CLASSIFICATION AND CORRECTION OF GLASSES-INDUCED EYE TRACKING ERRORS

by

Johannes Schweig

Master’s thesis

Technische Universität Berlin

Fakultät V: Verkehrs- und Maschinensysteme

Institut für Psychologie und Arbeitswissenschaft

Human Factors

Supervisors:

Prof. Dr. Matthias Rötting

M.Sc. Sarah-Christin Freytag

Berlin, 03.08.2017

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 31.07.2017 Johannes Schweig

Contents

List of Figures

List of Tables

List of abbreviations

AE	Artificial eye
GNB	Gaussian Naive Bayes
IR	Infrared
KNN	*k* nearest neighbor
LOGR	Logistic regression
LSVM	Linear support vector machine
MLP	Multilayer perceptron
SMI	SensoMotoric instruments
SVM	Support vector machine

1 Abstract

Eye tracking has proven to be a useful method in experimental setups and interactive systems. Unfortunately, certain user characteristics (drooping eyelids, eyeglasses) and environmental factors (illumination) jeopardize the quality of eye tracking data (Holmqvist et al., 2011). This master's thesis proposes an approach to identifying and eliminating errors related to the wearing of eyeglasses. Previous studies show that the refractive strength of worn glasses has a modifying effect on the eye tracking signal. Therefore, eye tracking data was collected in this thesis using a head model with adjustable artificial eyes (AEs) and different pairs of glasses in order to obtain a data set with the systematic effects of different refractive strengths under controlled conditions. The recorded data was then used to solve a classification and a regression task using different algorithms. For the first task, models were built to classify eye tracking data based on the refractive strength of the originating eyeglasses. The second task involved the use of that classification to eliminate errors caused by eyeglasses considering the refractive strength. Different models based on a variety of algorithms were trained and evaluated on independent data sets. The most promising models are examined, and strengths and weaknesses of the approach are discussed.

Zusammenfassung

Eye-Tracking ist eine vielversprechende Methode, die in wissenschaftlichen Experimenten und interaktiven Systemen eingesetzt wird. Leider beeinträchtigen bestimmte Nutzermerkmale (herabhängende Augenlider, Brillen) und Umweltfaktoren (Beleuchtung) die Qualität von Eye-Tracking Daten (Holmqvist et al., 2011). Diese Masterarbeit stellt einen Ansatz vor um Fehler, die durch das Tragen von Brillen entstehen, zu identifizieren und zu beheben. Studien haben gezeigt, dass die Brechkraft von getragenen Brillen das Eye-Tracking Signal verändert. Um dies zu untersuchen werden Daten mit einem Kopfmodell mit verstellbaren, künstlichen Augen und unterschiedlichen Brillen erhoben. Die erhobenen Daten werden dann benutzt um eine Klassifikations- und eine Regressionsaufgabe mithilfe verschiedener Algorithmen zu lösen. Für die erste Aufgabe werden Modelle erzeugt, die Eye-Tracking Daten anhand der Brechkraft der verwendeten Brille klassifizieren. In der zweiten Aufgabe wird diese Klassifikation genutzt um durch Brillen verursachte Fehler angepasst an die Brechkraft zu beheben. Verschiedene Modelle, die auf unterschiedlichen Algorithmen basieren, wurden trainiert und ihre Leistung auf unabhängigen Datensets eingeschätzt. Die vielversprechendsten Modelle werden untersucht und Stärken wie Schwächen dieses Ansatzes diskutiert.

Classification and correction of glasses-induced eye tracking errors

1.1 Motivation

Eye tracking is widely used in research and interactive systems and it promises big advances in recent technologies, such as the foveated rendering[1] in Virtual Reality (Constine, 2016) and the adaptive editing of photographs (DeCarlo & Santella, 2002; Santella, Agrawala, DeCarlo, Salesin, & Cohen, 2006). However, certain user characteristics (eyeglasses, lenses, mascara, drooping eyelids) and environmental factors (lighting) make the employment of eye tracking in everyday technical systems (ATMs, ticket vending machines) difficult. While the exclusion of users with problematic characteristics is reasonable in a research setting, it is not feasible in practice (Feit et al., 2017).

This thesis addresses the problem associated with the use of eyeglasses in eye tracking by exploring the relationship between eyeglasses and resulting change in the eye tracking signal. Therefore, eye tracking data is collected with different eyeglasses in a controlled laboratory setting and analyzed using a variety of methods from statistics and machine learning. The results of this thesis could greatly improve the real-world applicability of eye tracking in interactive systems with a variety of users.

1.2 The visual system

Basic knowledge of the structure and function of the visual system, and the eye in particular, is crucial for understanding the functionality of eye tracking.

1.2.1 Structure of the eye. The eyes are "one of the most salient features of the human face" (Hansen & Ji, 2010) and their movements play an important role in expressing a person’s desires, needs, cognitive processes, emotional states, and interpersonal relations (Underwood, 2005). This is why they are widely used in technical systems to gain information about behavioral or cognitive processes (e.g.

[1] In foveated rendering, the gaze position is used to reduce the image quality rendered in the peripheral vision of the users, thus greatly reducing rendering workload (*Foveated rendering*, 2017).

attention) of eye tracking users and therewith study humans' behavior and mind (research) or use as input in an interactive system (e.g. Virtual Reality).

The eye is a complex optical system composed of several biological components (see Figure 1) including the lens. The lens refracts and focuses light onto the retina, a thin layer of specially adapted sensory receptor cells positioned at the back of the eye (Tovée, 1996). The retina is populated with light-sensitive cells, called cones and rods, which convert the incoming light into electrical signals, which are then sent through the optic nerve to the visual cortex in the brain (Holmqvist et al., 2011, p. 21). Cones are photoreceptors concentrated in the fovea, a small area of the retina which accounts for less than 2° of the visual field, and provide high acuity, color vision (Tovée, 1996). Rods, which are concentrated away from the fovea, are very sensitive to light (Holmqvist et al., 2011) and provide low acuity, monochrome vision (Tovée, 1996). Visual acuity is greatest in the fovea, where cones are most highly concentrated (Holmqvist et al., 2011). Thus, in order to see a selected object sharply, a person must move his or her eyes so that the light falls directly onto this region of the retina (Holmqvist et al., 2011).

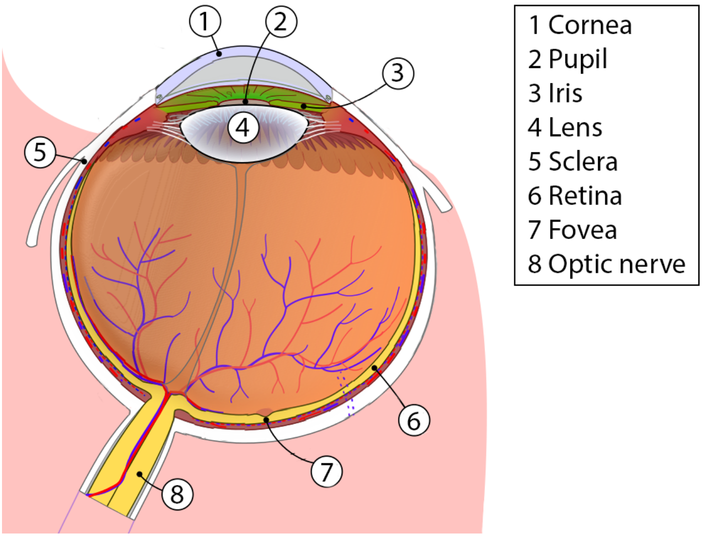


Figure 1. The human eye (own work adapted from Rhcastilhos, 2007)

1.2.2 Eyeglasses. The first appearance of lenses worn in front of the eyes for the modification of vision was reported in Northern Italy in the 13th century (Morgan, 1976). Major advances in technology have been made since then (achromatic lenses in

1758 by John Dollond; bifocals in 1784 by Benjamin Franklin) (Bach & Neuroth, 1998) and glasses are now widely used in modern society. According to Schiefer, Kraus, Baumbach, Ungewiß, and Michels (2016), 63% of persons in Germany over the age of 16 wear glasses.

Today, glasses are mainly used to correct for refractive errors in vision, such as myopia (nearsightedness), hyperopia (farsightedness), astigmatism (irregular corneal curvature) and presbyopia (aging eye condition) (Bach & Neuroth, 1998). Glasses are classified by the dioptric power of their lenses. A lens with positive power will cause incident rays of light to converge, while a lens with negative power will cause them to diverge (Bach & Neuroth, 1998). The common unit for optical power is the inverse meter (m^{-1}), which is referred to as the diopter. The refractive power of a lens is defined as the reciprocal of its focal length (Bach & Neuroth, 1998):

$$D = \frac{1}{f} \tag{1}$$

In myopia, parallel light rays are brought into focus in front of the retina and thus a negative or diverging lens is needed to counterbalance the effect (Bach & Neuroth, 1998) (see Figure 2). In hyperopia, the focal point of the unaccommodated eye lies behind the retina, and thus a positive or converging lens is used for correction (Bach & Neuroth, 1998).

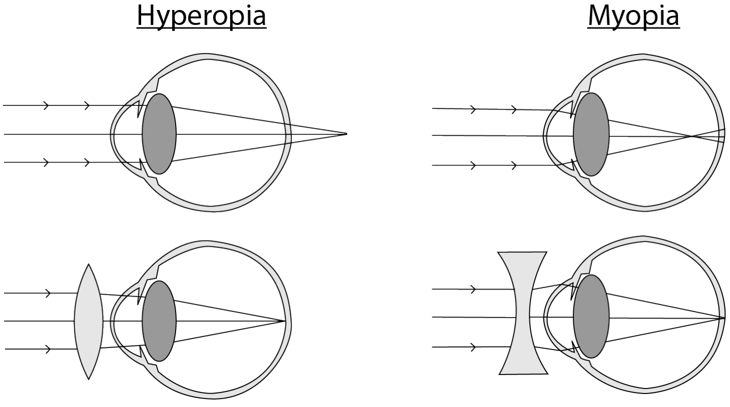


Figure 2. Effects of hyperopia (left) and myopia (right) on focal point with and without corrective lenses (own work adapted from CryptWizard, 2007a, 2007b)

Astigmatism refers to an asymmetric cornea in which the curvature, and therefore the refractive power, is unequal across its surface. If the meridians of the cornea are perpendicular, the astigmatism is regular and thus correctable. Correction is more difficult in cases of irregular astigmatism. Cylindrical lenses are used to correct astigmatism, while spherical lenses are used to correct for myopia and hyperopia. So-called "toric" lenses, which combine spherical and cylindrical surfaces, are used to compensate for both astigmatic and myopic or hyperopic defects (Bach & Neuroth, 1998).

Lenses are also classified as either single vision, multifocal or progressive lenses (Japan, 2012). Single vision lenses have the same refractive power over the whole lens. Multifocal lenses, which are built as either bifocal or trifocal lenses, are made by fusing two or three lens segments of different refractive powers together into a single lens (Artal, 2017). The difference in dioptric powers between the segments is called the addition (Artal, 2017). Progressive lenses address the abrupt change in dioptric power between the fused segments of multifocal lenses. Here, the dioptric power varies in a smooth fashion across the surface of the lens (Artal, 2017). This seamless progression of refractive power is said to closely mimic natural vision. Multifocal glasses more significantly distort the eye tracking data when compared to single-vision glasses, yet they are also widely used by the population (e.g. older people)[2].

1.3 Eye tracking

Understanding the technical processes of eye tracking is essential for identifying the causes of glasses-related eye tracking errors. The following section will introduce a variety of eye tracking methods and methods for gaze estimation.

Eye tracking is defined as a process in which a subject's point of gaze or eye motion relative to the head is measured (*Eye tracking*, 2017). Eye tracking is carried out using technical devices known as eye trackers. Eye trackers can be categorized according to their employed method for gaze detection. Duchowski (2009) distinguishes between the methods of Electro-OculoGraphy, scleral contact lens/search coil,

[2] In a study with 270 older persons (>50 years) 38% said they used varifocal glasses for refractive correction (Sivardeen, 2015)

Photo-OculoGraphy or Video-OculoGraphy, and video-based combined pupil and corneal reflection.

The last method, video-based combined pupil and corneal reflection, is currently the most popular. It utilizes inexpensive cameras and image processing hardware to calculate the gaze point in real-time (Duchowski, 2009) and is appropriate for table-mounted systems used in static environments and head-mounted systems used in interactive systems.

These systems illuminate the eye with an infrared (IR) light source to avoid all natural light reflections (Holmqvist et al., 2011). The IR light is reflected back from different layers of the eye (cornea, sclera, lens) and then identified as four "Purkinje reflections" (see Figure 3): The first and second reflection originate from the front and rear surface of the cornea, the third and forth from the front and rear surface of the lens (Duchowski, 2009). Whereas the Purkinje reflections are relatively stable, the eyeball, and thus the pupil, rotates when a subject changes its direction of gaze. Modern eye trackers use the location of the pupil center, in conjunction with the first Purkinje reflection, to calculate the point of gaze (Holmqvist et al., 2011). An advanced variant is the dual-Purkinje eye tracker, which utilizes two (the first and fourth) Purkinje reflections together with the pupil center.

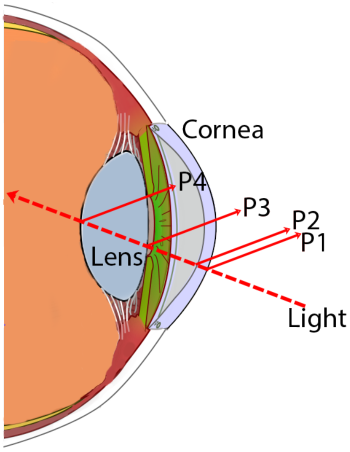


Figure 3. The four Purkinje reflections (own work adapted from Rhcastilhos, 2007)

In addition, there are dark (more common) and bright pupil systems. Both systems illuminate the pupil using IR light, which is reflected back from the retina through the pupil (Holmqvist et al., 2011). In a bright pupil system, the IR illumination must be co-axial with the view from the eye camera which causes the pupil to appear "bright" in the image of the eye. The main motivation is "to compensate for poor contrast sensitivity in the eye camera by increasing the difference in light emission between pupil and iris" (Holmqvist et al., 2011, p. 25). In a dark pupil system, the illumination source is offset which directs the IR reflection away from the eye camera. This causes the pupil to appear "dark".

The process of inferring gaze from an image of the eye is called gaze estimation. First, the image is analyzed, and image features such as pupil center and Purkinje reflections are extracted (Cerrolaza, Villanueva, Villanueva, & Cabeza, 2012). Next, gaze is deduced as a function of image features (Cerrolaza et al., 2012). The methods for modeling the connection between image and gaze can be subdivided into two main groups: Geometry-based models and interpolation-based methods (Cerrolaza et al., 2012).

Geometry-based models construct gaze as a function of the 3D configuration of the system and the human, while also taking into account geometry and physiology (Cerrolaza et al., 2012). According to recent research (Hansen & Ji, 2010; Kübler et al., 2016), no geometrical model accounting for the wearing of glasses has been devised, and gaze estimation techniques are typically evaluated on healthy-sighted subjects. As the geometrical model calculates a ray from the pupil center towards the camera, refractions by eyeglasses are not considered (Kübler et al., 2016).

Interpolation-based methods describe the gaze point as a function of image features (Cerrolaza et al., 2012). The unknown coefficients of the mapping function are then deduced for each participant via calibration. Mapping techniques include polynomials (most common), splines, nonparametric regressions, rational functions, and neural-networks (Cerrolaza et al., 2012). Relevant design aspects for a mapping function include the image features to be used as input, the degree of the polynomial, and the number of terms (Cerrolaza, Villanueva, & Cabeza, 2008). The determination of these design aspects remains unknown (Cerrolaza et al., 2008). Kübler et al. (2016)

assume that eyeglasses do not have a notable effect on the accuracy of interpolation-based methods, as their fitted function can be adjusted thus incorporating the effect of the glasses. Overall, the performance of geometry-based models and interpolation-based methods has proven to be very similar (Ramanauskas, 2015).

1.4 Eye tracking data

The data collected during an eye tracking experiment usually includes a timestamp in milliseconds and the X and Y coordinates of the estimated gaze point. Data files may also contain the number of trials, the onset of experimental stimuli, and horizontal and vertical pupil dilation.

The quantity of collected data is dependent upon the sampling rate of the eye tracker. If the eye tracker has a listed sampling rate of 50 Hz, it is capable of recording the gaze direction 50 times per second. Higher sampling rates are desirable for recording high-frequency eye movements, but generally also more expensive (Holmqvist et al., 2011).

After data collection, denoising is applied to eliminate excessive noise in the eye movement signal. Registered noise is often a result of the inherent instability of the eye and blinking. Blinks can be filtered out through the exclusion of those data samples which fall outside of a rectangular range (e.g. the spatial extent of the display) (Duchowski, 2009).

Distances can be reported in px, but are more typically reported in visual degrees ($^\circ$) or minutes ($60' = 1^\circ$). The visual angle v in degrees of arc can be calculated using the viewing distance, d, and the size of the object, x (Kortum, 2008):

$$v = arctan(\frac{x}{d}) \tag{2}$$

1.4.1 Data quality. Two important indicators of data quality are accuracy and precision (illustrated in Figure 4). Accuracy is the average difference between the true gaze position and the recorded gaze position (*offset*). It is calculated as the average angular offset θ_i ($^\circ$) between measured fixation locations and the corresponding locations of the fixation targets (Holmqvist et al., 2011):

$$\theta_{offset} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\theta_i^2} \tag{3}$$

Precision is the ability of an eye tracker to reliably reproduce a measurement (Holmqvist et al., 2011). Precision can be calculated using the standard deviation

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4}$$

or the root-mean-square error (RMSE) of successive data samples using angular distances (°) (Holmqvist et al., 2011):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\theta_i^2} \tag{5}$$

Reported RMSE values range from 1° for poorer eye trackers to 0.10° for high-end eye trackers (Holmqvist et al., 2011).

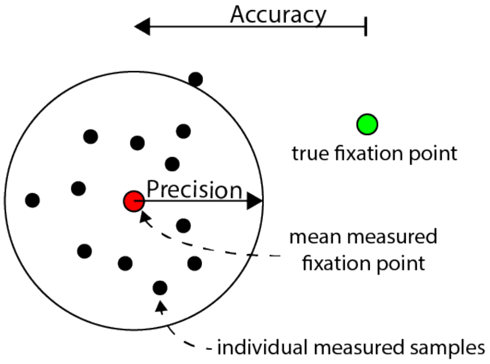


Figure 4. Precision and accuracy (adapted from Holmqvist et al., 2011)

Other concepts related to precision and accuracy include (Holmqvist et al., 2011):

- Offset: Distance between calculated fixation location and the location of the intended fixation target.
- Drift: Gradually increasing offset with time.
- System-inherent noise: Noise inherent to the eye tracking system.
- Optic artifacts: Physiologically impossible, high-speed movements caused by the interplay between the optical situation (glasses, reflections, shadows, varying ambient light conditions) and the gaze estimation algorithm.

