch kernels (and sparsity is conserved or even stronger), the *n*-gram has no representation of direction anymore. Furthermore, mismatches such as *AAA* and *AAB* cannot be merged. The method is therefore not suited for a scan pattern representation, where order is of essential importance.

6.2.2 RepeatScout

A mixture of mismatch kernel and the t-pattern algorithm, which finds long and gapped temporal patterns, is implemented by a bioinformatics tool called RepeatScout [219]. It is designed for the de-novo identification of repetitive genome sequences. Short patterns as identified by SubsMatch are used as seeds and extended sequentially. This extension step allows for a limited number of mismatches and continues as long as the extended pattern is still relevant, i.e., it occurs significantly more frequently in the sequences as a random baseline model.
In several evaluations, an adaptation of the RepeatScout algorithm for eye-tracking data did not improve over the SubsMatch approach: as for the mismatch kernels, allowing for mismatches in the patterns did not provide a relevant performance increase. The pattern length could not be increased a lot over the optimal *n*-gram length.

6.3 Evaluation

One of the major problems with existing scanpath comparison algorithms is that they lack proof of generality. While each algorithm is well suited for a specific use case and the data set it was developed for, it is not easy to assess which algorithm is suited for a new experimental design. Furthermore, little is known about the sensitivity of individual algorithms to specific experimental factors, nor is their sensitivity to different sources of noise explored. In order to overcome this, the applicability of the string kernel method to a wide variety of experiments is demonstrated in this section. We will find that it generalizes well over experimental settings ranging from simple image viewing to complex real-world driving tasks. Further, different post-processing steps are shown, e.g., the extraction of patterns that allow for an efficient distinction between scanpaths groups.
For this purpose, we provide an extensive eye-tracking data set collected in four representative experiments that cover a wide range of typical eye-tracking applications. It contains some classical eye-tracking experiments, such as the ones by Yarbus [8] and Land [56], but also new and challenging data, e.g., visual scanpaths during real-world driving. The selection of experiments is driven by the idea that the degree of freedom, which the subject is allowed to perform within, is probably the most important factor for which comparison algorithm is applicable.
Eye-tracking is traditionally employed in a wide field of applications: some experiments require head restraint, timed stimulus presentations of only few seconds duration and up to 1000 Hz sampling rate at controlled illumination. On the other side there are interactive real-world scenarios with sampling rates of 25-60 Hz and highly dynamic, uncontrollable

stimuli. These factors impose consequences for the signal quality of the eye tracker and the calibration, for the feasibility of ROI annotation and finally for scanpath comparison.
Short trial duration and repeated stimulus display in a laboratory setting is likely to lead to highly similar scanpaths and little noise. Real-world experiments are always associated with a high level of pupil detection failures and identical experimental conditions are hard to reproduce (e.g., the same amount traffic while driving). Scanpaths of such experiments are likely to be more heterogeneous and noisy.
For all the following evaluations, the goodness of the string kernel approach is judged by the share of correct scanpath classifications, i.e. the number of labels that were assigned to a scanpath correctly relative to the false assignments. Further, the statistical baselines for a classifier that always predicts the majority class are provided. Since for some experiments class sizes are not balanced, it might be easy to achieve high correct classification rates by just guessing the class label that occurs most in the data. In order to test whether the resulting proportion of correctly classified scanpaths is above the chance level, a binomial test can be performed (testing the number of correct predictions versus chance).

Classification accuracy was assessed by 10-fold cross-validation. Cross validation is the process of repeatedly splitting the available data into so-called training and test sets. The training set is used to train the SVM. The model's accuracy is then assessed on the test data. By repeating this procedure with different splits of the data, variability is reduced. 10-fold cross-validation stands for a split of the data where 1/10th of the data is used for validation, 9/10th for training. By permutation of which scanpaths belong to the training and which to the test set, 10 repetitions are performed with a different test fold in each permutation.
It should be noted that due to the adjustable parameters in the algorithm, a considerable amount of SVMs are trained (i.e. n-gram lengths 1 to 10, alphabet sizes 2 to 26 and different string encoding patterns). For most experiments, classification accuracy is clearly above the statistical baseline. No correction for multiple testing was performed on the binomial test. In those cases where classification rate and baseline are close to each other, one has to be careful with the interpretation of the results. As this extensive evaluations spans multiple experiments, correcting for all performed tests would diminish comparability of the results to other studies (as for the Yarbus experiment in Section 6.3.2).
The SVM C-parameter (for the soft margin) was not optimized (unless stated otherwise). SVM parameter optimization is usually done in a grid search approach. A validation data set would be required in order to test the effects of the optimization adequately. For some experiments however the creation of data is very expensive and requires massive effort. In order to avoid over-fitting of the classifier, no parameter optimization was done.
But an adjustment for unequal group sizes was performed by weighting a misclassification of the minority class stronger than in the majority class. In the following, results for each of the four evaluation experiments are presented.

6.3.1 Conjunction search task

The data acquisition process and setup of this experiment were introduced in Section 5.3.1. Subjects were instructed to count all objects displayed on a screen with the requested

properties, namely *color*, *shape* and the *conjunction* of both.
The conjunction search task can be performed under standardized conditions and has a well defined setup. Therefore, it is suited to study the effect of specific experimental factors on the results of the string kernel method. Furthermore, scanpath results can be correlated with a performance measure (search time and error rate).
Classification results for distinguishing scanpaths by the number of targets, the total number of objects and the individual stimulus screen are shown in Table 6.1. Recordings of minor quality (as reported in Section 5.3.1) were excluded from the analysis. Therefore, guessing chance might deviate from the intuitive chance level of a completely balanced design (e.g., 25% for a 4-class problem), even though the experiment was designed to be balanced in the number of trials per condition.

Table 6.1: Classification accuracies (10-fold cross-validation) for the conjunction search task with different research questions. Statistically significant results ($p < 0.05$) of a binomial test (without correction for multiple testing) are reported in bold. Different string encoding patterns are displayed: circular encoding, horizontal and vertical axis percentile binning, horizontal and vertical axis regular binning. The baseline column provides the guessing chance level and the last column the result of the best classifier

Research Question						**Baseline**	**Best**
Search task	**61.6**	**61.7**	**57.5**	**54.2**	**54.5**	37.2	61.7
# objects	27.2	27.1	27.1	26.7	27.9	33.3	27.9
# targets	45.8	45.5	**51.5**	46.6	48.2	45.8	51.5
Stimulus	**8.2**	**7.2**	**17.3**	**14.3**	**8.7**	2.9	17.3

The number of fixations performed differs significantly between the three tasks and the three total object counts. Overall, the performance measures indicate that it is very easy to perform the pop-out color task, but difficult to distinguish targets by shape.
This effect is also resembled in the confusion matrix (Figure 6.3). The diagonal of the confusion matrix shows the percentage of scanpaths that were assigned the correct label, i.e., the correct task. Elements that are not on the diagonal represent a misclassification, e.g., a color task scanpath wrongly classified as a conjunction task. We can assume that scanpaths from tasks that are often misclassified are more similar to the classifier.
Classifying the shape task is possible at a high accuracy of 72%. The confusion between color and conjunction task is much higher (31% / 33%) than towards the shape task (11%/15%). The tasks that are more similar in terms of task performance also result in more similar scanpaths. The more demanding shape task results in more unique, distinguishable scanpaths. The overall classification accuracy of 62% (against a 37% baseline) supports Yarbus' finding: classification of the observer's task from his scanpath is possible. This finding holds for the conjunction search task with its much simpler stimulus complexity than the image viewing experiment.
Surprisingly, the string kernel method did not reliably succeed in separating trials of different stimulus count or different target count from each other. We speculate that this is the case since the method cannot take different scanpath lengths into account but normalizes the

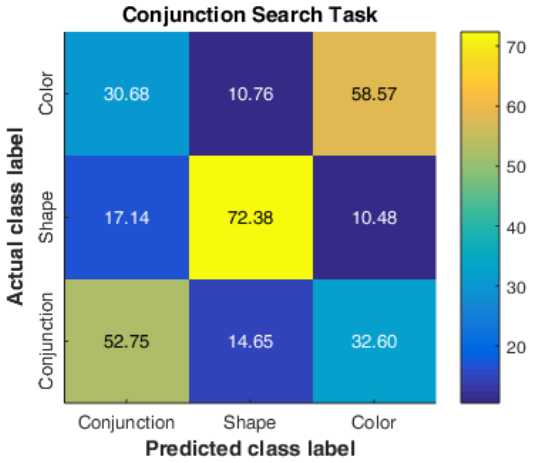


Figure 6.3: Confusion matrix of the conjunction search task. Diagonal entries from the lower left to the upper right show the percentage of correctly classified scanpaths. High correct classification rates of the shape task are visible. Off-diagonal elements are wrong classifications between the classes. Most wrong classifications mix up the color and the conjunction search task

transition frequencies in order to eliminate this effect. We can assume that visual search behavior occurred for the whole trial duration. A long search may appear very similar to a short search after normalization.
Especially remarkable is the high accuracy (17% versus 3% chance level) of assigning gaze recordings to a specific stimulus (see Table 6.1). Since the same stimuli were presented to all subjects, exact positional information is very sensitive and a high spatial resolution (represented by a large alphabet size) can capture this effect.
We can conclude that it is possible to detect high level effects such as the task given to a subject, but also to maintain high spatial sensitivity for image classification. Viewing times do not seem to have a relevant influence on scanpath similarity, as long as viewing behavior is similar for the whole duration. This implies scanpath length normalization is working well in eliminating the duration factor from the scanpath.

6.3.2 Yarbus' Unexpected Visitor

Yarbus claimed in his influential book about the effect of different tasks assigned to his subject during still image viewing [8]: *"depending on the task facing the subject, the eye movements varied."* What at first glance seems like an intuitive assumption, has a huge impact.
By turning this statement the other way round, its implications come to light: eye movements can give us insight on cognitive processes. When monitoring a subject's eye movements, we can (to some extent) draw conclusions on his or her thoughts. By now we know a lot more about eye movements and overt as well as covered attention. It has become clear that this inference is very limited. E.g., we are able to attend to regions and objects without ever looking at them directly. Consequently, reproducing Yarbus' finding has proven to be much

Table 6.2: Three replications and variations of the Yarbus experiment used for evaluation

	Subjects	Tasks	Stimuli	View time	Eye tracker [Hz]
OWN	20	2	2 paintings	30/120 s	EyeTribe (30 Hz)
RY [10]	17	4	20 images	60 s	Eyelink (1000 Hz)
DY [220]	21	7	15 images	30 s	Eyelink (1000 Hz)

harder than his intuitive findings suggest.

In this section three replications and variations of the original experiment are presented. One of them was recorded to evaluate different aspects of scanpath comparison algorithms. It was also used in Section 3.2.2.2 for the assessment of vigilance during still image viewing.

The two more data sets were provided by other research groups: Greene replicated parts of Yarbus' experiment and found a prediction of the observer's task was not possible above chance level [10]. Borji et al. replied [220] by employing a more sophisticated classification method that involved spatial characteristics of the scanpath. They were able to perform classification above chance level.

There are other studies on this topic [9, 221] but their data is not publicly available, nor did they evaluate their approaches on a common data set. Therefore, no objective comparison to their results is possible.

- **(OWN)**: our *own replication* with Ilya Repin's paintings at four different conditions: task variation (free-viewing and estimating the age of the people in the painting), stimulus change (different paintings by Repin) and different viewing times. The data acquisition process of this experiment was introduced in Section 3.2.2.2. Ten participants wore glasses, one contact lenses (age range 20-57, 6 male and 14 female).


- **(RY)**: the *Reconsidering Yarbus* [10] data contains 17 observers at four tasks (memorize, determine the decade when the picture was taken, estimate how well the people know each other, estimate the wealth of the people). The block design resulted in four to five observers viewing the same image with the same task assigned.

- **(DY)**: the *Defending Yarbus* [220] data contains recordings of 21 subjects performing seven different tasks on 15 images of natural scenes (free-viewing, estimate wealth, age, what was the family doing before the arrival, memorize the clothes, memorize positions of people and objects, estimate how long the visitor had been away). The block design resulted in three observers with the same task for each image.

On the (OWN) data the conditions task, image, and viewing time were varied in order to demonstrate the effects of these parameters on the scanpath similarity measure. A four-class SVM was trained in order to distinguish between scanpaths from each of the four stimulus screens. The in-detail classification results are presented as a confusion matrix in Figure 6.4. The confusion matrix shows that the extremely long viewing time of three minutes has the strongest influence on the scanpath measure, i.e., scanpaths of category free-viewing (3 min) can be assigned their label in 93% of the cases. From the previous experiment and the theory behind the algorithm we can conclude that this is unlikely caused by an imperfect scanpath length normalization but represents an actual change in the scanpath. Together with our previous finding on viewer's behavior during the extended viewing time (i.e., staring at certain image regions), and a heatmap analysis (not depicted) we can conclude that there are two ways of dealing with the task: getting bored and staring at objects or moving towards examining previously non-attended details of the painting. These temporal and spatial scanpath alterations were easy to distinguish.
The next strongest effect is exhibited by changing the stimulus image. Even though the paintings are very similar to each other, spatial image characteristics are changed. The algorithm can learn, which regions are attended to for each image and thereby separate the scanpaths quite well. We can further conclude that the separation between the two tasks, free-viewing and age estimation, is possible above chance level with 52-59% of correct assignments (Table 6.3), given the four-class problem. But there are more misclassification between these two classes (29% / 18%) than for the second or fourth stimulus.

