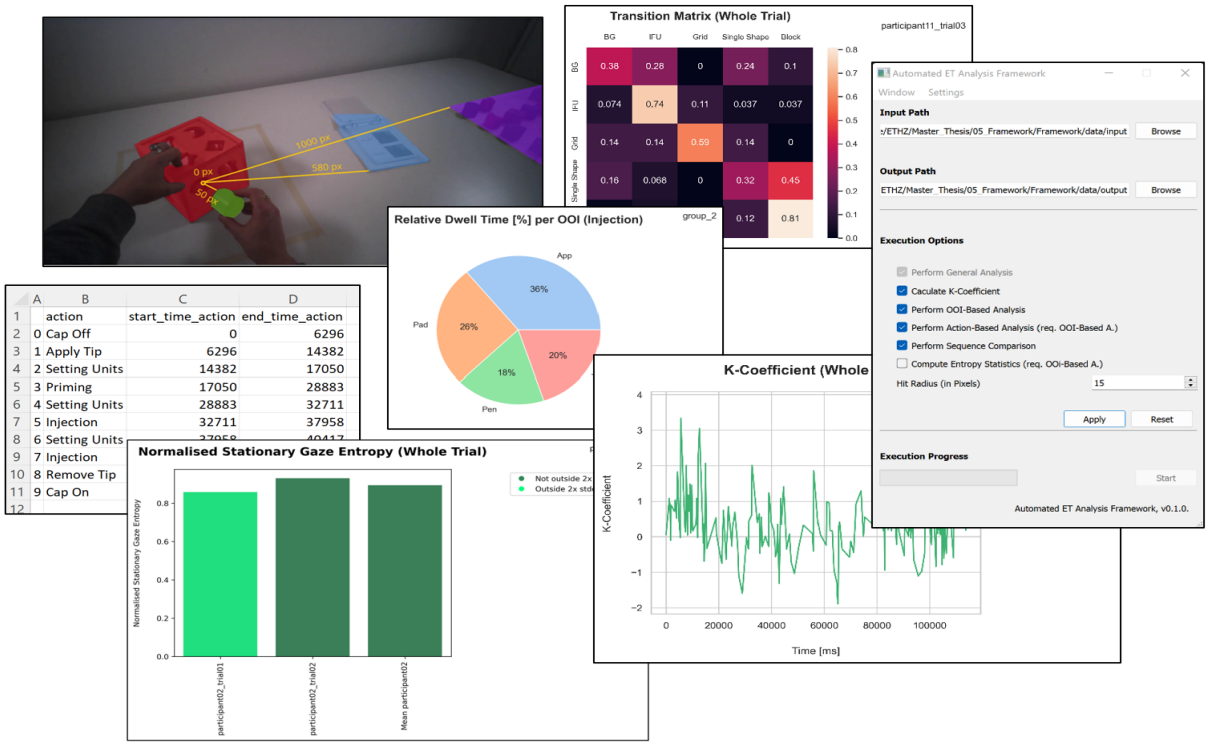
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

pd|z Product Development Group Zurich
Produktentwicklungsgruppe Zürich

Master Thesis

Implementation of an Automated Gaze-based Evaluation Framework for Medical Device Usability Testing and Surgical Workflow Assessment



	action	start_time_action	end_time_action
0	Cap Off	0	6296
1	Apply Tip	6296	14382
2	Setting Units	14382	17050
3	Priming	17050	28883
4	Setting Units	28883	32711
5	Injection	32711	37958
6	Setting Units		
7	Injection		
8	Remove Tip		
9	Cap On		

Author

Renée Muriel Sæthre

Supervisor

Felix Wang

Professor

Prof. Dr. Mirko Meboldt

Nr.1599

Zurich, December 4, 2022

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

pd|z Product Development Group Zurich
Produktentwicklungsgruppe Zürich

Master Thesis

Implementation of an Automated Gaze-based Evaluation Framework for Medical Device Usability Testing and Surgical Workflow Assessment

Dates

Start date 03.06.2022

End date 04.12.2022

Declaration of Originality

I hereby declare that I have written the present thesis independently and have not used other sources and aids than those stated in the bibliography.

_____

Renée Muriel Sæthre

Confirmation

This thesis was written at and accepted by *pd*|**z** Product Development Group Zurich of ETH Zurich.

Supervisor

_____

Professor

_____

Acknowledgement

I would like to sincerely thank my project advisor Felix Wang for his valuable time and support whenever I encountered problems or uncertainties during the project. My thanks also goes to the Product Development Group Zurich of ETH for letting me use their infrastructure, including the Tobii Pro Glasses 2. Further, my sincere thanks go to Prof. Dr. Mirko Meboldt for his time and for examining my Master Thesis.

Moreover, I would like to acknowledge Alexandra Elbakyan, the founder of Sci-Hub, for making scientific papers accessible to the public.

Finally, my thanks goes to my brother Jens Eirik Sæthre for his tireless help in setting up the GUI, as well as to my dear friends Cooper Harshbarger, Marco von Atzigen, Miro Giobbi, and Ronja Senn for proof-reading my written thesis.

Renée Muriel Sæthre, December 4, 2022

Abstract

Medical device usability testing and surgical skill assessments are vital to ensure patient safety. However, both are limited by labour-intensive and costly performance evaluations and are often kept at a minimum. There have been attempts to enhance their efficiency by employing eye tracking, which however has not found its way into practice since the analysis of dynamic gaze data is a tedious process bottlenecked by the manual annotation of objects and actions. To overcome these limitations, this project introduces a gaze analysis pipeline that applies deep learning-based object-gaze mapping and action recognition to eye tracking data in order to calculate a broad range of metrics suitable for a performance assessment. To test the framework on its functionality and the implemented metrics on their ability to make qualitative statements on a performed task, a proof-of-concept study was conducted with a shape sorter toy. Since this study did not include the automated action recognition, the framework's functional ability to process the action-based input was tested on an already existing data set from a medical device usability testing. The two validation studies proved that the proposed pipeline clearly facilitates the analysis of eye tracking data. The extensive output on different levels in the form of tables and graphs enables easy and efficient exploration of the outcome. The integrated GUI makes the program user-friendly and the auto-generated reports that include guidance on the interpretation of the results make the program accessible for people with limited experience in eye tracking. Further, the selected metrics can evaluate various aspects of a performance, i.e. cognitive workload, focus, visual entropy, attention, and efficiency. In addition to the computation of all metrics per action, a sequence comparison module was added to compare each trial to a given template sequence, which is very useful regarding the program's intended use. However, all outcomes must be interpreted in the respective context and do not enable a direct qualitative statement. Moreover, no single metric on its own will suffice to give comprehensive feedback, but only the combination of multiple ones may allow to do so. In conclusion, the pipeline is largely automated on the input side, but not on the output side. If a fully automated assessment as an output is desired for the future, a promising approach would be to provide information on the respective task to the pipeline as an additional input.

Contents

Abbreviations

AOI	Area Of Interest
BG	Background
cGOM	computed Gaze-Object Mapping
ET	Eye Tracking
ETH	Eidgenössische Technische Hochschule
FDA	U.S. Food and Drug Administration
GTE	Gaze Transition Entropy
GUI	Graphical User Interface
HAR	Human Action Recognition
HMM	Hidden Markov Model
IFU	Instructions For Use
IoU	Intersection over Union
IQR	Interquartile Range
ISO	International Organization of Standardization
MDR	Medical Device Regulation
MET	Mobile Eye Tracking
OGD	Object-Gaze Distance
OOI	Object Of Interest
pd|**z**	Product Development Group Zurich
PVHMM	Peripheral Vision-based Hidden Markov Model
SGE	Stationary Gaze Entropy
TG2	Tobii Pro Glasses 2

1 Introduction

1.1 Medical Errors

In the year 2000, the Institute of Medicine of the US published its infamous report 'To Err Is Human', breaking the longstanding silence surrounding medical errors and reporting that yearly up to 100 000 iatrogenic deaths occur in the country alone (Donaldson et al. (2000)). Accordingly, substantial efforts have been made to combat patient harm by building a safer health care system not only in the US but around the globe (Slawomirski et al. (2017), Clancy (2009), Dzau & Shine (2020)). Proposed strategies include lowering barriers to reporting errors, reducing punishment for mistakes, and educating the staff on the matter through structured initiatives (Rodziewicz et al. (2022). However, more recent studies suggest that deaths from medical errors have increased instead, estimating at least 250 000 cases per year in the US and declaring it the third leading cause of death in the country (Makary & Daniel (2016), James (2013)).

A medical error is defined as a preventable adverse effect of medical care that unintentionally harms the patient (Donaldson et al. (2000)). Its severity can vary from overlooking a torn ligament to amputating the wrong leg to administering a deadly dose of medication. Besides being harmful to the patient, they substantially contribute to health care costs (Wachter (2012)). Some of these errors are attributable to the usage of a medical device. The U.S. Food and Drug Administration (FDA) divides these incidents into two categories. On one hand, if an adverse event is caused by the failure of a medical device used for treatment, it is referred to as a medical device hazard. On the other hand, so-called use errors occur if the handling of the medical device was different from that expected by the manufacturer (FDA (2011)). While certain medical errors cannot be directly attributed to a person's failure, others are indeed linked to the level of skill and experience of a health care professional, or simply occur because humans make mistakes (Branaghan et al. (2021)).

1.2 Usability Testing of Medical Devices

While human error will always remain inevitable to a certain degree, many adverse events in medicine could be prevented by improving the usability of medical devices by making their handling as easy, efficient, and as safe as possible (Vincent et al. (2014), Michael et al. (2015)). In fact, inadequate consideration of human factors has been described as one of the root causes of medical errors (Hegde (2013)). Having too recognised its importance, the responsible regulatory bodies in the US and the EU (namely the FDA and the EU Medical Device Regulations (MDR)) made usability testing mandatory for market access of medical devices (FDA (2011), EU (2017)). Usability testing comes in various forms, from simple questionnaires to cognitive walkthroughs, and depends on the devices' properties and usage, e.g. the invasiveness or the intended user (Ravizza et al. (2019)), and is assessed with a risk-based approach. The goal is to identify and mitigate user-related problems early in the development process. The International Organization of Standardization (ISO) defines usability as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO (2018)). Inherent safety by design is regarded to be the highest mitigation strategy to reduce use-related risk IEC (2015)). Thus, incorporating this principle into the design of medical devices comes with multiple benefits including increased safety, ease of use and improved patient outcome, as well as reduced training time and product liability risks (Hegde (2013)).

Despite the uncontested benefits of usability testing for medical devices, they are often kept at the minimum required by regulatory measures. This is mainly due to the high cost, time and human resources needed for traditional usability assessments (Michael et al. (2015)). Since there are no clear rules regarding applicable standards, assessments vary greatly across devices, making them incomparable and subjective (Polisena et al. (2020)). Automating these processes would be a promising way to reduce these resources, increase the data that can be processed and enhance objectivity. However, no such attempts were found in the literature except from the field of health informatics, where various devices have been described as error-prone (Bastien (2010)). Ergo, a cost-efficient, automated method to objectively analyse the usability of a medical device is still missing.

1.3 Surgical Skill Assessment

Another approach to increase patient safety is to make training of health care professionals both more viable and more effective by improving workflow and proficiency assessments. Especially in the field of surgery, where rapid innovation of new techniques like minimally invasive or robot-assisted operations require a growing set of skills, there is an increasing demand for quality measurements by society, hospitals, as well as insurance companies (van Hove et al. (2010)). This is underlined by the fact that studies in the US and Australia have shown that more than half of all surgical complications are preventable (Gawande et al. (1999), Kable et al. (2002)). Commonly, assessments of a surgeon's technical skills are conducted in the operation room following standardized evaluation procedures, observed by an expert who provides feedback (Levin et al. (2019)). However, besides being prone to subjectivity and bias, these methods are also costly, time-consuming and labour-intensive. Moreover, they rely on the availability of a reviewer trained in the specific assessment procedure, usually an experienced surgeon (Levin et al. (2019)). To this end, there has been a substantial effort in automating surgical skill assessment. This does not only lead to more objective evaluations but also saves time and money by allowing novices to train in the absence of a qualified supervisor (Funke et al. (2019)). Literature on this matter is getting increasingly prevalent and was summarized by Levin et al. (2019). Current methods to automatically assess surgical skills extract data through kinetics, e.g. by using a force transducer attached to a surgical simulator, or computer vision, where a pre-trained algorithm detects objects. The extracted data is subsequently used for motion tracking of tools, hands or eyes. Further analysis is supported by machine learning algorithms that, in most cases, classify a trial as either novice or expert. Estrada et al. (2014) tracked the kinematic movement of a catheter tip during the simulation of an endovascular operation with electromagnetic sensors. They found that this tool motion tracking was able to accurately differentiate between skill levels and correlated to the conventional grading assessment. To name another example, Watson (2013) discovered that the complexity of hand motions during a simulated venous anastomosis task correlated with the surgical experience. Levin et al. (2019) concluded that many of these current approaches have great potential to dramatically change how surgeons are both trained and assessed in the future. However, these works use very specific analyses tailored to the needs of the respective task, which greatly limits their applicability to other tasks. Hence, with the goal to enable the translation of these procedures into practice, a method that more generally analyses performances on different levels and that is applicable to different types of tasks would be desirable.

1.4 Eye Tracking in Usability and Skill Assessment

By quantifying eye movements, eye tracking (ET) allows to gain insight into the cognitive processes and for the assessment of various aspects of a performance like efficiency, focus, attention and sequence of actions (Lohmeyer et al. (2019), Duchowski & Krejtz (2017), (Wang et al. (2022a), under review). Hence, ET has the potential to overcome the challenges in both surgical skill assessment and usability testing. Ahmidi et al. (2010) have successfully employed ET to evaluate expertise in laparoscopic surgery in combination with tool motion tracking and Wu et al. (2020) predicted the perceived workload in robotic surgery by tracking the participants' eye movements. In their review on the employment of ET in surgical training, Merali et al. (2019) report that ET could be a valuable tool for training future surgeons. Next to surgical workflow assessments, ET has also been employed in the usability testing of medical devices. In their study, Wegner et al. (2020) conducted user tests with two different patient assistance devices for peritoneal dialysis therapy. Besides identifying critical tasks in the handling of both devices, their gaze analysis provided objective, quantitative feedback on individual user interface features, thereby enabling a more detailed usability assessment compared to traditional methods. Koester et al. (2017) have conducted conventional ethnographic methods as well as ET-based usability testing on a medical device. Although they concluded that ET can indeed provide valuable additional insights that are otherwise not detectable, e.g. detect visual distractions during use, they warn the reader about additional time, cost and expertise needed for the analysis.

Indeed, recording a person's gaze and field of view generates a massive amount of data, and its manual analysis still is an iterative, tedious, and labour-intensive process that requires a lot of expertise. This restricts such ET studies to only a small number of subjects and greatly limits their application (Koester et al. (2017)). Thus, simplifying the analysis of ET data is needed to counter this limitation.