Accuracy loss on the edges of the screen. Accuracy is usually highest in the center of the screen and deteriorates towards the edges. This is due to the larger gaze angles, which impair accuracy (Tobii, 2011). Nonetheless, high accuracy at large gaze angles is important when testing large stimuli (Tobii, 2011).

Blignaut, Holmqvist, Nyström, and Dewhurst (2014, p. 95) reported definite trends "to the direction and magnitude of the offsets across the display" even though "the direction of the offset could be in opposite directions in different areas of the display." Dahlberg (2010) reported higher errors with increased distance between the calibration point and the mounted eye tracker. Errors were higher in the upper corners of the screen and lower towards the middle horizontally and the bottom vertically.

1.4.2 Eyeglasses and eye tracking.

Induced noise by eyeglasses. Users wearing glasses impose a challenge for both the eye tracker and the researcher. These effects can be further specified:

1. *Darkening of eye image*: Eyeglasses or their surface treatments darken the eye image and reduce contrast between pupil and iris. This poses a challenge for dark-pupil systems (Holmqvist et al., 2011), which rely on the detection of the pupil without a dedicated light source.

2. *Occlusion or shadows from the frames*: Glasses frames may cast a shadow on the eye, affecting corneal reflection and the detection of the pupil. Additionally, the frames may occlude parts of the eye in certain eye positions (Holmqvist et al., 2011).

3. *Reflections*: Apart from the corneal reflection additional, fainter reflections can appear if the light from the corneal reflection is reflected back to the eye by the eyeglass lenses. Additionally, reflections may appear if the glasses are old, scratched, or treated to reflect sunlight off their surface. If these reflections appear near to the pupil or the corneal reflection in any gaze direction, the data may be jeopardized (Holmqvist et al., 2011).

4. *Alteration of the eye image*: If a light ray moves from one optical medium (e.g.

air) to another (e.g. glass), reflection and refraction appear at the boundary surface (Bach & Neuroth, 1998) (see Figure 5). "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" (Kübler et al., 2016, p. 144). The refraction is different for different wavelengths of light. This effect is called dispersion (Kübler et al., 2016). Refraction alters the image of the eye nonlinearly (Hansen & Ji, 2010), distorting the observed pupil position and pupil size (Hammoud, 2008) and introducing an error that depends on the strength of the optical medium (Kübler et al., 2016). Exactly how the effect of refraction alters the eye tracking signal remains unclear. Both the alteration of the eye image and the refraction of the light from the IR light source may have substantial effects on the eye tracking signal.

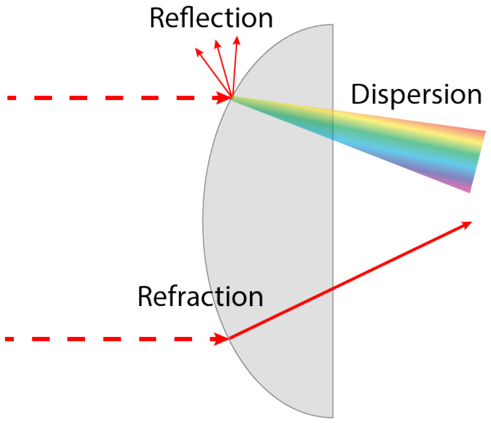


Figure 5. Refraction and reflection of a light ray on an eyeglass lens (own work adapted from Kübler et al., 2016)

In his study on the effect of glasses on eye tracking, Dahlberg (2010) reported a 20% higher error rate for participants with glasses than for those without. There was a positive relationship between magnitude of error and eyeglass strength (diopters), but only without the use of calibration. He concluded that the calibration procedure most likely compensated for some of the error.

Drewes, Masson, and Montagnini (2012) studied the effect of pupil size on eye tracking accuracy using a search-coil and camera-based method. They observed a systematic shift in gaze with the camera-based method during the constriction of the pupil. This led them to conclude that the shift occurred due to a change in the alignment between the true optical axis of the eye and the measured optical center of the pupil. "As the pupil constricts/dilates, the visible center of the pupil moves and causes a deviation in the computed pupil position" (Drewes et al., 2012, p. 210). This could also be a potential source of error, as the size of the pupil is also altered by the refraction of the eyeglasses.

Kübler et al. (2016) studied the effect of glasses on eye tracking using simulated, synthetic eye images. The eye tracking signal was deduced using either a polynomial fitting model or a geometrical model. Their simulations revealed that the polynomial fitting model causes a minor decrease in accuracy across all eyeglasses independent of their refractive strength. The employment of the geometrical model, on the other hand, is heavily influenced by the refractive strength of the eyeglasses.

Hohlfeld, Pomp, Link, and Guse (2015) reported no difference in accuracy between glass-wearing subjects ($704 \pm 84px$) and non-glass-wearing subjects ($707 \pm 137px$) in their experiments, concluding that gaze tracking performance is only affected by glasses "when they prevent visual properties from being detected" (Hohlfeld et al., 2015, p. 430).

The problem of low tolerance towards eyeglasses by eye tracking systems is a problem that has been solved only partially (Hansen & Ji, 2010).

Approaches for eliminating issues with glasses. Researchers have employed a number of approaches intended to eliminate problems in eye tracking which arise from the wearing of glasses. Proposed practical approaches to eliminating the problem include excluding participants who wear glasses, adjusting the IR light source so that reflections and shadows appear far away from the pupil and corneal reflection (Holmqvist et al., 2011), or providing participants with eye tracker-friendly glasses (Holmqvist et al., 2011).

In their study, Jo, Cho, Lee, Cha, and Kim (2013) addressed the problem of IR reflections on eyeglasses overlapping with the pupil region. They used a dual IR source eye tracker and turned one of the sources off, when large glints obstructing the view of the pupil were detected. With this method, they could preserve the accuracy of the system.

Ebisawa (1994) used two light sources to develop a method, which eliminated reflected light in eyeglass lenses. The effectiveness of the method was demonstrated using participants and an imaging processor, however, no performance figures were reported.

Gwon, Cho, Lee, Lee, and Park (2014) built a new system featuring four IR illuminators. The system was capable of determining whether a user was wearing glasses and would switch off an illuminator to find a position with minimal error from glasses reflection.

Hansen and Ji (2010) point out that the method of employing extra light sources comes with shortcomings: This method might not work outdoors or in situations where light conditions are less easily controllable (Hansen & Ji, 2010). If eye trackers ought to be used outdoors, a proper modeling of the glass itself may be needed (Hansen & Ji, 2010).

Huang, Kong, and Li (2013) developed an algorithm for differentiating between real corneal reflections and those from reflections caused by eyeglasses. In non-critical cases, the algorithm classified 85% of the recorded samples correctly.

Zhu and Ji (2005) built an eye tracking system featuring an appearance based pattern recognition method (Support vector machine [SVM]) and object tracking with a bright-pupil eye tracker. They concluded that their eye tracker could "still detect and track eyes robustly and accurately for people with glasses" (Zhu & Ji, 2005, p. 150).

1.5 Eye tracking with artificial eyes

AEs are used by manufacturers (SensoMotoric Instruments [SMI], Tobii, SR Research) (Holmqvist, Nyström, & Mulvey, 2012) of eye tracking equipment and researchers to measure the spatial and temporal precision of eye trackers (Holmqvist et al., 2011). Most AEs resemble the eyes of a doll, featuring an artificial pupil and a reflective surface that generates a first surface reflection which corresponds to the corneal reflection (Wang, Mulvey, Pelz, & Holmqvist, 2017). While it is relatively simple to produce AEs for dark pupil systems, it is more difficult to engineer them for bright pupil systems (Holmqvist et al., 2012).

AEs are advantageous because they completely exclude artifacts caused by human eye movements providing a more accurate estimate of a system's precision (Holmqvist et al., 2011). Wang et al. (2017) compared different AEs and eye trackers and demonstrated that precision was higher when compared to data collected from human participants. Furthermore, AEs are useful because the artificial pupil maintains a constant size and does not vary with changes in lighting conditions. As might be expected, AEs' primary drawback is their dissimilarity from the human eye in terms of iris, pupil, and corneal reflection features (Holmqvist et al., 2012). Finally, they "may be easier or more difficult for the image analysis algorithms in the eye tracker to process" (Holmqvist et al., 2012, p. 49). For these reasons, precision tests using AEs cannot fully replace precision tests performed on human test subjects (Holmqvist et al., 2012).

As the models used in this thesis rely on accurate data, the data quality is of vital importance for this thesis. AEs provide a suitable and popular technique to record accurate and stable eye tracking data. Human participants pose the thread of confounding the data with artifacts through eye movements and with variance caused by interindividual differences. Therefore, AEs will be used in this thesis for the data collection.

1.6 Computational approaches for better data quality

Researchers have devised approaches for improving the quality of eye tracking data by means of mathematical computation and correction. This often involves comparing the observed eye position (by the eye tracker) with the true eye position (by a valid source). Researchers collect potentially faulty eye tracking data together with data by an external source (e.g. a calibration grid). The offsets found between observed and true eye tracking data (offsets) are then used to shift the faulty eye tracking data and/or future incoming data with mathematical transformations.

Cherif, Nait-Ali, Motsch, and Krebs (2002) required that subjects ($n = 3$) look at two calibration grids (5x5, 4x4). Following the first calibration grid, polynomial transformations of order 1 to 5 were computed, thus minimizing the mean squared error. The polynomial transformations were then applied to the second calibration grid. The highest order (5^{th}) polynomial demonstrated the lowest mean squared error for the first calibration grid, but polynomials of lower orders were better suited for the second calibration grid. The authors were successfully able to improve mean squared error from 6.9° to 2.2°.

Blignaut et al. (2014) used a regression-based approach to improve the accuracy of their eye trackers in real time. Following the use of a calibration grid, a regression formula based on the five nearest calibration points was applied to the gaze location of the participants. This method improved the accuracy of the two eye trackers used in the study from 0.8° to 0.5° and 1.1° to 0.5°, respectively.

In their 2002 study, Hornof and Halverson observed that each participant exhibited a relatively consistent pattern of horizontal and vertical deviations across all trials, which they named the participant's "error signature." *Implicit required fixation locations* were used to automatically initiate recalibration if precision dropped below a certain threshold (2° from implicit required fixation locations). A weighted average of the four nearest error vectors was subsequently applied to the data to further minimize the effect of the participant's error signature on the data.

Zhang and Hornof (2011) used an annealed mean shift algorithm to extract and correct systematic error in their eye tracking data. The algorithm identified the global mode of disparities (a more sophisticated measure of central tendency), which they used to shift the eye tracking data. This reduced their mean absolute horizontal deviation from 0.3° to 0.1°.

Vadillo, Street, Beesley, and Shanks (2015) used probable fixation locations to correct their eye tracking data offline using an algorithm. This algorithm moved their coordinates using a 2 × 2 transformation matrix and also stretched or contracted the fixations' space. Most fixations from the experiment were between 0.7° to 1.4° away from the target, and the correction algorithm reduced this distance to 0.6° to 0.8°.

Table 1 contains an overview over the studies mentioned here.

Table 1

Overview over studies using computational approaches for better data quality

Authors	External source	Mathematical transformation
Cherif et al.	Calibration grid	Polynomial transformations with degree 1 to 5
Blignaut et al.	Calibration grid	Regression based on five nearest calibration points
Hornof and Halverson	Implicit required fixation locations	Weighted average of four nearest error vectors
Zhang and Hornof	Required fixation locations	Shift with error vector generated from all data
Vadillo et al.	Probable fixation locations	Transformation matrix

1.7 Objective

Researchers have addressed many of the issues presented by glasses by engineering new systems or making changes to the experimental procedure. These changes typically involve the repositioning of the IR light source in order to prevent the appearance of obstructing shadows or reflections. Although many researchers acknowledge refraction as a source of error, strategies for correcting this error have not been devised or investigated in a laboratory setting.

Existing research suggests that the most promising approach would be to examine the relationship between refractive strength and eye tracking signal using AEs in a controlled laboratory environment and then to apply methods from statistics and machine learning to isolate and eliminate the effect of refraction on the eye tracking signal. The statistical analysis can be divided in two tasks: In the first task, different pairs of glasses are recognized by eye tracking signal. For this, classification models will be used. In the second task, this information is used to correct incoming eye tracking data by applying regression models as in previous studies (Blignaut et al., 2014; Hornof & Halverson, 2002).

2 Data recording

2.1 Methods

2.1.1 Experimental setup. The aim of the experiments was to gather gaze data for different types of glasses in different regions of the screen. 18 glasses (see Table 2) and 15 calibration points (see Figure 9) were examined in two experimental recordings. The recordings differ mainly in the calibration points examined and glasses used. The second recording was conducted to check for the reproducability of the experimental setup and to evaluate the models' performance on unknown data (new calibration points and glasses). Both experiments were conducted using a head model with adjustable AEs, different pairs of glasses, a SMI red-m eye tracker, and a computer connected to two monitors (see Figure 6).

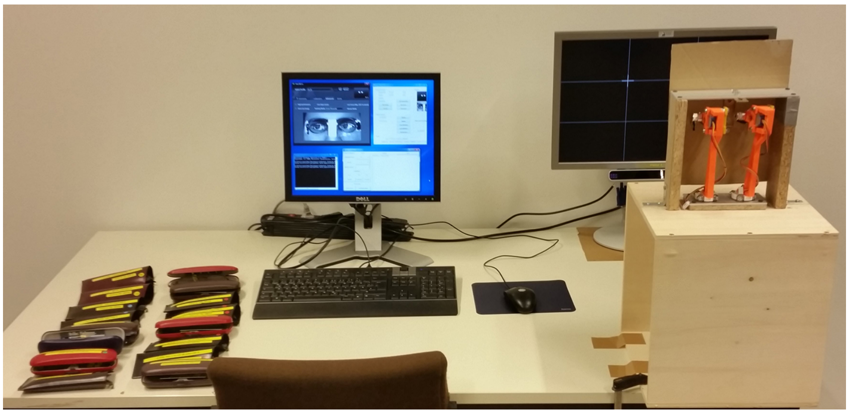


Figure 6. Setup for the first recording with 16 pairs of glasses (left), secondary screen with monitoring software (center), primary monitor with eye tracker (right), and head model mounted on a box (right)

Head model with artificial eyes. The head model features adjustable, blue AEs, a picture of a face printed on cardboard, and a nose for securing the glasses (see Figure 8). The AEs are held by ball bearings and a supporting structure of plastic (see Figure 7), which allow for free movement but also keep their position stable over time.

During the experiments, the eyes were positioned at a horizontal distance of 61 cm from the monitor (diagonal distance: 65 cm). The viewing distance for humans is dependent upon the size of the display and the type of visual task, however a minimum

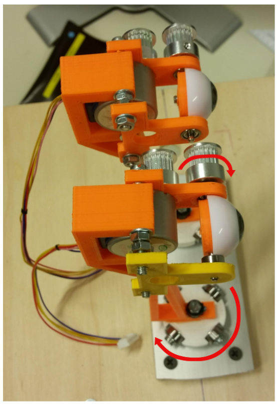


Figure 7. Supporting structure with adjustable AEs

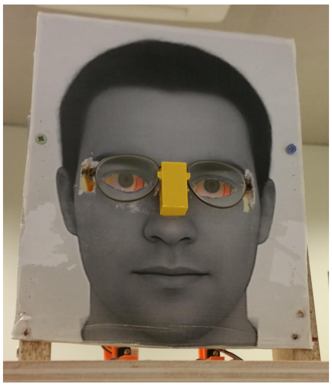


Figure 8. Head model with mounted glasses from the perspective of the eye tracker

of 50 cm is generally recommended (Kroemer & Kroemer, 2016). Tobii (2011) used a distance of 65 cm when testing their TX300 eye tracker. The gaze of the eyes was positioned 10 cm above the center of the monitor.

During the experiment, different pairs of glasses were mounted on the nose of the head model (see Figure 8). The glasses' temples and nose pads were removed prior to mounting. The 18 glasses used in the experiments are depicted in Table 2 (a more comprehensive overview can be found in Table 17 of the appendix). Glasses 1 through 16 were also used in previous, unpublished experiments conducted at the chair. They differ in lens type (multifocal, short-sighted, long-sighted, computer glasses, single-vision glasses), diopters ("-4" to "+5"), and other eyeglass-related parameters. All pairs with the exception of 17 and 18 feature the same rim (see Figure 27 in the appendix). Glasses 1 through 16 were used in the first experiment, whereas glasses 12, 13, 15, 16, 17 and 18 were used in the second. A wide range of glasses was chosen for the first experiment to build a good data set for future research. The most promising and stable glasses were chosen for the second experiment, with glasses 17 and 18 being added in order to test the generalization ability of the models (more in section 2.2.1). Glasses 17 and 18 are real glasses lent to the researcher by colleagues.

The cardboard face features a black-and-white picture of an adult male, which is a composite of 32 male faces (*Beautycheck - average faces*, n.d.). Two holes in the

Table 2

Overview of the utilized glasses

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Lens type	MF	MF	MF	MF	MF	MF	MF	MF	C	C	C	C	SV	SV	SV	SV	R	R
Diopters[a]	-3.00	-3.00	-3.00	-3.00	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	+1.00	+2.00	+5.00	-1.00	-4.00	-2.25 -1.50	+2.00 +3.00

Note. MF: multifocal lenses, C: computer glasses, SV: single vision lenses, R: real glasses

[a] single diopter values apply to left and right lenses. Double values specify left top and right bottom.

cardboard face offer free view to the AEs. The face allows for better detection of the head model by the eye tracker and hides the supporting structure of the eyes avoiding additional reflections from metal components (e.g. screws). According to Wang et al. (2017) the use of AEs in combination with a printed face is also used by eye tracking manufacturers. The nose used on the head model is a plastic mount produced specifically for the glasses. The small pocket and two wings allowed for easy mounting and dismounting of the glasses, and ensured that the position of the glasses remained fixed and consistent over all trials.

Hardware.

SMI red-m eyetracker. The SMI eye tracker was mounted under the monitor using a mounting bracket. The sampling rate was set to 120 Hz. The device is a dark pupil eye tracker that supports monitors from 10" to 22" (SMI, 2012). Users should sit between 50 to 75 cm from the eye tracker (SMI, 2012). The eye tracker has a reported gaze position accuracy of 0.5° and a spatial resolution (RMSD) of 0.1° (SMI, 2012). It is said to be compatible with most glasses and lenses (SMI, 2012).

Monitor with computer. A 22" Dell monitor with a width of 47.4 cm, a height of 29.6 cm and a resolution of 1680 px (width) × 1050 px (height) was used in the experiments. The computer used was a Dell Optiplex 9020 with Windows 7 Professional

Service Pack 1 (64-bit) running a i7-4790 CPU, 32 GB RAM and a NVIDIA GeForce GTX 745. 1 Pixel spans about 0.028 cm in this setup, or 0.025° visual angle ($1° = 1.13cm = 40.20px$). The origin ($0px/0px$) is in the top left corner of the screen.

Software. The experiment used the *iView* RED-m software provided by SMI, a C#-application for streaming the data, and a custom python application.