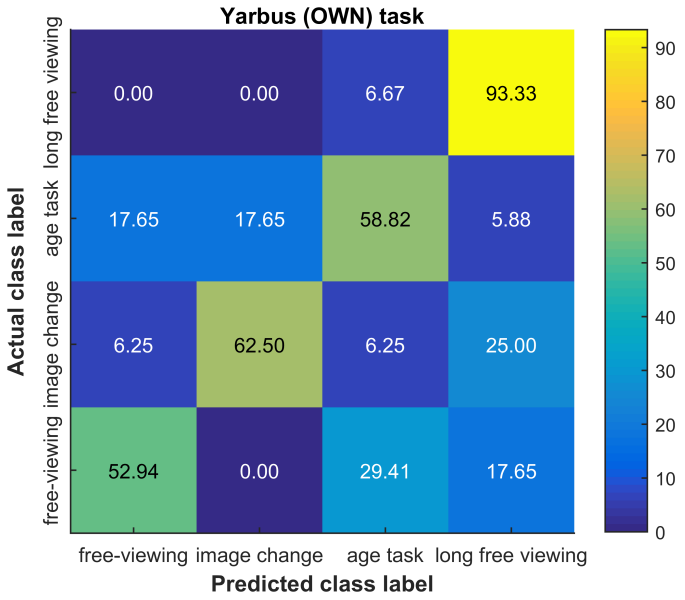


Figure 6.4: Confusion matrix of the Yarbus (OWN) experiment. Diagonal entries from the lower left to the upper right show the percentage of correctly classified scanpaths, off-diagonal elements are wrong classifications

This finding strengthens the claim of the importance to monitor attention and vigilance also during image viewing. As we can see from the results, the scanpath difference between recordings of the same free-viewing task performed on the same image (but for a different duration) is larger than the one caused by different tasks. In a nutshell, task classification is relatively hard compared to a change in the stimulus material or extended viewing durations – but possible.

In order to verify this, the string kernel method was also applied to data provided by Greene (RY) and Borji (DY). This data is especially interesting as it is possible to compare the string kernel method to related work. Both studies provide scanpath classification methods designed for their specific data set.

On the (RY) data, chance level is at 25%. Borji et al. reach a classification accuracy of 34.1% using a Boosting classifier [220]. Kanan et al. achieve 33% by a SVM trained on the parameters of a HMM [197]. The idea behind this is that the HMM's parameters change between tasks. Gaussian emissions and a probabilistic model are an elegant approach, but the Markov assumption cannot easily be circumvented. Each subsequent fixation is modeled to depend only on its single predecessor. The proposed string kernel reaches a classification accuracy of 34.4% with SVM parameter optimization and 32.1% without optimization.

On the (DY) data provided by Borji et al., the proposed string kernel achieves a classification

Table 6.3: Classification accuracies (10-fold cross-validation) for all experiments and research questions. Statistically significant results ($p < 0.05$) of a binomial test (without correction for multiple testing) are reported in bold. Different string encoding patterns are displayed: circular encoding, horizontal and vertical axis percentile binning, horizontal and vertical axis regular binning

Research Question						**Baseline**	**Best**
OWN: condition	**58.5**	**50.8**	**66.2**	**58.5**	**43.1**	26.2	66.2
RY : task	**30.3**	**32.1**	**30.9**	28.8	**17.4**	25.0	32.1
RY : image	**19.1**	**17.1**	**8.5**	**10.6**	3.2	5.0	19.1
DY : task	**21.6**	**17.8**	**18.7**	**19.0**	**18.7**	14.3	21.6
DY : image	**16.8**	**14.0**	**49.8**	**22.2**	**16.5**	6.7	49.8

accuracy of 24.2% with optimization, 21.6% without optimization. This is clearly above the chance level of 14.3% and very similar to the accuracy of 24.2% achieved by Borji [220]. Our results support the findings of related studies that scanpaths contain information useful for the prediction of the observer's task.
furthermore, the very similar classification results of the string kernel and the approach by Borji et al. indicate that with the extracted features and the current state-of-the-art classifiers an upper limit is reached. Observer's task identification can be performed above chance, but only moderately. It is remarkable that the limits are so close to each other, even though quite different scanpath features were extracted. Reliably predicting the task of one individual scanpath remains therefore a hard problem.
In a second run, the stimulus image that caused a specific scanpath was classified. The relatively high correct classification rates of up to 50% accuracy are caused by the good spatial resolution of the string kernel method. More specifically, for large alphabet sizes and a percentile binning approach, the method basically performs fixation clustering and thereby learns the positions and relevance of objects in each image. The structure of the images, i.e., the axis that best separates the relevant objects in the image, has a major influence on scanpath classification accuracy.
To conclude, the string kernel provides very similar prediction accuracy to other methods for the image viewing tasks.

The influence of n-gram length and alphabet size

As a general rule of thumb we can conclude that binning, either horizontally or vertically, shows good performance for stimuli presented on a computer screen. A real-world task adds additional complexity, i.e., relatively high measurement inaccuracies and calibration drift call for a more robust string encoding. When using the percentiles of data samples, one has to be aware of the severe limitations induced by this step. For example, a rightwards bias of one scanpath compared to another one would not be visible in this encoding.
There are certain factors to consider when choosing the n-gram length, since this method was designed to examine frequency differences. Therefore, n-grams have to occur multiple times in the scanpath. Choosing a large n will produce many unique n-grams and requires a long scanpath to allow subsequence counts to pile up. On the other hand, a low n might not

be able to capture characteristic patterns. For example, $n = 1$ corresponds to simple ROI frequencies, $n = 2$ represents ROI transition frequencies as in a Markov chain. The higher the n, the more specific the gaze patterns.
The choice of an alphabet size is highly stimulus dependent. The experiments demonstrated that for static stimuli and an expected high scanpath similarity a large alphabet is advantageous, for more abstract tasks with an expected low scanpath similarity a small alphabet size is preferable. For large alphabet sizes the method basically learns image properties, while low alphabet sizes seem to facilitate a better generalization to represent more abstract factors.

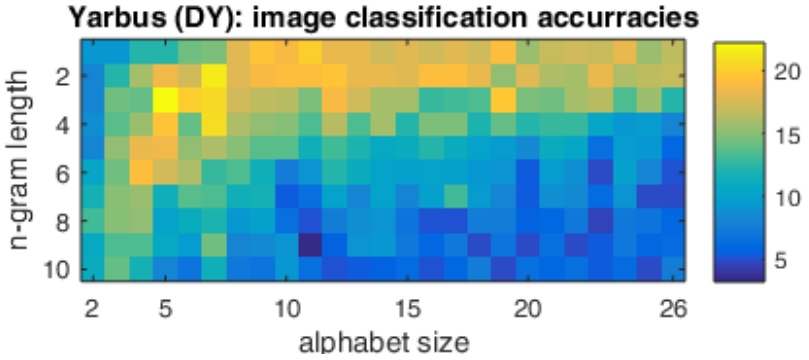


Figure 6.5: Visualization of a grid-search to determine the optimal combination of n-gram length and alphabet size for the Yarbus experiment. High correct classification rates can be achieved in the yellow areas. Continuous high scoring areas stand for a high robustness of the method towards parameter choice, small speckled areas would indicate a high importance of a specific parameter choice. In this example a relatively robust optimum can be found around n=3 and alphabet size 5

Figure 6.5 visualizes the process of a grid-search for the optimal parameter combination. We can conclude that there seems to be a trade-off between long patterns and a high spatial resolution in form of a large alphabet size. The longer the subsequences, the smaller the alphabet size needs to be chosen in order to achieve overlapping frequency counts. Therefore, both parameters are dependent on each other and have to be optimized in conjunction. It is worth a notice that for some applications classification accuracy varies strongly for small alphabet sizes, depending on whether a central region is encoded (i.e., the alphabet size is an odd number) or not.
We found that for short scanning sequences marking the first fixation n-gram uniquely can improve the results. This underlines the importance of the first few fixations in a static context. As this effect would also be visible in a multiple sequence alignment and becomes negligible for long scanpaths and dynamic scenarios, it was not included in the final version of the algorithm.

6.3.3 Neurosurgery under the operating microscope

An introduction to eye movements in microsurgery was given in Section 5.3.3. In this section, we apply the string kernel to a data set provided by Sharam Eivazi that contains

recordings of 23 novice and 16 expert surgeons while viewing a video (not images, as before) of a surgery. 20 subjects viewed a focused, 19 a blurred version of the video. The blur corresponds to a non-optimal focus of the operating microscope. Two experts had a tracking rate of 27% and 88% and were excluded from the analysis. All other subjects had a tracking rate above the cut-off criterion of at least 90%. Table 6.4 shows that the classification of both, expertise and blur, is possible high above chance level.

Table 6.4: Classification accuracies (10-fold cross-validation) for all research questions. Statistically significant results ($p < 0.05$) of a binomial test (without correction for multiple testing) are reported in bold. Different string encoding patterns are displayed: circular encoding, horizontal and vertical axis percentile binning, horizontal and vertical axis regular binning

Research Question						**Baseline**	**Best**
expertise	71.8	**76.9**	71.8	66.7	66.7	59	**76.9**
blur	**76.9**	**66.7**	64.1	**69.2**	56.4	51.3	**76.9**

6.3.3.1 Feature selection

For the data at hand, the n-gram data representation produces features matrices of sizes up to $k \times 4,000,000$ with k as the number of subjects. The matrix is usually extremely sparse, so that machine learning can still be handled computationally. However, the need for a feature selection step is obvious. In the following, only those features with a relevant impact on prediction accuracy will be extracted.

For general SMVs, this could be tackled by systematically deleting individual features from the matrix and observing whether the prediction gets worse. But for linear kernels a more elegant solution exists: the SVM learns feature coefficients that correspond directly to an importance weighting. For the following evaluation only the ten features that discriminate best between the tested groups are chosen based on the training data. The SVM is then retrained on only those ten features. The same features are extracted from the test data and classification accuracy is tested. Classification accuracy for the expertise label can be increased to 89% through this procedure.

As the number of features was just reduced dramatically (from up to four million to just ten), it is possible to merge features with a different string encoding. Classification accuracy of the expertise label can be increased to 94% by merging the two best scoring feature sets as determined by the earlier grid search (x-Percentile (n=3, a=13) and y-Percentile (n=3, a=4)). A high horizontal and a low vertical resolution turned out to be the optimal combination for this task. Note that the used patterns of length 3 would not be captured by transition matrices nor by Markov chains.

In medical decision making and other safety-relevant applications we might also be interested in the confidence of the label assignment. SVMs provide so-called decision values that correspond to the distance of a feature towards the separating hyperplane. The higher this distance, the more likely is a correct classification of a sample. Data points that come to lie close to the hyperplane are harder to classify and at a higher risk of being misclassified.

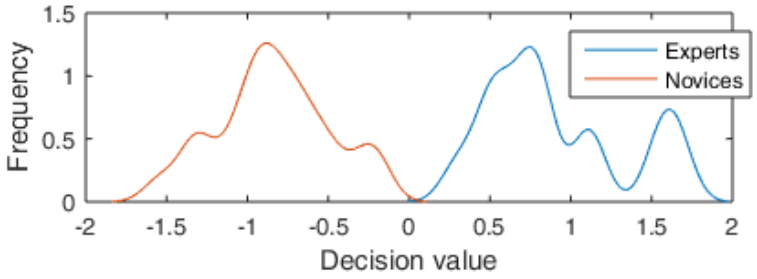


Figure 6.6: Distance distribution to the decision boundary for the expert and novice group

Figure 6.6 shows the decision values for an almost perfect classification: novice scanpaths are assigned decision values smaller than zero, expert scanpaths greater zero. We can observe two peaks in the experts' decision values, suggesting one group of clearly expert scanpaths (with decision values around 1.5) and another group that is closer to the decision boundary and, therefore, more similar to the novice scanpaths (centered around 0.5).
Feature selection and fusion of features with different encoding has a clear potential to improve classification accuracy. The level of expertise of surgeons viewing video material can be predicted from the scanpath at high accuracy.
The primary target of expert versus novice studies is usually to improve training. By analyzing and teaching the visual scanning patterns that distinguish experts from novices, know-how might be transferred faster than by traditional teaching methods. However, the real-time analysis enabled by the proposed algorithm extends the window of possibilities: based on the detection of partial patterns, gaze guidance could be employed to automatically pronounce areas that an expert would look at next. Furthermore, by distinguishing blurred from focused images based solely on gaze behavior, it might be possible to detect a non-optimal configuration of the operating microscope and to adjust it automatically.

6.3.4 Video gaming

Studying eye movements during video gaming is intriguing due to the dynamic, interactive environment. Few seconds after the game starts, every subject is confronted with a substantially different screen content. In each replication, the game will produce different states and screen contents depending on player interaction and some random events. Alike during real-world driving, some semantic entities change their position over time while others stay constant (examples would be the scenery outside of the car and the steering wheel or an in-game score display and opposing players). Yet, the virtual world is a controllable laboratory condition.
In this section the applicability of scanpath comparison algorithms to eye-tracking data derived during a racing game, namely Mario Kart for the Wii console, is studied. The racing game was used before to explore the influence of bottom-up saliency and top-down, task-driven attention [222]. It is demanding in terms of cognitive processing and visual load. Thus, various patterns of visual exploration can be expected. Furthermore, an objective performance measure is available, the average lap completion time.

In this setting the following factors' influence on the scanpath are examined:

- performance in terms of lap completion time.
- two tracks of different difficulties and attention requirements.

Methods

Twenty-one subjects were recorded using an EyeTribe tracker at 30 Hz. Head position was fixed with a chin and head rest. A video game (Mario Kart) was played on a Nintendo Wii and the video material transferred to a 24" monitor (Fujitsu Display B24T-7 LED with 1920×1080 px) and recorded in parallel via a frame grabber. Capturing the video digitally lead to a small delay between the Wii controller and a visible response. Nine-point calibration (as provided by the EyeTribe server 0.9.36 software) was performed for each of the two tracks. Median tracking rate was 91%. Subjects drove three laps per route, on two different tracks.
In addition, a questionnaire on regular video gaming and specifically the Mario Kart game was completed by the participants. The experiment was performed by Colleen Rothe as part of her Master thesis dealing with scanpath comparison algorithms [204].
Data quality is evaluated separately for both tracks as a separate calibration was performed. The distribution of measurement quality for Track 1 has its median at 92% valid samples, for Track 2 at 91%. For the further analysis all data were used, including two low quality recordings (< 50% valid points).