Modern video-based eye trackers are usually accompanied by software that already processes the recorded data to a certain extent, e.g. by automatically extracting the duration and gaze position of fixations (Punde et al. (2017)). More recently, tools such as automated human action recognition (HAR) have been developed to reduce the human workload and enable a more objective assessment (Wang et al. (2022a), under review). In Chapter 2,

the current state of the art of ET and its automation will be addressed in more detail. Despite numerous analysis tools being developed, an automated evaluation framework is still missing in both usability testing and surgical skill assessment.

1.5 Goal of the Thesis

To this end, the goal of this Master's thesis is to develop a pipeline that facilitates the analysis of gaze data in order to get valuable and objective feedback on a performed task. This is achieved by utilizing more traditional ET measurements like dwell time per object of interest or fixation count, more advanced metrics like entropy and K-coefficient, as well as the state-of-the-art analysis tools developed by the Human Behaviour group at Product Development Group ($pd|\mathbf{z}$), ETH. The proposed framework should be able to provide insights into different aspects of the performed task, from cognitive processes to attention-based elements, and should be able to handle both, surgical skill assessments and usability testing of medical devices. The selected metrics will be thoroughly discussed in Chapter 2. Finally, a proof-of-concept study will be conducted to test if the created pipeline with its implemented metrics is suitable for a performance assessment. Finally, the evaluation framework will be tested by applying it to more complex, already existing data from usability studies of a medical device.

2 Theory

This chapter starts by briefly introducing the reader to the physiology of eye movement in Section 2.1, followed by an insight into ET technology in Section 2.2. Section 2.3 gives an overview of selected ET metrics that will be implemented in the framework and Section 2.4 highlights existing efforts to automate ET data analysis.

2.1 Physiology of Eye Movements

Humans' binocular vision relies on perfectly coordinated motor and sensory functions, where both eyes look at the same point in space in order to create one single fused image (Jain (2019)). Light enters the eye through the pupil and is focused onto the retina to create an inverted image that is perceived by two types of photoreceptor cells, namely the rods and the cones. The rods are responsible for low-light vision, and the cones are tuned for bright lights and colours. The rods are highly concentrated in our periphery, whereas cones are densely packed in the centre of the retina (Davson (1990)). The central visual field of the retina is divided into three regions. In the centre of it lies the fovea centralis, where colour vision-enabling photoreceptor cells are most densely packed, creating a sharp image wherever the person is currently looking. This region is surrounded by the parafovea and the perifovea, as can be appreciated in Figure 2.1. Both visual acuity and contrast sensitivity peak at the fovea centralis and rapidly decline towards the periphery. Nevertheless, parafoveal and peripheral vision are essential for many aspects of human vision and can be exploited for HAR, as explained in Section 2.4 (Sakurai (2016)).

Eye movements are tightly coordinated by six extraocular muscles and have been studied for many years due to their ability to give objective, otherwise not accessible insight into human thought processes. Moreover, the method can be used in many different disciplines, from human-computer interaction to sports performance (Stuart (2022)). Most commonly, eye movements can

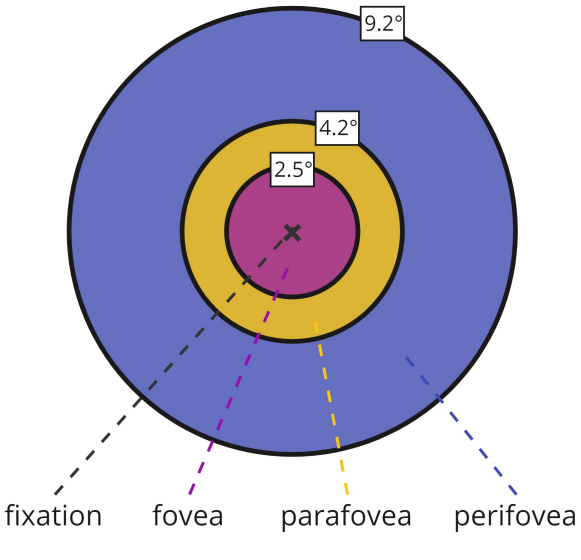


Figure 2.1: Adapted from Sakurai (2016). With a radial visual angle of around 2.5°, the fovea centralis is quite small. It is surrounded by the so-called parafovea that extends to up to 4° radial and the outermost layer, the perifovea, that extends to around 9.2° radial.

be categorized into fixations and saccades. Fixations are time periods of milliseconds up to multiple seconds where the eyes are fixated on a visual target and do not move, while the shorter saccades are rapid movements in between the fixations (Davson (1990)) which last around 70 ms on average Devillez et al. (2020)). During this short time, the eye is effectively blinded and is not able to perceive any visual information. This phenomenon is referred to as saccadic masking (Stuart (2022)). These eye movements are illustrated in Figure 2.2. Their relevance will be highlighted in Section 2.3.

2.2 Eye Tracking Technology

ET is the procedure of following eye movements across time. Being practised for dozens of years utilizing numerous different methods (Stuart (2022)), it has been subject to great technological advancements in recent years, making proficient eye trackers more accessible (Carter & Luke (2020)).

Most modern eye trackers are video-based devices, they are non-invasive and relatively easy to operate. As opposed to remote ET, where the ET sensor stays in place, e.g. mounted to a computer screen, mobile ET (MET) technology are wearable devices that track a user's gaze in an unobtrusive

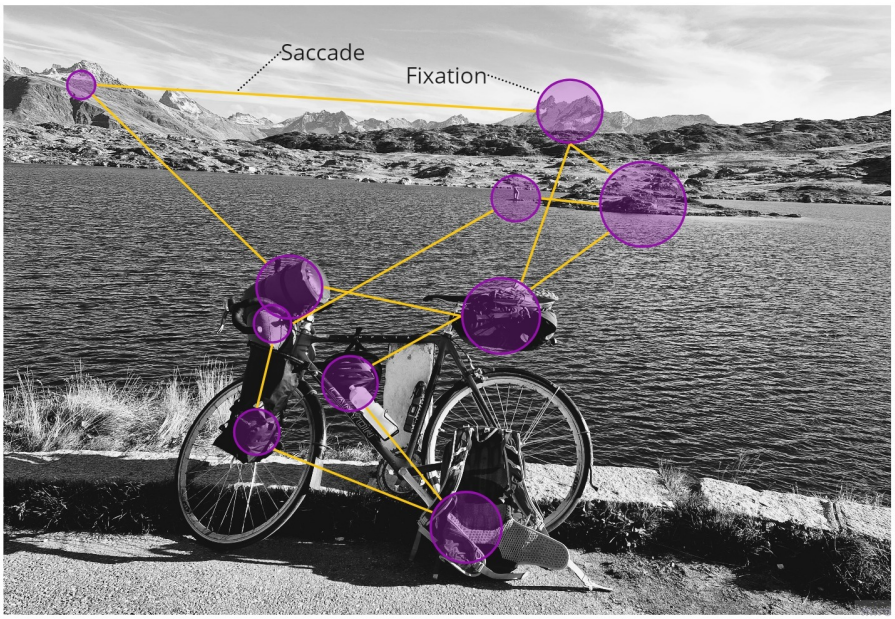


Figure 2.2: Illustration of fixations and saccades. Fixations are represented by purple circles whose size corresponds to the time the eyes rested on a target, while the yellow lines show the saccades in between.

manner. This allows the user to freely look around and interact with its environment without constraints (Stuart (2022)). This is essential to reliably assess both surgical workflow and usability of a medical device and is achieved by ET glasses equipped with a camera recording the wearer's field of view, i.e. scene video, and cameras recording each of the wearer's eyes. Computer vision enables to detect the centre of the pupils as well as corneal reflections created by the infrared light emitted by the tracker. After calibration, these quantities allow for estimating the viewing angle and in turn, where the person is looking at. This so-called gaze point can be displayed on the scene video, enabling to detect where the person is looking (Duchowski (2017)). Modern ET software supports the researcher by preprocessing the recorded data which will be further elaborated in Section 2.4.

2.3 Eye Tracking Metrics

Along with the improved ET devices, also the analysis of the collected gaze data has advanced and has proven useful in various disciplines (Stuart (2022)). Naturally, the employed measures differ depending on their use. So, different metrics are used to e.g. diagnose neurological disorders (Tao et al. (2020)),

study human psychology (Rahal & Fiedler (2019)), or support decision-making in marketing (Wedel (2018)). Certain metrics are employed with the intention to reveal meaningful information about a performed task and hence, meet the expectations of the here proposed framework. The metrics that best fit the proposed goal of the project and that are calculable with the given input were chosen to be implemented in the automated analysis tool. The selection is presented below, grouped according to what data input is required for the calculation. Some of the metrics can be defined in more than one way. Within this project, they will be calculated according to the definition provided in the manuscript. Potential interpretations of each metric as well as some previous findings are provided where available.

2.3.1 Fixation and Saccade-Based Metrics

As mentioned in Section 2.2, state-of-the-art ET software extracts fixations and saccades from the raw recordings. This provides the user, among others, with the temporal and spatial information of the fixations and saccades that can be used to calculate the ET metrics presented below. They come with the advantage of being independent of the inherent video content and therefore rather easy to determine. However, they tend to be less informative and have to be combined with other methods (Bylinskii et al. (2017)).

Number of Fixations

Naturally, the total fixations during a trial strongly correlate with the trial duration. Therefore, it is only useful if it is either normalised or if all recordings of a study have the same length. Wang et al. (2022b) have concluded that in most studies, the number of fixations decreases with increasing expertise.

Total Duration

In certain cases, the total duration of the trial can be a straightforward measure of the overall efficiency Lohmeyer et al. (2019).

Average Fixation Duration

The average duration per fixation is calculated over all fixations of a trial. Higher fixation times have been associated with higher complexity, engagement and interest. This has been explained by the fact that more difficult

tasks require more time to look at a stimulus, extract visual details and process the information (Bergstrom et al. (2014), Goh et al. (2009)). Accordingly, with growing expertise in a specific activity, this processing time and in turn, the fixation time decreases. Hence, higher fixation duration has been found in novices compared to experts in various settings like surgery (Dalveren & Cagiltay (2018)) and playing chess (Sheridan & Reingold (2017)). However, this relation was also non-significant in many other instances (Skaramagkas et al. (2021)).

Number of Saccades

Since saccades occur in between fixations, their total count strongly correlates with *Number of Fixations*.

Average Saccade Duration

The average duration per saccade is calculated over all saccades of a trial. In healthy subjects, there is a linear relationship between saccade duration and amplitude (Baloh et al. (1975)). In their review, Skaramagkas et al. (2021) have found saccade duration to be both positively and negatively correlated with the cognitive workload in different tasks. Moreover, since the distribution of saccade duration has been found to be bimodal, the mean has to be treated with caution (Devillez et al. (2020)).

Fixation/Saccade Time Ratio

The fixation/saccade ratio is calculated according to the following equation:

$$r = \frac{t_{sac}}{t_{fix}}$$

,

where t_{fix} is the total fixation time and t_{sac} is the total saccade time.

The ratio states the total time the person has spent fixating compared to travelling from one fixation to the next. The higher the ratio, the more time is spent processing compared to searching (Goldberg & Kotval (1999)).

K-coefficient

According to Unema et al. (2005), "short duration fixations followed by long saccades are characteristic of ambient processing, while longer duration fixations followed by shorter saccades are indicative of focal processing". Following this finding, Krejtz et al. (2016) created a dynamic coefficient K that distinguishes between focal and ambient vision for each fixation-saccade pair over the course of a trial and that is defined as follows:

$$K_i = \frac{d_i - \mu_d}{\sigma_d} - \frac{a_{i+1} - \mu_a}{\sigma_a},$$

such that

$$K = \frac{1}{n} \sum_n K_i,$$

where d_i and a_{i+1} are the fixation duration of fixation i and its succeeding saccade amplitude, μ_d and μ_a are the mean fixation duration and saccade amplitude, σ_d and σ_a are the fixation duration and saccade amplitude standard deviations, and K is the mean of all K_i of the n fixation-saccade pairs.

In other words, K_i is calculated for each fixation-saccade pair as the difference between the z-scores of each fixation d_i and its succeeding saccade amplitude a_{i+1}, enabling to monitor visual attention over time. K constitutes the mean K-coefficient over all fixation-saccade pairs for a given number of pairs i.

Accordingly,

$K > 0$ indicates relatively long fixations succeeded by short saccades, implying focal vision

$K < 0$ indicates relatively short fixations succeeded by long saccades, implying ambient vision

$K = 0$ represents an unusual case, where either rather short fixations are followed by short saccades, or long fixations are followed by long saccades

μ_d, μ_a, σ_d and σ_a are computed over all participants and all conditions such that differences amongst groups can be identified Krejtz et al. (2016).

2.3.2 Object of Interest-Based Metrics

In contrast to the fixation and saccade-based metrics, the object of interest (OOI) analysis makes use of the scene video recorded by the ET glasses, whereas the gaze point is mapped onto predefined objects. It thereby enables to analyse the subject's attention towards the selected objects and adds semantic meaning to the analysis.

An important concept of OOI-based ET analysis is the so-called Dwell Time which sets the foundation of many of the presented OOI-based metrics below. Dwell Time is defined as the time between the gaze entering an OOI and leaving it again (Holmqvist et al. (2011). A detailed explanation is given in Figure 2.3.

Hits per OOI

Hits per OOI, also called fixation count per OOI, is the total number of fixations that were placed onto an OOI (Doherty et al. (2010)). More hits can demonstrate increased interest, high complexity or low search efficiency (van Kasteren (2019)).

Total Hits

The more hits were counted on all OOIs together, the more attention was generally paid to the chosen objects. This value can be similarly interpreted as *Hits per OOI*.

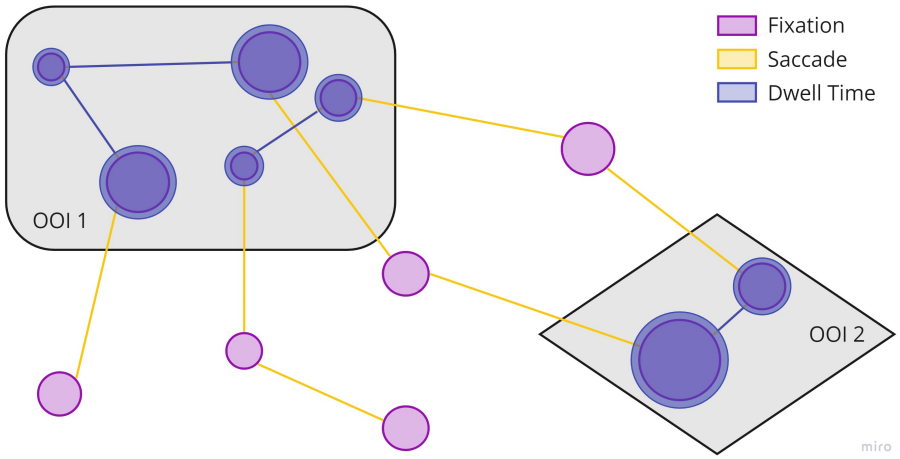


Figure 2.3: Illustration of dwell time calculation for two OOIs. One dwell is temporally defined from the entrance to the exit of the gaze into, respectively out of an OOI. Since saccades are extremely fast (2.7 ms/degree (Baloh et al. (1975))), it is not practically feasible to capture the exact time point of the gaze entering an OOI. Thus, in this work, the dwell time is calculated from the first to the last fixation within an OOI, illustrated in blue. The total dwell time incorporates all dwells, which would be two for OOI 1 and one for OOI 2. Thus, the Total Dwell Time per OOI is calculated as the sum of the dwell times of all revisits in that OOI.