The *iView* software was used to specify the physical position of the eye tracker in relation to the monitor and to set calibration settings and calibration mode. The settings used are as follows (see Figure 28 in appendix):

- "Depth" (Horizontal distance between monitor and eyetracker): 47 mm
- "Height" (Vertical distance between monitor and eyetracker): 9 mm
- "RED Angle" (Angle between monitor and eye tracker): 22°
- Screen dimensions: 47.9 cm width, 29.9 cm height
- Tracking mode: Smart Binocular
- Calibration: 0 Point

The 0 Point calibration is an automatic calibration without measuring the physical characteristics of the participant (SMI, 2012). It "is not recommended for highest accuracy, but it is suitable for users who have difficulty achieving a successful calibration with 2, 5, or 9 points" (SMI, 2012, p. 76). The 0 Point calibration was chosen because the current state of the head model does not allow for a calibration using calibration points. The only method to tell the position of the AEs is by using the eye tracking signal[3].

The *iView* software displays an eye image of the tracked user's eye area with crosses depicting the center of the pupil (black) and the Purkinje reflection (white) (SMI, 2009) (see Figure 29 in appendix). An eye tracking monitor that illustrates the

[3] A variant which rotates the eyes with stepping motors is currently being worked on. This will allow the AEs to be moved to a fixed position (e.g. 20° to the right) without using the eye tracking signal as a source of information for the AEs' position.

position of the eyes, their tracking status (not found/tracked), and estimated distance is also displayed.

The *C# iView X API Sample* was used to stream data from the *iView* software to a local server so that it could be easily accessed by the python application.

A custom python application based on work from previous, unpublished experiments at the chair was programmed for the study. The application contains the logic for collecting the data streamed by the *C# iView X API Sample*, and hence that from the *iView* application. It also displays the calibration points (see Figure 9) along with the gaze position on screen. This allows the researcher to rotate the AEs to a position where the position of the gaze data overlaps with the position of a calibration point. Using a graphical user interface (see Figure 30 in appendix) programmed in Tkinter, various experimental parameters can be set and the recording can be started. When the recording is finished, the application saves a log file containing the gaze data.

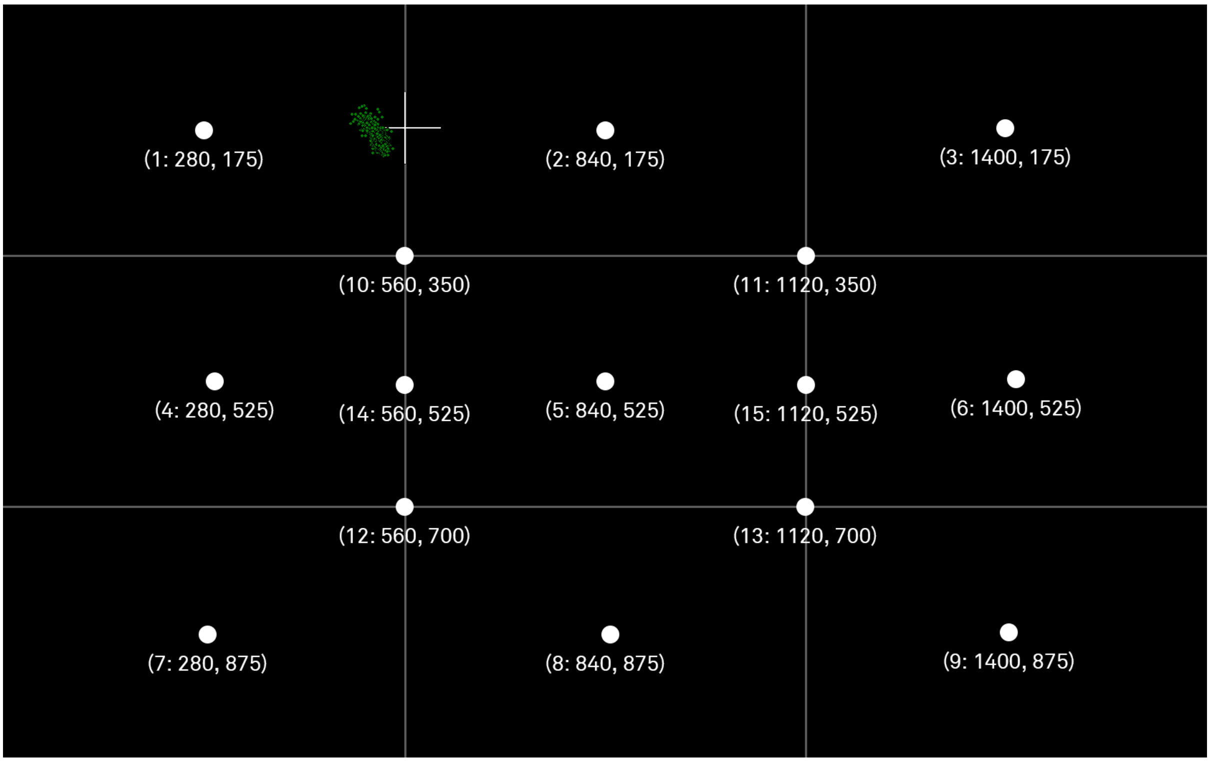


Figure 9. Position and coordinates of the 15 calibration points (own work). The white cross and green dots show the calibration cross and gaze data displayed during recording. Nomenclature: (index: X-coordinate, Y-coordinate)

In Figure 9 the 15 calibration points and their screen coordinates are depicted. Nine of the points (1-9) were also used in previous, unpublished experiments at the chair. Six (10-15) were added to study the effect of screen size more in depth. In the first experiment, calibration points 1-13 were recorded, while points 1, 3, 5, 8, 12, 13, 14 and 15 were recorded in the second. New calibration points (14, 15) were included in the second recording in order to investigate the generalization ability of the models. In the second experiment only 8 calibration points were examined (first experiment: 13) to keep the effort for data recording to a reasonable extent for this thesis.

2.1.2 Procedure of the experiment. Following each trial with glasses, a baseline without glasses was recorded to control for errors of drift or accidental shifting of the eyes. The procedure of the experiment can be described as follows:

1. Start software and initialize with correct parameters: *iView*, *C# iView X API Sample*, custom python application

2. Repeat for each calibration point:

   (a) Hide right eye, adjust left eye to calibration cross displayed on screen, save gaze point (L) with custom python application

   (b) Hide left eye, adjust right eye to calibration cross displayed on screen, save gaze point (R) with custom python application

   (c) Uncover both eyes, save gaze point (B) of both eyes with custom python application

   (d) Check if eyes are correctly detected by the *iView* software

   (e) Record baseline trial

   (f) Repeat for each pair of glasses:

      i. Mount glasses, record glasses trial

      ii. Unmount glasses, record baseline trial

   (g) Switch to new calibration point

Each trial lasted for approximately 30 seconds (inaccuracies in the range of 0.5 seconds could occur due to the different computational loads of the computer). At the end of a trial, a measure for the distance between eye tracker and AEs was saved. For the first experiment, the total number of trials was 429, i.e. 13 calibration points * (16 pairs of glasses + 17 baseline trials). The first experiment was conducted in three sessions on separate days (06-08.03.2017). The total number of trials for the second experiment was 104, i.e. 8 calibration points * (6 glasses + 7 baseline trials). It was conducted in one session on 31.05.2017. To minimize interference from direct sunlight, the experiments were carried out in complete darkness in a room with closed blinds.

The eyes were adjusted as good as possible. Occasionally, difficulties with the positioning of the eyes were encountered. In the event of such a difficulty, the position with the most stable recognition was chosen. The glasses were cleaned prior to mounting to avoid unwanted reflections. In some trials, reflections from the rims of the glasses or the face led to false detections of the pupil or Purkinje reflections. Such reflections were avoided through the application of black tape to the affected areas. In some cases, the trial was repeated. Efforts were made to ensure that the tape did not obstruct the visibility of the pupil.

The procedure of the experiment was slightly optimized in the second experiment. The gaze point (L, R, B) was no longer saved for each calibration point, for its assessment provided no value to the analysis.

2.1.3 Recorded data. The recorded data features X and Y coordinates over a 30 second interval for the 13 calibration points and 16 pairs of glasses. For the sake of brevity, trials are referred to by the coding scheme of c.g.bl, in which c represents the calibration point, g the pair of glasses, and bl the running number of the baseline trial.

Selection of glasses for analysis. The first experiment contained 16 different glasses with similar diopters. One pair of glasses was chosen for each diopter value available ("-4", "-3", "-1", "+1", "+3", "+5"). Therefore, for each trial with glasses the standard deviation for X and Y values was computed. These values should be below the reported gaze position accuracy of the SMI RED-m, 0.5° (see section

"Hardware"). Glasses with overall low standard deviations per trial should be suited best for the data analysis.

For the second experiment all recorded glasses were used for the analysis.

Data cleaning. The aim of the data cleaning phase was to achieve better data quality by stripping the data off "outliers." Outliers are "values that fall outside the normal range of measurements" (Holmqvist et al., 2011, p. 88). They "need to be handled with care, as they may exert a disproportionately large influence on the results of the final analysis" (Holmqvist et al., 2011, p. 88). While outliers may be legitimate rare observations, they may also be the result of errors in data recording (Holmqvist et al., 2011). In the latter case, Holmqvist et al. (2011) suggest that they be excluded or corrected, although there is no strict guideline as to how they should be addressed.

Outliers can be identified by examining standardized values and excluding all observations that exceed a certain range (Holmqvist et al., 2011). As a criterion, the outlier definition commonly used in box plots was chosen (Stocker & Steinke, 2017). The upper border was set to the sum of the third quartile and 1.5 interquartile range ($Q_3 + 1.5IQR$), the lower border to the difference between the first quartile and 1.5 interquartile range ($Q_1 - 1.5IQR$). These borders were calculated for each trial. An observation was excluded if its X or Y value exceeded these borders values.

To further improve data quality every trial in which standard deviation for either the X or Y values exceeded the 0.5° criterion, was examined visually and manually cleaned of artifacts.

2.2 Results

The data was imported using the statistics software *R*.

2.2.1 Selection of glasses for analysis.

One pair of glasses for each unique diopter value available (dpt: "-4", "-3", "-1", "+1", "+3", "+5") is chosen from glasses 1 through 16. Glasses 13-16 were chosen because they feature a unique diopter value. One pair of glasses from 1 to 4 (dpt: "-3") and one pair of glasses from 5 to 12 (dpt: "+1") has to be chosen. The standard deviations of all trials with glasses were computed and it was counted how often the standard deviation from a specific pair of glasses exceeded the criterion of 0.5° (see Figure 10). The pair of glasses exceeding the criterion in the least number of trials was glasses 1 for diopter value "-3" (trials: 2) and, glasses 6, 11 and 12 for diopter value "+1" (trials: 1). For glassses 12 the average standard deviation for X and Y was lower than for glasses 6 and 11 $(sd_X = 0.102°, sd_Y = 0.258°)$. Therefore, glasses 1 and 12 together with glasses 13-16 were chosen for the analysis (see points in red in Figure 10).

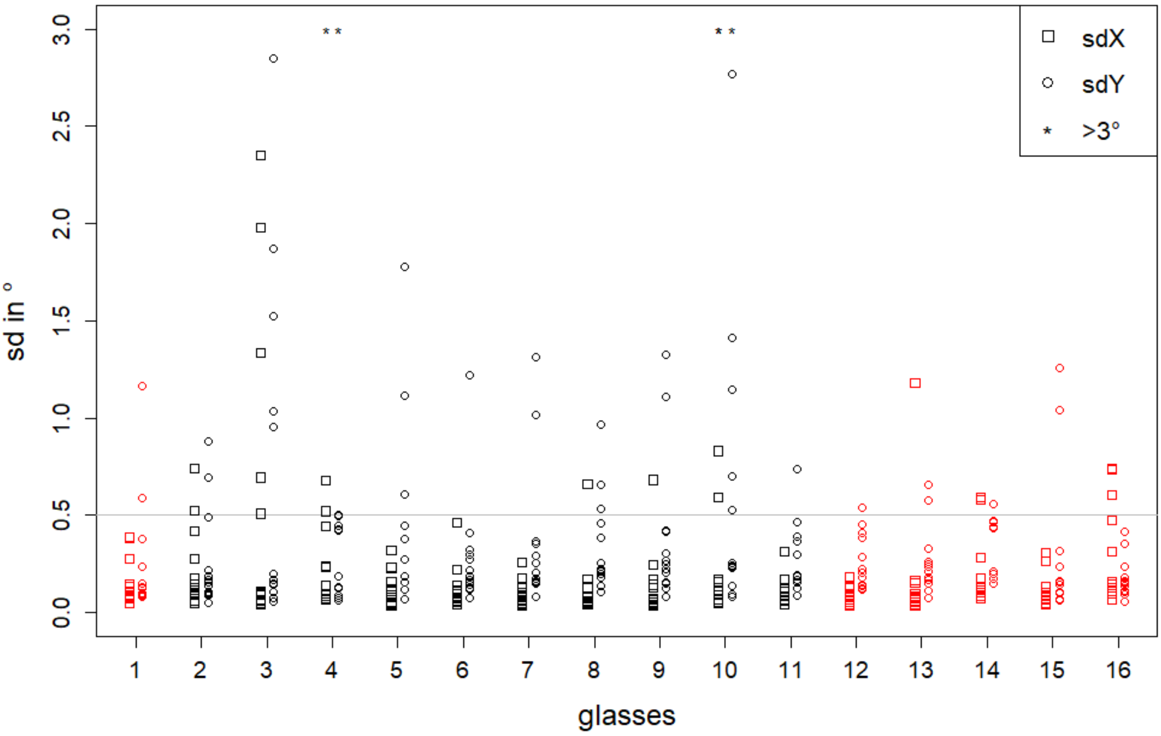


Figure 10. Standard deviations in ° for glasses 1-16 for X (squares) and Y (circles). Values higher than 3° are represented by an asterisk. The grey line depicts the 0.5° criterion.

2.2.2 Result of data cleaning. One source of outliers during the trials was false detection of the pupil and/or the first Purkinje reflection (as seen in Figure 11). Most of these false detections did only appear in a fraction of the trial recording time and differed noticeably in X and Y from the remainder of the data. As a result, the gaze data scatterplots often displayed two separate point clouds (see Figure 12).

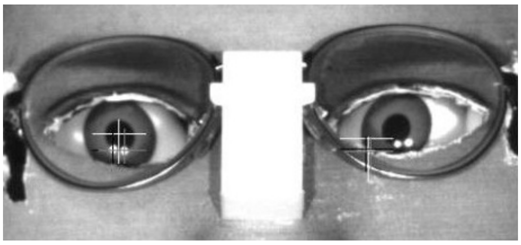


Figure 11. Screenshot of the last eye image from 2.14.14 (experiment 1) with a false detection of the right eye

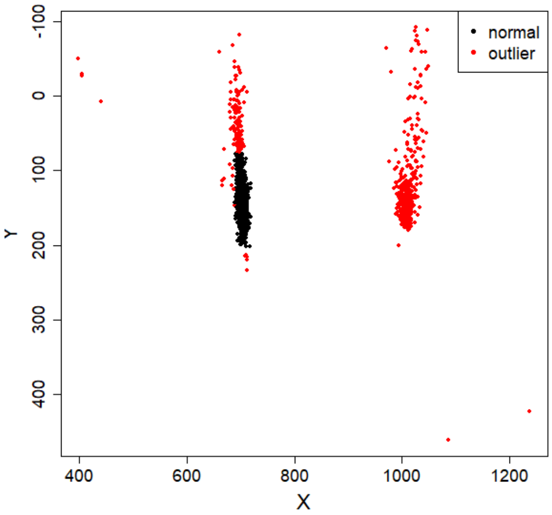


Figure 12. Scatterplot with outliers for trial 2.14.14 (experiment 1)

Manual cleaning. In the first recording, 21 trials showed an excessively high standard deviation in either X, Y, or both directions ($sdX/sdY > 0.5°$). These trials were examined more closely and attempts were made to clean them of unwanted artifacts. In five of the 21 trials (13.0.13, 7.1.1, 7.15.15, 9.13.13, 13.16.16), the high variance was caused by two clouds of points. In these cases, the cloud with fewer observations was excluded (e.g. 7.1.1 $Y < 1{,}000$ 792 obs., $Y < 1{,}000$ 1,839 obs.). In Figure 13, the results of the data cleaning and manual cleaning for trial 7.1.1 are presented. The dark red cloud of points in the middle ($X \sim 210; Y \sim 1040$) was excluded from the analysis during the manual cleaning phase. The light red cloud of points in the lower right corner of the plot ($X \sim 260; Y \sim 1150$) was excluded during the data cleaning phase. If the data could not be visually discriminated, no action was taken. In the second recording, 9 trials showed an excessively high standard deviation. Two of these trials were cleaned of outliers.

Finally, missing values and trials containing insufficient data ($<$ 10 observations, trials from the first experiment: 1.14.14, 3.14.14, 6.14.14) were excluded from the analysis.

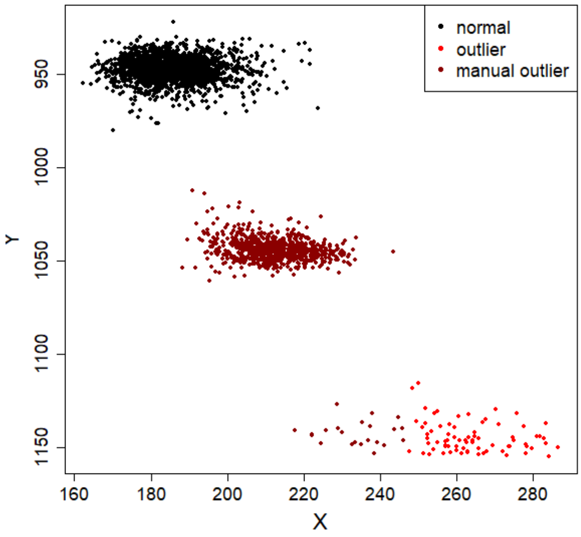


Figure 13. Scatterplot with outliers for trial 7.1.1

Excluded observations. The results of the data cleaning can be seen in Table 3. In the first experiment, 53.3% from a total of 1,026,841 observations were omitted, as they were the product of glasses that were ultimately not used in the analysis (see section 2.2.1).

Table 3

Number and percentage of excluded observations after data cleaning

	Missing data	Omitted glasses	Insufficient obs. in trial	Manual cleaning	Outlier	Rest
First experiment						
n	27,150	547,405	3	5,855	60,030	386,398
%	2.6	53.3	0.0	0.6	5.8	37.6
Second experiment						
n	2,610	0	0	955	15,161	227,328
%	1.0	0.0	0.0	0.0	6.1	92.4

Nine percent of the samples were excluded from the analysis due to missing data, insufficient observations in their trials, manual cleaning or identification as outliers. This left a remainder of 386,398 observations in 104 baseline trials and 75 glasses trials for the analysis. The number of observations per trial after the cleaning phase ranged from 306 to 2,952 (M = 2,159, sd = 370). The average standard deviation per trial was higher in direction Y ($sd_Y = 9.6px = 0.24°$) than in direction X ($sd_X = 4.6px = 0.11°$).