Results

57% (12 subjects) specified playing video games regularly, less than two hours a day. Nine subjects do not play video games regularly. 67% (14 subjects) were familiar with the video game Mario Kart, whereas seven subjects never had played Mario Kart before.
Lap completion time was averaged over all three laps for each track. We found that regular players (as determined by the questionnaire) completed the laps faster than non-regular players on average. However, subjects who rated themselves as non-regular players varied a lot in their performance: two subjects of the non-regular player group completed both tracks faster than any regular player. But the group of regular players shows more homogeneous, consistently good results with a lower spread.
Figure 6.7 shows the performance by players familiar with Mario Kart and players new to the specific game. Having played Mario Kart before clearly has a strong influence on task performance time. The huge impact of Mario Kart (but not general video gaming) experience suggests there are specific game elements that players attend to and interact with in order to get an advantage.
I decided to separate the groups by the objective lap completion time criterion to avoid any subjective self-estimation biases, even though this means the assignment of few experienced but badly performing regular players to the group of *slow drivers* and few novice players to the *fast drivers* group. Thereby, the effect of novices learning the important game elements quickly can be captured.

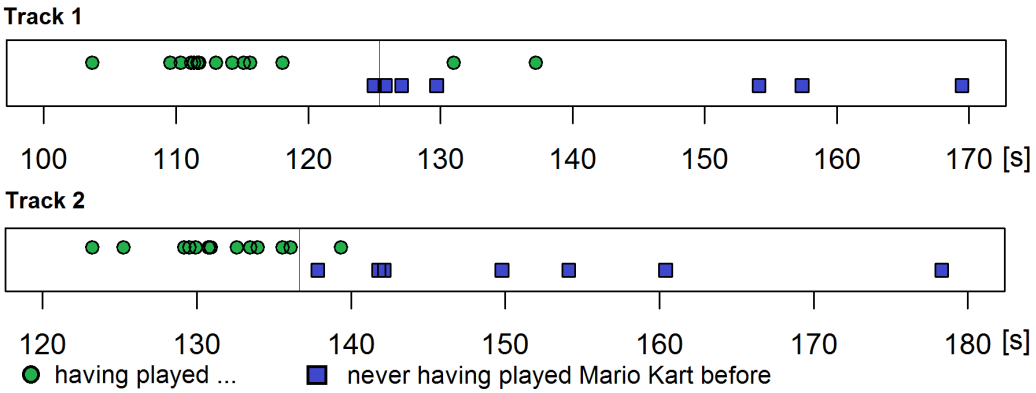


Figure 6.7: Average lap time of players familiar with the Mario Kart game specifically, and players not familiar with the game. Both tracks are shown, with the separation line between the groups of *fast* and *slow* drivers

Based on the percentage of experienced players according to the questionnaire (which also fits the distribution of lap completion times quite well), the fastest 60% (the 13 subjects with < 125 s for Track 1 and < 136 s for Track 2) are considered the group of fast drivers, whereas the remaining 40% (8 subjects) are the group of slow drivers.

Figure 6.8 demonstrates that several scanpath measures are able to detect differences in the scanning behavior of slow and fast drivers. The SubsMatch measure can only detect a difference for the more demanding Track 2. Track 1 (also the first level of the game) might not be difficult enough to challenge good players.

Figure 6.9 shows two MDS plots of scanpath distances. The FuncSim algorithm found differing scanpath distance distributions between and within the groups of slow and fast drivers. The plot for Track 1 visualizes that this fact does not necessarily imply a good separability between the groups. In fact, the most important influence could originate from a completely different factor, such as participants' age. But it is already a good indicator, as we can see for Track 2, where the group centroids obviously differ - but mostly in the second component of the multidimensional scaling. iComp and SubsMatch are unable to detect a significant difference between scanpath groups for Track 1, the track where the multidimensional scaling plot of FuncSim did not show a good group separability. But a difference for Track 2 can be found.

Figure 6.9(b) might inspire to use a classifier, such as a SVM, on the final distance matrix. New scanpaths could then be assigned to a group based on their similarity to other scanpaths. For Track 2 this would work quite well, but distinguishing players for Track 1 will turn out difficult.

By this example the strength of using the SVM directly on the subsequence frequencies instead of on the final distance measure becomes clear:

Figure 6.10 shows the confusion matrix for the classification of fast and slow drivers for both tracks. 86% of scanpaths were assigned the correct label. Table 6.5 reveals that there is no difference in the classification rates between the first and the second track. As for the

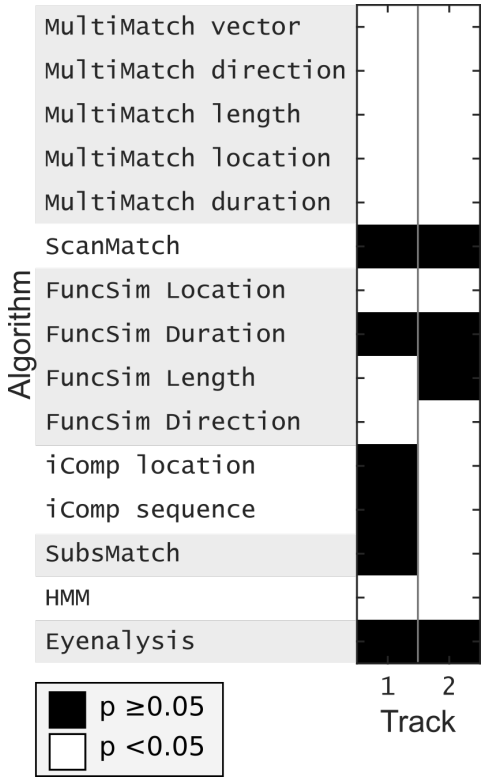


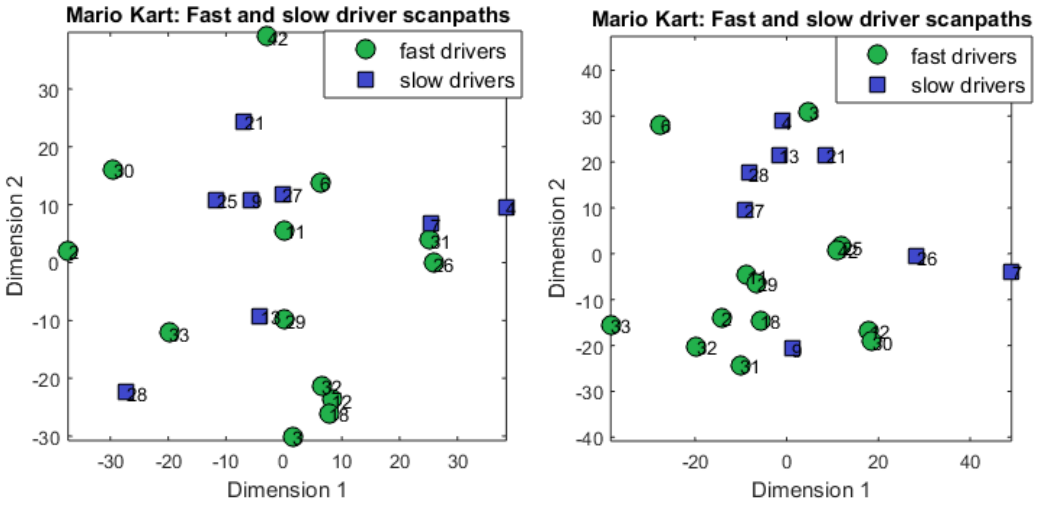
Figure 6.8: Wilcoxon rank sum test results for the difference in scanpaths between slow and fast drivers for two different tracks. Significant differences between scanpath distance distributions ($p < 0.05$) are shown in white

Conjunction search, Yarbus, and surgeon experiments, the distinction between scanpaths of different stimuli (here the first and second track) is easier than inferring task or expertise (here the lap time).

6.3.5 Driving fitness and compensatory eye movements in patients with visual field defects

In this section, the scanpath comparison approaches are applied to eye-tracking data from a simulated as well as an on-road driving experiment. They were originally recorded to assess the driving fitness of patients with visual field defects and to study their eye movements. The simulated drive was described in detail in Section 3.3. Some patients' driving safety was found to be indistinguishable from healthy subjects', whilst other patients did not react adequately to hazardous situations and were judged as unsafe drivers.

For the simulator experiment, a break-down of normal visual exploratory behavior was found in the group of unsafe drivers. Patients rated as fit-to-drive showed either no change in gaze behavior or a potentially compensatory viewing pattern with increased attention distributed towards peripheral areas. Consequently, one can expect that unsafe drivers are



(a) Track 1 (b) Track 2

Figure 6.9: Multidimensional scaling of the inter-scanpath distances as determined by the FuncSim (Direction) measure. For both tracks the distances within and between classes follow different distributions. But only for Track 2 a convincing separability between groups can be observed

Table 6.5: Classification for the Video game experiment (10-fold cross-validation). Statistically significant results ($p < 0.05$) of a binomial test (without correction for multiple testing) are reported in bold. Different string encoding patterns are displayed: circular encoding, horizontal and vertical axis percentile binning, horizontal and vertical axis regular binning

Research Question						**Baseline**	**Best**
Lap time Track 1	76.2	81.0	76.2	76.2	**85.7**	61.9	**85.7**
Lap time Track 2	76.2	81.0	71.4	76.2	85.7	61.9	**85.7**
Track 1 / Track 2	**92.9**	**71.4**	**88.1**	**73.8**	**76.2**	50.0	**92.9**

relatively easy to classify by their distinct scanpaths. Safe-to-drive patients might either differ from the control group in case some special compensatory patterns are employed, or be indistinguishable from the control group. In the latter case, the compensatory movements would consist of an upkeep of normal visual exploration.

Furthermore, I applied the algorithm to data of an on-road experiment described by Kasneci et al. [117]: three out of ten patients with homonymous hemianopia and five out of eight glaucoma patients failed this driving test. Control groups of eleven and seven participants, respectively, took part in the experiment and three of them failed the driving test. These driving sessions were conducted in real traffic.

Table 6.6 shows that it is generally possible to distinguish safe from unsafe drivers, but also patients and control subjects above chance level.

We can go more in-depth by looking at the confusion matrices in Figure 6.11. Especially for the simulated driving scenario, the distinction of unsafe drivers stands out (71%). Coherently

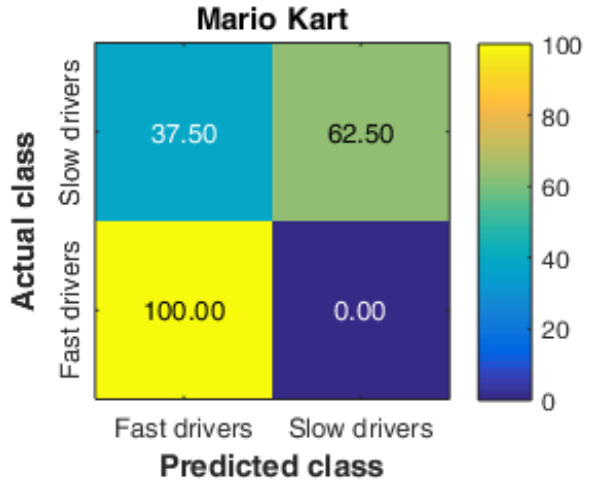


Figure 6.10: Confusion matrix of Track 1 in the Mario Kart video game. Fast and slow drivers are separated from each other with 85.7% accuracy using the circular string features

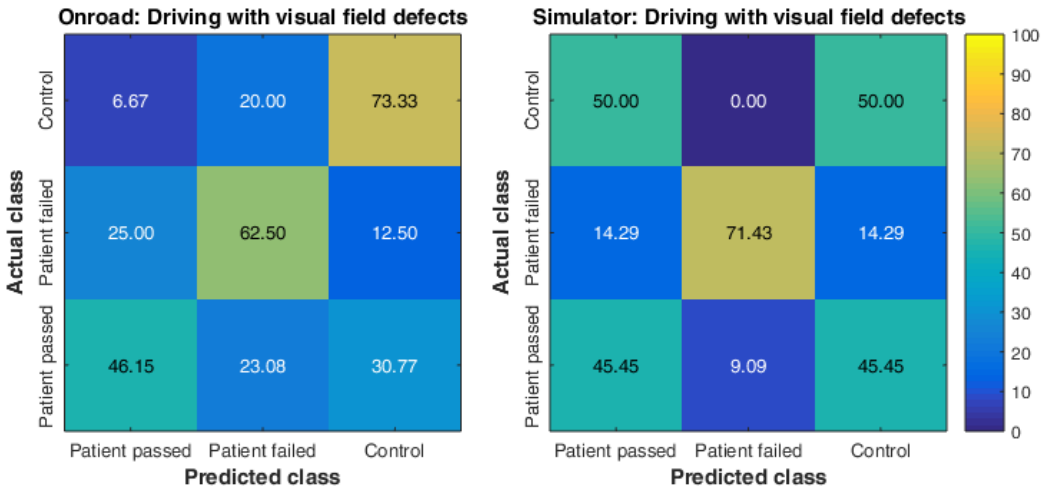
Table 6.6: Classification for the driving experiments (10-fold cross-validation). Statistically significant results ($p < 0.05$) of a binomial test (without correction for multiple testing) are reported in bold. Different string encoding patterns are displayed: circular encoding, horizontal and vertical axis percentile binning, horizontal and vertical axis regular binning

Research Question						**Baseline**	**Best**
SIM fitness	59.4	**84.4**	75.0	65.6	78.1	71.9	**84.4**
SIM Patient/Control	68.8	**75.0**	71.9	56.3	62.5	56.3	**75.0**
ROAD fitness	72.2	75.0	**77.8**	**77.8**	66.7	69.4	**77.8**
ROAD Patient/Control	75.0	**83.3**	69.4	69.4	61.1	58.3	**83.3**

with our previous finding, there seems to be a change in visual scanning for unsafe drivers. Patients fit to drive are indistinguishable from the control group (they get classified as either fit-to-drive or control group by 50% chance). Only very few mix-ups between safe and unsafe drivers do occur.