Total Fixation Time per OOI

The total fixation time is the sum of all fixation durations of a trial. A high total fixation time can indicate high complexity of an OOI, but can also mean that the OOI is more engaging and interesting to the subject. Often, the Total Fixation Time per OOI correlates strongly with the Total Hits per OOI (Just & Carpenter (1976)).

Average Fixation Duration per OOI

The average fixation time per OOI is the mean duration of all fixations placed onto a particular OOI. See *Average Fixation Duration*.

Time to First Fixation per OOI

This metric states how quickly an OOI captures the participant's attention. In general, the less time passes until the object is noticed, the higher its importance or the more noticeable it is (Guo et al. (2016)).

Total Dwell Time per OOI

The total dwell time per OOI is the sum of all dwell times onto a particular OOI and thus, how much time the participant has spent looking at the OOI. Again, longer dwell times can correspond to either high interest or high cognitive effort. In the case of usability studies, the latter could reflect lower ease of use of a device (Mussgnug et al. (2017), Lohmeyer et al. (2019)). As opposed to total fixation time, where solely the fixation times are summed up, the total dwell time also includes the saccades in between. Since saccades are very short, these two values only slightly differ from each other, particularly in short trials.

Relative Dwell Time per OOI

The relative dwell time of a particular OOI is calculated as the total dwell time of all OOIs divided by total dwell time of the particular OOI and reflects how much time the subject has spent looking at this OOI compared to all the other OOIs.

Average Dwell Time per OOI

OOIs with a higher average dwell time may keep up the attention for longer individual time periods.

Average Dwell Time

The average dwell time is calculated as the mean of all dwell times of all chosen OOIs. See *Average Dwell Time per OOI*.

Revisits per OOI

Revisits are the number of times a particular OOI was revisited. Again, more revisits can indicate high engagement but also difficulty with performing the task or with comprehension.

Total Revisits

The total number of revisits is defined by the number of times that all the defined OOIs together were revisited.

Stationary Gaze Entropy

Stationary Gaze Entropy (SGE) was first introduced by Krejtz et al. (2014). It represents the stationary distribution of the gaze over all OOIs and is calculated using Shannon's equation for entropy H (Shannon (1948)):

$$H = -\sum_{i=1}^{n} p_i * \log_2(p_i) \tag{2.3.1}$$

In a first step, the observed entropy SGE_{obs} is calculated using Equation 2.3.1, whereas

$$p_i = \frac{n_i}{n_{tot}},$$

where n_i is the number of hits of the i-th OOI, n_{tot} is the Total Hits, and n is the number of OOIs.

In a second step, again using Shannon's equation 2.3.1, the maximum entropy H_{max} is calculated, i.e. the entropy in case of an even distribution of the hits amongst all OOis, whereas

$$p_i = \frac{1}{n_{OOI}},$$

where n_{OOI} is the number OOIs.

In a third step, the calculated maximum entropy H_{max} is used to normalise the observed entropy SGE_{obs}:

$$SGE_{norm} = \frac{SGE_{obs}}{H_{max}}$$

Thus, a higher SGE implies a more equal distribution of visual attention between the OOIs. A lower value reflects when fixations tend to be concentrated on specific OOIs, either because they are more complex or more interesting to the subject (Krejtz et al. (2014)).

For a comprehensive explanation of the calculation of the SGE, please refer to the supplementary material that Shiferaw et al. (2019) provided in their review.

Gaze Transition Entropy

The second entropy measurement that Krejtz et al. (2014) introduced was the Gaze Transition Entropy (GTE). In contrast to the SGE which utilises the static distribution of hits only, the GTE takes the order of the hits into account, making it a more dynamic metric.

Firstly, a first-order Markov chain is modelled from the sequence of hits. For this metric, the background is counted as an additional OOI to accurately represent the transitions of the gaze. From there, a probability transition matrix is constructed, including within-state transitions, as seen in Table 2.1. It states the probability $p_{i,j}$ of the gaze moving from one specific OOI (prior state i) to the same or another specific OOI (current state j).

Table 2.1: Example of a transition probability matrix. The value in row i and column j represents the probability of the next hit being on the j-th OOI, if the current hit was on the i-th OOI (Krejtz et al. (2014)).

P(i,j)	1	2	3	4
1	0.35	0.15	0.26	0.24
2	0.21	0.21	0.37	0.21
3	0.27	0.38	0.33	0.02
4	0.51	0.06	0.13	0.3

Again, Shannon's equation 2.3.1 is employed in order to calculate the observed gaze transition entropy GTE_{obs} of the transition matrix:

$$GTE_{obs} = -\sum_{i=1}^{n} p_i \sum_{j=1}^{n} p(i,j) * \log_2(p(i,j))$$

As a next step, GTE_{obs} is normalised by dividing it by the maximum entropy H_{max} that is computed in the same manner as explained in *Stationary Gaze Entropy*:

$$GTE_{norm} = \frac{GTE_{obs}}{H_{max}}$$

In short, the GTE represents the level of uncertainty or randomness in the pattern of visual scanning, where a higher entropy means less predictability (Ellis & Stark (1986)). This could imply more randomness in the visual scanning pattern and in turn, less focus and efficiency. For a comprehensive explanation of the calculation and interpretation of the GTE, please refer to Shiferaw et al. (2019)'s review.

2.3.3 Action-based metrics

To conduct an action-based analysis, each time point of the trial is assigned to one of the predefined actions. The above-introduced metrics are then calculated over each action, allowing for a much deeper look into the performed task compared to the calculations over the entire trial. Additionally, for each trial, a sequence of actions along with the duration can be extracted. Next to the already presented metrics, two additional ones are introduced.

Average Duration per Action

Similar to *Total Duration*, the average duration per action can be used to assess efficiency for each of the predefined actions.

Levenshtein Distance

The Levenshtein distance is a common measure to quantify how similar two strings are to each other. The algorithm computes the minimum number of operations required to change one string into another string through three operations on the individual characters; insertion, deletion, and replacement Levenshtein et al. (1966). This approach can be applied to compare how different the observed sequence of actions of a trial is from a predefined template sequence. This template sequence could be e.g. the action sequence of an expert user or the sequence stated in the instructions for use (IFU) of a medical device. The higher the calculated distance, the bigger the difference between the two sequences.

2.4 Automation of Eye Tracking Analysis

In remote ET, the recorded gaze data is linked to the screen and thus, fixations can easily be mapped onto objects on the screen, enabling an efficient OOI analysis. However, in MET, this link does not exist, and in turn, mapping the gaze onto the objects is a lot more challenging. Thus, numerous methods have been suggested to overcoming this bottleneck of MET studies (Wegner (2021)). One approach is the manual fixation-by-fixation analysis, where for each fixation, the researcher compares the gaze point to the OOIs in the scene video and notes the hits (Ooms et al. (2015)). Despite being a tedious and costly process, this method represents the current state of the art, as it is applicable to almost all types of ET studies (Wolf et al. (2018)). A

suggested procedure to speed up this process is the manual assignment of each OOI on the scene video (e.g. with a bounding box) on a number of key frames, whereas the algorithm creates estimates for the frames in between through interpolation. However, this method has been found not suitable for interactive MET studies due to unpredictable user behaviour (Ooms et al. (2015)). Another example is the placement of markers on OOIs which indeed has been shown to accelerate the mapping process. However, the markers come with the disadvantage of distracting the subject and since they have to be readable in the scene video at all times, they are limited in their application (Evans et al. (2012)). Thus, Vansteenkiste et al. (2013) conclude that it is almost inevitable to manually assign the gaze point onto OOIs in studies that aim for a natural setting.

Multiple solutions have been presented in order to accelerate this inefficient process. Deane et al. (2022) and Wolf et al. (2018) both used a deep learning-based algorithm, the computational Gaze-Object Mapping (cGOM) and the deep-learning-based system for automatic gaze annotation (deep-SAGA), respectively, that automatically maps the gaze point coordinate onto OOIs. This is achieved by employing the state-of-the-art object detection and segmentation algorithm Mask R-CNN and extending it with a gaze-mapping feature. A gaze mapping occurs when the fixation coordinates lay within the detected OOI mask. In contrast to these two-step procedures that are applied in a frame-by-frame manner, Uppal et al. (2022) introduced an end-to-end deep-learning-based method that incorporates gaze data into the network which is combined over multiple consecutive frames. However, this approach was outperformed by the two-step solutions.

Nonetheless, these methods are completely neglecting the individual's peripheral vision even though its importance in decision-making and task execution is undisputed (Krupinski et al. (2006), Reingold & Sheridan (2011)). Therefore more recently, Wang et al. (2021) extended this method by introducing the Object-Gaze Distance (OGD) algorithm, where for each fixation the 2D Euclidean pixel distance between each detected OOI mask and the gaze coordinates is calculated. Thereby, OOI hits can be extended from one single pixel to a higher number that more accurately represents the foveal vision. Moreover, next to more realistic gaze mapping, it allows to include the peripheral vision in the analysis. Further, the OGD algorithm has been used for HAR (Wang et al. (2022a), currently under review). The newly introduced Peripheral Vision Based Hidden Markov Model (PVHMM) algorithm for HAR exploits the OGD data to predict actions. Again, this comes with the benefit of including peripheral vision. The OGD and PVHMM algorithms will be described in more detail in Section 3.1.

3 Methods

The goal of this project is to combine the previous works of *pd*|**z** at ETH, namely the OGD and the PVHMM algorithm (Wang et al. (2021), Wang et al. (2022a)), with different types of ET-based performance analyses outlined in Section 2.3. In the first Section of this chapter (3.1), the framework is thoroughly discussed. Afterwards, in Section 3.2, the proof-of-concept study to test it is described, followed by the validation with ET data from a medical device usability study in Section 3.3.

3.1 Framework

The source code developed and used during this thesis can be found in Appendix A.1. The framework was developed in Python (version 3.6.12) using pandas (version 1.1.5), numpy (1.19.5), matplotlib (version 3.3.4), and seaborn (version 0.11.2), among others.

3.1.1 Pipeline

A simplified overview of the pipeline is depicted in Figure 3.1. In the first step, the recorded ET data is processed in three different ways to generate the data input for the framework, which will be described in Section 3.1.2. Additionally, user input is required in order to provide the paths to the data input and output, as well as to configure the settings for the analysis. This will be outlined in Section 3.1.3. Once these two inputs are collected, the analysis can be started. The analysis is divided into six modules that will be presented in Section 3.1.4. Within these modules, the calculated metrics in the form of tables and graphs will get exported into the output directory, as described in Section 3.1.5. Additionally, easy-to-understand reports that summarize the output are exported as .pfd files.

The directory tree of the framework can be found and is commented on in

Appendix A.2.

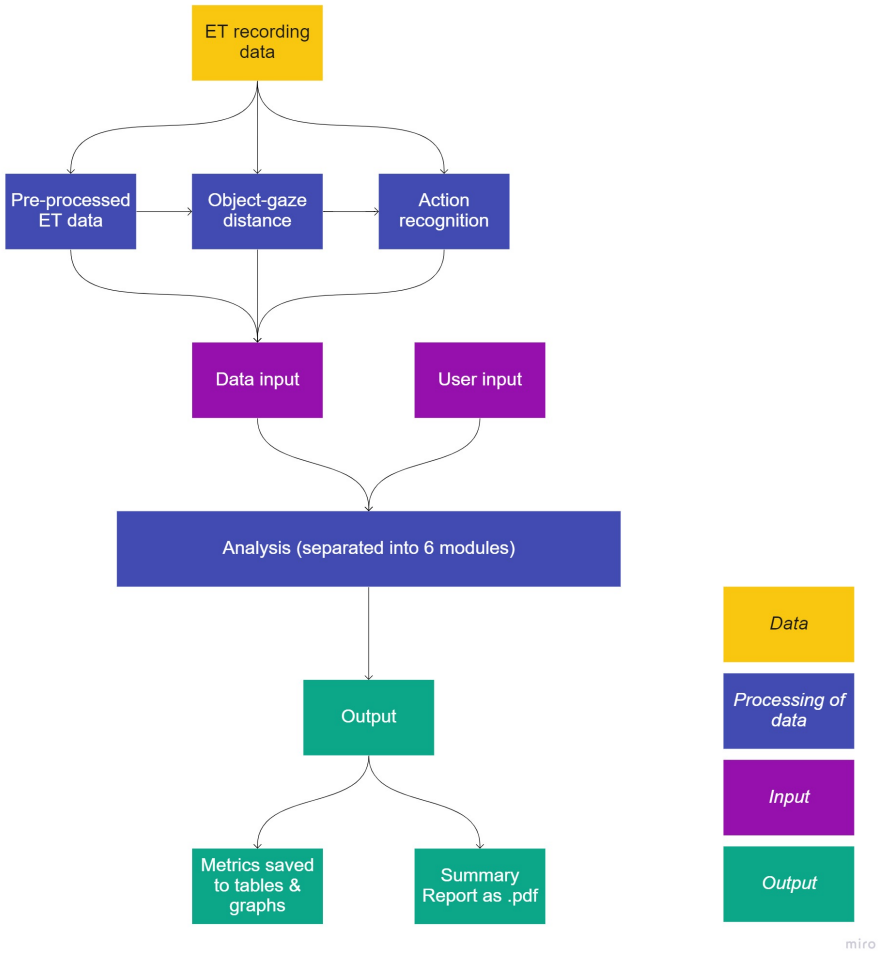


Figure 3.1: Proposed pipeline showing the way from the recorded ET data through the framework to the output.

3.1.2 Data Input

Three distinct data input files exist for the six modules. The generation of each of these input files is discussed below and an overview is presented in Figure 3.2.