In the second experiment, even fewer samples than in the first experiment (7.6% vs. 9.0%) were excluded from the analysis, leaving 227,328 observations. All trials were used for further analysis. The number of observations per trial ranged from 1,689 to 2,984 (M = 2,366, sd = 423). The average standard deviation per trial was similar to that of the first experiment ($sd_X = 4.4px = 0.11°$; $sd_Y = 9.0px = 0.22°$).

2.2.3 Gaze data.

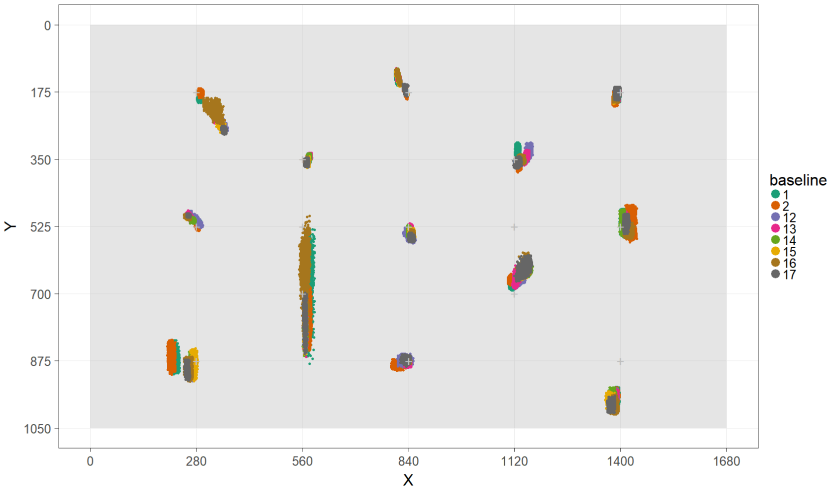


Figure 14. X and Y coordinates of the gaze data for each baseline trial

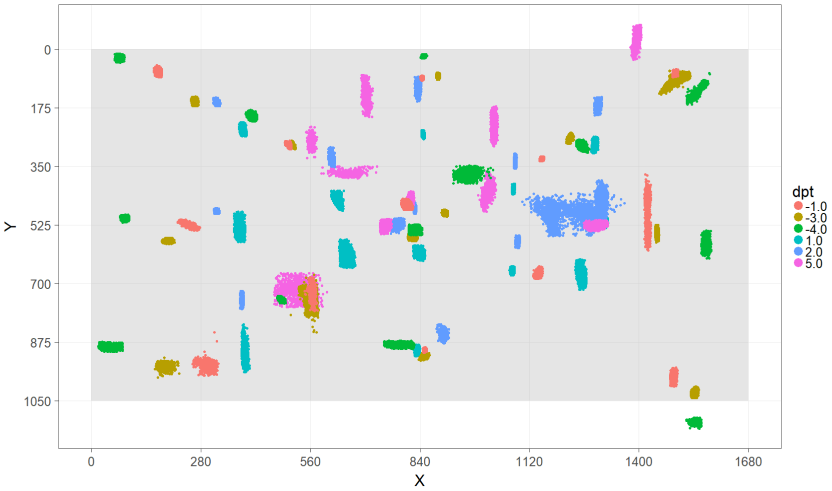


Figure 15. X and Y coordinates of the gaze data for each diopter

Training and testing data set. Figure 14 shows the gaze data for all baseline trials and Figure 15 shows the gaze data for all glasses trials. As the origin (X: 0 px/Y: 0 px) is located in the topleft corner of the screen the Y axis is inverted (Y top: 0 px, Y bottom: 1050 px) in Figure 15. The grey rectangle marks the area of the computer screen, grey crosses depict the position of the calibration points. Observations outside the area of the screen were reported to be either smaller than 0 or larger than 1,680 in X position or smaller than 0 or larger than 1,050 in Y position.

The baseline trials are closely positioned for each calibration point and their shape is similar to the calibration points at which they were captured. The position of the baseline trials has a small offset, as the adjustable AEs could not always be perfectly aligned to the original calibration points.

The glasses trials show a notably wider variation. The trials are clearly offset and show varying distributions (X_{sd} : 0.035° to 1.183°, Y_{sd} : 0.055° to 1.255°). Note that some data is positioned "off the screen" ($X \notin [0, 1680]$, $Y \notin [0, 1050]$).

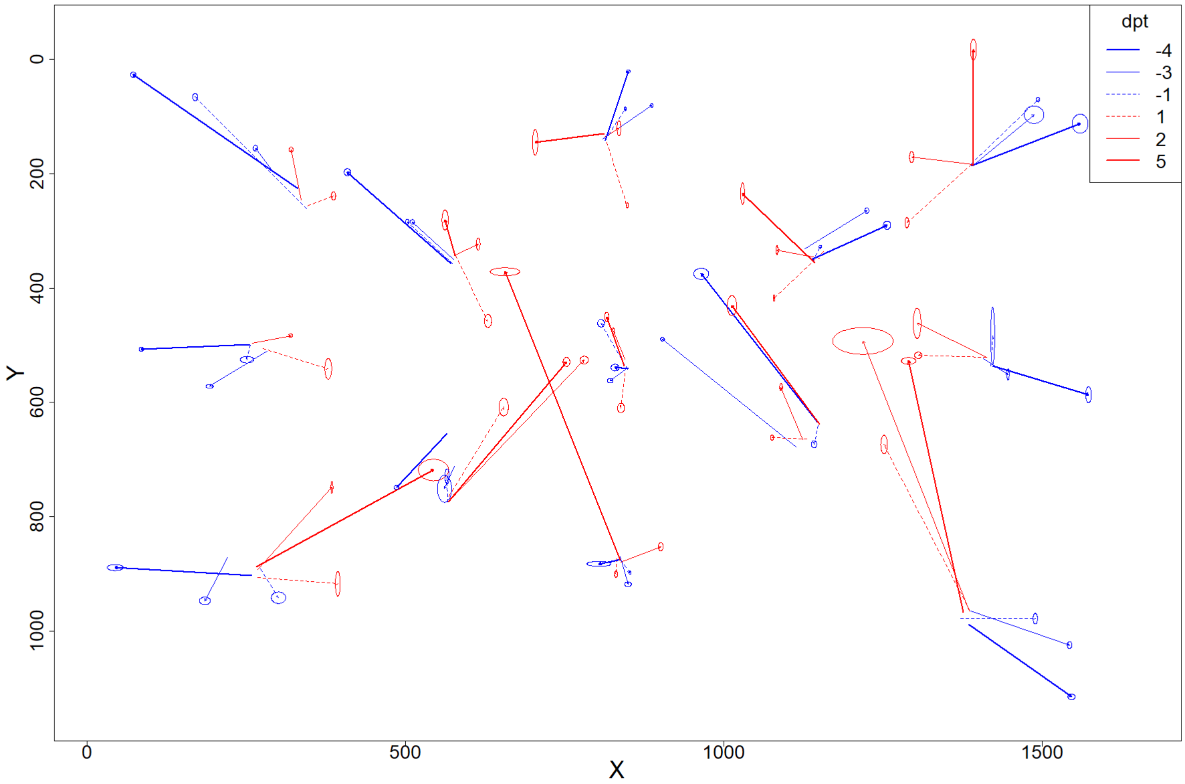


Figure 16. Glasses-induced offsets in training and testing data set

The glasses-induced offsets are illustrated in Figure 16. Each arrow represents the offset from a single trial. The blue arrow in big stroke located in the upper left-hand corner represents trial 1.16.16. The arrow is drawn from the mean position of the preceding baseline trial (here 1.0.16, $mean_X = 331$, $mean_Y = 226$) to the mean position of the glasses trial (here 1.16.16, $mean_X = 73$, $mean_Y = 27$). The ellipse at the end of the arrow represents the variance in the glasses trial (here 1.16.16, $sd_X = 4.5$, $sd_Y = 4.6$).

The offsets of the leftmost calibration points (1, 4, 7) are similar to the offsets of the rightmost calibration points (3, 6, 9), simply flipped vertically. The offsets appear to be smaller in the middle of the screen (cal: 2, 5, 8) than at the edges of the screen. A regression between the offset (absolute value of X and Y combined) and the distance from the calibration point showed the following form for angular distances ($F_{1,73} = 6.903, p = .01, R^2 = .087$):

$$off = 1.57 + 0.17 * dist \tag{6}$$

This means that at a distance of 0° (in the very center of the screen), the predicted offset (off) is 1.57°, growing by .17° every 1°. At the largest distance (16.4°, i.e. calibration points 1, 3, 7, 9), the regression predicts an offset of 4.4°.

Validation data set. In Figure 17, the centroids of each glasses trial from experiment 1 (triangles) are depicted together with the centroid of each glasses trial from experiment 2 (circles). Grey crosses depict the calibration points. One standard deviation is shown as a grey dashed line from the centroid of the trial in the X and Y directions. Only the calibration points and glasses from experiment 1 that also appear in experiment 2 are shown (g: 12-16, cal: 1, 3, 5, 8, 12, 13).

The glasses-induced offsets are shown in Figure 18. Note that only the calibration points 1, 3, 5, 8, 12, 13, 14 and 15 were recorded. Additionally, it should be noted that the highest positive diopter in the validation data set is "+3" (compared to "+5" in training and testing data set) and that "-3" has been replaced with "-2".

The new calibration points (14 center left, 15 center right) show a similar pattern to calibration points 12 (lower left) and 13 (lower right). Offsets for positive

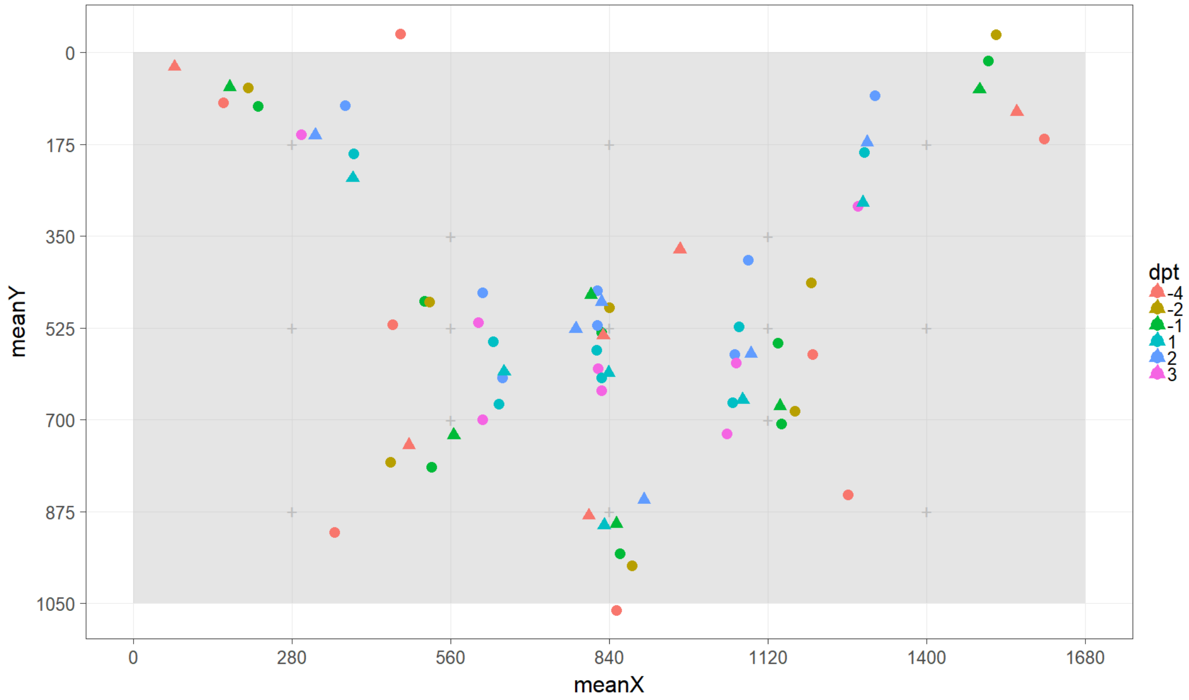


Figure 17. *X* and *Y* coordinates of the gaze data for all glasses trials in experiment 1 (triangle) and experiment 2 (circles)

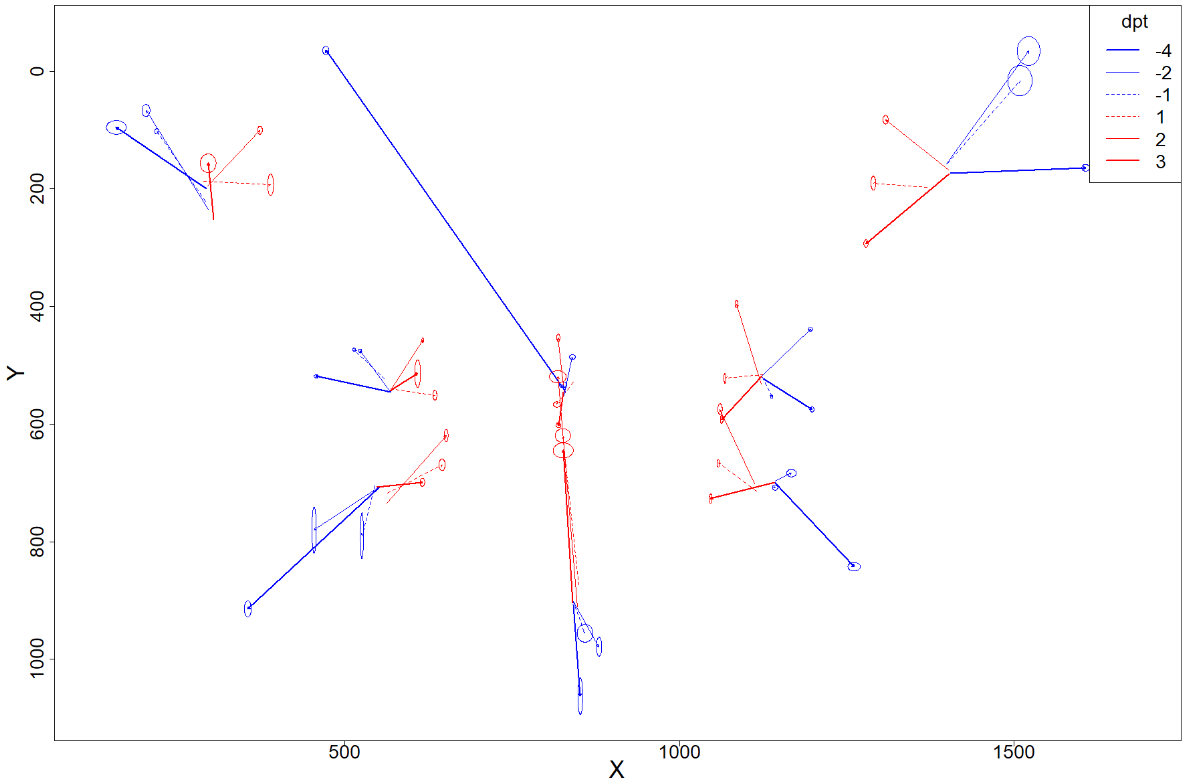


Figure 18. Glasses-induced offsets in validation data set

diopters tend to point towards the center. "+2" points more towards the north of the screen, "+1" towards the center, and "+3" towards the south of the screen. The negative diopters generally point outward, but the single diopters do not follow a pattern or form of a similar relationship. Moreover, several noticeable high offsets in magnitude in the validation data set can be observed: The "-4" diopter offset from calibration point 5 right in the center and the offsets of positive diopters from calibration point 8 pointing toward the center.

These high offsets in magnitude are especially interesting when compared to the offsets in the training and testing data sets (see Figure 16). They are not reported in the training and testing data sets at the corresponding positions (cal: 5, 8). The general tendency, however, seems similar: Offsets caused by positive diopters pointing toward the center of the screen and offsets caused by negative diopters pointing to the edges of the screen.

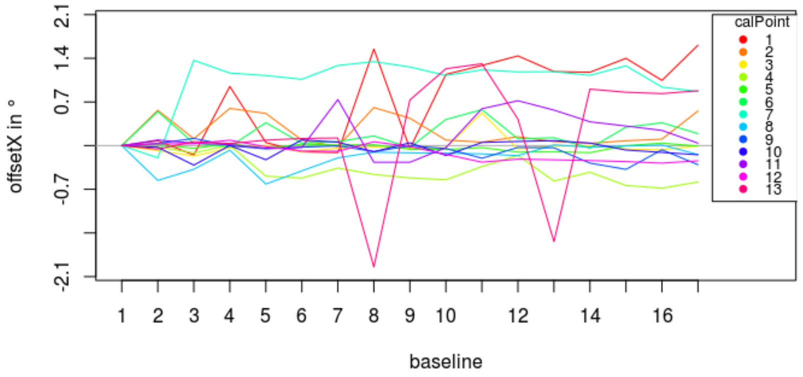


Figure 19. Shift in X-coordinate in baseline trials per calibration point (colors). The black line depicts the X-coordinate of the first baseline.

Baseline shift. As mentioned in section 2.1.2, a baseline trial was recorded in between each pair of glasses to control for errors of drift and accidental shifting of the eyes. This shift can be seen in the recorded data from the first experiment (see Figures 19 and 20). Based on visual inspection, it can be concluded that there is no systematic shift over time in the data. For most calibration points, the shift in X and in Y direction is random and small in magnitude (X: $Q_1 = -0.10, Q_3 = 0.14$, Y:

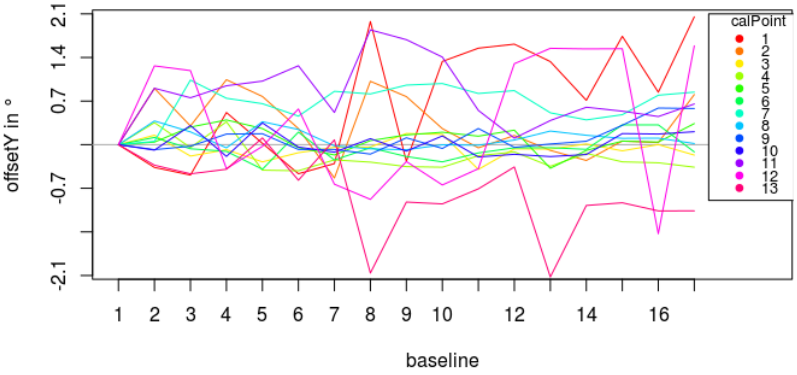


Figure 20. Shift in Y-coordinate in baseline trials per calibration point (colors). The black line depicts the Y-coordinate of the first baseline.

$Q_1 = -0.14, Q_3 = 0.40$), and therefore negligible. Addionally, there is probably no systematic influence of the used glasses, e.g. the baseline does not shift more strongly after a specific pair of glasses is tested. Instead, the precision remains at a constantly high level, with an average standard deviation of $4.5px$ (0.11°) in X and $11.0px$ (0.27°) in Y direction for the baseline trials. Only for calibration point 12 does the precision drop to $11.3px$ (0.29°) in X and $41.3px$ (1.03°) in Y direction.