For the on-road driving experiment the effect is not that pronounced: control subjects, and the patient groups are mixed up more often, e.g., 25% of the unsafe drivers get classified as fit-to-drive patients.

A possible explanation is that the on-road driving experiment featured less standardized traffic conditions and a pass or fail in the driving test depended not exclusively on perception performance due to the visual field defect, but also on the encountered traffic situations. The considerable failure rate of 17% in the control group underpins this. The on-road driving test filters not only for failures due to the visual field defect, but many other factors (such as age, experience and mere chance). This results in several subjects, who might be rated safe drivers in a simulator experiment, to fail the on-road test. The relatively high failure rate in



(a) On-road driving in real traffic

(b) Driving simulator session

Figure 6.11: Confusion matrices of safe and unsafe drivers with HVFD and a control group for both the on-road and the simulated driving experiment

the control group underpins this assumption.
These subjects with intact visual scanning are included in the training data of the SVM and labeled as unsafe drivers. The SVM can therefore not properly learn the features of altered visual scanning behavior. That results in a less accurate distinction between the groups. The high level of control group members failing the driving test would even justify the relatively high misclassification rate of 20% of control subjects towards the not-fit-to-drive patient group, if other, perhaps age-rated, alterations of the scanpath are learned.
In both, on-road and simulator driving, there is a high level of agreement for the frequent classification of fit-to-drive patients to the control group. Gaze behavior of fit-to-drive patients is indistinguishable from the control group. Unsafe drivers exhibit a change in visual exploratory behavior (and thereby facilitate the distinction). We can therefore state that unsafe drivers alter their viewing behavior. But disproving a behavioral alteration in safe-to-drive patients group is not possible: the employed method might just not be sensitive to that kind of alteration.
As shown in this section, the string kernel method is able to adapt to a large variety of settings. This is possible through a combination of powerful scanpath features that can be compiled at very different detail levels (i.e., alphabet sizes and choices of n) together with a machine learning step that is effective in separating a few relevant from a multitude of irrelevant features.

7 Smart Ocular Motility Analysis

This section demonstrates an application of some of the methods introduced in this work to the measurement of ocular motility.

Lesions in one or more of the nerves that innervate the extraocular muscles result is a motility disorder where the eyes are unable to move in alignment. Figure 7.1 shows the extraocular muscles and nerves responsible for the movement of the eyeball. A failure of one eye to move with the other eye results in diplopia, double vision caused by the misalignment of the visual axes [95].

Neurogeneous ocular motility disorders have a huge impact on the quality of life [223]. Besides restrictions in visual perception there are also cosmetic issues. A stable motility disorder, i.e., one that does not change over time, can be treated by surgery. Essential for the success of such a surgery is the exact measurement and progress control of the motility disorder.

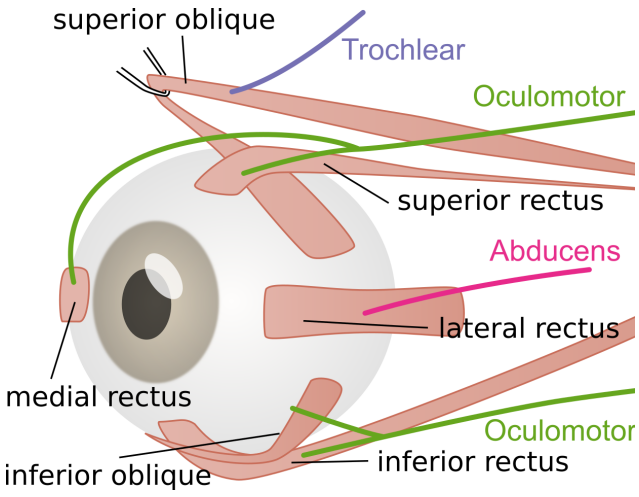


Figure 7.1: Extraocular muscles relevant for ocular rotation and their innervating nerves (note that actual interactions are more complex, e.g., nerves influence each other). A deficiency at one point results is a movement restriction in the direction in which the affected muscle pulls. Such a defect is only visible if an effect of the muscle is required in order to achieve the current eye rotation. Otherwise the remaining fully functional muscles can still adjust the eyeball properly. (Figure after https://vector.childrenhospital.org/2011/02/the-neurology-resident-that-could, 12.2016)

7.1 HARMS tangent screen

One of the most accurate non-invasive ways of measuring ocular motility is the HARMS tangent screen: the subject is placed in a distance of 2.5 m to a tangent screen. This screen contains a rectangular grid in 5° steps and an overlaid diagonal grid tilted by 45° . A fixation target (i.e., a light source) is placed at the center of the tangent screen.
The subject places a dark red glass in front of one eye. Through this glass only the fixation target is visible as a red dot. The fusion of images from both eyes is thereby disrupted. The non-occluded eye maintains fixation to the central fixation target and the occluded eye cannot correct its position through an image fusion feedback loop.
The experimenter moves the subject's head into nine different orientations. These orientations can be controlled by a forehead crosshair projector and its projection onto the tangent screen (the projector is calibrated to point *straight forward* in primary position). At each orientation, the subject points at the location where the *red dot* is perceived. The deviation of this location towards the center of the tangent screen is measured.
In a second step, the rotational component of ocular motility is measured. Therefore, the subject's head is tilted by 45° towards the left and right shoulder and a lever is adjusted. The rotational deviation is measured by the rotation of the lever.
The examination at the HARMS tangent screen is both time consuming and difficult to perform, as the experimenter and the subject have to attend to multiple tasks at the same time. It is difficult for the experimenter to check all the conditions required for a good measurement and to read the tangent screen scale in parallel. Furthermore, the equipment is expensive and requires a lot of space. Aim of this section is to demonstrate how the complex HARMS tangent screen examination can be simplified by the use of an eye-tracker, how measurements can be validated during the examination process and how the results can be further processed automatically.

7.2 Automation of the HARMS tangent screen

The central idea is the replacement of an indirect measurement method, i.e., the step where the subject points at the red dot, by a direct measurement of ocular alignment via an eye tracker.
SMI (Teltow, Germany) developed a device called 3D VOG [224]. It consists of an eye tracker and two displays, one per eye. Eye movements are generated by the presentation of fixation targets at different locations on the screen. Thereby, an examination similar to that of the HARMS wall can be performed and ocular motility can be measured directly. But eliciting eye movements by such a stimulus presentation is not necessarily identical to inducing them via head movements, as the vestibular system is not involved. Furthermore, the device weights 440-750 g (depending on the version) and is able to measure only within the central 20° (the HARMS screen usually spans 30°). Measuring through eyeglasses is not possible.
The approach proposed here is based on an adaptation of the HARMS tangent screen that retains the induction of eye movements through a change in head orientation but records

validated, direct ocular measurements automatically. The examination work-flow in shown in Figure 7.2. In Section 3.1.3, on the accuracy of gaze mapping approaches through eyeglasses, only the polynomial calibration method was found suited to compensate for the optical effects of eyeglasses. As we are measuring in 2.5 m distance to the subject, wearing optical correction is non-optional for some subjects. This method requires a calibration step in which we need to determine point correspondences between the eye image and the scene (in this case the gaze position at the HARMS tangent screen). For this reason, the examination work-flow requires always measuring both eyes. The measurement of the fellow eye is used as a calibration step for the non-covered eye.

The created recording software includes a calibration routine that assumes a constant fixation to the fixation target and uses samples recorded while the head is continuously moved towards the measurement points. Thereby, a large amount of densely sampled calibration points is generated.

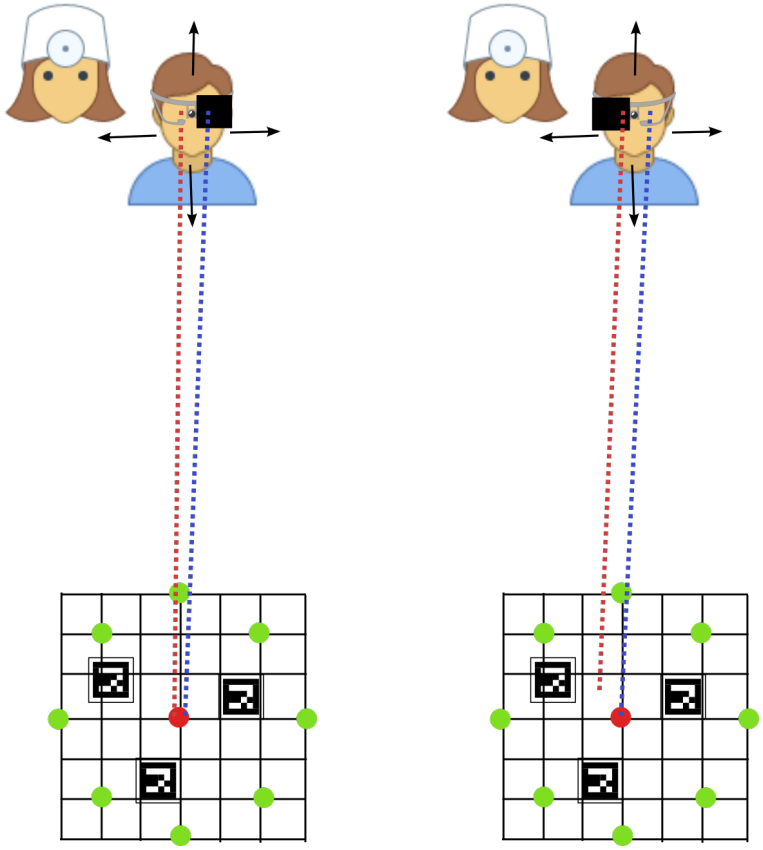


Figure 7.2: Measurement process. (a) left eye is covered, right eye fixating the target. By changing the head orientation towards the measurement points (green circles), a gaze shift towards the opposite direction is induced. The right eye is calibrated, the left eye measured. (b) repetition with the fellow eye covered. The measurement is the gaze deviation from the fixation target

The Else [25] algorithm for pupil detection was included in a recording tool that provided the original code base for the EyeRec [26] tool. Furthermore, the Dikablis professional (Ergoneers, Manching/Germany) eye tracker was altered in order to meet the requirements for the examination: foldable covers for both eyes were added and custom lenses placed in front of the eye cameras in order to achieve an image that is focused on the iris structure (see Figure 7.3).

Figure 7.3: Measurement device used for this study. An adaptation of the binocular Dikablis professional with custom lenses to focus the eye cameras on the iris structure and foldable covers (with one cover folded downwards in the right image) for each eye

Measurement of head position and orientation

Head position and orientation relative to a central fixation target need to be measured, as the induced ocular movement of the non-covered eye can be assumed to be a counter-rotation to the head orientation. Furthermore, the central position of the subject's head with respect to the fixation target and the examination distance of 2.5 m need to be validated.
To achieve this, the position and orientation of the eye tracker's scene camera can be computed from the produced image of computer vision markers. An example of such a marker is shown in Figure 7.2. Markers were detected by the Aruco library [225]. Theoretically, an exact location and orientation can be determined from the four corners of only one marker, but in practice the estimation gets more accurate and robust the more points are available. Therefore, the detected markers were post-processed to determine the exact sub-pixel position of all corners within the marker and to merge information from all currently detected markers in the scene camera's image. The OpenCV library [226] can then be used to estimate the exact camera position.

Iris rotation analysis

The inferior and superior oblique muscles are primarily (but not exclusively) responsible for torsional eye movements. These movements are relatively difficult to detect via an eye tracker, as they do not result in a change in the location of the pupil. Instead, much more subtle changes in the rotation of the iris structure need to be detected. This process

is difficult for a variety of reasons: the appearance of the iris changes with pupil diameter. The ring muscle responsible for adjusting the pupil also alters the iris structure; parts of the iris are almost always covered by the eyelid or eyelashes; refraction at the cornea may result in a different appearance of the iris structure to the camera that depends on gaze orientation. Further, reflections of the IR illumination (such as the glint) may occlude parts of the iris.

In order to tackle this, the first step is to extract an iris structure patch from the image. Therefore, samples on an ellipse that follows the elliptic shape of the pupil are extracted over 360° in 1° steps. The sampling process is shown in Figure 7.4 as black lines, originating shortly outwards of the iris-pupil boundary and stopping before the iris-sclera boundary. The number of ellipses with gradually decreasing radii used for sampling and the degree steps of the samples (here 1°) determines how accurate the iris structure is represented in the patch (capped by the quality of the original image). The result is a rectangular iris patch, with each column of pixels (from the top to the bottom) representing data of one ellipse. Each pixel row (left to right) corresponds to a rotation by 1°. The vertical offset between two such iris patches relates directly to the rotation angle of the iris. Therefore, all that is left is finding the optimal vertical offset between two such patches.