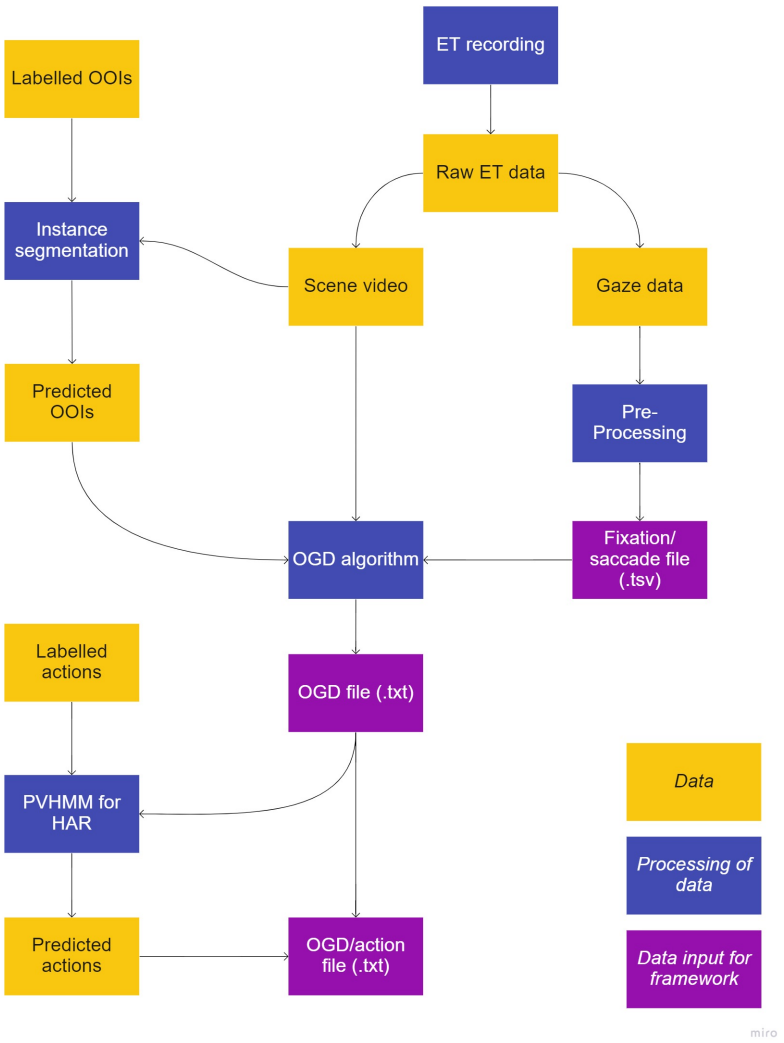


Figure 3.2: Data input structure of the framework. The recorded ET data is processed in multiple procedures to generate three distinct data input files that are fed to the framework: *Fixation/saccade file*, *OGD file*, and *OGD/action file*.

ET Recording Data

In this work, the Tobii Pro Glasses 2 (TG2, Figure 3.3) and the accompanying software Tobii Pro Lab are used to collect and pre-process gaze data. The TG2 provide robust state-of-the-art MET technology. It features two infra-red cameras per eye to detect the direction and orientation of each pupil by using the corneal reflection technique. The intersection of the two gaze vectors of both eyes is used to compute the gaze point coordinates for each fixation. The front-facing camera has a resolution of 1920 x 1080 and records the user's

field of view at 25 frames per second (Tobii AB (2017)). The recorded data, in particular the scene video and the gaze coordinates, can be loaded into the accompanying software Tobii Pro Lab where they are superimposed, as seen in a screenshot in Figure 3.4. The software supports the user by extracting fixations and saccades by applying a filter, namely the *Tobii I-VT (Attention)* filter. It was created for MET data recorded by the TG2 under dynamic conditions where either the subject, the targets, or both are moving (Tobii AB (2022)). The exported .tsv document (*Fixation/saccade file*) provides the start and end time of fixation and saccades, the fixation coordinates, as well as the saccade amplitudes. These are used to conduct the *General analysis* and the *K-coefficient analysis* afterwards.

Figure 3.3: The Tobii Pro Glasses 2 connected to the recording unit.

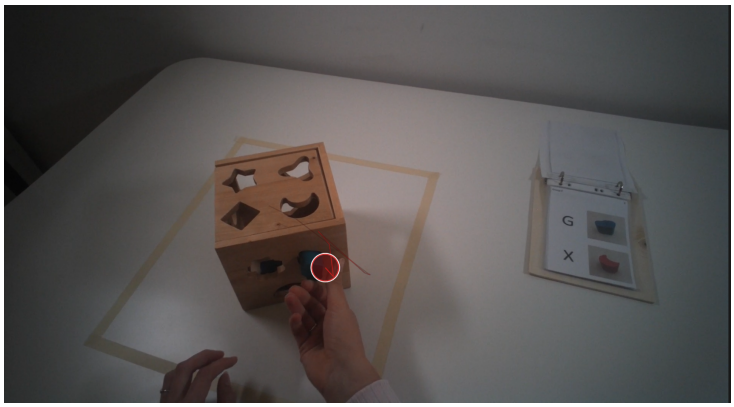


Figure 3.4: Screenshot of the Tobii Pro Lab where the gaze path is laid over the scene video.

OGD Data

The OGD algorithm uses instance segmentation (e.g. Mask R-CNN, see Section 3.2.3) of the OOIs in the scene video and the fixation coordinates of the *Fixation/saccade file* to detect the OOIs and calculate the Euclidean pixel distance from each OOI to the gaze point (Wang et al. (2021)). This is done for every fixation and exported as the *OGD file*, more precisely a .txt document where each row corresponds to one fixation, as shown in Table 3.1. It is important to note that the instance segmentation algorithm requires a number of labelled images as the ground truth to learn from (*Labelled OOIs* in Fig. 3.2).

Table 3.1: Format of the *OGD file*. The numeric values indicate the distance of each OOI to the fixation point in pixels at the given time point.

start_time	end_time	App	Cap	Gauge	Pad	Pen	Safety	Tip
0.85	0.95	300.8	24.7	183.8	551	0	368.1	1600
0.95	1.05	302.4	30	182.9	552	0	380.7	1600
1.05	1.1	305.4	31.8	182.4	560	0	387.4	1600
1.119	1.219	279.7	65.8	151.1	533	0	431.7	1600
1.269	1.369	265.7	98.7	120.2	503.1	0	468.4	1600

OGD Data with HAR

If labelled actions are available, the *OGD file* can subsequently be used for HAR by employing the PVHMM algorithm (Wang et al. (2022a)). Simply put, Hidden Markov models (HMM) have the ability to establish stochastic relations between observable states (i.e. peripheral vision gaze data) and hidden states (i.e. human actions) by taking their temporal relation into consideration. For this, HMM classifiers require the observable state to be one single numerical entry. Thus, for each fixation, the object-gaze distances of all OOIs are categorized into distinct vision areas in physiological accordance (foveal, parafoveal, perifoveal and peripheral, see Section 2.1). By using a dictionary transformation, they are transformed into one single number that represents the vision area of each OOI and by this means, the observed state. The temporal sequence of these states, along with labelled actions, are used to train a model that predicts the action of each fixation of the *OGD file* input. The file is extended by a column stating the predicted action to generate

the *OGD/Action file*. An extract of such a file is provided in Table 3.2. For a more detailed explanation of the PVHMM algorithm, please refer to the Master Thesis of Kreiner (2021).

Table 3.2: Format of the *OGD/Action file*. The numeric values indicate the distance of each OOI to the fixation point in pixels at the given time point. The Action column denotes the detected action by PVHMM.

start_time	end_time	App	Cap	Gauge	Pad	Pen	Safety	Tip	Action
0.85	0.95	300.8	24.7	183.8	551	0	368.1	1600	Cap Off
0.95	1.05	302.4	30	182.9	552	0	380.7	1600	Cap Off
1.05	1.1	305.4	31.8	182.4	560	0	387.4	1600	Cap Off
1.119	1.219	279.7	65.8	151.1	533	0	431.7	1600	Cap Off
1.269	1.369	265.7	98.7	120.2	503.1	0	468.4	1600	Cap Off

Data Input Format

In order for the program to be able to correctly summarise the trials per participant and per group, a specific input format has to be followed, which can be found in the Appendix A.3

3.1.3 User Input

Input Variables

The following input elements can be set manually in the configuration file, in particular a .yaml file, or via the graphical user interface (GUI):

- File paths. The user has to provide the location of the input files, i.e. the *Fixation/saccade file* and, depending on availability, the *OGD file* or the *OGD/Action file*, as well as the output directory.
- Selection of modules. The aforementioned modular structure of the framework enables the user to specify which of the six parts of the analysis (introduced in Section 3.1.4) they want to run, depending on the available data or the research goals.

- List of actions. If the module *Action analysis* (see 3.1.4) is selected, the user is asked to provide all available actions in the form of a list.
- Template sequence. In case the user wants to conduct a sequence comparison to calculate the Levenshtein distance, they must enter a template sequence.
- Hits radius. Additionally, thanks to the *OGD file*, OOI hits can be extended from one pixel to the desired number in order to more accurately capture the foveal vision. This can be done by changing the hits radius.

Graphical User Interface

In order to make the program more user-friendly, a GUI was designed with Qt Creator. In the "Main Window" (Figure 3.5, left), the user can provide the path to the input and output folder. Also, they can select the desired sub-analyses and click "Apply". If *Action-based analysis* was selected, another window opens; "Action List" (Figure 3.5, top right). There, the user has to provide a list of actions. This can be done by either entering individual actions into the upper text box and clicking 'Add' or by typing the entire list into the lower text box. If on top of the *Action-based analysis*, also the *Sequence comparison* box was checked, another window named "Provide Sequence" will appear that asks for the template sequence in the form of a list (Figure 3.5, bottom right). There, the list from the previous window appears to inform the user about the available actions. Now, the actions have to be added in the correct order, in the same manner as before. Once the user proceeds with 'Next', the analysis can be started. As an additional feature, in the main window, the configuration can be saved into a .yaml file to later be loaded into the GUI again.

3.1.4 Modules

As stated above, the framework is divided into six individual parts, each of which will be introduced in this Section. An overview of the modules and their dependencies can be found in Figure 3.6. Each module corresponds to a separate .py file that is called by the main file *app.py* if specified in the user input. Please refer to Appendix A.2 for further details on the directory structure.

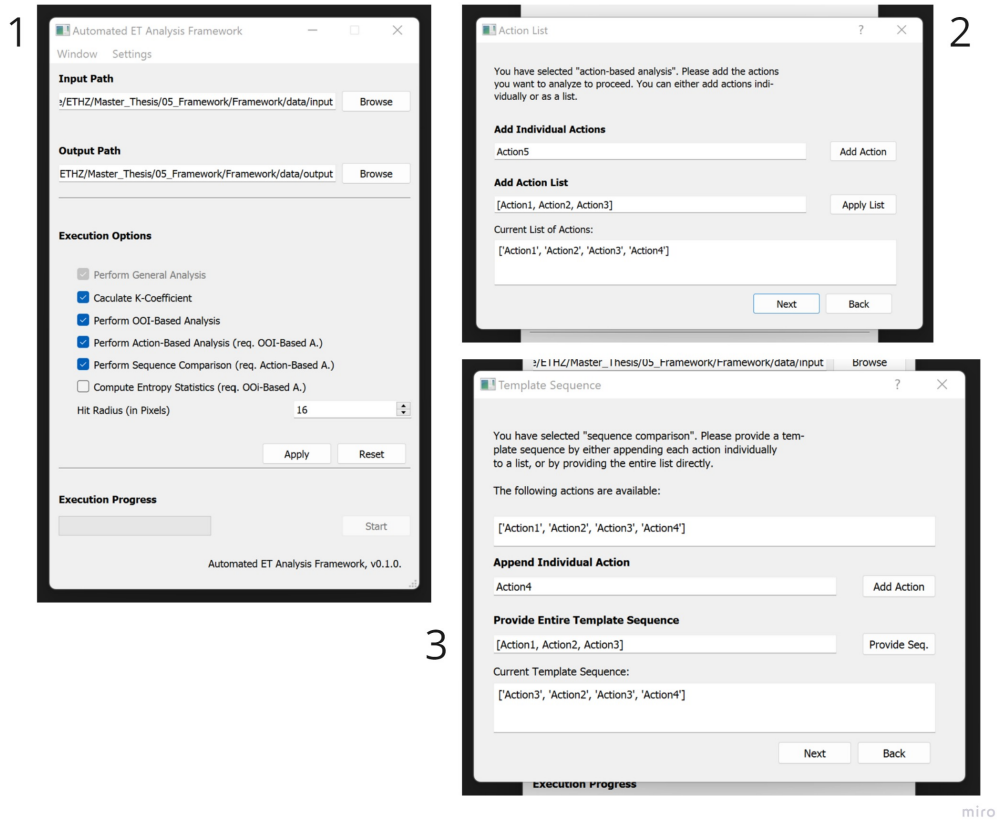


Figure 3.5: Screenshots of the "Main Window" (1), the "Action List" window (2), and the "Provide Sequence" window (3).

General Analysis

Since the *Fixation/saccade file* sets the basis for the generation of the two other inputs file, it has to be created in any case. Therefore, the *General analysis*, which only requires this file as data input, is always executed and the user cannot choose to exclude it from the analysis.

Initially, the *Fixation/saccade file* is split up into two new files that are filtered by either fixations or saccades only. In the next step, the following fixation and saccade-based metrics are calculated from one of the two newly created files:

- Number of Fixations
- Number of Saccades
- Total Duration [ms]

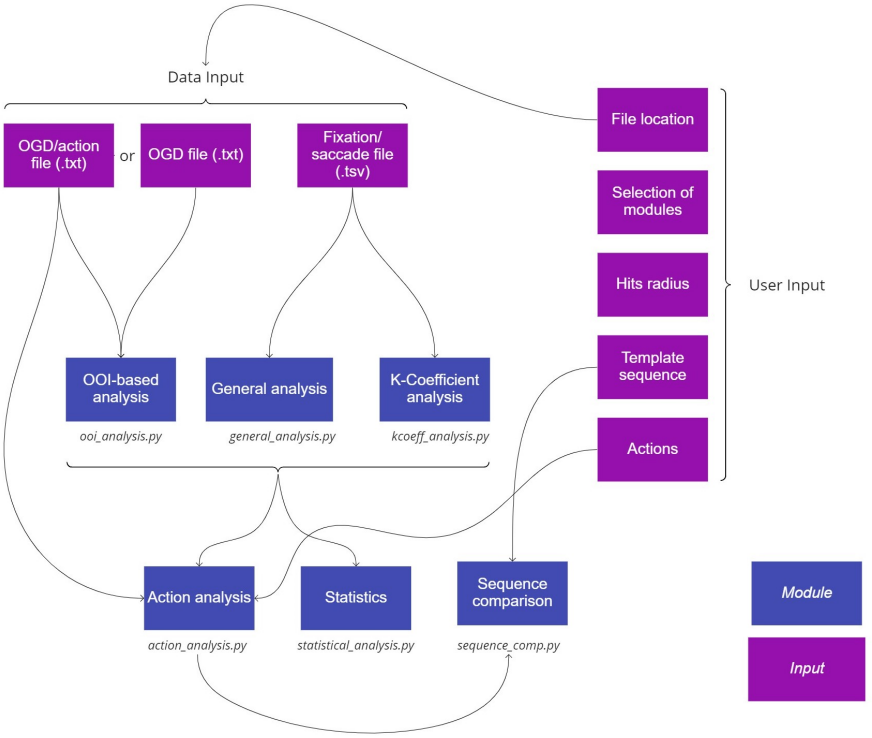


Figure 3.6: Overview of the six modules and their dependencies on the data input and the user input.

- Average Fixation Duration [ms]
- Average Saccade Duration [ms]
- Fixation/saccade Ratio

The calculation and an interpretative approach are given in Section 2.3.1. The analysis is conducted by running *general_analysis.py*.

K-coefficient Analysis

As stated in Section 2.3.1, the K-coefficient is calculated via the fixation duration and saccade amplitude. Thus, the *Fixation/saccade file* serves as an input document. Further, the computation necessitates the mean and the standard deviation of the fixation durations and saccade amplitudes over all trials across all conditions or groups. Hence, the program first loops through all trials to calculate the standard deviations, μ_d and μ_a, and the means, σ_d and σ_a. In a second step, the K-coefficient K_i is determined for each

fixation-saccade pair i, providing insight into how the ambient/focal vision changed over the time course of a trial. Additionally, the K-coefficients over time are saved as a .csv to provide the user with the exact numbers over time, but also to be able to determine the K-coefficient per action, if the user wishes to do so.