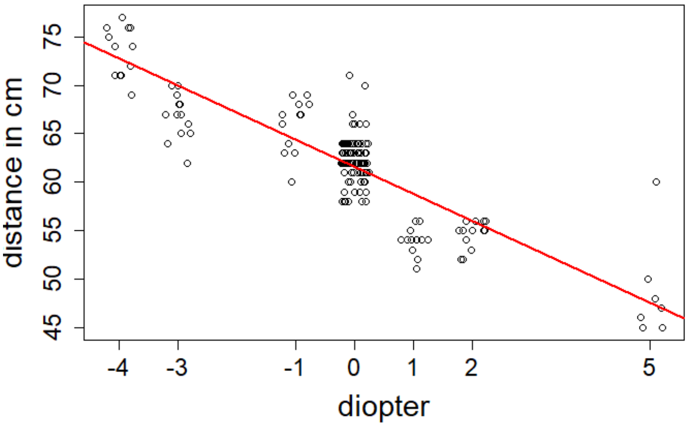


Figure 21. Estimates of distance depending on diopters. The red line shows a regression.

Distance measures. The distance measures show an overestimation for negative diopters and an underestimation for positive diopters. The eye tracker using the size of the pupil for estimations of distance could be a possible cause for this effect. The pupil appears magnified in the eye image for positive diopters and demagnified for negative diopters (see Figure 31 in appendix).

The relation between diopters and distance can be examined in Figure 21. A regression has been fitted to the observations and shows the linear relationship between distance and diopters (measured distance: 61 cm, regression: $distance = 60.4 - 3.2 * dpt$). Observations have been slightly jittered on the X-axis to avoid overplotting.

3 Classification

3.1 Methods

3.1.1 Aim. In a classification task, a model specifies which of various categories ("classes") an input belongs to (Goodfellow, Bengio, & Courville, 2016). The aim of the classification task in this thesis is to classify labeled data points and build a system for classifying new data points. This classification task is intended to solve a multi-class problem, i.e. classify an input as only one of several non-overlapping classes (Sokolova & Lapalme, 2009). This is a supervised learning problem, as both input (X, Y) and output data (*dpt*) are available. The aim of supervised learning is "to learn a mapping from the input to an output whose correct values are provided by a supervisor" (Alpaydin, 2010, p. 11). By contrast, in unsupervised learning only input data is available and the task is to find regularities in the input variable (Alpaydin, 2010). It should be noted, that no assumption about the distribution of the data can be made. Data distribution is most likely dependent upon the noise inherent in the eye tracker and the refraction of the glasses, although other factors may also play a role (mapping function, head model).

3.1.2 Features and label. An input consists of multiple numeric values called features (e.g. weight and height of a person). These features are then used to predict the membership of a data sample (e.g. a person) to a class, also called label (e.g. man/woman).

The features have to be chosen such that they can be applied easily in real-world situations. Therefore, features should be easily computable in a running system and only include data that is sent by the eye tracker in real time. This would allow an interactive application (e.g. ticket vending machine) to utilize the data as user input. With these limitations and requirements in mind, the following features have been chosen:

- A reference point for the actual gaze point: For glasses trials, the reference point is constructed by taking the mean of the X and Y values of the baseline that was recorded right before the trial (e.g. trial 1.2.2 -> bl trial 1.0.2). For baseline trials it is constructed by taking the mean of all X and Y values in the trial itself. The average of the corresponding X values creates the feature ref_X, while the average of the corresponding Y values creates the feature ref_Y.

- Offsets in X and Y direction: The offsets represent the deviation from the actual gaze point and are directly related to the reference point. They are computed by subtracting the raw X (and Y) value from the X (and Y) value of the reference point (ref_X, ref_Y). This creates the following two features $offset_X$ and $offset_Y$.

The reference point is the same for all observations in a trial, for it is assumed that the physical setup of the AEs, glasses and eye tracker remains constant over the course of a single trial.

As mentioned in the section "Induced noise by eyeglasses", the extent of refraction is largely determined by the refractive strength of the material. Thus, the numeric value in diopters ("+5", "+3", "+2", "+1", "0", "-1", "-2", "-3", "-4") is chosen as a label. For glasses 17 and 18 an approximate value from left and right eye (17: -2.25/-1.50 → "-2"; 18: +2/+3 → "+3") is chosen as a label.

3.1.3 Choice of algorithms. The following algorithms can be employed to solve the classification task: K nearest neighbors (KNN) classifiers, support vector machines (SVMs), logistic regressions (LOGRs), naive Bayes classifiers and multilayer perceptrons (MLPs).

K nearest neighbors. The KNN was introduced by Fix and Hodges (1989) and is a nonparametric[4] method for classification and regression. For a new, unclassified observation the classes of the KNN from the training set are considered and the new observation is classified as the most common class among these neighbors (Denoeux, 1995). The KNN algorithm is quite popular in the pattern recognition community due to its good performance in practical applications (Denoeux, 1995).

[4] "Parametric models learn a function described by a parameter vector whose size is finite and fixed before any data is observed. Non-parametric models have no such limitation" (Goodfellow et al., 2016, p. 115).

Support vector machines. SVMs are supervised learning algorithms. They excel at the classification of input data representing highly complex linear relationships. SVMs transform this input data "to a high dimensional feature space in which the input data becomes more linearly separable compared to the original input space" (Olson & Delen, 2008, p. 111). Thus, it is also possible to work with highly complex nonlinear relationships in the data (Olson & Delen, 2008). High dimensional feature spaces are computationally costly, but kernel functions allow SVMs to operate efficiently (Goodfellow et al., 2016). After the transformation, the maximum-margin hyperplane, which separates the classes in the training data, is constructed. This is done by maximizing the distance between the hyperplane and the nearest observation from each class (Olson & Delen, 2008). Although, SVMs suffer from high computational costs of training when the data set is large (Goodfellow et al., 2016), they show highly competitive performance in various real-world applications, including medical diagnosis, bioinformatics, face recognition, image processing and text mining (Olson & Delen, 2008). They are one of the most popular, state-of-the-art tools for knowledge discovery and data mining (Olson & Delen, 2008). In the eye tracking domain, Rello and Ballesteros (2015), for example, used SVMs to detect readers with dyslexia using eye tracking data.

Two variants are used in this thesis: A linear support vector machine (LSVM) with a linear classifier, and a nonlinear SVM with a nonlinear kernel, abbreviated simply as SVM.

Logistic Regression. LOGR can be used for the estimation of non-continuous data (Olson & Delen, 2008). It is derived from the logistic sigmoid function $\frac{1}{1+e^x}$ (Goodfellow et al., 2016). In its simplest form, it can produce an output as a binary class "0"/"1", but is also capable of tackling multi-class problems ("multinomial logistic regression").

Gaussian Naive Bayes. The Gaussian Naive Bayes (GNB) classifier is a simple probabilistic classifier based on Bayes' theorem (*Naive Bayes classifier*, 2017). Its key assumption is that, conditioned on the class, the distributions of the input variables are independent (Bishop, 2006). When given an observation's feature vector, it computes the conditional probabilities for each class. These probabilities are computed as frequency counts using a "master" decision table (Olson & Delen, 2008).

The key assumption of this model is a strong one and may degrade performance due to poor representations of the class-conditional densities (Bishop, 2006). Nevertheless, the model may still demonstrate good performance in practice if the assumption is not satisfied (Bishop, 2006; Olson & Delen, 2008).

Multilayer perceptrons. MLPs are feedforward neural networks. Their purpose in classification is to approximate a function, that maps an input X to a category y (Goodfellow et al., 2016). Feedforward networks form the basis of many important commercial applications (e.g. object recognition, natural language applications) (Goodfellow et al., 2016). Goodfellow et al. (2016) state that the deep learning renaissance in present times began when Hinton, Osindero, and Teh (2006) demonstrated that a neural network could outperform the SVM on the MNIST benchmark (large database of handwritten digits).

Overview. An overview of the chosen algorithms, with underlying assumptions and short explanations as to why they were selected, is provided in Table 4.

3.1.4 Finding the best model configuration.

Hyperparameters. Machine learning algorithms often feature one or more hyperparameter (e.g. k in KNN) that influence their prediction. Hyperparameters are settings that control the behavior of a learning algorithm (Goodfellow et al., 2016). The goal is to find a suitable value for the hyperparameter, so that it provides accurate prediction without overfitting the data. Several values must be tested (see section "Grid search") and the best configuration is chosen to breed the final model.

Cross-validation. "In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, one can use a methodology called k-fold cross validation." (Olson & Delen, 2008, p. 141). In a k-fold cross-validation, the data set is divided into equal-sized k parts. One set is used as the test set and the other $k-1$ sets as the

Table 4

Overview of the chosen algorithms

Models	Parametric	Assumptions	Selection decision
KNN	no	Close samples are similar	Popular algorithm with good performance in practical applications
SVM & LSVM	no	No distributional assumptions (Lee, 2010)	State-of-the-art tool for real-world applications
LOGR	yes	Absence of perfect multicollinearity[5] (Allison, 2012)	Simple algorithm for multi-class classification
GNB	yes	Values associated with each class are distributed according to a Gaussian distribution.	Highly scalable and efficient classifier
MLP	no	No distributional assumptions	Widely used in commercial applications

training set. This is done k times, so that every part of the data set is used for training and testing (Alpaydin, 2010). k is typically 10 or 30 (Alpaydin, 2010), but Olson and Delen (2008) mention that 10 seems to be the most common and optimal number for folds. For training and validating the algorithms, a 10-fold cross-validation is used. A fixed seed is used to allow for reproduction.

Grid search. To find the optimum hyperparameters for the models Alpaydin (2010) suggests the use of factorial design ("grid search"), in which "factors are varied together, instead of one at a time."

KNN. The following values for k are tested: $k = 1^2, 2^2, ..., 35^2$. The square root of the number of samples (n) is said to be a good value for k. For this reason, the range $1 \leq \sqrt{n} \leq 2\sqrt{n}$ will be examined ($35^2 = 1225 \sim 1244.4 = 2\sqrt{n}$). k can be in the range of 1 to $n - 1$.

SVM. Hsu, Chang, and Lin (2003) suggest using the radial basis function kernel, commonly abbreviated as RBF kernel, for SVMs. Two hyperparameters will be optimized during grid search: C - a penalty parameter - and γ - a kernel coefficient. Good parameters for C and γ can be determined by trying an exponentially growing sequence for $C = 2^{-5}, 2^{-3}, ..., 2^{15}$ and $\gamma = 2^{-15}, ..., 2^{3}$ (Hsu et al., 2003).

SVM / LOGR. Additionally, LSVM and LOGR models were examined with the same values for C as for the SVM.

GNB. A grid search is not needed for the GNB because it features no hyperparameters.

MLP. The hyperparameters to be tuned for the MLP are the number of hidden layers and the number of neurons in those layers. A MLP is able to solve linear problems without hidden layers. As a pure linearity of the problem cannot be assumed in the data, the number of hidden layers is set to 1. A further examination of additional hidden layers is outside the scope of this thesis.

The following sequence of neurons in the hidden layer was examined: $i = 1, 2, 4, 8, 16, 32, 64$. *Adam* was used as an optimization algorithm following a recommendation from sklearn for large data sets (sklearn, n.d.).

3.1.5 Testing and evaluating.

Data sets. The training of the models for the classifier was carried out with python 3.6 and the sklearn-library (Pedregosa et al., 2011). The complete data set ($n =$ 386,398) was separated into training and test sets. The training set was used for cross-validation and grid search, and later for the training of the final model. For the training set, 80% of all data samples ($n =$ 309,118) were randomly selected. The test set contained the remaining 20% of the data set ($n =$ 77,280). For evaluation of the final model, the test and validation sets were used, guaranteeing that the final model is evaluated using data, which it has not previously been trained. The validation set was exclusively composed of data from the second experiment ($n =$ 227,328). In contrast to the training and test set, it also contained gaze data from "new" calibration points and glasses. This was to ensure that the models' ability to generalize on new data is tested.

The i.i.d assumptions have to be made for the training, test, and validation sets, meaning that the samples in each data set are independent from each other, and that all three sets are identically distributed and drawn from the same probability distribution (Goodfellow et al., 2016). Furthermore, many machine learning algorithms expect data in a certain format (SVM (Hsu et al., 2003), KNN (Alpaydin, 2010)). The data format also decreases training time and increases performance. Therefore, the data used here was generally z-standardized on the training data, except when otherwise stated (e.g. Naive Bayes). The training of the algorithms did mostly differ in the choice of hyperparameters that were examined in the grid search.

Performance measures. The measures utilized to rate the performance of a model are accuracy, precision[6], recall and the F-score.

The overall accuracy of a model "is estimated by dividing the total correctly classified positives and negatives by the total number of samples" (Olson & Delen, 2008, p. 138). It represents the "average per-class effectiveness of a classifier" (Sokolova & Lapalme, 2009, p. 430). Precision is "the number of correctly classified positive examples divided by the number of examples labeled by the system" (Sokolova & Lapalme, 2009, p. 430). Recall is "the number of correctly classified positive examples divided by the number of positive examples in the data" (Sokolova & Lapalme, 2009, p. 430). The F-score is a combination of precision and recall.

For example, a precision of .6 in our class "+2 dpt" would show that 60% of all observations identified as "+2 dpt" by the system were correctly classified. A recall of .6, on the other hand would show that 60% of all "+2 dpt" observations were correctly classified. The F-score unites these two measures. For the sake of clarity, it is the decisive factor for model performance in this thesis and will be primarily reported. Values for precision and recall can be found for each class in the confusion matrices in

[6] Accuracy and precision in the context of classification models are not to be mistaken with the definition introduced in section 1.4.1. For raw eye tracking data the accuracy and precision are values derived from the "physical" position of the recorded data points (px, °). In this context they describe the performance of a classification model in terms of correct/incorrect classifications and are in the range of 0% to 100%.

the digital appendix.

Overfitting and underfitting. Goodfellow et al. (2016) claim that, the central challenge in machine learning is good performance on new, previously unseen inputs, rather than simply on those, which the model was trained. This is called generalization. Thus, the algorithm must not only be optimized for a training set, but also capable of performing well on an independent test set. This is achieved by maximizing performance on both the training and the test sets. Underfitting occurs "when the model is not able to obtain a sufficiently low error value on the training set" (Goodfellow et al., 2016, p. 111), while overfitting occurs "when the gap between the training error and test error is too large" (Goodfellow et al., 2016, p. 111).

Referencing mapping functions, Cerrolaza et al. (2008, p. 260) state that "it is common practice to make use of the most complete mathematical expression available. There is, however, no statistical basis for this practice. Furthermore, the systematic inclusion of terms to the mapping expression can lead to overfitting, a consequence which is too often ignored". When evaluating the performance of different mapping functions, they also discovered that higher order polynomials did not noticeably enhance accuracy (Cerrolaza et al., 2008).

To address the issue of under- and overfitting, the performance of the models is measured using the training, test, and validation sets. This provides necessary insights into performance on known data and generalization ability on unknown data (validation). The general aim is to achieve a proper trade-off between model complexity and model performance.

3.1.6 Dependence of features and label. The dependence of the four features and the label can be examined in a correlation matrix (see Table 5). The medium correlation between $offset_X$ and $offset_Y$ ($r = .24$) is notable. Higher offsets in the X direction generally have higher offsets in the Y direction, and vice versa. The medium negative correlation between $offset_Y$ and the label ($r = -.302$) should also be mentioned. This means that an increase in *dpt* is associated with a decrease in vertical offset ($offset_Y$).

Table 5

Correlation matrix of the features and label

	ref_X	ref_Y	$offset_X$	$offset_Y$	dpt
ref_X	-				
ref_Y	-.034	-			
$offset_X$	-.028	.066	-		
$offset_Y$	-.096	-.112	.240	-	
dpt	.031	.033	.050	-.302	-

3.2 Results

The results of the grid search with 10-fold cross-validation and performance measures for training, testing, and validation for each algorithm are reported in the following subsections.

3.2.1 *K* nearest neighbors.

Grid search. All models showed very high accuracies ranging from .97 to 1.0 for training accuracy and from .97 to .99 for test accuracy (see Figure 22). The model with $k = 4$ (surrounded by grey box) had the highest accuracy on the test set.

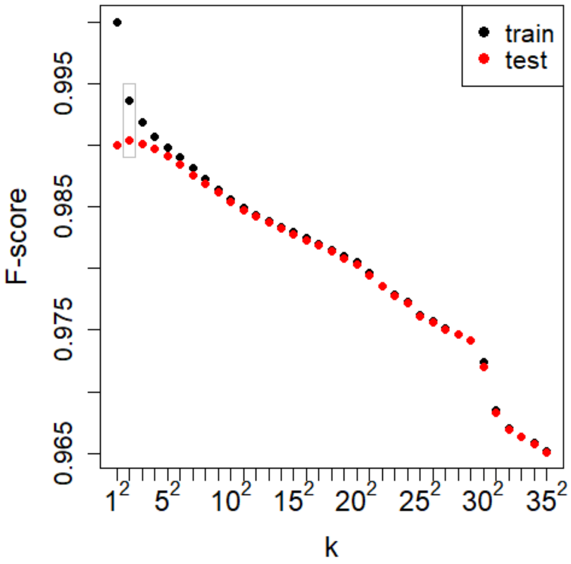


Figure 22. Performance of the KNN during grid search on training and test set depending on *k*

It should be noted that the training and test sets were derived from the "original" training set. In a 10-fold cross-validation, 90% of the data from the "original" training set is taken for training and 10% for testing. The reported values were averaged over the 10 cross-validation values.

Final model. The final model ($k = 4$) achieved high overall F-scores on the training (1.00) and the test sets (.99) and high F-scores for each class on the test set (.96-1.00, see Table 6).

Table 6

Performance (F-score) of the KNN on the test and validation set for each class and overall

dpt	-4	-3	-1	0	+1	+2	+5	total
test	1.00	.99	.96	.99	1.00	1.00	1.00	.99
validation	.38	-	.13	.91	.55	.16	-	.69

Performance on the validation set was split into two parts because the classifiers are unable to predict a class label they have not been trained with (classification vs. regression problem). The glasses that the algorithm has been trained with (12, 13, 15, 16) are referred to as "old glasses" opposed to the glasses the algorithm has not been trained with (17, 18) which will be referred to as "new glasses". The final model reached an overall F-score of .69 on the validation set for old glasses. Performance was highest for dpt "0" and lower for "-4", "-1", "+1" and "+2" (see Table 6). When the performance for "0" is not considered, overall performance drops to .31.

For new glasses, the model classified the "-2" and "+3" in the old class labels. If the model generalizes well, it would classify them as classes that are close in diopter ("neighboring classes"). So for class "-2" classifications as "-3" and "-1" would be desirable and for class "+3" classifications as "+2" would be desirable. Of the 18,597 "-2" samples, 45% were classified as a neighboring class ("-3": 29%, "-1": 16%). The remaining 55% were classified as "-4" (14%), "0" (22%), "2" (18%) and rarely as "5" (.02%). Fourteen percent of the "+3" were classified as "+2". Forty and 45% were classified as "0" and "+1".

3.2.2 Support Vector Machine.

Grid search. The results of the grid search for the SVM can be seen in Figure 23. Performance was highly dependent upon the hyperparameter configuration. The ideal hyperparameter configuration was $C = 32{,}768$ and $\gamma = 2$ (test-F-score final configuration: .99).