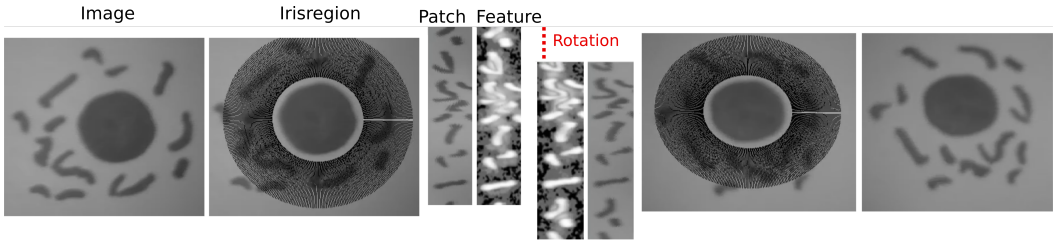


Figure 7.4: Extraction and circular sampling of the iris region of an artificial eye to create the rectangular iris patches. Feature maps are calculated from the patches by contrast enhancement. The vertical offset between feature maps corresponds to the rotation angle of the iris structure

But before the patches can be compared to each other, artifacts caused by the eyelid and eyelashes have to be eliminated. Therefore, a threshold approach is applied. As the skin appears very bright in the IR illumination, it can be distinguished from the iris relatively easily. To make sure that continuous regions are masked, the threshold image is dilated and eroded, filling in small gaps in the mask area.

It is also necessary to equalize illumination over the whole iris region. As the Dikablis device has only one IR-LED, illumination differs significantly for different gaze orientations and areas of the iris. For this purpose, a median blurred version of the iris patch is calculated. The calculation of the median results in an image that contains only a (low-frequency) background intensity. By subtracting the background from the iris patch, local intensity changes with a clear edge, i.e., the iris structure features, remain. Histogram equalization is performed on the image parts that are not masked as either eyelid or eyelashes in order to visualize the feature patterns over the whole color spectrum, as shown in Figure 7.4. The equalization step is not necessary for a mathematical structure comparison, but aids the visibility of the structure for visualization purposes.

Result visualization and classification

Twelve subjects, one with confirmed NIII, two with NIV and nine with NVI palsy, were measured both with the new and the standard HARMS screen methods. As the exact same measurement locations were used, the ocular deviations can directly be compared to each other. Figure 7.5 shows the average measurement error over the whole system, including head- and eye-tracking. For practical usage a measurement error below 3° would be required. For most measurement positions this high accuracy demand cannot be met with the current device.
One major reason is the flexible design of the head-mounted tracker combined with only one IR light. As most of the time no glint is visible in the eye image, a compensation of any movement of the device on the subject's head cannot be compensated for. The examination procedure includes several steps where the foldable covers need to be changed and the head orientation is guided by the experimenter. These steps are often associated with a small displacement of the device.

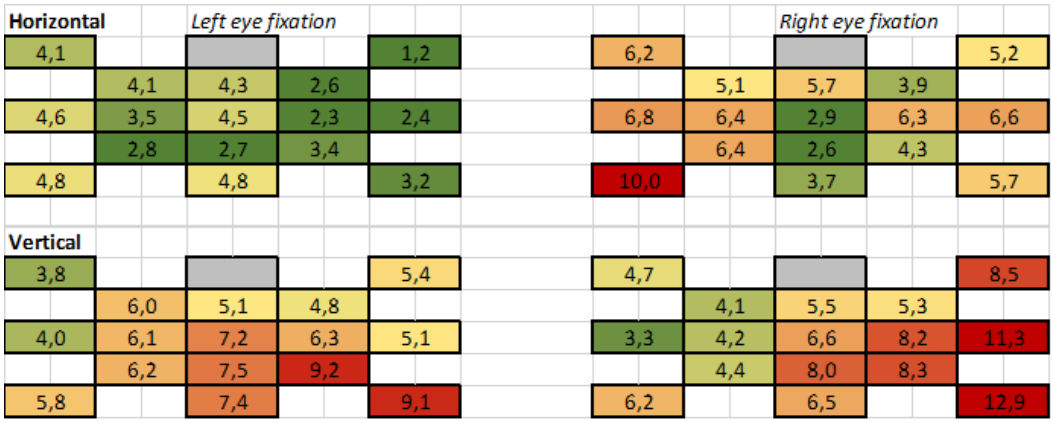


Figure 7.5: Average measurement error for each of the measurement locations, when compared to the HARMS tangent screen results. Horizontal and vertical measurement error are reported separately. The colors encode measurement accuracy from green corresponding to good to red, corresponding to worse accuracy

Several automated classification approaches for strabismus do exist and reach high accuracies, for example for the Hirschberg test [227] or the prism cover-test a neural network can be trained [228, 229]. For the HARMS screen the classification problem consists of the three classes oculomotor nerve (NIII), trochlear nerve (NIV) and abducens nerve (NVI) palsy. We can further distinguish between the left and the right eye, as the motility disorder cannot simply be mirrored. However, one has to keep in mind that we always have knowledge about the measured eye available. Therefore, a misclassification between an abducens palsy left eye and an abducens palsy right eye is not a practically relevant false classification but can easily be resolved.
The results of a radial basis function (RBF) kernel SVM trained on data retrieved from standard HARMS screen examinations is shown in the confusion matrix in Figure 7.6. The

Table 7.1: Data used for the training and testing of the classifier

Palsy	Eye	n
Oculomotor N.III	left	16
Oculomotor N.III	right	18
Trochlearis N.IV	left	12
Trochlearis N.IV	right	11
Abducens N.VI	left	36
Abducens N.VI	right	36

data used for training and evaluation is shown in Table 7.1. Classification accuracy was determined by leave-one-out training and classification [230]. When adding up left and right eye classifications, the three classes can be classified with over 90% accuracy.

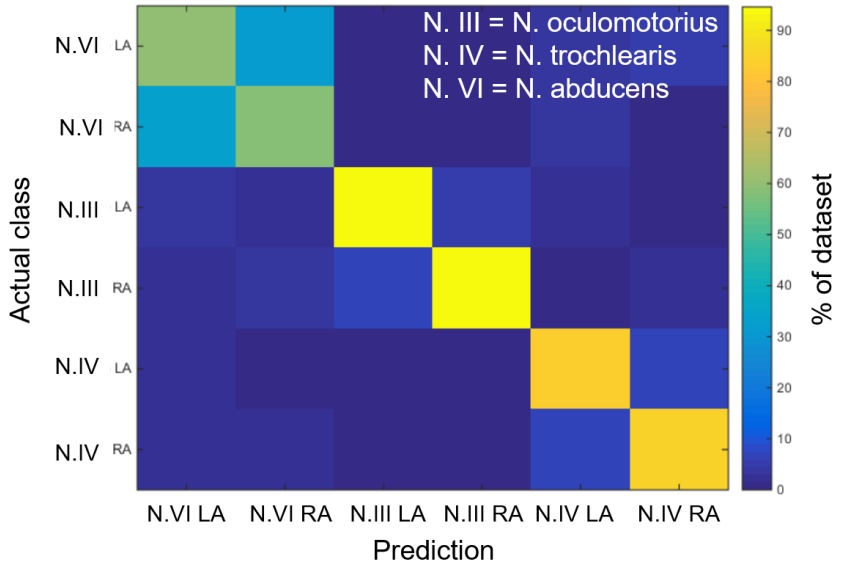


Figure 7.6: Confusion matrix of the ocular motility disorder classes

Although the high accuracy demands of the examination could not be met for this prototype, the shortcoming could be compensated by a more accurate eye-tracking system that allows for the compensation of displacement during the measurement. Therefore, more IR LEDs would be required so that the glint can reliably be used as a reference point. Furthermore, the foldable eye covers should be replaced either by electronic shutters or specially crafted plastic covers that could be applied more easily and without tampering with the device as much as with the paper covers. Ideally, a dark red glass would be used so that the standard HARMS wall examination could be applied with the same device.

From an ophthalmologist's point of view one of the major advantages of the new approach is the automated measurement and validation of the head position and orientation. Moving the subject's head into a specific orientation and keeping it axis aligned with regard to the

other orientations (i.e., not tilting the head) is difficult and can easily result in a false vertical deviation being measured. Theoretically this module could be used independent of the eye-tracker and could also be used in conjunction with a detection of the pointer directed towards the HARMS wall to capture data automatically.

8 Conclusion

Understanding the quality of an eye-tracking recording, how it is influenced by the device used and the choice of algorithms for recording and analysis, is essential to derive reliable findings from the data. This thesis presents a novel method for the identification of smooth pursuit movements as well as checking the quality of a fixation identification filter via PCA and ellipse fitting. The method performs especially well in a dynamic driving scenario with mostly linear motion components. Further, features derived from a Mixture of Gaussian model proposed for fixation and saccade identification were shown to contain subject-specific information in a biometry challenge.
It is essential to have information about measurement characteristics and the influence of individual error sources to distinguish between signal and noise. Rendering of artificial eyeglasses generates ground truth that helps to improve gaze mapping and pupil detection algorithms, but also helps to better understand the shortcomings of current methods. It is possible to exclude or systematically vary error sources (such as dust and reflections) to measure an accurate, objective performance without the need for extensive manual annotation.

Besides mere gaze location, eye trackers record a variety of other signals (such as the pupil diameter) and some more complex measures that can be derived from the signal (such as the vergence angle). These parameters can be used for the assessment of directed attention and vigilance. Example applications to perimetry and an image viewing experiment were demonstrated. Biosensors can further complement the eye-tracking signal. This study showed how sensor fusion can be utilized to disambiguate hazard perception in a driving scenario and to provide a deeper understanding of visual exploration, perception, and adequate behavioral reaction.

Novel visualization techniques reveal how saccade trajectories aggregate into gaze trails. These trails are characteristic to the stimulus material and correspond to composition principles in the fine arts. Instead of a fully automated approach to ROI labeling, which is hardly within reach given the current state of computer vision algorithms, a semi-automated approach was demonstrated. ROI labeling can be performed with a speed-up factor of $2.8\times$ when compared to manual annotation.

I introduced a new technique for the comparison of scanpaths, which is based on the occurrence frequency of small subsequences. I also demonstrated a way to encode these subsequences by percentile binning in order to compensate for typical measurement errors, and to circumvent the length normalization problem. These n-gram features are sensitive to a variety of different factors and contain information that allows to detect differences

between groups of scanpaths.
Machine learning was used to distinguish relevant features within an overall high noise level. By learning these features from a training set, the method is able to adapt to a large variety of different settings and detail levels. Discriminating features between groups of scanpaths can be identified and individual scanpaths classified on the basis of the occurrence of these features. The generalization of the method over a vast field of eye-tracking applications ranging from laboratory conditions to real-world scenarios was demonstrated.
While the proposed scanpath comparison method is already useful for practical applications, such as the classification of a driver's secondary task, there are some issues that could be tackled in order to improve performance. We found that the distinction of different stimulus material was generally possible at a high accuracy, while the prediction of the observer's task and expertise is a much harder problem, but still possible significantly above chance level. This might be due to the amount of change in the scanning patterns, but also to characteristics of the patterns that the string kernel method is sensitive to.
Most importantly, the encoding of scanpath features is very coarse. It might be possible to perform an unsupervised clustering of the scanpath snippets in order to determine a better encoding. Therefore, two-dimensional positions and timing relative to the first fixation in the snippet or angles between and lengths of saccades could be used and aggregated to a higher level (e.g., spiral- or grid-shaped scanning patterns). Furthermore, the fusion of features with different encoding after the feature selection step remains mostly unexplored and has the potential to boost classification performance significantly.
While several improvements are possible, the methods introduced in this thesis are the first that are applicable and adaptable to a general scanpath comparison problem and have already led to a practical and useful metric for distinguishing eye movement sequences.

Bibliography

[1] Raichle, M. E. Two views of brain function. *Trends in Cognitive Sciences* **14**, 180–190 (2010).

[2] Dagnelie, G. (ed.) *Visual prosthetics: physiology, bioengineering, rehabilitation* (Springer US, Boston, MA, USA, 2011).

[3] Jacobs, R. Visual resolution and contour interaction in the fovea and periphery. *Vision Research* **19**, 1187–1195 (1979).

[4] Deubel, H. Attention and Action. In Nobre, K. & Kastner, S. (eds.) *The Oxford handbook of attention*, chap. 30, 865–869 (Oxford University Press, Oxford, UK, 2013).

[5] Tatler, B. W., Kirtley, C., Macdonald, R. G., Mitchell, K. M. & Savage, S. W. The active eye: Perspectives on eye movement research. In *Current Trends in Eye Tracking Research*, 3–16 (Springer, Berlin, D, 2014).

[6] Noton, D. & Stark, L. Eye movements and visual perception. *Scientific American* **224**, 35–43 (1971).

[7] Buswell, G. T. How people look at pictures: a study of the psychology and perception in art. *University of Chicago Press, Chicago, USA* (1935).

[8] Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M. & Velichkovsky, B. M. Yarbus, eye movements, and vision. *i-Perception* **1**, 7–27 (2010).

[9] DeAngelus, M. & Pelz, J. B. Top-down control of eye movements: Yarbus revisited. *Visual Cognition* **17**, 790–811 (2009).

[10] Greene, M. R., Liu, T. & Wolfe, J. M. Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research* **62**, 1–8 (2012).

[11] Griffin, J. R. & Borsting, E. J. *Binocular anomalies : Theory, Testing & Therapy* (Optometric Extension Program Foundation, Santa Ana, CA, USA, 2010), 5th edn.

[12] Kübler, T. C. *et al.* Driving with glaucoma: Task performance and gaze movements. *Optometry & Vision Science* **92**, 1037–1046 (2015).

[13] Kübler, T. C. *et al.* Driving with homonymous visual field defects: Driving performance and compensatory gaze movements. *Journal of Eye Movement Research* **8**, 1–11 (2015).

[14] Schag, K. *et al.* Impulsivity in binge eating disorder: Food cues elicit increased reward responses and disinhibition. *PloS one* **8**, e76542 (2013).

[15] Kübler, T. C., Eivazi, S. & Kasneci, E. Automated visual scanpath analysis reveals the expertise level of micro-neurosurgeons. In *International Conference on Medical Image Computing and Computer Assisted Intervention: Workshop on Interventional Microscopy* (2015).