In the next step, the mean K-coefficients per trial K (of all fixation-saccade pairs) are determined. From there, the overall mean K-coefficient K_{all} and its standard deviation σ_K over all trials are resolved. Finally, for each mean K-coefficient K, it is determined whether it is inside or outside two standard deviations σ_K from K_{all}. These K are marked in both the exported tables and bar plots. An example of the latter is shown in Figure 3.7.

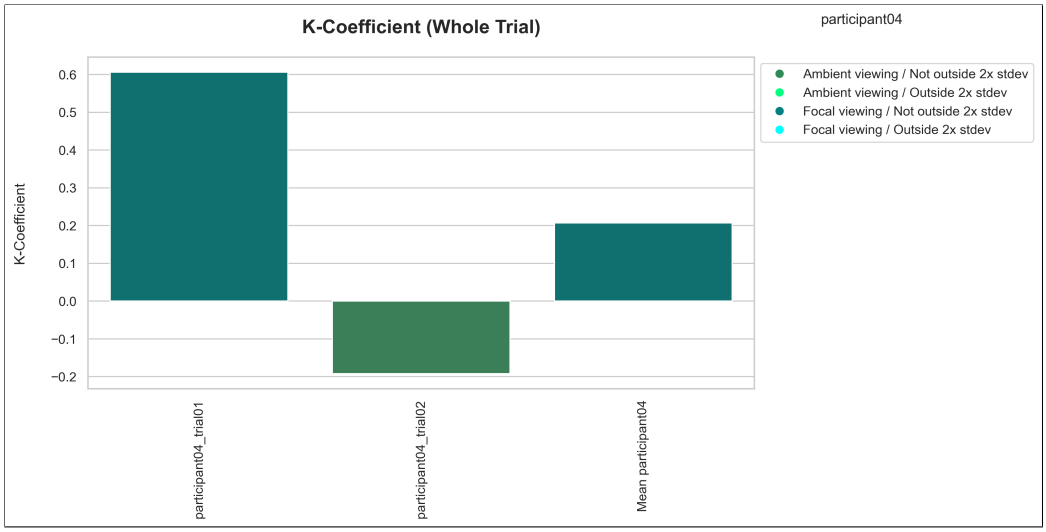


Figure 3.7: Example of a bar plot showing the mean K-coefficient across all trials of a group compared to the mean K-coefficient per participant. K > 0 indicates relatively long fixations and short saccades, implying focal viewing, whereas K < 0 indicates relatively short fixations and long saccades, implying ambient viewing. For either group, the value is still within two standard deviations from the overall mean.

OOI-based Analysis

Depending on the availability of information on actions, the *OGD file* or the *OGD/Action file* is used as data input for this module. The *OOI-based Analysis* is separated into two parts, both are carried out by the execution of *ooi_analysis.py*.

The first part incorporates the metrics that are calculated per OOI, namely:

- Hits per OOI
- Total Fixation Time per OOI [ms]
- Average Fixation Duration per OOI [ms]
- Time to First Fixation per OOI [ms]
- Total Dwell Time per OOI
- Relative Dwell Time per OOI
- Average Dwell Time per OOI
- Revisits per OOI

The second part contains metrics that are calculated for all OOIs together:

- Total Hits
- Average Dwell Time [ms]
- Total Revisits
- Stationary Gaze Entropy
- Gaze Transition Entropy (with Transition Matrix)

Action-based Analysis

At first, according to the *OGD/Action file*, a sequence of actions with the corresponding start and end times is created and saved. This list will be used for the *Sequence Comparison* if the user decided to conduct one. According to this list, the previously ran modules (*General analysis, K-coefficient analysis, OOI-based Analysis*) are separated per action. In other words, the metrics are calculated over the time frame at which the respective action took place. Next to this, there is solely one action-specific metric calculated:

- Average Duration per Action [ms]

Sequence comparison

Since this module (*sequence_comparisons.py*) makes use of the previously created list of actions, it can only be conducted if *Action Analysis* has been executed too. The observed sequence is then compared to the template sequence that is delivered via the user input. Up to now, there is only one sequence comparison algorithm that can be carried out:

- Levenshtein Distance

However, this can readily be extended in a future version of the program.

Statistics

The statistical analysis of the computed metrics has not been the primary focus of the pipeline, therefore it has only been implemented to a very limited extent and solely incorporates two metrics to date, namely the SGE and the GTE. The algorithm iterates through all trials in order to calculate the overall mean and standard deviation across all conditions. Subsequently, each (mean) SGE and GTE (calculated for all levels, described in Section 3.1.5) that are outside two standard deviations of the mean are marked as such in the respective bar plot, similar to the *K-coefficient analysis*. An example of such a graph is depicted in Figure 3.8. When the plot is automatically exported, it replaces the corresponding figure that was previously generated by the *OOI-based analysis* that lacks this marking. Again, the module is built in a way that enables efficient extension in the future, i.e. to add more metrics.

3.1.5 Output

Levels

The metrics presented above are calculated and summarized on different levels:

- per trial
- per participant
- per group
- for all groups

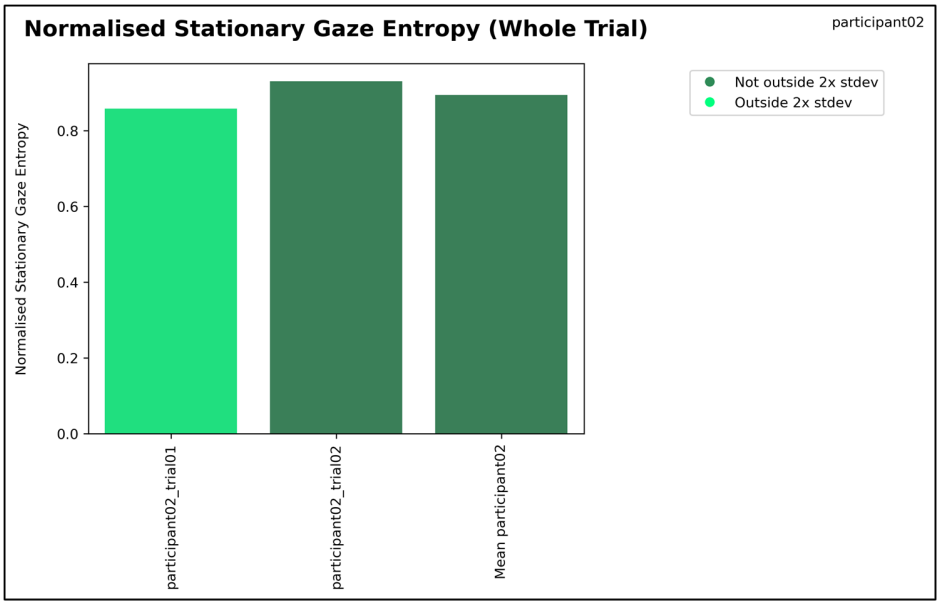


Figure 3.8: Bar plot showing the normalized mean SGE across all trials of one participant compared to the mean SGE per trial. The higher the mean SGE, the more equally distributed the visual attention between different OOIs. Values within or outside two standard deviations of the mean are shown in different colours.

Consistent with this structure, the resulting files are exported to a directory tree, as depicted in Figure 3.9. In this example, only the *General analysis* and the *K-coefficient analysis* were run.

The mean and the standard deviation of each level are calculated from the means of the level below in order to ensure equal weights of each of these elements. For instance, the mean and standard deviation of a group is calculated from all the participant means, and not from the trial means, so that all participants have a similar influence, even if they conducted a different amount of trials.

Output Files

All the calculated metrics are saved in data frames and exported as .csv files to the respective directory so that they can be reviewed and analysed in detail. Additionally, many different graphs are created and saved in a separate *visualisations* folder in the corresponding analysis directory (e.g. *general_analysis*, see Figure 3.9). Figure 3.10 provides an extract of an output generated by running the *OOI analysis*.

```
output/
├ general_analysis/
├ k-coefficient_analysis/
├ group1/
├ group2/
|  ├ general_analysis/
|  ├ k-coefficient_analysis/
|  ├ participant04/
|  ├ participant05/
|  ├ participant06/
|  |  ├ general_analysis/
|  |  ├ k-coefficient_analysis/
|  |  ├ participant06_trial01/
|  |  ├ participant06_trial02/
|  |  |  ├ general_analysis/
|  |  |  ├ k-coefficient_analysis/
```

Figure 3.9: Directory tree of an example output generated by running the *General analysis* and the *K-coefficient analysis* in a study with two groups and six participants with two trials each.

Results Summary for Novices

As seen in Section 2.3, for the majority of the selected metrics it is not possible to provide a straightforward statement on the performance, since it is highly task-specific. However, to support novice users with little expertise in ET data analysis, the output is summarised in a short report that is created for each participant, each group and all groups. It is saved as a .pdf file in a separate sub-directory named "Summary Report". The report is divided into three categories; Efficiency, Focus and Attention/OOI-based Analysis. For each category, the results of one to four metrics are graphically presented, depending on the executed modules. Further, for each metric, there is a suggestion on how to interpret these values that are based on the literature summarized in Section 2.3. A snippet of an example report is shown in Figure 3.11.

	A	B	C	D	E	F	G	H
1		App	Cap	Gauge	Pad	Pen	Safety	Tip
2	Hits	141	39	78	20	72	88	56
3	Total Fixation Time [ms]	13727	3600	7538	1770	6510	8476	5198
4	Average Dwell Time [ms]	7302	656.8571429	1595.4	599.5	422.1111111	551.4705882	381.2
5	Total Dwell Time [ms]	14604	4598	7977	2398	7598	9375	5718
6	Revisits	2	7	5	4	18	17	15
7	Average Fixation Time [ms]	97.35460993	92.30769231	96.64102564	88.5	90.41666667	96.31818182	92.82142857
8	Time to First Fixation [ms]	20978	2039	14382	33311	959	6296	7066
9	Relative Dwell Time [%]	27.94061376	8.796969465	15.26172802	4.587893166	14.53661896	17.93640468	10.93977194

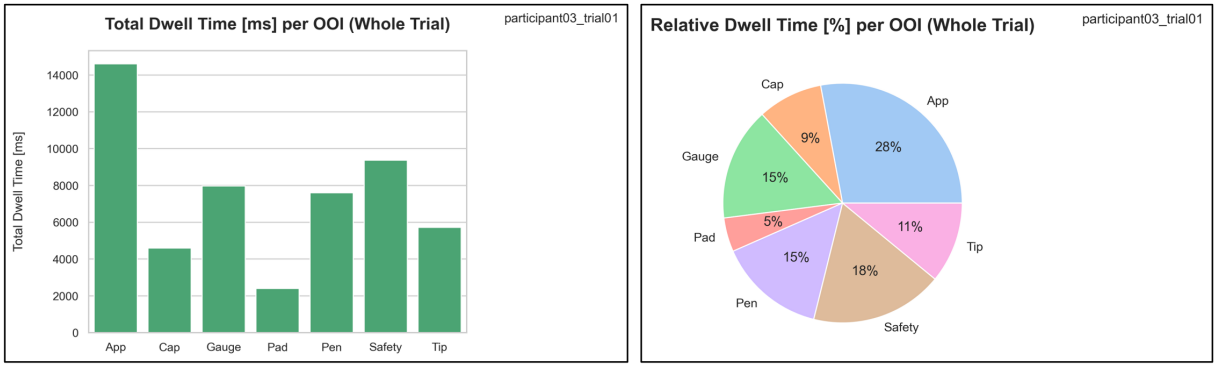


Figure 3.10: Example .csv output of the metrics calculated per OOI and two associated graphs that are automatically created by the framework, namely the Relative Dwell Time per OOI and the corresponding Total Dwell Time.

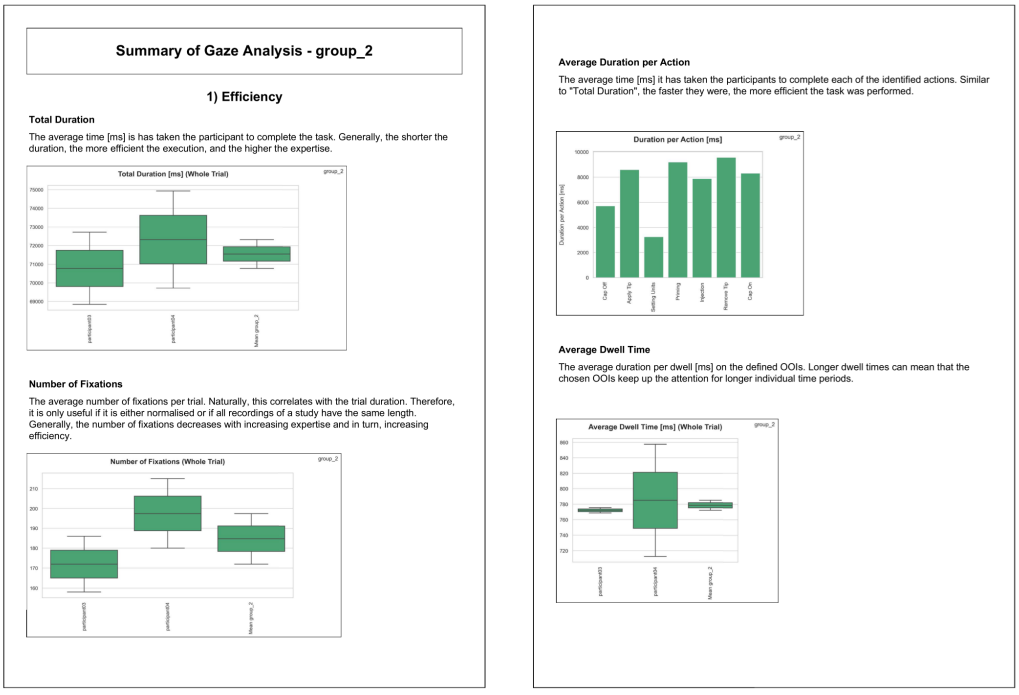


Figure 3.11: Two first pages of an example Summary Report showing results regarding efficiency along with a suggestion on how to interpret the values.

3.2 Study

The goal of the study is to test if the framework is suitable for a performance assessment with the chosen metrics presented in Section 2.3 by utilizing the *Fixation/saccade file* and the output generated by the OGD algorithm of Wang et al. (2021), namely the *OGD file*. Due to time constraints, HAR by *PVHMM* had to be omitted for this study. Hence, the modules *Action-based analysis* and *Sequence comparison* could not be executed. Two tasks of distinct difficulty were designed in order to assess if the implemented metrics can reveal relevant differences between them. Firstly, the experimental setup and the materials are described, followed by the study design and the hypotheses. Lastly, the data analysis workflow of the recorded ET data is presented.

3.2.1 Experimental Set-Up and Materials

Subjects

30 subjects, mostly students, aged between 20 and 31 years, participated in the study, of which two served as pilot tests. All participants reported normal or corrected-to-normal vision and no neurological conditions. Each participant provided informed consent prior to testing and received monetary compensation afterwards.