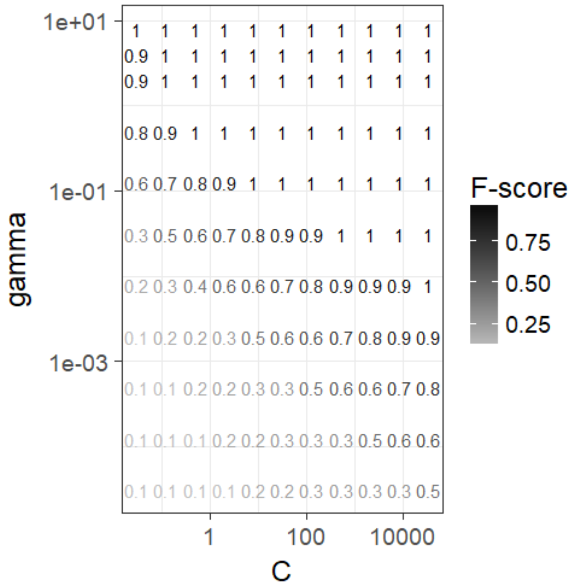


Figure 23. Performance of the SVM on the test set depending on C and γ

Final Model. The final model reached high F-scores for all classes on the test and training sets (.98 - 1.00), averaging 1.00 for both sets.

The performance on the validation set was lower than on the training and test set, with an F-score of .54 for old glasses. Predictions were mostly accurate for class "0" (.77), but less accurate for the other classes ("-4": .00, "-1": .17, "+1": .09, "+2": .27). Performance drops to .17 when only glasses (not "0" dpt) are considered.

The results of the new glasses are as follows: Twenty-three percent of the "-2" samples were classified as neighboring classes ("-3": 14%, "-1": 9%) and 34% of the "+3" samples were classified as the neighboring class ("+2").

3.2.3 Linear Support Vector Machine.

Grid search. Results for the grid search are shown in Figure 24. Performance was generally weak for F-scores ranging from .16 to .21. The best configuration ($C = 8$) achieved an F-score of .209.

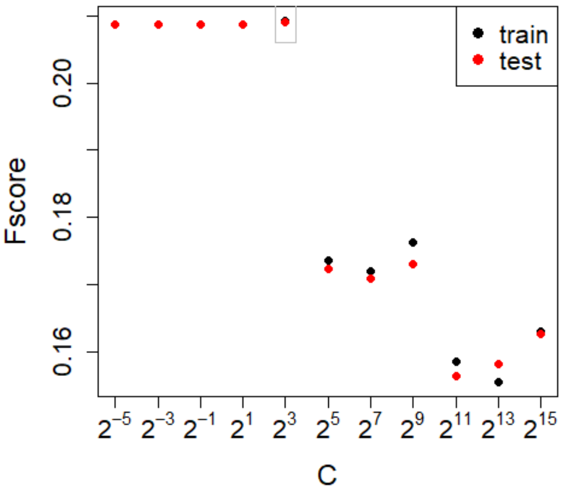


Figure 24. Performance of the LSVM on training and test set depending on γ

Final model. After the final training with more data, performance reached an F-score of .48 on the training as well as the test sets. Performance was highest for classes "0" (F-score: .77) and "5" (F-score: .53). The model was unable to classify all other classes (.00). Performance on the test set drops to a low F-score of .07 if only glasses trials are considered (excluding diopter values of "0").

Performance on the validation set showed a similar pattern, with a F-score of .49 (without "0": .00). Non-zero performance could only be reported for class "0" (.79).

New glasses in the validation set were completely misclassified ("-2": 79%"0" + 21% "+5", "+3": 85% "0", 15% "+5").

3.2.4 Logistic Regression.

Grid search. The LOGR showed low performance in the grid search. F-scores fluctuated around .21 ± .01, with the best configuration ($C = 32$) reaching an F-score of .212 on the test set.

Final model. After the final training, the final configuration ($C = 32$) reached an F-score of .49 on the training and test sets. The algorithm showed moderate performance for classes "-4", "0", and "+5", but failed in the prediction of classes "-3", "-1", "1", and "2" (see Table 7). When the "0" dpt class is not considered, performance on the test set drops to an F-score of .10.

Table 7

Performance (F-score) of the LOGR on the test and validation set for each class and overall

dpt	-4	-3	-1	0	+1	+2	+5	total
test	.19	.00	.00	.77	.00	.00	.52	.62
validation	.00	-	.00	.79	.00	.00	-	.49

The LOGR showed a similar result to that of the LSVM on the validation set: A medium F-score with the old glasses (.49, without "0": .00) caused by high performance for "0" (.79), but low performance for the other classes (.00). The new glasses ("-2", "+3") were misclassified as "0" and "5", respectively.

3.2.5 Gaussian Naive Bayes classifier.

The GNB does not include any hyperparameters, and as such, no grid search was performed. The GNB showed medium performance over all classes (F-score: .11 – .49) and good performance for dpt "0" (F-score: .91) in the training and test sets (see Table 8). This resulted in an overall F-score of .66 on the test set (without "0": .30).

Table 8

Performance (F-score) of the GNB on the test and validation set for each class and overall

dpt	-4	-3	-1	0	+1	+2	+5	total
test	.48	.13	.25	.91	.38	.11	.48	.66
validation	.5	-	.18	.96	.00	.25	-	.68

Medium performance of .68 for the old glasses was achieved on the validation set (without "0": .22). Highest performance was seen in classes that also showed the highest performance during training ("0", "-4"), while the remainder of the classes showed medium to low performance (see 8). Samples from "-2" were classified 35% as a neighboring class ("-3" 5%, "-1": 29%). Samples from "-3" were loosely classified as every other class (.03 - .29), except for "-4" and the target class "+2".

3.2.6 Multilayer perceptron.

Grid search. The MLP demonstrated increasing performance as hidden layer size increased (see Figure 25). The best configuration ($h = 64$) had a high F-score of .98.

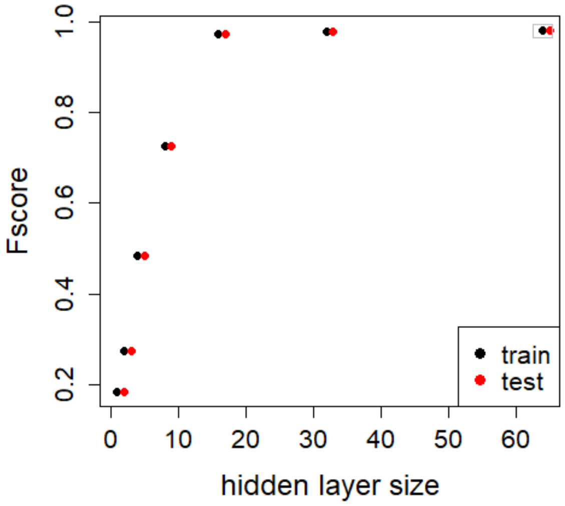


Figure 25. Performance of the MLP on training and test set depending on the hidden layer size

Final model. In the final training performance could be increased even further - compared to the results in the grid search - to reach an F-score of .99 on the training and test sets. F-score values for the individual classes ranged from .93 to 1.00 for the training and test sets.

The MLP showed good performance on the validation set for old glasses (.73, without "0": .38). The classes "0" and "-4" showed the best performance (.94, .60). The other classes showed medium to low performance ("+1": .41, "-1": .36, "+2": .17). Classification of the new glasses was accurate (neighboring class) for 52% of the "-2" samples and for 39% of the "+3" samples.

3.2.7 Overview. The following table (Table 9) contains the key performance metrics for each classifier for easy comparison. The highest value in each column is highlighted in bold.

Table 9

Key performance metrics for each classifier

Model	Configuration	Grid search	Final	\"0"	Validation	\"0"	"-2"	"+3"
KNN	$k = 4$	.99	.99	.99	.69	.31	45%	14%
SVM	$C = 32{,}768$ $\gamma = 2$	**.99**	**1.0**	**1.0**	.54	.17	23%	34%
LSVM	$C = 8$	.21	.48	.07	.49	.00	0%	0%
LOGR	$C = 32$	.21	.49	.10	.49	.00	0%	0%
GNB	-	-	.66	.30	.68	.22	35%	0%
MLP	$h = 64$	.98	.99	.98	**.73**	**.38**	**52%**	**39%**

4 Error correction

4.1 Methods

The aim of the error correction task was to construct a model that would output X and Y-values as offsets when given a new sample of raw X and Y coordinates and the corresponding diopter of the glasses used. The output will not be of ordinal scale as it was in the classifier (*dpt*: -4, -3, -1, 0, 1, 2, 5), but instead of metric scale (corrected X/Y: $-\infty$ to $+\infty$). This regression problem must be addressed with a nonparametric method and a regression-based approach.

4.1.1 Nonparametric approach. The collected data can be represented as offset values for X and Y that are related to a specific position on the screen and to a certain type of glasses. An observation produced using glasses "-1" at the X/Y-position of 500/500 with a reference point of 300/300 would produce the offset values 200/200. To generate a more reliable estimation, the offset values are averaged for each trial, resulting in offset values for 13 calibration points and 6 types of glasses. A new data point can be corrected by the offset values of the nearest calibration point. This works well when the data collected features enough calibration points to provide a fine-grained grid of offset values. The amount of calibration points needed is dependent upon the complexity in the data. Hornof and Halverson (2002) and Blignaut et al. (2014) used an interpolation of the four/five nearest calibration points to shift their data, but they did not justify their choice. In this thesis, all available data with a weighting based on the distance to the new data point was used. The offset values for a new data point should be determined largely by the calibration points closest to it, and less so by the calibration points which are farther away. The general equation is written as

$$offset_{new} = \frac{\frac{offset_1}{dist_1} + \frac{offset_2}{dist_2} + ... + \frac{offset_{13}}{dist_{13}}}{dist_1 + dist_2 + ... + dist_{13}} = \frac{\sum_1^{13} \frac{offset_c}{dist_c}}{\sum_1^{13} dist_c} \tag{7}$$

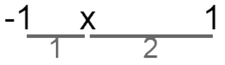
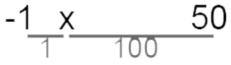
where $offset_c$ is the offset at a calibration point c ($offset_c \in \mathbb{R}$), $offset_{new}$ is the offset combined from all calibration points combined that can be applied to the sample for error correction, and $dist_c$ is the distance between the new sample and reference point of c ($dist_c \in \mathbb{R}^{>0}$). The distance $dist_c$ is a Euclidian distance that cannot be zero, as such value would lead to division by zero. In this rare case, only the offset at the exact point of the sample is considered. The equation is used to generate both the X-and Y-offset. This can also be seen as a derivative of a weighted mean.

The distance ($dist_c$) alters the effect of the offset ($offset_c$). Large distances diminish the impact of the individual offset, while small distances increase the impact of the individual offset. To globally alter the effect of distance on the offset computation ($offset_{new}$) positive and negative exponents can be applied to all $dist_c$ in the equation. A positive exponent enlarges the distance, giving closer points a larger leverage on the result of the equation. A negative exponent reduces the distance, giving further points a larger leverage on the result of the equation.

These effects are illustrated in two examples in Table 10. The x depicts the position of the new data point, the black numbers depict the position and magnitude of the offset, and the grey numbers show the distance between data point and offsets.

Table 10

Different parametrization of the nonparametric model and the effect on offset computation

Examples		
Neutral	$\frac{\frac{-1}{1}+\frac{1}{2}}{1+0.5} = -\frac{1}{3}$	$\frac{\frac{-1}{1}+\frac{50}{100}}{1+\frac{1}{100}} = -0.495$
Squared ($dist^2$)	$\frac{\frac{-1}{1^2}+\frac{1}{2^2}}{1+\frac{1}{2^2}} = -0.6$	$\frac{-1+\frac{50}{100^2}}{1+\frac{1}{100^2}} = -0.99$
Squarerooted ($\sqrt{dist}$)	$\frac{\frac{-1}{\sqrt{1}}+\frac{1}{\sqrt{2}}}{1+\sqrt{0.5}} = -0.17$	$\frac{-1+\frac{50}{\sqrt{100}}}{1+\frac{1}{\sqrt{100}}} = 3.6$

4.1.2 Regression approach. The nonparametric approach is unable to predict offsets for diopters other than those used in the model. It cannot, for example, predict possible offsets for "-2" dpt glasses. The regression model addresses this shortcoming. Regression models, also in their polynomial form, have been used by Hornof and Halverson (2002) and Cherif et al. (2002) to correct data. Hansen and Ji (2010) point at the nonlinearity between refraction and the alteration of the eye image. Therefore, incorporation of polynomials in the regression seems reasonable. The regression equation has the form of a generalized linear model:

$$y = X\vec{\beta} + \varepsilon \tag{8}$$

It can be written as a multiple linear regression:

$$\begin{aligned} offset_{X/Y} &= \beta_0 + \beta_1 X + \beta_2 Y + \beta_3 dpt + \beta_{12} XY + \beta_{13} X dpt + \beta_{23} Y dpt + \beta_{123} XY dpt + \varepsilon \\ &= (1 + X + Y + dpt + XY + X dpt + Y dpt + XY dpt) * \vec{\beta} + \varepsilon \end{aligned} \tag{9}$$

$\vec{\beta}$ in the generalized linear model is the parameter vector containing all β_j. These will be estimated by the model. X in the generalized linear model is the design matrix containing all factors: X (depicts the X coordinate), Y, dpt, and their interactions XY, $X dpt$, $Y dpt$, $XY dpt$. ε is the error of the model. Two separate regressions are used to predict $offset_X$ and $offset_Y$.

Higher powers of the predictors may be included in the model to approximate non-linear types of relationships (Harrell, 2015). "Polynomial interpolation is still one of the main tools of numerical analysis, mainly thanks to its simplicity of realization and the good quality of the interpolation obtained from it" (Cerrolaza et al., 2012, p. 10:3). Polynomials of degree d can be included in the model as X^d, Y^d and dpt^d. This, however, leads to a massive increase in predictors, as all degrees with lower d and interactions (such as X^2Y^2dpt for $d = 3$) can also be included. For example, a full model of the second degree will feature 11 predictor terms and a model of the tenth degree will contain 286.

The aim is to specify a parsimonious model by choosing the highest-performing predictors. This model should provide good fit on the data and should perform well on future data. If too complex of a model is specified (too many predictors), performance will be exaggerated (R^2), and it will perform poorly on new data (Harrell, 2015).

4.2 Results

4.2.1 Nonparametric approach. Four different variants of the nonparametric approach were constructed:

1. weighted model, $dist^1$

2. weighted model, distances are squared, $dist^2$
3. weighted model, distances to the power of 10, $dist^{10}$
4. only nearest calibration point is considered (no weighting)

Model 4 corrects data by looking at the nearest calibration point only, meaning no weighting is applied. Model 4 was constructed to illustrate the effect of weighting on performance and generalization ability.

The models were trained using the training set (80% of data from experiment 1). Their performance was assessed using the test set (20% of data from experiment 1) and the validation set (data from experiment 2). Trials from the validation set with new glasses (17, 18) were not used for the nonparametric models, as the models are unable to predict offsets for diopters other than they have been trained with.

To gauge the amount of offset in the data, the mean absolute offset for all glasses trials from their respective reference points was calculated. The offset averaged to 85.2 px (2.12°) in X-direction and to 93.8 px (2.33°) in Y-direction on the training set. When the model's correction is applied to the different data sets, the offset decreases to the levels reported in Table 11 for the training, test and validation sets.

Table 11

Average absolute offset in ° in X and Y direction for the nonparametric models

	Training		Test		Validation		New cals[a]	
model	X	Y	X	Y	X	Y	X	Y
none	2.12	2.33	2.12	2.34	1.87	2.58	1.40	1.25
1	1.42	1.57	1.42	1.57	1.33	2.20	0.86	**1.04**
2	0.89	1.15	0.88	1.15	1.02	**2.25**	0.86	1.41
3	0.54	0.84	0.54	0.86	**1.02**	2.53	1.18	1.90
4	**0.003**	**0.024**	**0.014**	**0.044**	1.29	2.48	**0.64**	2.07

[a] A set containing only the trials with unknown calibration points (14, 15) from the validation set.

The offsets for the training and test sets are very similar because the data is drawn from the same experimental recording. Performance is shown for the new calibration points to gauge the models' ability to generalize on new data.

Model 4 revealed the best performance for the training and test sets. Model 3 (X) and 2 (Y) showed the best performance for the validation set and model 4 (X) and 1 (Y) showed the best performance for the new calibration points. Each model had its strengths and weaknesses in performance for the different data sets, but model 2 was deemed most suitable due to its strong performance on the validation set (for more information on this determination, see section 5.3.2 in the discussion).

4.2.2 Regression model.

Model complexity and performance. To study the effect of model complexity on performance and potential overfitting, the full models for degrees 1 through 20 were trained and their performance examined (see Figure 26). Performance is shown in coefficient of determination R^2 ($0 \leq R^2 \leq 1$). Higher values indicate a better fit to the data (Fahrmeir, Kneib, & Lang, 2007). The performance of the $offset_Y$-regression is lower overall compared to the $offset_X$-regression. There was a steep increase in performance from degrees 1 to 2 and only gradual increases from then

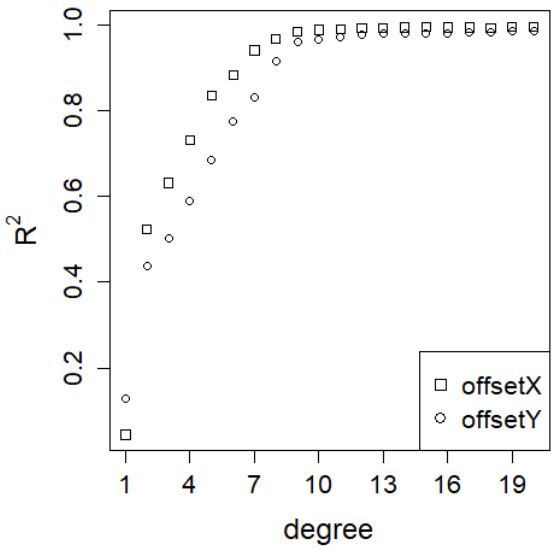


Figure 26. Performance of the regression models (R^2) on the training set depending on their maximum polynomial degree

on. For $offset_X$, a low R^2 of .04 for degree 1 was followed by a major increase for degree 2 (.52) and a lower increase for degree 3 (.63). $offset_Y$ showed a similar relationship (R^2 degree 1: .13; degree 2: .44; degree 3: .50).