[16] Pernice, K. & Nielsen, J. How to conduct eyetracking studies. Tech. Rep., Nielsen Norman Group, Fremont, CA, USA (2009).

[17] Leonards, U. *et al.* Mediaeval artists: Masters in directing the observers' gaze. *Current Biology* **17**, R8–R9 (2007).

[18] Kübler, T. C. *et al.* Stress-indicators and exploratory gaze for the analysis of hazard perception in patients with visual field loss. *Transportation Research Part F: Traffic Psychology and Behaviour* **24**, 231–243 (2014).

[19] Braunagel, C. *et al.* Exploiting the potential of eye movements analysis in the driving context. In *15. Internationales Stuttgarter Symposium Automobil- und Motorentechnik*, 1093–1105 (Springer, Berlin, D, 2015).

[20] Kasneci, E., Kasneci, G., Kübler, T. C. & Rosenstiel, W. The applicability of probabilistic methods to the online recognition of fixations and saccades in dynamic scenes. In *Proceedings of the 2014 Symposium on Eye Tracking Research and Applications*, 323–326 (ACM, New York, NY, USA, 2014).

[21] Tafaj, E., Kübler, T. C., Kasneci, G., Rosenstiel, W. & Bogdan, M. Online classification of eye tracking data for automated analysis of traffic hazard perception. In *Artificial Neural Networks and Machine Learning*, 442–450 (Springer, Berlin, D, 2013).

[22] Kasneci, E., Kasneci, G., Kübler, T. C. & Rosenstiel, W. Online recognition of fixations, saccades, and smooth pursuits for automated analysis of traffic hazard perception. In Koprinkova-Hristova, P., Mladenov, V. & Kasabov, N. K. (eds.) *Artificial Neural Networks*, vol. 4 of *Springer Series in Bio-/Neuroinformatics*, 411–434 (Springer International Publishing, Berlin, D, 2015).

[23] Santini, T., Fuhl, W., Kübler, T. C. & Kasneci, E. Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the 2016 Symposium on Eye Tracking Research and Applications* (ACM, New York, NY, USA, 2016).

[24] Kübler, T. C., Rittig, T., Kasneci, E., Ungewiss, J. & Krauss, C. Rendering refraction and reflection of eyeglasses for synthetic eye tracker images. In *Proceedings of the 2016 Symposium on Eye Tracking Research and Applications*, 143–146 (ACM, New York, NY, USA, 2016).

[25] Fuhl, W., Santini, T. C., Kübler, T. C. & Kasneci, E. Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the 2016 Symposium on Eye Tracking Research and Applications* (ACM, New York, NY, USA, 2016).

[26] Santini, T., Fuhl, W., Kübler, T. C. & Kasneci, E. Eyerec: An open-source data acquisition software for head-mounted eye-tracking. In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 3, 386–391 (2016).

[27] Fuhl, W., Kübler, T. C., Sippel, K., Rosenstiel, W. & Kasneci, E. Excuse: Robust pupil detection in real-world scenarios. In *Computer Analysis of Images and Patterns*, 39–51 (Springer, Berlin, D, 2015).

[28] Kübler, T. C. *et al.* Towards automated comparison of eye-tracking recordings in dynamic scenes. In *European Workshop on Visual Information Processing*, 1–6 (IEEE, 2014).

[29] Kübler, T. C. & Kasneci, E. Automated comparison of scanpaths in dynamic scenes. In Pfeiffer, T. & Essig, K. (eds.) *Proceedings of the 2nd International Workshop on Solutions for Automatic Gaze Data Analysis 2015* (2015).

[30] Fuhl, W., Kübler, T. C., Sippel, K., Rosenstiel, W. & Kasneci, E. Arbitrarily shaped areas of interest based on gaze density gradient. In *European Conference on Eye Movements* (2015).

[31] Sippel, K. *et al.* Eyetrace2014 - eyetracking data analysis tool. In *Proceedings of the International Conference on Health Informatics*, 212–219 (2015).

[32] Kübler, T. C. *et al.* Analysis of eye movements with Eyetrace. In *International Joint Conference on Biomedical Engineering Systems and Technologies*, 458–471 (Springer, Berlin, D, 2015).

[33] Kübler, T. C., Fuhl, W., Rosenberg, R., Rosenstiel, W. & Kasneci, E. Novel methods for analysis and visualization of saccade trajectories. In Hua, G. & Jégou, H. (eds.) *Computer Vision – ECCV 2016 Workshops, Proceedings, Part I*, 783–797 (Springer International Publishing, Cham, 2016).

[34] Kübler, T. C., Kasneci, E. & Rosenstiel, W. SubsMatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the 2014 Symposium on Eye Tracking Research and Applications*, 319–322 (ACM, New York, NY, USA, 2014).

[35] Kübler, T. C., Rothe, C., Schiefer, U., Rosenstiel, W. & Kasneci, E. Subsmatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behavior Research Methods* 1–17 (2016).

[36] Holmqvist, K. *et al. Eye tracking: A comprehensive guide to methods and measures* (Oxford University Press, Oxford, UK, 2011).

[37] Ooms, K., Dupont, L., Lapon, L. & Popelka, S. Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental set-ups. *Journal of Eye Movement Research* **8**, 1–24 (2015).

[38] Tafaj, E. *et al.* Vishnoo – an open-source software for vision research. In *International Symposium on Computer-Based Medical Systems*, 1–6 (IEEE, New York, NY, USA, 2011).

[39] Li, D., Winfield, D. & Parkhurst, D. J. Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *Computer Vision and Pattern Recognition - Workshops*, 79–79 (IEEE, Washington, DC, USA, 2005).

[40] Rayner, K., Li, X., Williams, C. C., Cave, K. R. & Well, A. D. Eye movements during information processing tasks: Individual differences and cultural effects. *Vision Research* **47**, 2714–2726 (2007).

[41] Martinez-Conde, S., Macknik, S. L. & Hubel, D. H. The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience* **5**, 229–240 (2004).

[42] Bremmer, F., Kubischik, M., Hoffmann, K.-P. & Krekelberg, B. Neural dynamics of saccadic suppression. *Journal of Neuroscience* **29**, 12374–12383 (2009).

[43] Ross, J., Morrone, M. C., Goldberg, M. E. & Burr, D. C. Changes in visual perception at the time of saccades. *Trends in Neurosciences* **24**, 113–121 (2001).

[44] Kapoula, Z., Robinson, D. & Hain, T. Motion of the eye immediately after a saccade. *Experimental Brain Research* **61**, 386–394 (1986).

[45] Bahill, A. T., Clark, M. R. & Stark, L. Glissades—eye movements generated by mismatched components of the saccadic motoneuronal control signal. *Mathematical Biosciences* **26**, 303–318 (1975).

[46] Salvucci, D. D. & Goldberg, J. H. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research and Applications*, 71–78 (ACM, New York, NY, USA, 2000).

[47] Komogortsev, O. V. & Karpov, A. Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior Research Methods* **45**, 203–215 (2012).

[48] Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P. & Itti, L. Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision* **9**, 19–19 (2009).

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

[50] Vidal, M., Bulling, A. & Gellersen, H. Detection of smooth pursuits using eye movement shape features. In *Proceedings of the 2012 Symposium on Eye Tracking Research and Applications*, 177–180 (ACM, New York, NY, USA, 2012).

[51] Caldara, R. & Miellet, S. iMap: a novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods* **43**, 864–878 (2011).

[52] Privitera, C. M. & Stark, L. W. Evaluating image processing algorithms that predict regions of interest. *Pattern Recognition Letters* **19**, 1037–1043 (1998).

[53] Santella, A. & DeCarlo, D. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the 2004 Symposium on Eye tracking Research and Applications*, 27–34 (ACM, New York, NY, USA, 2004).

[54] Privitera, C. M. & Stark, L. W. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 970–982 (2000).

[55] Vansteenkiste, P., Cardon, G. & Lenoir, M. Dealing with head-mounted eye-tracking data: comparison of a frame-by-frame and a fixation-based analysis. In *Proceedings of the 2013 Conference on Eye Tracking South Africa*, 55–57 (ACM, New York, NY, USA, 2013).

[56] Land, M., Mennie, N. & Rusted, J. The roles of vision and eye movements in the control of activities of daily living. *Perception* **28**, 1311–1328 (1999).

[57] Olsen Anneli, M. R. Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In *Proceedings of the 2012 Symposium on Eye Tracking Research and Applications*, 212, 317–320 (ACM, New York, NY, USA, 2012).

[58] Andersson, R., Nyström, M. & Holmqvist, K. Sampling frequency and eye-tracking measures: how speed affects durations, latencies, and more. *Journal of Eye Movement Research* **3**, 1–12 (2010).

[59] Berger, C., Winkels, M., Lischke, A. & Höppner, J. GazeAlyze: a MATLAB toolbox for the analysis of eye movement data. *Behavior Research Methods* **44**, 404–419 (2012).

[60] Komogortsev, O. V., Karpov, A., Price, L. R. & Aragon, C. Biometric authentication via oculomotor plant characteristics. In *2012 5th IAPR International Conference on Biometrics* (IEEE, 2012).

[61] Komogortsev, O. V. & Rigas, I. Bioeye 2015: Competition on biometrics via eye movements. In *International Conference on Biometrics Theory, Applications and Systems*, 1–8 (IEEE, New York, NY, USA, 2015).

[62] Rigas, I. & Komogortsev, O. V. Current research in eye movement biometrics: An analysis based on BioEye 2015 competition. *Image and Vision Computing* (2016).

[63] Holland, C. D. & Komogortsev, O. V. Complex eye movement pattern biometrics: Analyzing fixations and saccades. In *International Conference on Biometrics*, 1–8 (IEEE, New York, NY, USA, 2013).

[64] Braunagel, C., Geisler, D., Stolzmann, W., Rosenstiel, W. & Kasneci, E. On the necessity of adaptive eye movement classification in conditionally automated driving scenarios. In *Proceedings of the 2016 Symposium on Eye Tracking Research and Applications*, 19–26 (ACM, New York, NY, USA, 2016).

[65] Villanueva, A. & Cabeza, R. Models for gaze tracking systems. *Journal on Image and Video Processing* **2007**, 4 (2007).

[66] Świrski, L. & Dodgson, N. A. A fully-automatic, temporal approach to single camera, glint-free 3D eye model fitting. In *Proceedings of the European Conference on Eye Movements* (Lund, SE, 2013).

[67] Heyman, T., Spruyt, V. & Ledda, A. 3D Face tracking and gaze estimation using a monocular camera. In *Proceedings of the 2nd International Conference on Positioning and Context-Awareness*, 23–28 (Brussels, BE, 2011).

[68] Morgan, I. & Rose, K. How genetic is school myopia? *Progress in Retinal and Eye Research* **24**, 1–38 (2005).

[69] Schaeffel, F. Myopia: the importance of seeing fine detail. *Current Biology* **16**, R257–R259 (2006).

[70] Wang, J.-G., Sung, E. & Venkateswarlu, R. Estimating the eye gaze from one eye. *Computer Vision and Image Understanding* **98**, 83–103 (2005).

[71] Świrski, L. & Dodgson, N. Rendering synthetic ground truth images for eye tracker evaluation. In *Proceedings of the 2014 Symposium on Eye Tracking Research and Applications*, 219–222 (ACM, New York, NY, USA, 2014).

[72] Böhme, M., Dorr, M., Graw, M., Martinetz, T. & Barth, E. A software framework for simulating eye trackers. In *Proceedings of the 2008 symposium on Eye tracking research and applications*, 251–258 (ACM, New York, NY, USA, 2008).

[73] Diepes, H. & Blendowske, R. *Optik und Technik der Brille* (DOZ Verlag, Optische Fachveröff. GmbH, Heidelberg, D, 2002).

[74] Zeiss. Wissenswertes rund ums Brillenglas. Tech. Rep., Zeiss (2002). URL http://www.vision.zeiss.de/4125680F0055C122/EmbedTitelIntern/Kap_A/$File/KAP_A.pdf. As of 11.2016.

[75] Wood, E., Baltrušaitis, T., Morency, L.-P., Robinson, P. & Bulling, A. Learning an appearance-based gaze estimator from one million synthesised images. In *Proceedings of the 2016 Symposium on Eye Tracking Research and Applications*, 131–138 (ACM, New York, NY, USA, 2016).

[76] Hospach, D., Mueller, S., Rosenstiel, W. & Bringmann, O. Simulation of falling rain for robustness testing of video-based surround sensing systems. In *2016 Design, Automation & Test in Europe Conference & Exhibition*, 233–236 (IEEE, 2016).

[77] Ungewiss, J. *Parameters for Vigilance, Attention and Cognitive Workload within Eye Tracking Recordings*. Master thesis, University of Applied Sciences Aalen (2015).

[78] Henson, D. B. & Emuh, T. Monitoring vigilance during perimetry by using pupillography. *Investigative Ophthalmology & Visual Science* **51**, 3540–3543 (2010).

[79] Wilhelm, B. *et al.* Daytime variations in central nervous system activation measured by a pupillographic sleepiness test. *Journal of Sleep Research* **10**, 1–7 (2001).

[80] Beatty, J. & Lucero-Wagoner, B. The pupillary system. In Cacioppo, J. T., Tassinary, L. G. & Berntson, G. (eds.) *Handbook of psychophysiology*, vol. 2, 142–162 (Cambridge University Press, Cambridge, UK, 2000).

[81] James, W. *The principles of psychology*, vol. 1 (Henry Holt and Company, New York, NY, USA, 1890).

[82] Berntson, G. & Cacioppo, J. *Handbook of Neuroscience for the Behavioral Sciences* (Wiley, Hoboken, NJ, USA, 2009).

[83] Pomplun, M., Ritter, H. & Velichkovsky, B. Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception* **25**, 931–948 (1996).