Recording Set-Up

The TG2 introduced in Section 3.1.2 along with the Tobii Pro Glasses Controller software were used to record the trials. After gently mounting the glasses onto the subject, it was checked if they fit securely and if the pupils are clearly visible in the eye-facing cameras. If necessary, the nose pad was exchanged. After a successful calibration, the recording was started.

Stimuli

In order to test the framework in a proof-of-concept manner, a simple task that allows for two different difficulty levels was desired. For this purpose, a shape sorter toy was used as the study stimulus, depicted in Figure 3.12. The toy is made for children from twelve months on, and the goal is to put the coloured small shapes (referred to as *shapes* in the following) into

the corresponding slot in the big wooden block (referred to as *block* in the following). For the study, the subjects were asked to toss ten shapes into the block in a specific order and orientation according to a printed instruction pad. The shapes were placed in a marked 5x5 grid where each shape had its designated spot. The instructions consisted of five pages with two shapes pictured, one above the other, on each page. The subjects were asked to work through the instructions from top to bottom and from the first to the last page. Additionally, two letters in the form of stickers were attached on two opposite sides of the shapes. On the instructions, next to the depicted shapes, one or multiple letters (depending on the difficulty) were listed, indicating which letter on the shape should point towards the opening. The task was completed once all ten shapes have been tossed into the block. The two different levels of difficulty will be explained in detail in Section 3.2.2.

Figure 3.12: Photo of the shape sorter toy (Legler OHG small foot company). The goal of the game is to toss the coloured shapes into the corresponding slot in the wooden block.

The task was performed on a table, whereas the wooden block was placed on the left, the instructions in the middle, and the coloured shapes on the right, as depicted in Figure 3.13. The TG2 can only record a person's gaze when watching through the glasses, and not below them. This can be avoided by making sure the objects the participants are interacting with are placed at a certain minimum distance. Thus, the task was conducted in a standing position, whereas the table's height was adjusted to the height of the subject's belly button. Additionally, the block was placed in a marked zone and the participants were asked to not move it out of there, as seen in Figure 3.13.

3.2.2 Study Design

The subjects were randomly assigned a participant number from one to 30. Thereby, they got evenly divided into two groups; participants 1 to 15 make

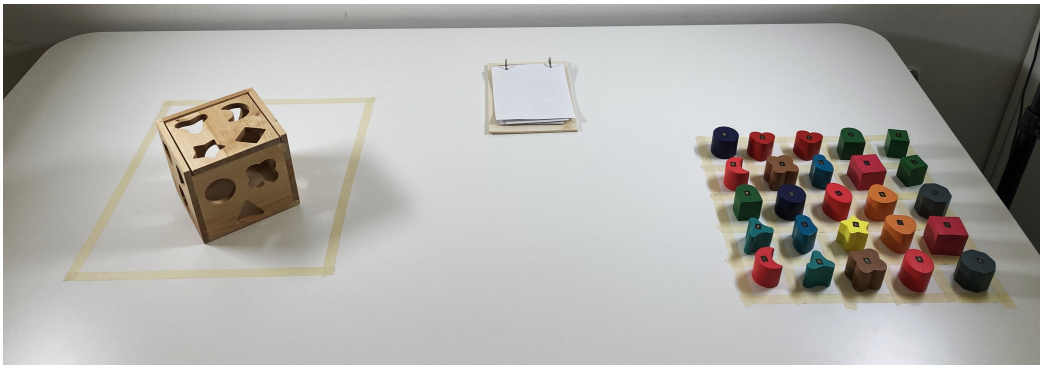


Figure 3.13: Photo of the set-up encountered by the participants at the start of the experiment.

up *Group Easy* and 16 to 30 belong to *Group Difficult*. Each participant repeated the task three times. The trials of the first participant of either group served as pilot tests and were not included in the analysis. In particular, these were *participant02* and *participant22*. Hence, in total, 84 recordings made up the preliminary data set, 42 per group. The two tasks differed in the instructions on sequence and orientation, as well as in the selection of shapes in the grid. Thus, once the correct shape was discovered and the correct orientation was identified, finding the correct slot on the block was similar in both levels of difficulty. The two tasks are described in the following. Further, a detailed comparison of the two tasks can be found in Table 3.3.

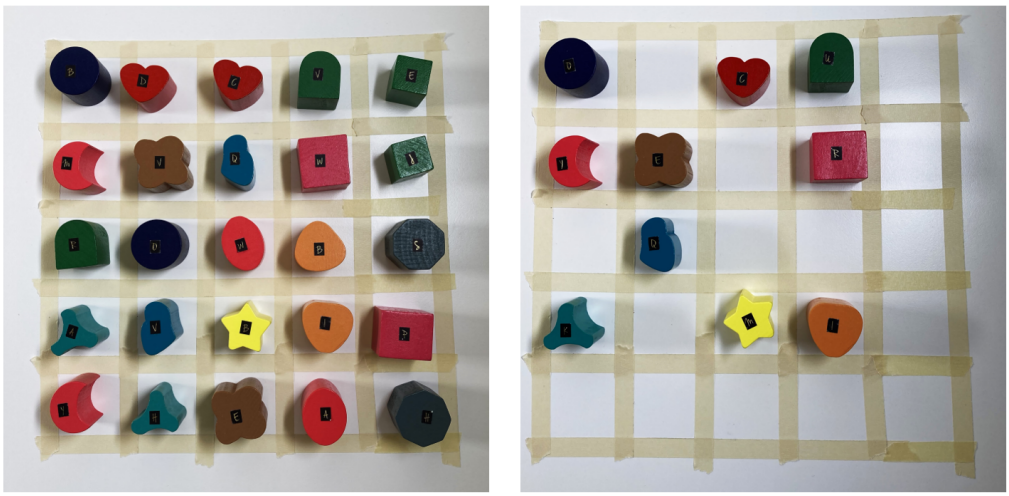


Figure 3.14: The 5x5 grid with ten easily distinguishable shapes for Group Easy (right) and 25 shapes for Group Difficult (left).

Table 3.3: Table comparing the task of Group Easy to Group Difficult by dissecting the task properties.

Task properties	**Group Easy**	**Group Difficult**
Sequence of shapes	Sequence A: 10 easily distinguishable shapes	Sequence B: 10 similar shapes
Selection of shapes on grid	Only the 10 shapes that must be tossed into the block (see Figure 3.14)	25 shapes, all 10 shapes that need to be tossed occur twice, but have different letters on them (see Figure 3.14)
Depiction of shapes on instructions	Coloured photo (see Figure 3.15)	Greyscale photo with lowered brightness, enhanced contrast and "chalk sketch" effect (see Figure 3.15)
Letters on instructions	1 letter on the instructions that match 1 of the 2 letters on the correct shape (see Figure 3.15)	4 letters on the instruction, one of which matches 1 of the 4 letters found on the two correct shapes (see Figure 3.15)

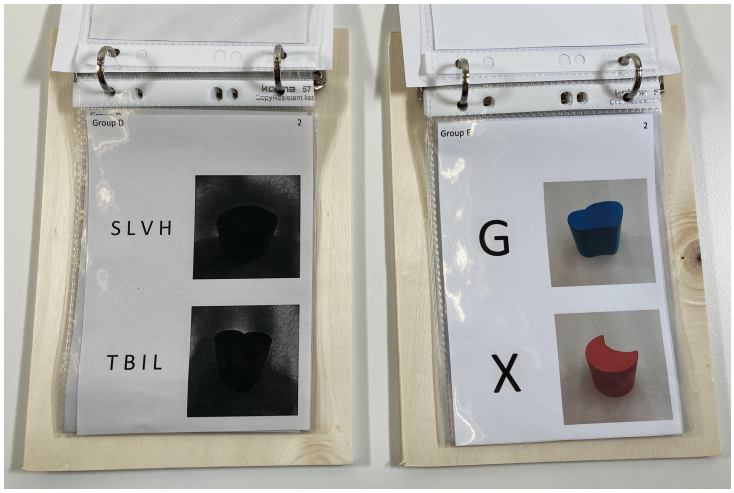


Figure 3.15: Comparison of the instructions given to the two groups. The shapes are much harder to identify in the photos of Group Difficult (left) than Group Easy (right).

Task of Group Easy

Participants of *Group Easy* were given straightforward instructions with coloured images of the shapes and only one letter denoting the orientation (see Figure 3.15). In the grid, only the ten shapes that actually had to be used were present, without any duplicates (see Figure 3.14). Thus, the subject had to grab the correct shape, find the corresponding slot, make sure the correct letter is pointing towards the opening, and toss it in. Further, the sequence (named Sequence A) consisted of easily distinguishable shapes.

Task of Group Difficult

Finding the correct shape and its orientation was markedly more difficult in *Group Difficult*. Firstly, editing the photos (greyscale, lowered brightness, enhanced contrast, "chalk sketch" effect) made it harder to identify the shape in it (see Figure 3.15). Secondly, the sequence (named Sequence B) included some very similar shapes that were hard to distinguish from each other (see Figure 3.14). Thirdly, there were always two of each shape listed in the sequence, so both had to be checked. Lastly, there were four letters per shape listed in the instructions, whereas one of them matched one of the four letters on the two corresponding shapes. The goal was to make the search task a lot more complex than in *Group Easy*.

Hypotheses

To test the automated assessment framework and its applicability to an eye tracking study, the following hypotheses were constructed for *Group Difficult* (*GD*) and *Group Easy* (*GE*):

- The more pronounced search behaviour in *GD* will draw the attention towards the grid and the instruction. In turn, this will lead to:
  - H1: Higher Total Fixation Duration on the instructions and the grid in *GD* than in *GE*
  - H2: More Revisits on the instructions and grid in *GD* than in *GE*
- To identify the correct shape, *GD* will have to look back and forth between the grid and instruction. This will lead to a less random distribution of fixations and transitions between fixations, and thus:
  - H3: A lower GTE in *GD* than in *GE*
  - H4: A lower SGE in *GD* than in *GE*
- *GD* needs more cognitive effort and focus to perform the task than *GE*, which will manifest in:
  - H5: A higher Average Fixation Duration in *GD* than in *GE*
  - H6: A higher Fixation/saccade Ratio in *GD* than in *GE*
  - H7: A higher K-coefficient in *GD* than in *GE*
- *GD* is less efficient in finding and inserting the ten shapes than *GE*, which will lead to:
  - H8: A higher Total Duration in *GD* than in *GE*
  - H9: Higher Number of Fixations and Saccades in *GD* than in *GE*

3.2.3 Data Analysis Workflow

This Section outlines the workflow from the TG2 recordings to the generation of the *Fixation/saccade files* and the *OGD files*.

Generation of Fixation/saccade file

Once all the trials were recorded, the gaze data was loaded into Tobii Pro Lab where the recordings were exported into one single .tsv file via the so-called "Metrics Export". For this, the aforementioned *Tobii I-VT (Attention)* filter was used. In a second step, this file was split into single recordings by running *split_tobi_output.py* in the *scripts* folder to generate the *Fixation/saccade files*.

OOIs

As explained before, the OGD algorithm requires an instance segmentation algorithm that masks OOIs in the video frames to subsequently calculate the object-gaze distances. In the present study, four OOIs were selected: Instructions, grid, block and single shape, as depicted in Figure 3.16. The single shape was defined as such if the participant had it in his hand, i.e. after it was lifted from the table. An example is shown in Figure 3.17.

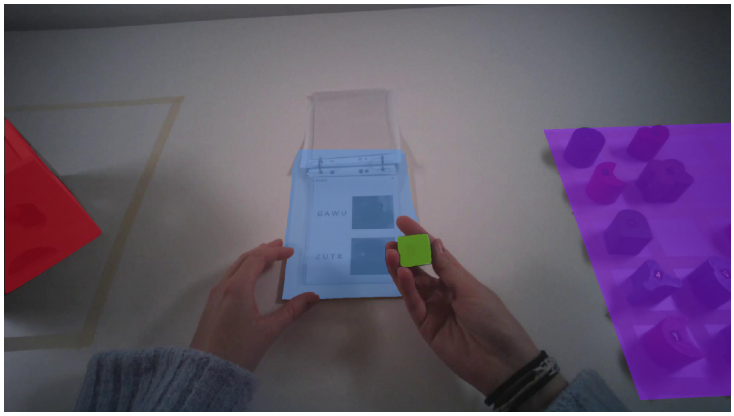


Figure 3.16: A video frame with the four defined OOIs: Block (red), instructions (blue), single shape (green) and grid (purple).

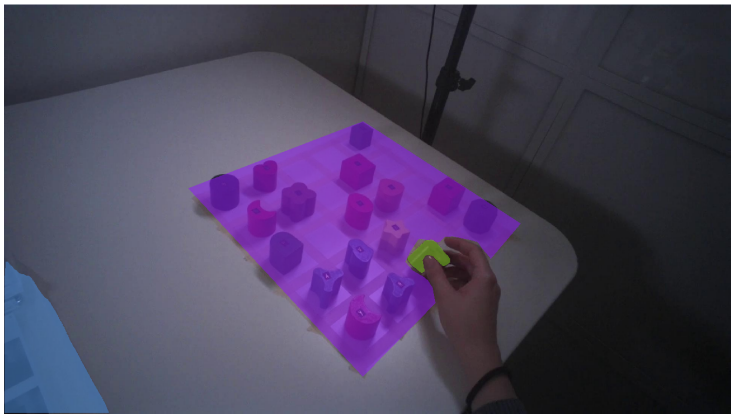


Figure 3.17: Demonstration of the case where a shape, when lifted off the table, immediately counts as a Single Shape (green).

Mask R-CNN

In this work, the segmentation algorithm was based on the Mask R-CNN architecture, pre-trained on the COCO dataset (Lin et al. (2014)). Mask R-CNN is a state-of-the-art Convolutional Neural Network (CNN) for object detection and instance segmentation (He et al. (2017)). The training was conducted on *supervise.ly*, an easy-to-use web-based platform that facilitates computer vision development for researchers (*supervisely* (2022)). Within this study, *supervise.ly* was used for labelling, image augmentation and training of the neural network.

Training Data

As any neural network, Mask R-CNN has to be trained with labelled images to obtain the weights of the model. Hence, around 30 images each were extracted from 18 recordings. The recordings were roughly evenly distributed in terms of group and trial number, and around half of the participants were represented. The extracted images were preselected to exclude very similar images, blurry images and images without any OOIs in them. A total of 376 images were uploaded to *supervise.ly* and labelled with the integrated annotation tool.

Image Augmentation

CNNs require large amounts of images to maximize their performance. Image augmentation is a technique to artificially increase the amount of training

data by slightly transforming the original images and adding them to the training data set. In this study, data augmentation was again conducted on *supervise.ly*, using the embedded augmentation tool called DTL. The transformations applied to the training images included crop, rotation, brightness, contrast, and Gaussian blur. For each of these transformations, ranges of specific parameters can be set. As an example, in the case of rotation, these parameters are a minimum and a maximum degree of rotation on either side. Within the given range, a value gets randomly selected (e.g. +15° rotation) and applied to the respective image. Further, DTL enables combining transformations in a tree-like structure, such that certain images will undergo more than one manipulation. The DTL structure used in this study is depicted in the Appendix A.4. After applying data augmentation, a total of 25 567 images were generated to be used for training.