The model with the best balance between performance and model complexity was chosen and the ratio of performance to predictors (terms) examined. The model with degree 2 showed a performance of $R^2 = .52/.44$ (X/Y) with 11 terms. It had the best ratio of $\frac{R^2}{terms}$ of all models (X: .05, Y: .04, other models $< .03$). The full models of degree 2 are written as:

$$offset_{X/Y} = (1+X+Y+dpt+X^2+Y^2+dpt^2+XY+Xdpt+Ydpt+XYdpt)*\vec{\beta}+\varepsilon \quad (10)$$

Further tuning with exhaustive search. All of the models' estimated coefficients ($\vec{\beta}$) were significantly different from 0 ($p < .05$). However, due to the large sample size (> 350,000 samples), even small differences are detected by the statistical analysis. Therefore, an exhaustive search was performed to zero in on the most effective subset of predictors. The algorithm returns the best model (regarding R^2_{adj}) out of all combinations of the predictors in the full model for a given target size. For example, for a target size of 2, the algorithm would examine all combinations of 2 predictors ($\binom{10}{2} = 45$, e.g. $X + Y$, $X + dpt$, $X^2 + XYdpt$,...) and choose the model with the highest R^2_{adj}. The exhaustive search was performed for target sizes 1 to 8. A portion of the results are shown in Table 12. The models of size 8 reached the highest R^2_{adj} value with .53 for $offset_X$ and .44 for $offset_Y$. The models with target size 2 were chosen

Table 12

Results of the exhaustive search for size 1 to 4

Model size	$offset_X$ model	R^2_{adj}	$offset_Y$ model	R^2_{adj}
1	$Xdpt$	.07	dpt^2	.20
2	$\boldsymbol{dpt + Xdpt}$	**.48**	$\boldsymbol{dpt^2 + Ydpt}$	**.38**
3	$dpt + Xdpt + Y^2$	.50	$dpt^2 + Ydpt + dpt$	.40
4	$dpt + Xdpt + Y^2 + XY$	.51	$dpt^2 + Ydpt + dpt + Xdpt$	.42

for further consideration.

Performance on training, test and validation sets. Performance on the training, test and validation sets for the final model of degree 2 (deg2) and for the models with degree 5 (deg5) and 9 (deg9) is provided in Table 13 for comparison. The models with degree 5 and 9 were not optimized (as explained in the section "Further tuning with exhaustive search"). The models were used in their complete form, including all polynomials and interactions (degree 5: 56 terms, degree 9: 220 terms).

Table 13

Average absolute offset in ° in X and Y direction for the regression models

	Training		Test		Validation		New cals[a]	
Model	X	Y	X	Y	X	Y	X	Y
None	2.12	2.33	2.12	2.34	1.73	2.45	1.39	1.35
deg2	1.53	1.81	1.52	1.82	**0.93**	**2.24**	**0.48**	**1.59**
deg5	0.76	1.11	0.76	1.11	1.78	3.14	1.26	2.46
deg9	**0.03**	**0.07**	**0.03**	**0.08**	11.55	16.37	5.84	6.83

[a] A set containing only the trials with unknown calibration points (14, 15) from the validation set.

Unlike the nonparameteric model, the regression model is capable of making predictions about unknown glasses. Therefore, the validation set contained all data from the new glasses (17, 18).

The second degree model showed good performance for the training, test and validation sets. When applied to the data, it reduced the amount of offset in all cases, with the exception of the $offset_Y$ for the new calibration points. The full models for degrees 5 and 9 showed superior performance for the training and test sets, but failed to reduce offset for the validation set.

4.2.3 Comparison of the nonparametric and regression model. The nonparametric and regression models have different validation data sets (regression also includes the new glasses), and thus cannot be as easily compared in absolute visual angle. For this reason, percentages are used in Table 14. The nonparametric model showed higher reductions for training and test sets, and the regression model outperformed it in some subsets of the validation set.

Table 14

Average absolute reduction in offset in X and Y direction after correction

	Training		Test		Validation		New cals[a]	
Model	X	Y	X	Y	X	Y	X	Y
2	-58%	-51%	-58%	-51%	-45%	-13%	-39%	+13%
deg2	-28%	-22%	-28%	-22%	-46%	-9%	-65%	+18%

[a] A set containing only the trials with unknown calibration points (14, 15) from the validation set.

5 Discussion

5.1 General discussion

5.1.1 Congruence between first and second recording. The replication of the first experiment in the second recording was successful. The more error-prone glasses were not examined (1, 14), thus improving data quality. The amount of missing data, which was often the result of the eye tracker losing track of the eye, dropped from 2.6% to 1.0% (see Table 3). The per-trial variance for both experiments was also comparable ($sd_{X_1} = 0.11°, sd_{X_2} = 0.11°; sd_{Y_1} = 0.24°, sd_{Y_2} = 0.22°$). It appears that the offsets follow a similar pattern in the second experiment: Trials for glasses with negative diopters were offset in the opposite direction compared to trials for glasses with positive diopters. Additionally, the offsets for the new calibration points (14, 15) were very similar to the offsets of their nearest counterparts ($12 \rightarrow 14; 13 \rightarrow 15$, see Figure 16 and 18).

5.1.2 Other measures of central tendency. In this thesis, the primary measure of central tendency was the mean, employed in scales such as $offset_X$, $offset_Y$, ref_X, and ref_Y. During the cleaning phase of data acquisition, attempts were made to clean the data of outliers (see section 2.2.2). To diminish the effect of outliers on the mean, additional, alternative measures of central tendency such as the median or the mode (with discrete values) could have been used. Zhang and Hornof (2011), for example, used the mode of disparities to correct their eye tracking data.

5.1.3 Offsets and refractive strength. Kübler et al. (2016) observed a consistent pattern for the geometrical gaze mapping in their data: Offsets for negative diopters were pointed towards the center of the screen and the magnitude of the offsets appeared larger for weak lenses ("-1") than for strong lenses ("-3", "-5"). An increase in refractive strength was also related to a decrease in gaze prediction accuracy (Kübler et al., 2016).

The pattern observed by Kübler et al. (2016) can only be partially reported in the conducted experiments. First, offsets of negative diopters were more pointed towards the edges, as seen in Figures 16 and 18. Second, the magnitude of the offsets was not clearly related to refractive strength. Precision appeared to deteriorate with increasing refractive strength for positive diopters (see Table 15). The correlation between the two factors was not significant ($r = .31, t_{34} = 1.87, p = .07$).

Table 15

Accuracy (offset) and precision (standard deviation) for all diopters in °

dpt	-4	-3	-1	0	+1	+2	+5
Accuracy	4.29	2.48	2.03	-	2.91	3.33	6.22
Precision	0.27	0.23	0.28	0.26	0.28	0.33	0.46

5.1.4 Number of observations per trial.

The number of observations per trial differed widely (1,087-2,952) and did not match the sampling rate (120Hz). For a 30s recording with a sampling rate of 120 Hz, one trial should yield a number of observations close to the maximum of 3,600. The number of observations depends upon the CPU load of the computer, the interaction between the applications streaming the data, and the quality of the data. During the cleaning phase, trials with an insufficient number of samples were excluded. The relatively long length of each trial, 30 seconds, was chosen in order to account for problems with sampling and data quality. Exclusion of samples due to outliers or missing data should not have been a problem because the data is captured from the same physical setup. For this reason, a significant loss of information is unlikely to have occurred.

5.1.5 Baseline shift.

Two possible causes of the shift in baseline (introduced in section "Baseline shift") are imaginable. First, parts of the head model may have become tilted when the glasses were changed or gradual drift may have occured over time. Second, the eye tracker runs an optimization algorithm that accounts for previous gaze data and then, alters the incoming gaze data. The head model has been built to prevent the alteration of the physical structure, however, this cannot be assured. As for the second reason, the exact gaze estimation technique of the eye tracker is a company secret, and therefore it is neither known nor given details of.

5.1.6 Validity of test set. Test sets are generally created in order to assess a model's ability to generalize and to serve as a secondary source of verification, as they include data not previously trained on the model. Over the course of the experiments, the test set yielded similar results to the training set. Performance was nearly identical (see Figures 22, 23, 25, for example). One possible explanation is that the generated data contained many observations carrying the same information (i.e. two gaze points at the same position). A 20% reduction in data (test set), for example, does not deprive the algorithm of relevant data. For this reason, the validity of the test set as a measure of independent performance must be questioned. The validation set showed a good estimate of generalization ability, so this deficit in experimental design does not appear to have had a major impact on model selection. Nonetheless, a second recording for the test set is possibly a solution. The second recording should include the same experimental setup and procedure (calibration points, glasses) as the first recording.

5.2 Classification

5.2.1 Selection of features. $offset_X$, $offset_Y$, ref_X and ref_Y were chosen as classification features. $offset_X$ and $offset_Y$ were well suited features because they are comparable for all calibration points. Raw X or Y values would be unsuitable features because generalization to other calibration points would be unlikely. ref_X and ref_Y were chosen over the initial X and Y coordinates of the calibration points in order to make up for a potential shift in the baseline. The indexes of the calibration points (1-13) would also have been inappropriate, as it would not represent the similarity in X and Y direction for the calibration points. Calibration points 1 and 4 (Both $X = 280$, $dist = 350px$), for example, are closer in Euclidian distance than calibration points 1 and 3 (Both $Y = 175$, $dist = 1{,}120px$).

5.2.2 LSVM and LOGR. Both the LSVM and LOGR algorithms exhibited a similar pattern in performance. Performance was low after the grid search (both: .21), but increased after the final training, possibly on account of more training data (LSVM: .48, LOGR: .49). The models in the cross-validation were trained with $\frac{9}{10}$ of the training set ($n_{cv} = 309{,}118 * 0.9 \approx 278{,}206$), while the final model was trained using the full training set ($n_{train} = 309{,}118$). Nonetheless, performance was only strong for the "0" dpt class in the test and validation sets. The model successfully predicted two classes ("-4", "+5") in the test set, but failed to predict all classes (.00) except "0" in the validation set.

The failure to predict classes aside from "-4" and "+5" could have multiple causes: For one, the absence of multicollinearity in the features must be assumed for the LOGR algorithm. A small (Cohen, 1988) correlation ($r = .24$) between $offset_X$ and $offset_Y$ can be exhibited in the data (see section 3.1.6) questioning the complete absence of multicollinearity. Furthermore, both algorithms attempt to perform a classification using linear methods: LOGR as a linear classifier and LSVM with its linear kernel. If strong non-linear relationships are present in the data, they can obstruct classification.

5.2.3 KNN, SVM and MLP. KNN, SVM and MLP are the most promising algorithms for classification. All three performed at near-optimum levels on the training and test sets, with the SVM leading by a slight margin. The MLP, however, had the best performance for the validation set when classifying samples from both old glasses (see Table 9).

5.2.4 Classification of "0" dpt. One drawback of the inclusion of the "0" dpt class in the classification is that the final F-scores of the models were overrated in some cases. The affected models (LSVM, LOGR, GNB) showed good performance on the "0" dpt class, but showed very poor performance for the other classes. Due to its high number of samples the performance of the "0" dpt class had a major impact on the overall F-score. To provide a more useful understanding of performance, an alternative overall F-score excluding the effect of "0" dpt was also reported ("without 0", as seen in Table 9).

The exclusion of the baseline recordings ("0") from classification would strip the algorithms of the ability to detect eye tracking data from users without glasses. In order to preserve the real-world applicability of the algorithms, the baseline recordings were not excluded from classification. If the classification of glasses is of primary interest (e.g. because another system handles the classification of glasses/no-glasses), such an exclusion might be a viable way to improve classification performance.

Another approach is balancing out the overrepresentation of "0" dpt samples. This approach was attempted creating a new GNB model. It was trained with a balanced training set (class "0": 20,007 samples) and produced the following results for the test set (see Table 16): Performance increased for classes "-3" and "+1" and declined for classes "0" and "+5", and overall performance dropped (.66 → .37) with a small decrease for all classes except "0" (.30 → .33). This approach did not yield promising results, and was therefore not pursued.

Table 16

Performance on test set for normal and balanced training set

dpt	-4	-3	-1	0	+1	+2	+5	total
F-score-normal	.48	.13	.25	.91	.38	.11	.48	.66
F-score-balanced	.48	.21	.26	.69	.44	.11	.49	.37

5.2.5 Concept of neighboring class. To examine the performance of the models on the classification of unknown glasses, the concept of "neighboring classes" was introduced. Because the models cannot classify samples belonging to classes with which they have not been trained with, it was assumed that performance was strong if they were able to classify a new sample as a class of a similar dioptric value (e.g. "-2"→"-1"/"-3"; "+3"→"+5"). It must be noted, that the assumption that samples generated by glasses with similar diopter values are similar in X and Y coordinates has not been proven. A pair of "-2" dpt glasses, for example, could potentially generate data similar to that produced by a pair of "+3" dpt glasses. Based on the conveyed in Figures 16 and 18, it seems reasonable to infer that gaze data from glasses with positive diopters is generally similar to that of other glasses with positive diopters, and that the same applies to glasses with negative diopters. As mentioned, this assumption was made to aid in the demonstration of the classification performance of new glasses and therefore should be treated with caution.

5.3 Correction

5.3.1 Nonparametric model 4. Nonparametric model 4 was constructed to illustrate the effect of weighting on performance. It showed best performance on the training and test sets, but very poor performance on the validation set and the new calibration points. This pattern in performance can be explained by the model's training: It was trained with the same data that it corrected afterwards, making it highly dependent on the training data. The model merely shifts each glasses trial to the mean position of the prior baseline. In the four model variants of the nonparametric approach the influence of distance on the offset computation has been altered. An increase in the influence of distance in a nonparametric model (by raising the power) will cause it to approach the performance of model 4. A model with a high enough power will only be affected by the nearest calibration point disregarding the influence of the other calibration points. While this may result in best performance on the training data, it will most likely result in poor performance for new data (new calibration points).

5.3.2 Choosing the most suited nonparametric model. The performance of the nonparametric models (1, 2, 3, 4) was different for each data set. However, performance on the validation set is the most important aspect because it shows the models' ability to generalize on new data. Accordingly, model 2's strong performance on the validation set resulted in its selectiona and in-depth examination. It showed (close to) the best performance on the validation set and acceptable performance for the new calibration points.

5.3.3 Overfitted regression models. For the purpose of comparison, the two regression models with polynomial degrees 5 and 9 were trained, and their performance was then examined in the results. It is clearly evident, that the high polynomial overfit the training and test data. The regression models showed the best performance on the training and test sets (deg9: 1-3px offset after correction), but massively worsened the eye tracking error on the validation set (deg9 $offset_Y$: $2.45° \rightarrow 16.37°$). The degree 9 model reduced the error to an even lower value than the reported gaze accuracy of the eye tracker (0.5°) (see the section "Hardware"). The pattern in performance for the degree 5 and degree 9 models on the training, test and validation sets is a clear example of how attempts to aim for the highest fit can lead to bad performance in real-world situations. It should be noted that a full model of polynomial degree d (e.g. 5) includes all terms from degree 1 to $d-1$ (e.g. 1 to 4). An increase in complexity will always increase fit, and therefore performance on the fitted data.

5.4 Outlook

5.4.1 Other features for classification. This thesis used the four features $offset_X$, $offset_Y$, ref_X and ref_Y as features in the classification process. Other possible features for use in such studies include the measured distance between the eye tracker and the subject or the size of the pupil. The employability of such features in an experiment depends upon the technical setup and the capabilities of the eye tracker. The data with the features needs to be transmitted at a rate fast enough to allow for online correction.

5.4.2 Recording of a 0 dpt lens. All baseline trials were recorded without glasses. The use of "0" dpt lenses during recording could serve as an alternative to this method. This method was rejected to keep the experiment more naturally valid: Persons wearing prescription-free eyeglasses are in the minority when compared to the number of persons who do not wear glasses whatsoever. An additional recording with 0 dpt lenses could be made in order to examine the effect of refractive strength more thoroughly.

5.4.3 Further tuning of the algorithms. Many algorithms were examined and their hyperparameters and additional settings were altered to achieve maximum performance on the training and test sets. More rigorous examination of the configuration of the algorithms could be undertaken in future studies (e.g. additional hidden layers for MLP). The weaknesses of the test set were mentioned in section 5.1.6. Given these weaknesses, a second, independent recording with different calibration points and the glasses from the first recording is recommended in order to provide a solid, foundational test set for optimization.

5.5 Conclusion

The aim of this thesis was to develop models capable of classifying users by eye tracking signal and to use this classification to apply error correction to future eye tracking data. In order to accomplish this aim, eye tracking data was collected in two experiments in a laboratory setting using a head model with AEs and different pairs of eyeglasses. The collected data was then cleaned off unwanted artifacts to prepare it for further analysis. Machine learning algorithms were employed to train models to classify glasses based on the refractive strength of their lenses. The performance of these models was evaluated on different, independent data sets, and the best performing models were then selected for further analysis. A regression-based approach and a nonparametric approach were chosen for error correction. A variety of models were created, trained and evaluated using these two approaches.

The results demonstrated that certain models are better suited to eye tracking samples classification than others. The higher performing models were capable of successfully classifying new samples and samples of unknown refractive strenghts. Overall, the MLP model proved to be the best performing model for the classification task, for it showed the best accuracy on training data as well as for the classification of new samples and samples of unknown refractive strengths. The models trained for the correction task were able to reduce the offset induced by eyeglasses. The most effective nonparametric model outperformed the most effective regression model on the training data and largly outperformed it on the validation set. Overall, the models demonstrated a good balance between specialization on the specific task and generalization on new and unknown data.

Nevertheless, there is room for improvement. All classification models suffered from imbalanced classification performance, favoring the "0" dpt class, while neglecting the other classes. Alternative approaches and possible solutions have been discussed and should be implemented in future studies to achieve higher degrees of classification performance. While the error correction models led to a decent reduction in error on the data, their performance should be improved to further increase the error reduction. Although the models are capable of reducing error in eye tracking data considerably, error is still significantly higher than the reported gaze position accuracy of the eye tracker. Moreover, eye tracking research would greatly benefit from insights into the interaction of light rays, refraction of eyeglasses, and eye tracking sensors and how this interaction affects eye tracking data. These topics require proper examination and their investigation will benefit not only research, but also future interactive systems.

6 References

Allison, P. D. (2012). *Logistic Regression Using SAS: Theory and Application, Second Edition.* SAS Institute.

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed ed.). Cambridge, Mass: MIT Press.

Artal, P. (Ed.). (2017). *Handbook of Visual Optics, Volume Two: Instrumentation and Vision Correction* (1edition ed.). Boca Raton: CRC Press.

Bach, H., & Neuroth, N. (Eds.). (1998). *The Properties of Optical Glass.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Beautycheck - average faces. (n.d.).

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

Blignaut, P., Holmqvist, K., Nyström, M., & Dewhurst, R. (2014). Improving the Accuracy of Video-Based Eye Tracking in Real Time through Post-Calibration Regression. In M. Horsley, M. Eliot, B. A. Knight, & R. Reilly (Eds.), *Current Trends in Eye Tracking Research* (pp. 77–100). Springer International Publishing.

Cerrolaza, J. J., Villanueva, A., & Cabeza, R. (2008). Taxonomic Study of Polynomial Regressions Applied to the Calibration of Video-oculographic Systems. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (pp. 259–266). New York, NY, USA: ACM. doi: 10.1145/1344471.1344530

Cerrolaza, J. J., Villanueva, A., Villanueva, M., & Cabeza, R. (2012). Error Characterization and Compensation in Eye Tracking Systems. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 205–208). New York, NY, USA: ACM. doi: 10.1145/2168556.2168595

Cherif, Z. R., Nait-Ali, A., Motsch, J. F., & Krebs, M. O. (2002). An adaptive calibration of an infrared light device used for gaze tracking. In *IMTC/2002. Proceedings of the 19th IEEE Instrumentation and Measurement Technology Conference (IEEE Cat. No.00ch37276)* (Vol. 2, pp. 1029–1033 vol.2). doi: 10.1109/IMTC.2002.1007096

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2Rev ed. ed.).