[84] Oken, B. S., Salinsky, M. C. & Elsas, S. Vigilance, alertness, or sustained attention: physiological basis and measurement. *Clinical Neurophysiology* **117**, 1885–1901 (2006).

[85] Nishiyama, J., Tanida, K., Kusumi, M. & Hirata, Y. The pupil as a possible premonitor of drowsiness. In *International Conference of the IEEE Engineering in Medicine and Biology Society*, 1586–1589 (IEEE, New York, NY, USA, 2007).

[86] Sweller, J. Cognitive load during problem solving: Effects on learning. *Cognitive Science* **12**, 257–285 (1988).

[87] Wickens, C. D., Hollands, J. G., Banbury, S. & Parasuraman, R. *Engineering psychology & human performance* (Psychology Press, Oxford, UK, 2012), 4th edn.

[88] Wooding, D. S. Eye movements of large populations: II. Deriving regions of interest, coverage, and similarity using fixation maps. *Behavior Research Methods, Instruments, & Computers* **34**, 518–528 (2002).

[89] Barbato, G. *et al.* Diurnal variation in spontaneous eye-blink rate. *Psychiatry Research* **93**, 145–151 (2000).

[90] Stern, J. A., Boyer, D. & Schroeder, D. Blink rate: a possible measure of fatigue. *Human Factors* **36**, 285–297 (1994).

[91] Schleicher, R., Galley, N., Briest, S. & Galley, L. Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics* **51**, 982–1010 (2008).

[92] Tanaka, Y. & Yamaoka, K. Blink activity and task difficulty. *Perceptual and Motor Skills* **77**, 55–66 (1993).

[93] Lowenstein, O., Feinberg, R. & Loewenfeld, I. E. Pupillary movements during acute and chronic fatigue a new test for the objective evaluation of tiredness. *Investigative Ophthalmology & Visual Science* **2**, 138–157 (1963).

[94] Russo, M. *et al.* Oculomotor impairment during chronic partial sleep deprivation. *Clinical Neurophysiology* **114**, 723–736 (2003).

[95] Kaufmann, H. & Steffen, H. *Strabismus* (Georg Thieme Verlag, Stuttgart, D, 2012), 4th edn.

[96] Vernet, M. & Kapoula, Z. Binocular motor coordination during saccades and fixations while reading: a magnitude and time analysis. *Journal of Vision* **9**, 2–2 (2009).

[97] Yuan, W. & Semmlow, J. L. The influence of repetitive eye movements on vergence performance. *Vision Research* **40**, 3089–3098 (2000).

[98] Rantanen, E. M. & Goldberg, J. H. The effect of mental workload on the visual field size and shape. *Ergonomics* **42**, 816–834 (1999).

[99] Haag-Streit AG. *Octopus900 brochure*, 2nd edn. URL http://www.haag-streit.com/product/perimetry/octopusr-900.html. As of 11.2016.

[100] Müller, M. *Vigilanzüberwachung während automatisierter, statischer Perimetrie*. Masterthesis, University of Tübingen (2013).

[101] Repin, I. Unexpected. Tretyakov Gallery in Moscow, Russia (1884).

[102] Repin, I. Unexpected visitors. Tretyakov Gallery in Moscow, Russia (1888).

[103] Marshall, S. P. The index of cognitive activity: Measuring cognitive workload. In *Proceedingsof the IEEE 7th Conference on Human Factors and Power Plants*, 7–5 (IEEE Computer Society Press, Los Alamitos, CA, USA, 2002).

[104] Brookhuis, K. A. & De Waard, D. The use of psychophysiology to assess driver status. *Ergonomics* **36**, 1099–1110 (1993).

[105] Brookhuis, K. A. & de Waard, D. Monitoring drivers' mental workload in driving simulators using physiological measures. *Accident Analysis & Prevention* **42**, 898–903 (2010).

[106] Coughlin, J. F., Reimer, B. & Mehler, B. Driver wellness, safety & the development of an AwareCar. *MIT AgeLab White Paper* (2009).

[107] Healey, J. A. & Picard, R. W. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* **6**, 156–166 (2005).

[108] Van Stavern, G. P., Biousse, V., Lynn, M. J., Simon, D. J. & Newman, N. J. Neuro-ophthalmic manifestations of head trauma. *Journal of Neuro-Ophthalmology* **21**, 112–117 (2001).

[109] Gilhotra, J. S., Mitchell, P., Healey, P. R., Cumming, R. G. & Currie, J. Homonymous visual field defects and stroke in an older population. *Stroke* **33**, 2417–2420 (2002).

[110] McGwin, G. *et al.* Is glaucoma associated with motor vehicle collision involvement and driving avoidance? *Investigative Ophthalmology & Visual Science* **45**, 3934–3939 (2004).

[111] Tant, M., Brouwer, W., Cornelissen, F. & Kooijman, A. Driving and visuospatial performance in people with hemianopia. *Neuropsychological Rehabilitation* **12**, 419–437 (2002).

[112] Wood, J. M. *et al.* On-road driving performance by persons with hemianopia and quadrantanopia. *Investigative Ophthalmology & Visual Science* **50**, 577–585 (2009).

[113] Szlyk, J. P., Brigell, M. & Seiple, W. Effects of age and hemianopic visual field loss on driving. *Optometry & Vision Science* **70**, 1031–1037 (1993).

[114] Tant, M., Cornelissen, F., Kooijman, A. & Brouwer, W. H. Hemianopic visual field defects elicit hemianopic scanning. *Vision Research* **42**, 1339–1348 (2002).

[115] Bowers, A. R., Ananyev, E., Mandel, A. J., Goldstein, R. B. & Peli, E. Driving with hemianopia: IV. Head scanning and detection at intersections in a simulatorhead scanning by drivers with hemianopia. *Investigative Ophthalmology & Visual Science* **55**, 1540–1548 (2014).

[116] Bowers, A. R., Mandel, A. J., Goldstein, R. B. & Peli, E. Driving with hemianopia, I: Detection performance in a driving simulator. *Investigative Ophthalmology & Visual Science* **50**, 5137–5147 (2009).

[117] Kasneci, E. *et al.* Driving with binocular visual field loss? A study on a supervised on-road parcours with simultaneous eye and head tracking. *PloS one* **9**, e87470 (2014).

[118] Papageorgiou, E. *et al.* Collision avoidance in persons with homonymous visual field defects under virtual reality conditions. *Vision Research* **52**, 20–30 (2012).

[119] Racette, L. & Casson, E. J. The impact of visual field loss on driving performance: evidence from on-road driving assessments. *Optometry & Vision Science* **82**, 668–674 (2005).

[120] Hamel, J. *et al.* Driving simulation in the clinic: testing visual exploratory behavior in daily life activities in patients with visual field defects. *Journal of Visualized Experiments* e4427–e4427 (2012).

[121] Alberti, C. F., Peli, E. & Bowers, A. R. Driving with hemianopia: III. Detection of stationary and approaching pedestrians in a simulator. *Investigative Ophthalmology & Visual Science* **55**, 368–374 (2014).

[122] Quigley, H. A. & Broman, A. T. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology* **90**, 262–267 (2006).

[123] Goldberg, I. *et al.* Assessing quality of life in patients with glaucoma using the Glaucoma Quality of Life-15 (GQL-15) questionnaire. *Journal of Glaucoma* **18**, 6–12 (2009).

[124] Haymes, S. A., LeBlanc, R. P., Nicolela, M. T., Chiasson, L. A. & Chauhan, B. C. Risk of falls and motor vehicle collisions in glaucoma. *Investigative Ophthalmology & Visual Science* **48**, 1149–1155 (2007).

[125] McCloskey, L. W., Koepsell, T. D., Wolf, M. E. & Buchner, D. M. Motor vehicle collision injuries and sensory impairments of older drivers. *Age and Ageing* **23**, 267–273 (1994).

[126] Szlyk, J. P., Mahler, C. L., Seiple, W., Edward, D. P. & Wilensky, J. T. Driving performance of glaucoma patients correlates with peripheral visual field loss. *Journal of Glaucoma* **14**, 145–150 (2005).

[127] Szlyk, J. P., Taglia, D. P., Paliga, J., Edward, D. P. & Wilensky, J. T. Driving performance in patients with mild to moderate glaucomatous clinical vision changes. *Journal of Rehabilitation Research and Development* **39**, 467 (2002).

[128] Crabb, D. P. *et al.* Exploring eye movements in patients with glaucoma when viewing a driving scene. *PloS one* **5**, e9710 (2010).

[129] Folstein, M. F., Folstein, S. E. & McHugh, P. R. "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* **12**, 189–198 (1975).

[130] Vanier, M. *et al.* Evaluation of left visuospatial neglect: norms and discrimination power of two tests. *Neuropsychology* **4**, 87 (1990).

[131] Zeeb, E. Daimler's new full-scale, high-dynamic driving simulator–a technical overview. *Actes INRETS, Institut national de recherche sur les transports et leur sécurité* 157–165 (2010).

[132] Healey, J., Seger, J. & Picard, R. Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural situations. *Biomedical Sciences Instrumentation* **35**, 193–198 (1999).

[133] Östlund, J. *et al.* Driving performance assessment-methods and metrics. (EU Deliverable, Adaptive Integrated Driver-Vehicle Interface Project (AIDE) No. D2. 2.5) (2005).

[134] Van Winsum, W., Brookhuis, K. A. & de Waard, D. A comparison of different ways to approximate time-to-line crossing (TLC) during car driving. *Accident Analysis & Prevention* **32**, 47–56 (2000).

[135] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67** (2015).

[136] Schulte, T., Strasburger, H., Müller-Oehring, E., Kasten, E. & Sabel, B. Automobile driving performance of brain-injured patients with visual field defects. *American Journal of Physical Medicine & Rehabilitation* **78**, 136–142 (1999).

[137] Papageorgiou, E. *et al.* The neural correlates of impaired collision avoidance in hemianopic patients. *Acta Ophthalmologica* **90**, e198–e205 (2012).

[138] Lövsund, P., Hedin, A. & Törnros, J. Effects on driving performance of visual field defects: a driving simulator study. *Accident Analysis & Prevention* **23**, 331–342 (1991).

[139] Kooijman, A. *et al.* Compensatory viewing training improves practical fitness to drive of subjects with impaired vision. *Visual Impairment Research* **6**, 1–1 (2004).

[140] Coeckelbergh, T. R., Brouwer, W. H., Cornelissen, F. W., Van Wolffelaar, P. & Kooijman, A. C. The effect of visual field defects on driving performance: a driving simulator study. *Archives of Ophthalmology* **120**, 1509–1516 (2002).

[141] Hardiess, G., Papageorgiou, E., Schiefer, U. & Mallot, H. A. Functional compensation of visual field deficits in hemianopic patients under the influence of different task demands. *Vision Research* **50**, 1158–1172 (2010).

[142] Bahill, A. T., Adler, D. & Stark, L. Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology & Visual Science* **14**, 468–469 (1975).

[143] Zangemeister, W., Meienberg, O., Stark, L. & Hoyt, W. Eye-head coordination in homonymous hemianopia. *Journal of Neurology* **226**, 243–254 (1982).

[144] Engström, J., Johansson, E. & Östlund, J. Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour* **8**, 97–120 (2005).

[145] Smith, N. D., Crabb, D. P., Glen, F. C., Burton, R. & Garway-Heath, D. F. Eye movements in patients with glaucoma when viewing images of everyday scenes. *Seeing and Perceiving* **25**, 471–492 (2012).

[146] Vega, R. P., van Leeuwen, P. M., Vélez, E. R., Lemij, H. G. & de Winter, J. C. Obstacle avoidance, visual detection performance, and eye-scanning behavior of glaucoma patients in a driving simulator: a preliminary study. *PloS one* **8**, e77294 (2013).

[147] Wood, J. M. *et al.* Hemianopic and quadrantanopic field loss, eye and head movements, and driving. *Investigative Ophthalmology & Visual Science* **52**, 1220–1225 (2011).

[148] Bowers, A. R., Mandel, A. J., Goldstein, R. B. & Peli, E. Driving with hemianopia, II: Lane position and steering in a driving simulator. *Investigative Ophthalmology & Visual Science* **51**, 6605–6613 (2010).

[149] Haymes, S. A., LeBlanc, R. P., Nicolela, M. T., Chiasson, L. A. & Chauhan, B. C. Glaucoma and on-road driving performance. *Investigative Ophthalmology & Visual Science* **49**, 3035–3041 (2008).

[150] Hardiess, G. & Mallot, H. A. Task-dependent representation of moving objects within working memory in obstacle avoidance. *Strabismus* **18**, 78–82 (2010).

[151] Machner, B. *et al.* Visual search disorders beyond pure sensory failure in patients with acute homonymous visual field defects. *Neuropsychologia* **47**, 2704–2711 (2009).

[152] Soto, D. & Humphreys, G. W. Stressing the mind: The effect of cognitive load and articulatory suppression on attentional guidance from working memory. *Perception & Psychophysics* **70**, 924–934 (2008).

[153] Fördős, G., Bosznai, I., Kovács, L., Benyó, B. & Benyó, Z. Sensor-net for monitoring vital parameters of vehicle drivers. *ACTA Polytechnica Hungarica* **4**, 25–36 (2007).

[154] Ellis, S. R. & Smith, J. D. Patterns of statistical dependency in visual scanning. In Groner, R. G., Mcconkie, G. W. & Menz, C. (eds.) *Eye Movements and Human Information Processing*, vol. 9, 221–238 (Elsevier Science Publishers BV, Amsterdam, NE, 1985).

[155] Arheim, R. *Art and Visual Perception: A Psychology of the Creative Eye* (University of California Press, Berkeley, CA, USA, 1974).

[156] Peysakhovich, V., Hurter, C. & Telea, A. Attribute-driven edge bundling for general graphs with applications in trail analysis. In *Pacific Visualization Symposium (PacificVis)*, 39–46 (IEEE, 2015).