Training

The augmented images were split into a training and a validation set with a ratio of 80:20, respectively. They were reshaped and padded with zeroes to get a square image with a size of 256 x 256 pixels. The Mask R-CNN model pre-trained on the COCO dataset was trained with a transfer learning approach for 25 epochs and a learning rate of 0.001. The epoch with the lowest loss value was used as the final model. Applying the neural network to images yet unseen to the network yielded promising results. Four example images of this inference process conducted on *supervise.ly* are shown in Figure 3.18.

Post-Training Modifications

Afterwards, the weights of the trained neural network were exported from *supervise.ly* and applied to both seen and unseen trial videos to predict the OOIs. In either, the network occasionally detected the single shape and the block within the grid. Moreover, it infrequently inferred the single shape in the instructions. Since these overlaps should not occur, the overlapping pixels were subtracted from the wrongly predicted OOIs within the grid or the instructions. The modifications were made for each frame of the video after the masks were inferred. Hence, the following masks were removed from single shape, respectively block:

- Single shape when overlapping with grid
- Block when overlapping with grid

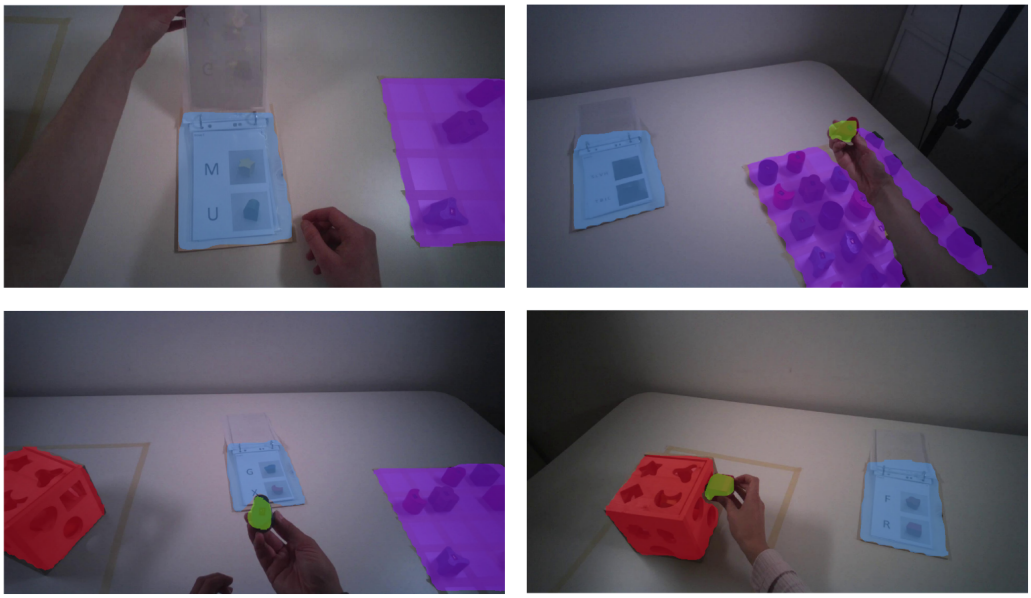


Figure 3.18: Four images showing the inference process, where the neural network is applied on new images that were not used for the training.

- Single shape when overlapping with instructions

OGD Algorithm

Finally, the newly modified neural network could be used to calculate the object-gaze distances from each trial. Next to the model, the recording without gaze visualization with the corresponding *Fixation/saccade file* were required as inputs (see Figure 3.2) to generate the *OGD file*.

3.2.4 Data Sets

Before the analysis was started, both the *Fixation/Saccade files* as well as the *OGD files* were checked on the fulfilment of different criteria in order to be included in the trial. As stated above, 84 recordings from 28 participants were captured for the automated ET analysis. To ensure an accurate representation of the gaze, the minimum gaze sample rate for a trial to be included in the analysis was set to 70%. All recordings except for one, namely *participant03_-trial02* of *GE*, fulfilled this criterion. Being the middle trial of the person, it was assumed its sole exclusion would not markedly affect the results, so trials 1 and 3 of *participant03* were kept in the data set. Thus, in total 83 recordings, 42 in *GD* and 41 in *GE*, were eligible for the analysis.

Since there are no additional criteria for the generation of the *Fixation/saccade file*, all of these 83 trials could be used to conduct the *General analysis* and the *K-coefficient analysis*. This data set will be referred to as *DS1* in the following. *DS1* was used to test hypotheses H5, H6, H7, H8, and H9.

For the generation of the *OGD file*, next to the 70% gaze sample rate, a reasonably good instance segmentation by the trained model is required. If the segmentation is faulty, the calculated object-gaze distances are so too, rendering calculations of OOI-based metrics useless. Thus, for each recording, instance segmentation was performed on each frame by the neural network and visually checked afterwards. All except the trials of two participants could be used for the *OOI-based analysis* and the subsequent *Statistics* calculation. In particular, these were *participant04* of *GE* and *participant20* of *GD*. While in *participant04* a different angle of view on the OOIs might be the reason for the inaccurate segmentation could be found, *participant20* was the only participant who always grabbed two shapes from the grid at the same time with both hands, which was never seen on the training data. After the exclusion of these trials, the data set will be referred to as *DS2*. The *DS2* was used to test hypotheses H1, H2, H3, H4, and H7. Table 3.4 summarises the included trials in *DS1* and *DS2*.

Table 3.4: Overview of the included trials in the two data sets used in the analysis of the study.

	DS1			DS2		
	GD	GE	Total	GD	GE	Total
Total Number of Included Trials	42	41	83	39	38	77
Total Number of Included Participants	14	14	28	13	13	26
Total Number of First Trials	14	14	28	13	13	26
Total Number of Second Trials	14	13	27	13	12	25
Total Number of Third Trials	14	14	28	13	13	26

To evaluate the quality of the instance segmentation applied to *DS2*, the mean intersection over union (IoU) was determined for 30 frames extracted from recordings that were not utilised in the training data. Thereby, the labelled masks, i.e. the ground truth, are compared to the masks predicted by the trained model, i.e. the prediction. Simply put, the IoU is the ratio of the correctly predicted pixels and the total number of pixels present across the ground truth mask and the predicted mask, as stated in Equation 3.2.1:

$$IoU = \frac{P \cap G}{P \cup G}, \tag{3.2.1}$$

where P represents the predictions and G the ground truth.

The mean IoU of *DS2* was 0.854. As stated in Table 3.5, the mean IoU for the classes grid, instructions, block, as well as the background were all around 0.9, denoting good segmentation. The mean IoU for the single shape was 0.278, denoting a rather low quality of detection. However, since the main differences in the two task difficulties evolve around the grid and the instructions, and the prediction quality was equally low for both groups, this will only have a minor influence on the hypothesis testing. Table 3.6 shows the matched pixels matrix of all OOIs, as well as the background.

Table 3.5: Pixel accuracy and mean IoU of all OOIs and background.

	Accuracy	Mean IoU
Block	0.948	0.922
Grid	0.94	0.915
IFU	0.901	0.883
Single Shape	0.3	0.278
BG	0.996	0.983

Table 3.6: Matched pixels matrix of the OOIs and the background (BG).

		Predicted Values				
		Grid	IFU	Single Shape	Block	BG
Actual Values	Grid	0.9396	0	0	0	0.0604
	IFU	0	0.9007	0	0	0.0993
	Single Shape	0.0007	0	0.2998	0.0574	0.6421
	Block	0	0	0	0.9479	0.0521
	BG	0.0023	0.0009	0.0002	0.0007	0.9959

3.2.5 Statistics

The hypotheses were statistically tested by comparing participant means of *GD* and *GE* extracted from the outputs generated by the framework. In the first step, the assumptions for an independent t-test (also called Student's t-test) were tested. Any clear outliers were removed from the statistical analysis. While the assumption of independence of the groups was always given, the following criteria were checked for each metric:

- Normality was tested with the Shapiro-Wilk-Test and visually by plotting a histogram and a Q-Q plot.
- Homogeneity of variance was tested using Levene's test.

If all of these assumptions were given, an independent t-test was performed to test if the two sample means differ from each other. Alternatively, if normality was not given, the non-parametric Wilcoxon rank sum test was performed to test for equal medians. If non-equal variances were found in combination with given normality, Welch's t-test was performed. $\alpha = 0.05$ was chosen as the significance threshold.

3.3 Action Recognition Validation

3.3.1 Purpose

In addition to the data that was specifically collected for the aforementioned study, the proposed framework was tested on already existing, more complex ET recordings from a medical device usability study conducted at $pd|\mathbf{z}$. The primary reason for this was that both the OGD algorithm, as well as the HAR by the PVHMM algorithm were already successfully carried out for these recordings. Thus, both the *Fixation/saccade file* and the *OGD/Action file* were ready to be used as input for the framework. This came with the advantage of testing the *Action-based analysis* and the *Sequence comparison* which was not possible in the shape sorter study. Thus, in contrast to the self-conducted study, the goal here was not to assess the framework's ability to find differences amongst experimental conditions, but rather to test the integration of action-based data in the pipeline, since this was neglected so far. Alternatively stated, this data only served as test input for the framework, whereas the interpretation of its output was largely omitted. Hence, the stimuli and set-up are only briefly reviewed.

3.3.2 Input Structure and Stimuli

The ET data were recorded during a usability study with the UnoPen, a self-injection device for insulin and other multidose therapies, coupled to the SmartPilot, a connectivity device that connects the UnoPen to a smartphone application (Ypsomed Holding AG, Burgdorf) (Kreiner (2021)). The device is shown in Figure 3.19. For more detailed information on the usability study with the UnoPen, please refer to the Master's Thesis of Kreiner (2021). A detailed overview of the chosen OOIs and actions can be found in the Appendix A.5.

Eight recordings from the usability study were selected. In order to simulate two groups, the recordings were pairwise assigned to four fictional participants

(two trials each), i.e. from *participant01_trial01* to *participant04_trial02*. *participant01* and *participant02* were assigned to *Group 1*, whereas *participant03* and *participant04* were assigned to *Group 2*. The corresponding input files (*Fixation/saccade file* and *OGD/Action file*) were renamed according to the data input format explained in Section 3.1.2. All modules were selected to be executed in the user input before starting the *main.py*.

Figure 3.19: SmartPilot add-on mounted on the UnoPen.

4 Results and Discussion

In this chapter, the results of the proof-of-concept study will be discussed first, whereas each of the hypotheses (H1-H9) will be reviewed (Section 4.1). In the second part, the outcomes of the validation run are presented, in particular, the modules that are not discussed in the study, namely *Action-based analysis* and the *Sequence comparison* (Section 4.2). While in these first two sections the utility of the selected metrics and the quality of the output are discussed, the third section reviews the pre-processing steps in terms of automation (Section 4.3). The last section summarises the limitations of this work (Section 4.4).

4.1 Study

4.1.1 Hypotheses

It is important to point out that the entire data sets (*DS1* and *DS2*) were included in these figures, whereas some outliers were removed in the statistical analysis. In the auto-generated figures of the program, the instruction is named IFU, short for instructions for use.

H1: Higher Total Fixation Duration on the instructions and the grid in *GD* than in *GE*

For the fixation time on the instructions, two outliers were removed from the sample of *GD*, since the durations were markedly higher in these participants (*participant16* and *participant23*). A Welch's t-test revealed a significant difference ($t(12.636) = 4.645$, $p > .001$) in the Average Fixation Duration on the instructions between *GD* (M = 19 871 ms) and *GE* (M = 8803 ms). Regarding the grid, a Welch's t-test showed a significant difference in the Total Fixation Duration on the grid between *GD* (M = 21 644 ms) and *GE*

(M = 3848 ms); t(12.628) = 9.080, $p > .001$.

The significantly higher Total Fixation Durations on both the instructions and the grid support the hypothesis that *GD* spent more time identifying and searching for the correct shape. On top of that, Figure 4.1 shows that the Total Fixation Duration on the other two objects, the block and the single shape, has been similar in both groups. Thus, the difference seen in the grid and the instructions can largely be attributed to the fact that *GD* spent more time identifying and searching for the correct shape, while finding the correct opening and tossing in the shape was equally challenging in both groups. Accordingly, the extra time needed in the search task is very likely to be the reason for the higher mean Total Duration of *GD* seen in H8. Naturally, *GD* also exhibited more Total Hits on those two objects. Thus, the Total Fixation Duration, as well as the Total Hits are suitable metrics to measure attention towards OOIs due to complexity. However, in other tasks, it could also denote higher interest.

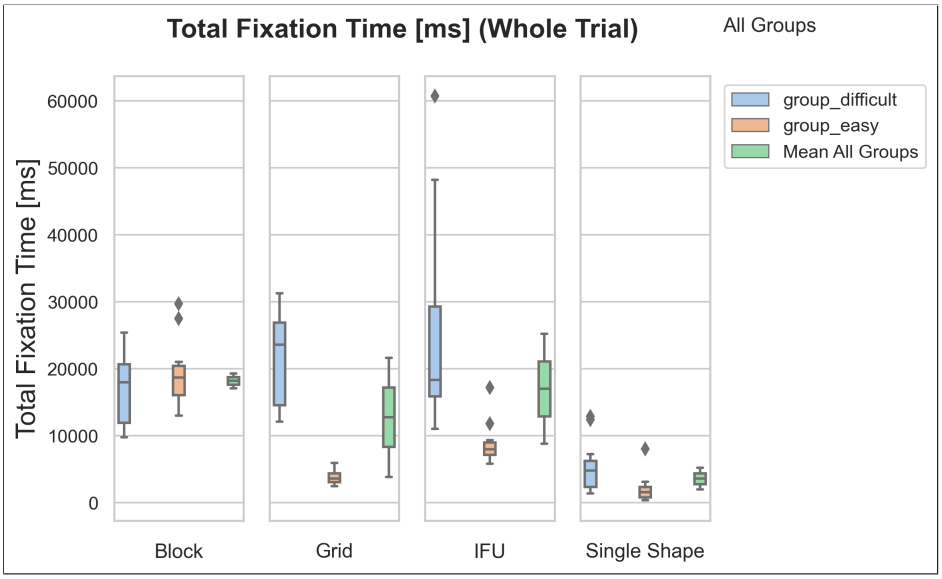


Figure 4.1: Box plots of the Total Fixation Duration [ms] per OOI of *GD*, *GE* and their means.

H2: More Revisits on the instructions and grid in *GD* than in *GE*

When analysing the Revisits on the instructions, again, *participant16* had to be removed as an outlier due to a markedly higher value. A Welch's *t*-test revealed a significant difference (t(13.185) = 3.505, $p = .004$) in the Revisits on the instructions between *GD* (M = 30) and *GE* (M = 18.56). Also in the evaluation of the Revisits on the grid *participant16* had to be removed from

the sample. A *t*-test showed that *GD* (M = 31.22) conducted significantly more Revisits on the grid than *GE* (M = 12.21); t(23) = 9.970, $p > .001$.