Hillsdale, N.J: Taylor & Francis Inc.

Constine, J. (2016, December). *Oculus acquires eye-tracking startup The Eye Tribe | TechCrunch.*

CryptWizard. (2007a). *Schematic representation of hypermetropia. Made by A. Baris Toprak MD. Vectorized by CryptWizard.* Retrieved 2017-06-14, from `https://commons.wikimedia.org/wiki/File:Hypermetropia.svg?uselang=de`

CryptWizard. (2007b). *Schematic representation of myopia.* Retrieved 2017-06-14, from `https://commons.wikimedia.org/wiki/File:Hypermetropia.svg?uselang=de`

Dahlberg, J. (2010). *Eye Tracking with Eye Glasses* (Unpublished master's thesis). Umeå universitet.

DeCarlo, D., & Santella, A. (2002). Stylization and Abstraction of Photographs. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 769–776). New York, NY, USA: ACM. doi: 10.1145/566570.566650

Denoeux, T. (1995, May). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics*, *25*(5), 804–813. doi: 10.1109/21.376493

Drewes, J., Masson, G. S., & Montagnini, A. (2012). Shifts in Reported Gaze Position Due to Changes in Pupil Size: Ground Truth and Compensation. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 209–212). New York, NY, USA: ACM. doi: 10.1145/2168556.2168596

Duchowski, A. (2009). *Eye Tracking Methodology: Theory and Practice* (2nd ed. 2007 ed.). London: Springer.

Ebisawa, Y. (1994, May). Improved video-based eye-gaze detection method. In *Conference Proceedings. 10th Anniversary. IMTC/94. Advanced Technologies in I M. 1994 IEEE Instrumentation and Measurement Technolgy Conference (Cat. No.94ch3424-9)* (pp. 963–966 vol.2). doi: 10.1109/IMTC.1994.351964

Eye tracking. (2017, June). Retrieved 2017-06-13, from `https://en.wikipedia.org/w/index.php?title=Eye_tracking&oldid=783608703`

Fahrmeir, L., Kneib, T., & Lang, S. (2007). *Regression: Modelle, Methoden und Anwendungen.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Feit, A. M., Williams, S., Toledo, A., Paradiso, A., Kulkarni, H., Kane, S., & Morris, M. R. (2017). Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design. In (pp. 1118–1130). ACM Press. doi: 10.1145/3025453.3025599

Fix, E., & Hodges, J. L. (1989). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *International Statistical Review / Revue Internationale de Statistique*, *57*(3), 238–247. doi: 10.2307/1403797

Foveated rendering. (2017, March). Retrieved 2017-07-14, from `https://en.wikipedia.org/w/index.php?title=Foveated_rendering&oldid=768146718` (Page Version ID: 768146718)

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Cambridge, Massachusetts: The MIT Press.

Gwon, S. Y., Cho, C. W., Lee, H. C., Lee, W. O., & Park, K. R. (2014, January). Gaze Tracking System for User Wearing Glasses. *Sensors*, *14*(2), 2110–2134. doi: 10.3390/s140202110

Hammoud, R. I. (2008). *Passive Eye Monitoring: Algorithms, Applications and Experiments.* Springer Science & Business Media.

Hansen, D. W., & Ji, Q. (2010, March). In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 478–500. doi: 10.1109/TPAMI.2009.30

Harrell, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (2nd ed.). Springer International Publishing.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006, July). A Fast Learning Algorithm for

Deep Belief Nets. *Neural Comput.*, *18*(7), 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hohlfeld, O., Pomp, A., Link, J. A. B., & Guse, D. (2015). On the Applicability of Computer Vision Based Gaze Tracking in Mobile Scenarios. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 427–434). New York, NY, USA: ACM. doi: 10.1145/2785830.2785869

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Weijer, J. v. d. (2011). *Eye Tracking: A comprehensive guide to methods and measures.* OUP Oxford.

Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye Tracker Data Quality: What It is and How to Measure It. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 45–52). New York, NY, USA: ACM. doi: 10.1145/2168556.2168563

Hornof, A. J., & Halverson, T. (2002, November). Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *34*(4), 592–604.

Hsu, C.-w., Chang, C.-c., & Lin, C.-j. (2003). *A practical guide to support vector classification.*

Huang, Y., Kong, W., & Li, D. (2013, April). Robust Feature Extraction for Non-contact Gaze Tracking with Eyeglasses. *ResearchGate*, *22*(2), 231–236.

Japan, T. C. S. o. (2012). *Advanced Ceramic Technologies & Products.* Springer Science & Business Media.

Jo, H., Cho, D.-C., Lee, H., Cha, J., & Kim, W.-Y. (2013, December). A Robust Gaze Tracking Method for Users Wearing Glasses. In (pp. 132–135). Science & Engineering Research Support soCiety. doi: 10.14257/astl.2013.43.28

Kortum, P. (Ed.). (2008). *HCI beyond the GUI: design for haptic, speech, olfactory and other nontraditional interfaces.* Amsterdam ; Boston: Elsevier/Morgan

Kaufmann.

Kroemer, A. D., & Kroemer, K. H. E. (2016). *Office Ergonomics: Ease and Efficiency at Work, Second Edition.* CRC Press.

Kübler, T. C., Rittig, T., Kasneci, E., Ungewiss, J., & Krauss, C. (2016). Rendering Refraction and Reflection of Eyeglasses for Synthetic Eye Tracker Images. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 143–146). New York, NY, USA: ACM. doi: 10.1145/2857491.2857494

Lee, Y. (2010). Support vector machines for classification: a statistical portrait. *Methods in Molecular Biology (Clifton, N.J.)*, *620*, 347–368. doi: 10.1007/978-1-60761-580-4_11

Menard, S. (2002). *Applied Logistic Regression Analysis.* SAGE.

Morgan, M. W. (1976). *The optics of ophthalmic lenses.* University of California Multimedia Center.

Naive Bayes classifier. (2017, March). Retrieved 2017-06-21, from `https://en.wikipedia.org/w/index.php?title=Naive_Bayes_classifier&oldid=772906859`

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques.* Berlin: Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ramanauskas, N. (2015, March). Calibration of Video-Oculographical Eye-Tracking System. *Elektronika ir Elektrotechnika*, *72*(8), 65–68. doi: 10.5755/j01.eee.72.8.10791

Rello, L., & Ballesteros, M. (2015). Detecting Readers with Dyslexia Using Machine Learning with Eye Tracking Measures. In *Proceedings of the 12th Web for All Conference* (pp. 16:1–16:8). New York, NY, USA: ACM. doi: 10.1145/2745555.2746644

Rhcastilhos. (2007, January). *Diagram of the human eye in English. It shows the lower* *part of the right eye after a central and horizontal section.* Retrieved 2016-12-02, from `https://commons.wikimedia.org/wiki/File:Schematic_diagram_of_the_human_eye_en.svg`

Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., & Cohen, M. (2006). Gaze-based Interaction for Semi-automatic Photo Cropping. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 771–780). New York, NY, USA: ACM. doi: 10.1145/1124772.1124886

Schiefer, U., Kraus, C., Baumbach, P., Ungewiß, J., & Michels, R. (2016). Refractive errors. *Dtsch Arztebl International*, *113*(41), 693–701. doi: 10.3238/arztebl.2016.0693

Sivardeen, A. (2015). *Determining which factors influence the optimum multifocal contact lens correction for presbyopia* (phd). Aston University.

sklearn. (n.d.). *Neural network models (supervised).* Retrieved 2017-05-09, from `http://scikit-learn.org/stable/modules/neural_networks_supervised.html#mlp-tip`

SMI. (2009, August). *iView X System Manual.*

SMI. (2012, September). *RED-m Eye Tracking System Manual.*

Sokolova, M., & Lapalme, G. (2009, July). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. doi: 10.1016/j.ipm.2009.03.002

Stocker, T. C., & Steinke, I. (2017). *Statistik: Grundlagen und Methodik.* Walter de Gruyter GmbH & Co KG.

Tobii. (2011). *Tobii TX300 Eye Tracker User Manual.*

Tovée, M. J. (1996). *An Introduction to the Visual System.* Cambridge University Press.

Underwood, G. (Ed.). (2005). *Cognitive Processes in Eye Guidance.* Oxford, New York: Oxford University Press.

Vadillo, M. A., Street, C. N. H., Beesley, T., & Shanks, D. R. (2015, December). A simple algorithm for the offline recalibration of eye-tracking data through best-fitting linear transformation. *Behavior Research Methods*, *47*(4), 1365–1376. doi: 10.3758/s13428-014-0544-1

Wang, D., Mulvey, F. B., Pelz, J. B., & Holmqvist, K. (2017, June). A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods*, *49*(3), 947–959. Retrieved 2017-06-17, from `http://link.springer.com/10.3758/s13428-016-0755-8` doi: 10.3758/s13428-016-0755-8

Zhang, Y., & Hornof, A. J. (2011, September). Mode-of-disparities error correction of eye-tracking data. *Behavior Research Methods*, *43*(3), 834–842. doi: 10.3758/s13428-011-0073-0

Zhu, Z., & Ji, Q. (2005, April). Robust real-time eye detection and tracking under variable lighting conditions and various face orientations. *Computer Vision and Image Understanding*, *98*(1), 124–154. doi: 10.1016/j.cviu.2004.07.012

7 Appendix

Table 17

Overview of the utilized glasses

index	type of glasses	diopters left	diopters right	addition
1	MF	-3.00	-3.00	2.75
2	MF	-3.00	-3.00	2.75
3	MF	-3.00	-3.00	2.00
4	MF	-3.00	-3.00	2.00
5	MF	+1.00	+1.00	2.00
6	MF	+1.00	+1.00	2.75
7	MF	+1.00	+1.00	2.75
8	MF	+1.00	+1.00	2.00
9	C	+1.00	+1.00	2.00
10	C	+1.00	+1.00	2.75
11	C	+1.00	+1.00	2.75
12	C	+1.00	+1.00	2.00
13	SV	+2.00	+2.00	-
14	SV	+5.00	+5.00	-
15	SV	-1.00	-1.00	-
16	SV	-4.00	-4.00	-
17	SV	-2.25	-1.50	-
18	SV	+2.00	+3.00	-

Note. MF: multifocal lenses, C: computer glasses, SV: single vision lenses

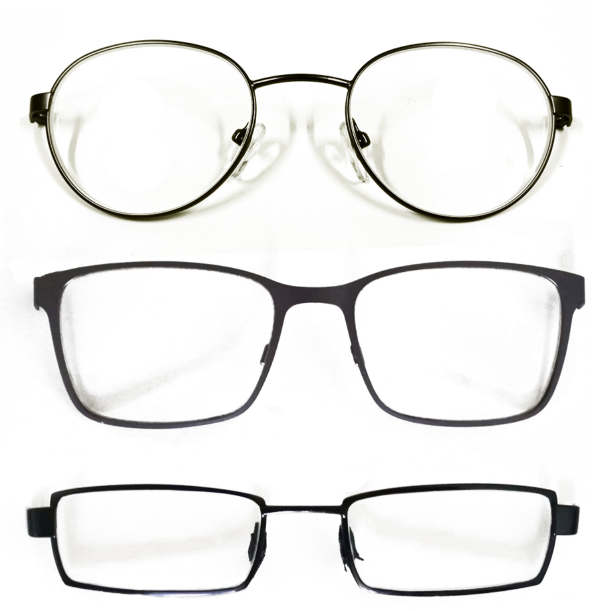
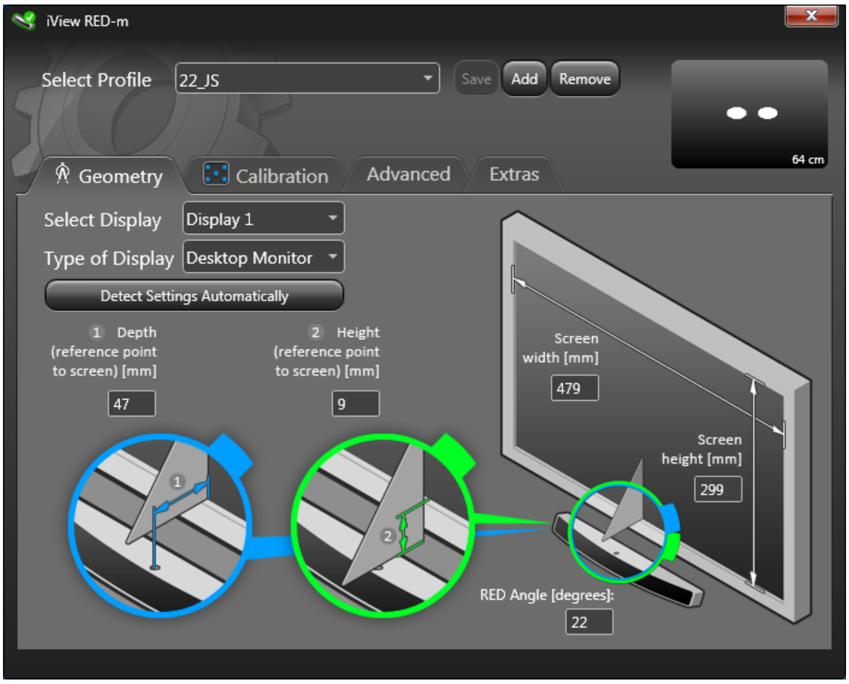


Figure 27. Eyeglass frames for glasses 1 to 16 (top), glasses 17 (center) and glasses 18 (bottom)



Select Profile 22_JS

Select Display Display 1

Type of Display Desktop Monitor

1 Depth (reference point to screen) [mm] 47

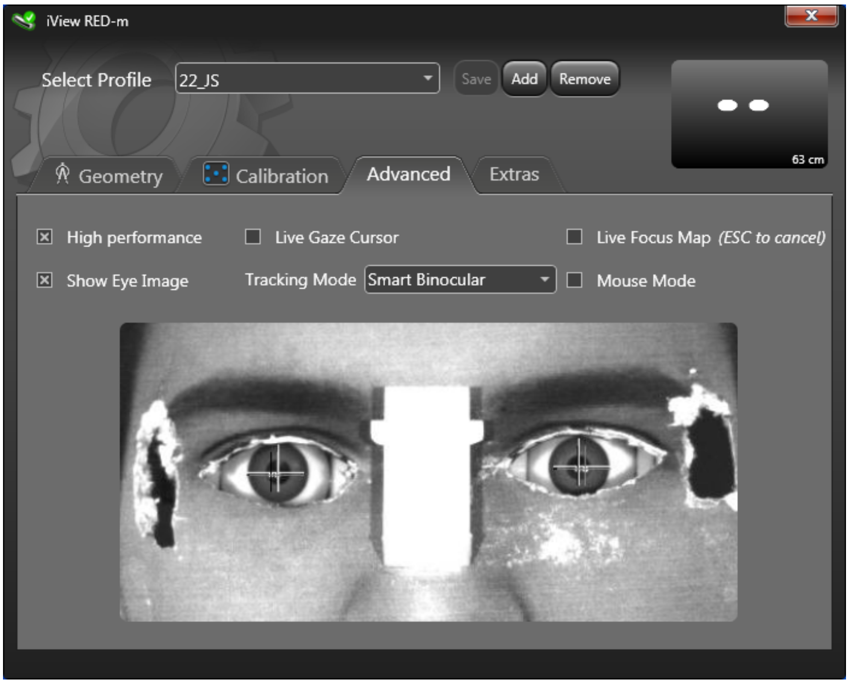
2 Height (reference point to screen) [mm] 9

Screen width [mm] 479

Screen height [mm] 299

RED Angle [degrees]: 22

Figure 28. Settings of the iView application with eye tracking monitor (top right)

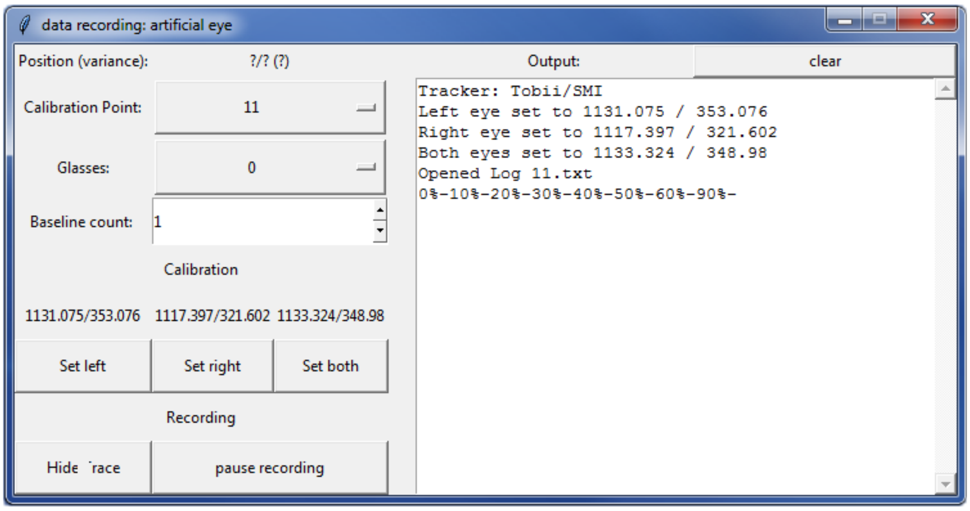


Select Profile: 22_JS

☒ High performance □ Live Gaze Cursor □ Live Focus Map *(ESC to cancel)*

☒ Show Eye Image Tracking Mode: Smart Binocular □ Mouse Mode

Figure 29. Eye image of the iView application with eye tracking monitor (top right)



Position (variance): ?/? (?)

Calibration Point: 11

Glasses: 0

Baseline count: 1

1131.075/353.076 1117.397/321.602 1133.324/348.98

```
Tracker: Tobii/SMI
Left eye set to 1131.075 / 353.076
Right eye set to 1117.397 / 321.602
Both eyes set to 1133.324 / 348.98
Opened Log 11.txt
0%-10%-20%-30%-40%-50%-60%-90%-
```

Figure 30. Graphical user interface of the custom python application

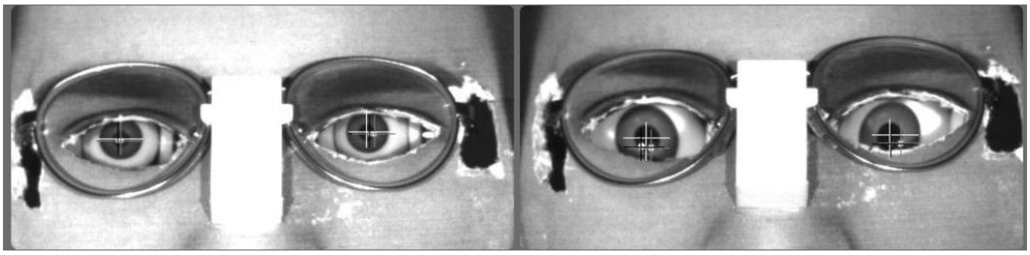


Figure 31. Demagnification for negative ("-4", left) and magnification of the pupil for positive diopters ("+5", right)