[157] Tafaj, E., Kasneci, G., Rosenstiel, W. & Bogdan, M. Bayesian online clustering of eye movement data. In *Proceedings of the 2012 Symposium on Eye Tracking Research and Applications*, 285–288 (ACM, New York, NY, USA, 2012).

[158] Engelbrecht, M., Betz, J., Klein, C. & Rosenberg, R. Dem Auge auf der Spur: Eine historische und empirische Studie zur Blickbewegung beim Betrachten von Gemälden. *IMAGE–Zeitschrift für interdisziplinäre Bildwissenschaft* **11**, 29 (2010).

[159] Rosenberg, R. & Klein, C. The moving eye of the beholder: Eye tracking and the perception of paintings. In Huston, J. P., Nadal, M., Mora, F., Agnati, L. F. & Conde, C. J. C. (eds.) *Art, Aesthetics, and the Brain*, chap. 5, 79–108 (Oxford University Press, Oxford, UK, 2015).

[160] Dong, W., Liao, H., Roth, R. E. & Wang, S. Eye tracking to explore the potential of enhanced imagery basemaps in web mapping. *The Cartographic Journal* (2014).

[161] Blascheck, T. *et al.* State-of-the-art of visualization for eye tracking data. In Borgo, R., Maciejewski, R. & Viola, I. (eds.) *EuroVis - STARs* (The Eurographics Association, 2014).

[162] Bojko, A. A. Informative or Misleading? Heatmaps Deconstructed. In Jacko, J. A. (ed.) *Human-Computer Interaction. New Trends: Proceedings of the 13th International Conference HCI 2009, Part I*, 30–39 (Springer, Berlin, D, 2009).

[163] Popelka, S. & Voženílek, V. Specifying of requirements for spatio-temporal data in map by eye-tracking and space-time-cube. In *2012 International Conference on Graphic and Image Processing*, 87684N–87684N (International Society for Optics and Photonics, 2013).

[164] van der Zwan, M., Codreanu, V. & Telea, A. CUBu: Universal real-time bundling for large graphs. In *Transactions on Visualization and Computer Graphics*, vol. 22, 2550–2563 (IEEE, 2016).

[165] Burch, M., Schmauder, H., Raschke, M. & Weiskopf, D. Saccade plots. In *Proceedings of the 2014 Symposium on Eye Tracking Research and Applications*, 307–310 (ACM, New York, NY, USA, 2014).

[166] Badt, K., Dittmann, L. & van Delft, J. V. *Modell und Maler von Jan Vermeer: Probleme der Interpretation: eine Streitschrift gegen Hans Sedlmayr* (DuMont, Cologne, D, 1997).

[167] Schievelbein, G. *Similarity Collect - Clustering of Saccades*. Bachelor thesis, Universidade federal do Rio Grande do Sul (2015).

[168] Härdle, W. & Simar, L. *Applied multivariate statistical analysis*, vol. 22007 (Springer, Berlin, D, 2007).

[169] Rosenberg, R. Blicke messen: Vorschläge für eine empirische Bildwissenschaft. *Jahrbuch der Bayerischen Akademie der Schönen Künste* **27**, 71–86 (2014).

[170] Kalal, Z., Mikolajczyk, K. & Matas, J. Tracking-Learning-Detection. In *Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, 1409–1422 (IEEE, 2012).

[171] Xu, C., Xiong, C. & Corso, J. J. Streaming hierarchical video segmentation. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y. & Schmid, C. (eds.) *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision Proceedings, Part VI*, 626–639 (Springer, Berlin, Heidelberg, D, 2012).

[172] Comaniciu, D. & Meer, P. Mean shift: A robust approach toward feature space analysis. In *Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, 603–619 (IEEE, 2002).

[173] Muja, M. & Lowe, D. G. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications*, 331–340 (INSTICC Press, 2009).

[174] Lowe, D. G. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision*, vol. 2, 1150–1157 (IEEE, 1999).

[175] Seber, G. A. *Multivariate Observations* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2008).

[176] Anderson, N. C., Anderson, F., Kingstone, A. & Bischof, W. F. A comparison of scanpath comparison methods. *Behavior Research Methods* **47**, 1377–1392 (2015).

[177] Noton, D. & Stark, L. Scanpaths in eye movements during pattern perception. *Science* **171**, 308–311 (1971).

[178] Cristino, F., Mathôt, S., Theeuwes, J. & Gilchrist, I. D. ScanMatch: a novel method for comparing fixation sequences. *Behavior Research Methods* **42**, 692–700 (2010).

[179] West, J. M., Haake, A. R., Rozanski, E. P. & Karn, K. S. eyePatterns: software for identifying patterns and similarities across fixation sequences. In *Proceedings of the 2006 symposium on Eye Tracking Research and Applications*, 149–154 (ACM, New York, NY, USA, 2006).

[180] Myers, C. W. & Schoelles, M. J. Protomatch: A tool for analyzing high-density, sequential eye gaze and cursor protocols. *Behavior Research Methods* **37**, 256–270 (2005).

[181] Duchowski, A. T. *et al.* Scanpath comparison revisited. In *Proceedings of the 2010 Symposium on Eye Tracking Research and Applications*, 219 (ACM, New York, NY, USA, 2010).

[182] Heminghous, J. & Duchowski, A. T. iComp: A tool for scanpath visualization and comparison. In *Proceedings of the 3rd Symposium on Applied Perception in Graphics and Visualization*, 152–152 (ACM, New York, NY, USA, 2006).

[183] Over, E. A., Hooge, I. T. & Erkelens, C. J. A quantitative measure for the uniformity of fixation density: the Voronoi method. *Behavior Research Methods* **38**, 251–261 (2006).

[184] Zangemeister, W. H. & Oechsner, U. Evidence for scanpaths in hemianopic patients shown through string editing methods. *Advances in Psychology* **116**, 197–221 (1996).

[185] Feusner, M. & Lukoff, B. Testing for statistically significant differences between groups of scan patterns. In *Proceedings of the 2008 Symposium on Eye-Tracking Research and Applications*, 43 (ACM, New York, NY, USA, 2008).

[186] Lao, J., Miellet, S., Pernet, C., Sokhn, N. & Caldara, R. iMap 4: An open source toolbox for the statistical fixation mapping of eye movement data with linear mixed modeling. *Journal of Vision* **15**, 793–793 (2015).

[187] Li, X., Çöltekin, A. & Kraak, M.-J. Visual exploration of eye movement data using the space-time-cube. In *International Conference on Geographic Information Science*, 295–309 (Springer, Berlin, D, 2010).

[188] Mannan, S. K., Ruddock, K. H. & Wooding, D. S. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision* **10**, 165–188 (1996).

[189] Mannan, S., Ruddock, K. & Wooding, D. Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision* **11**, 157–178 (1997).

[190] Mathôt, S., Cristino, F., Gilchrist, I. & Theeuwes, J. A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research* **5**, 1–15 (2012).

[191] Dewhurst, R. *et al.* It depends on how you look at it: scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods* **44**, 1079–100 (2012).

[192] Jarodzka, H., Holmqvist, K. & Nyström, M. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 Symposium on Eye Tracking Research and Applications*, 211–218 (ACM, New York, NY, USA, 2010).

[193] Foerster, R. M. & Schneider, W. X. Functionally sequenced scanpath similarity method (FuncSim): Comparing and evaluating scanpath similarity based on a task's inherent sequence of functional (action) units. *Journal of Eye Movement Research* **6**, 1–22 (2013).

[194] Ponsoda, V., Scott, D. & Findlay, J. M. A probability vector and transition matrix analysis of eye movements during visual search. *Acta Psychologica* **88**, 167–185 (1995).

[195] Stark, L. W. & Ellis, S. R. Scanpaths revisited: Cognitive models direct active looking. In Fisher, D., Monty, R. & Senders, J. (eds.) *Eye movements: Cognition and visual perception*, 193–226 (Lawrence Erlbaum Associates, 1981).

[196] Engbert, R. & Kliegl, R. Mathematical models of eye movements in reading: A possible role for autonomous saccades. *Biological Cybernetics* **85**, 77–87 (2001).

[197] Kanan, C., Ray, N. A., Bseiso, D. N., Hsiao, J. H. & Cottrell, G. W. Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the 2014 Symposium on Eye Tacking Research and Applications*, 287–290 (ACM, New York, NY, USA, 2014).

[198] Mast, M. & Burmester, M. Exposing repetitive scanning in eye movement sequences with t-pattern detection. In *Proceedings of IADIS International Conference on Interfaces and Human Computer Interaction*, 137–145 (Rome, IT, 2011).

[199] Magnusson, M. S. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers* **32**, 93–110 (2000).

[200] Lin, J., Keogh, E., Wei, L. & Lonardi, S. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery* **15**, 107–144 (2007).

[201] Rakthanmanon, T. & Keogh, E. Fast shapelets: A scalable algorithm for discovering time series shapelets. In *Proceedings of the 13th SIAM Conference on Data Mining* (2013).

[202] Rubner, Y., Tomasi, C. & Guibas, L. J. A metric for distributions with applications to image databases. In *Proceedings of the International Conference on Computer Vision*, 59–66 (IEEE, 1998).

[203] Rüschendorf, L. The Wasserstein distance and approximation theorems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **70**, 117–129 (1985).

[204] Rothe, C. *Evaluation of Scanpath Comparison metrics for static and dynamic tasks*. Master thesis, University of Applied Sciences Aalen (2015).

[205] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300 (1995).

[206] Eivazi, S. *et al.* Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In *Proceedings of the 2012 Symposium on Eye Tracking Research and Applications*, 377–380 (ACM, New York, NY, USA, 2012).

[207] Law, B., Atkins, M. S., Kirkpatrick, A. E. & Lomax, A. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *Proceedings of the 2004 Symposium on Eye Tracking Research and Applications*, 41–48 (ACM, New York, NY, USA, 2004).

[208] Nodine, C. & Kundel, H. Using eye movements to study visual search and to improve tumor detection. *Radiographics* **7**, 1241–1250 (1987).

[209] Pietrzyk, M. W., McEntee, M. F., Evanoff, M. E., Brennan, P. C. & Mello-Thoms, C. R. Direction of an initial saccade depends on radiological expertise. In Mello-Thoms, C. R. & Kupinski, M. A. (eds.) *Medical Imaging 2014: Image Perception, Observer Performance, and Technology Assessment* (SPIE-International Society for Optics and Photonics, San Diego, CA, USA, 2014).

[210] Reingold, E. M., Charness, N., Pomplun, M. & Stampe, D. M. Visual span in expert chess players: Evidence from eye movements. *Psychological Science* **12**, 48–55 (2001).

[211] Hembrooke, H., Feusner, M. & Gay, G. Averaging scan patterns and what they can tell us. In *Proceedings of the 2006 Symposium on Eye Tracking Research and Applications*, 41–41 (ACM, New York, NY, USA, 2006).

[212] Chang, Y.-W. & Lin, C.-J. Feature ranking using linear SVM. *Causation and Prediction Challenge Challenges in Machine Learning* **2**, 47 (2008).

[213] Braunagel, C., Kasneci, E., Stolzmann, W. & Rosenstiel, W. Driver-activity recognition in the context of conditionally autonomous driving. In *Proceedings of the 18th International Conference on Intelligent Transportation Systems*, 1652–1657 (IEEE, 2015).

[214] Rieck, K., Wressnegger, C. & Bikadorov, A. Sally: a tool for embedding strings in vector spaces. *Journal of Machine Learning Research* **13**, 3247–3251 (2012).

[215] Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**, 27:1–27:27 (2011).

[216] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* **9**, 1871–1874 (2008).

[217] Leslie, C. S., Eskin, E., Cohen, A., Weston, J. & Noble, W. S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20**, 467–476 (2004).

[218] Eskin, E., Weston, J., Noble, W. S. & Leslie, C. S. Mismatch string kernels for svm protein classification. In *Advances in Neural Information Processing Systems*, vol. 15, 1417–1424 (MIT Press, Vancouver, Canada, 2002).

[219] Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).

[220] Borji, A. & Itti, L. Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision* **14**, 29 (2014).

[221] Haji-Abolhassani, A. & Clark, J. J. An inverse Yarbus process: Predicting observers' task from eye movement patterns. *Vision Research* **103**, 127–142 (2014).

[222] Peters, R. J. & Itti, L. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8 (IEEE, New York, NY, USA, 2007).

[223] Hatt, S. R., Leske, D. A., Kirgis, P. A., Bradley, E. A. & Holmes, J. M. The effects of strabismus on quality of life in adults. *American Journal of Ophthalmology* **144**, 643–647 (2007).

[224] Wassill, K. & Kaufmann, H. Binokulare dreidimensionale Videookulographie. *Der Ophthalmologe* **97**, 629–632 (2000).

[225] Garrido-Jurado, S., Muñoz Salinas, R., Madrid-Cuevas, F. J. & Marín-Jiménez, M. J. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* **47**, 2280–2292 (2014).

[226] Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* **25**, 120–126 (2000).

[227] de Almeida, J. D. S., Silva, A. C., de Paiva, A. C. & Teixeira, J. A. M. Computational methodology for automatic detection of strabismus in digital images through Hirschberg test. *Computers in Biology and Medicine* **42**, 135–146 (2012).

[228] Fisher, A. C., Chandna, A. & Cunningham, I. P. The differential diagnosis of vertical strabismus from prism cover test data using an artificially intelligent expert system. *Medical & Biological Engineering & Computing* **45**, 689–693 (2007).

[229] Chandna, A., Fisher, A. C., Cunningham, I., Stone, D. & Mitchell, M. Pattern recognition of vertical strabismus using an artificial neural network (StrabNet©). *Strabismus* **17**, 131–138 (2009).

[230] Wirtz, T. *Klassifikation von neurogenen Augenmotilitätsstörungen*. Master thesis, University of Tübingen (2015).