In other words, the more pronounced search behaviour of *GD* could also be manifested in the significantly higher average Revisits on the grid and the instructions per trial, as depicted in Figure 4.2. Thus, the participants of *GD* did not only exhibit more hits on the grid while scanning it but also on the instructions while trying to recognise the depicted shape. Additionally, they had to look back and forth between the grid and the instructions, leading to a high number of Revisits on both objects. Similar to the statement in H1, the number of Revisits is a good indicator for attention on OOIs, whereas the underlying reason cannot be manifested from this value alone.

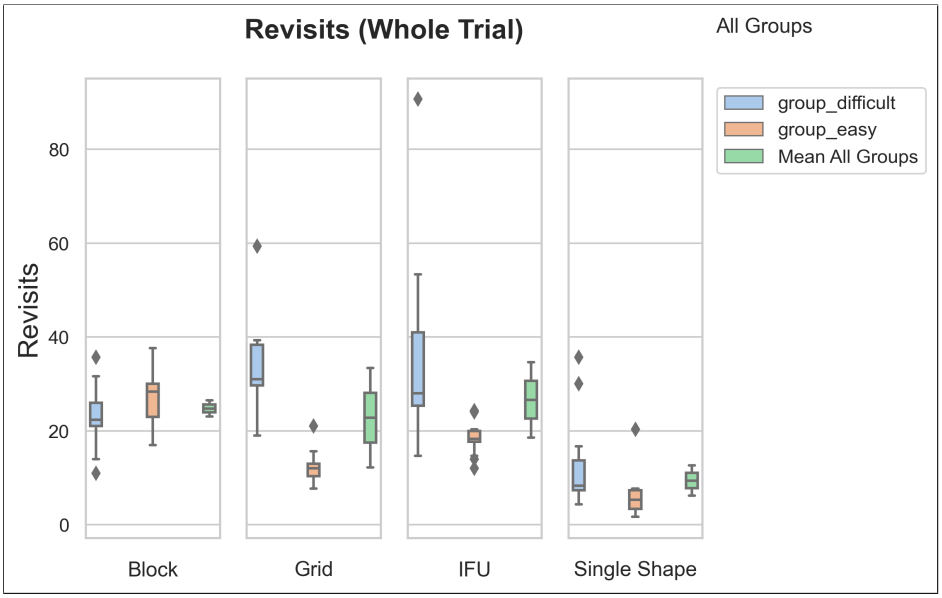


Figure 4.2: Box plots showing the Revisits per OOI of *GD*, *GE* and their means.

H3: A lower GTE in *GD* than in *GE*

A *t*-test reported a significantly higher normalised GTE in *GD* (M = 0.59) than in *GE* (M = 0.56); t(24) = 2.245, $p = .034$. This finding stands in contrast to our hypothesis and is depicted in Figure 4.3. In other words, *GE* has exhibited fewer random distributions of transitions between OOIs than GD. A possible reason for this could be that the difficult task led to more confusion and in turn, more randomness in the OOI transitions. Nonetheless, the average GTE does not seem to provide comprehensible feedback on the task.

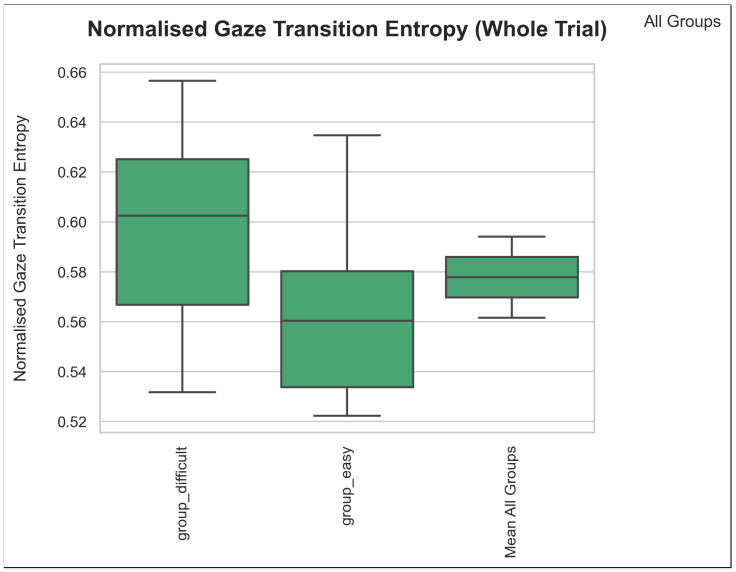


Figure 4.3: Box plot showing the normalised GTE of *GD*, *GE*, and their mean.

However, the transition probability matrix that is exported for each trial can still provide some more in-depth information. In most trials of the study, transitions within the same state were very frequent, as seen in the example of the first trial of *participant11* in Figure 4.4. Further, when *participant11* looked at the single shape, the chance was high that the block was fixated next. This most probably resembles the action of checking the shape before searching for the corresponding slot in the block. Moreover, the scanning of the grid is nicely represented with the high probability of a within-state transition of the grid. The same is true for the block, where the subject has to find the right slot, and the instructions, where the shape and the letter had to be captured.

As outlined in Chapter 2, also the background represents an OOI in this metric. Accordingly, all pixels that are not masked by the model are counted as background. Thus, each time the model failed to detect the OOI the subject was looking at, it is detected as background. This makes this metric very vulnerable to object detection uncertainty.

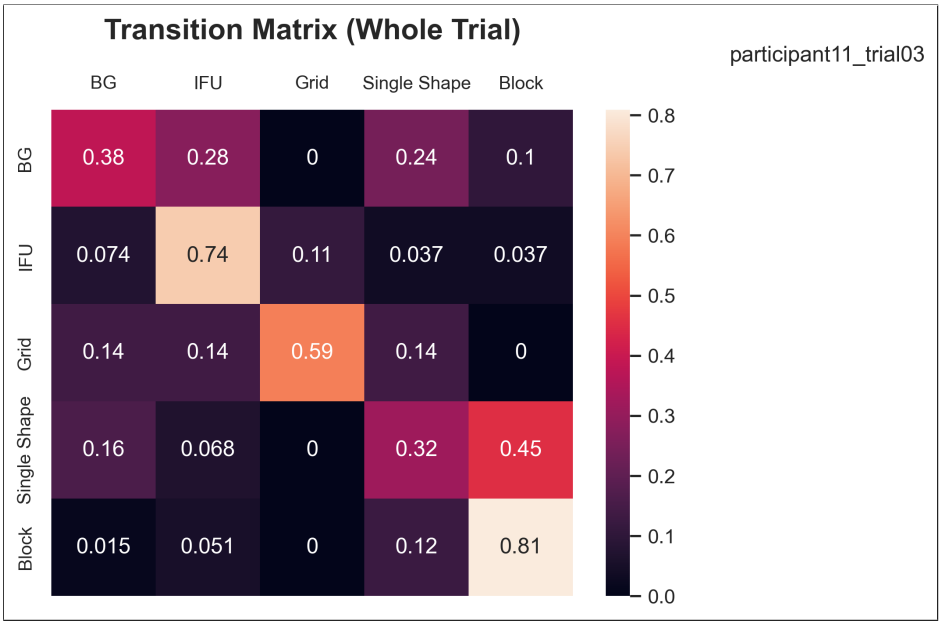


Figure 4.4: Transition probability matrix of *participant11_trial03*. It illustrates the probability of transitioning from the current OOI (represented by the rows) to the next OOI (represented by the columns). BG: background.

In the literature, both the SGE and the GTE are usually determined in a different manner than in this framework. Firstly, entropy calculations are mostly applied to remote eye tracking where the subject looks at a screen during their task. The image on the screen is then divided into content-driven areas of interest (AOIs) that capture hits which then are used to calculate the entropy. On the other hand, in the case of mobile eye tracking, the scene video is usually divided into a grid, whereas each parcel of the grid corresponds to an AOI (Shiferaw et al. (2019)). Only a few studies, e.g. Schieber & Gilland (2008), used content-based OOIs in a dynamic setting to calculate the entropy like it was done in this project. However, since in this framework we benefit from automated object-gaze mapping through the OGD algorithm, we decided to calculate both metrics based on our content-based OOIs, even in a dynamic setting. However, this promising novel approach comes with the disadvantage of lacking studies to compare the results to. To make the output more comparable, an entropy calculation based on a grid could be implemented in addition to the current method in the future. This comes with the additional advantage of allowing the calculation with only the *Fixation/saccade file*.

H4: A lower SGE in *GD* than in *GE*

A *t*-test found a significantly lower SGE in *GE* (M = 0.78) than in *GD* (M = 0.86); t(24) = 6.921, $p > .001$. Hence, similar to H3, the more complex search task of *GD* was thought to lead to less random distributions of fixations between the OOIs. However, also in the SGE, the opposite was the case, as shown in Figure 4.5.

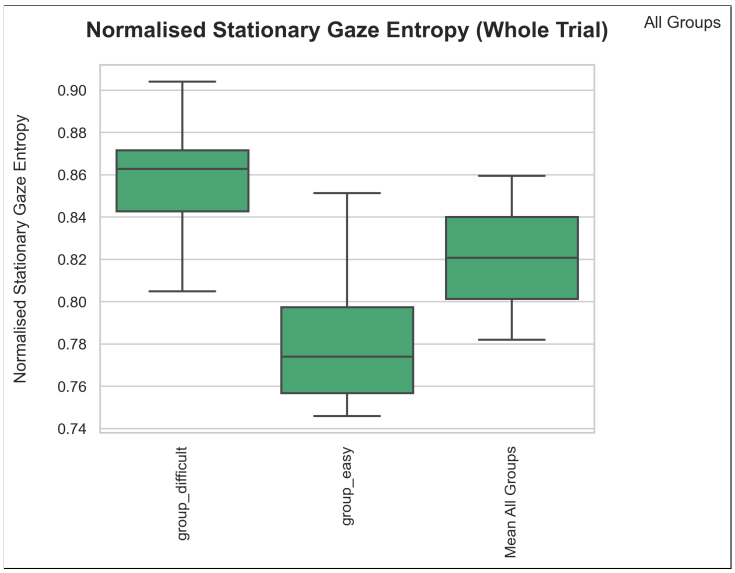


Figure 4.5: Box plot showing the normalised SGE of *GD*, *GE*, and their mean.

This outcome can be explained when considering the Relative Dwell Time per OOI in per cent, depicted in Figure 4.6. In the easy task, subjects spent the majority of the time (58%) looking at the block in order to find the correct opening, resulting in an uneven distribution of hits. This, in turn, leads to lower SGE. In the difficult task, subjects had a much higher dwell time on the instructions and the grid, leading to a more even distribution of hits and in turn, a higher SGE. This shift of the distribution is nicely illustrated in the pie charts of Figure 4.6. Summed up, the Relative Dwell Time on the block was underestimated in this hypothesis. Hence, similar to the other discussed metrics, the SGE has to be interpreted in the given context.

This example illustrates how the automatically produced output widely supports the user in exploring the results and revealing underlying reasons.

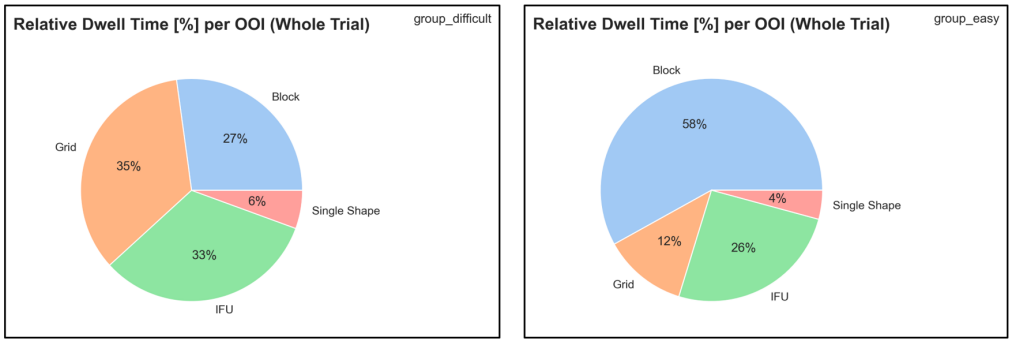


Figure 4.6: Pie charts showing the Relative Dwell Time per OOI in percent of *GD* (left) and *GE* (right).

When examining the Relative Dwell Time in Figure 4.6, the share of the single shape is very small in both groups. On one hand, this is attributable to the aforementioned low-quality object detection. On the other hand, when visually examining the recordings with the overlaid gaze point, it was found that the subjects indeed spent very little time looking at the individual shape that they are holding in their hand. One reason for this is that they have already examined it within the grid. Additionally, once the subject has decided on a shape, it is oftentimes barely fixated until it is placed in the opening of the block. Hence, many subjects carry out the alignment and orientation of the single shape towards the block with the shape in their near-peripheral vision, as depicted in Figure 4.7. This could be further examined by exploiting the object-gaze distances in the *OGD file* but was out of the scope of this thesis. However, incorporating the peripheral vision in the analysis of OOIs could be a promising novel method.

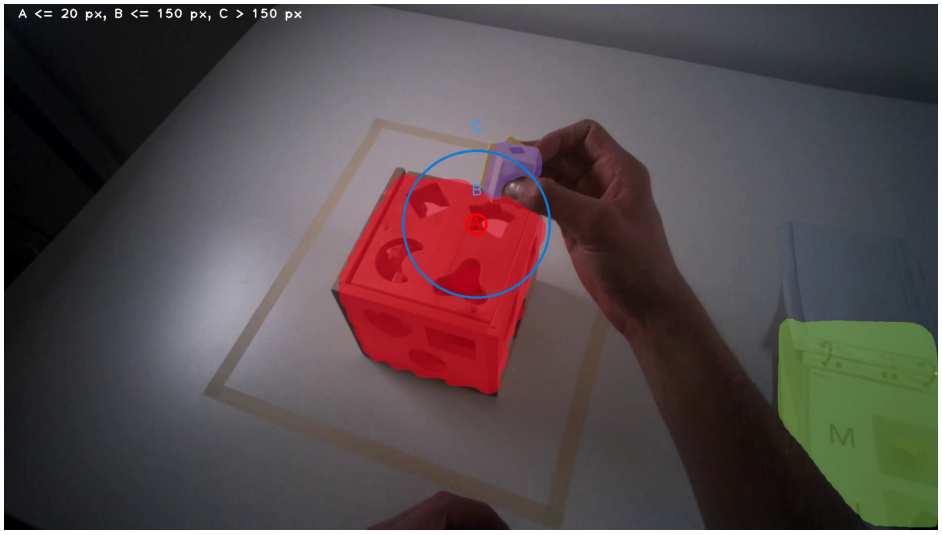


Figure 4.7: Scene video of *participant09_trial01* with overlaid gaze point and vision areas. The radius in pixels of the areas is indicated on the top left. With a radius of 20 pixels, the innermost circle corresponds to the hit radius of this analysis. In other words, masks lying within this circle get counted as a hit.

H5: A higher Average Fixation Duration in *GD* than in *GE*

Since the task of *GD* was considerably more difficult, it required more cognitive effort which in turn has been associated with longer fixations, as stated in Section 2.3. However, a Wilcoxon rank sum test reported a non-significant difference (Z = -1.347, p = .178) between *GD* (Mdn = 248.20 ms) and *GE* (Mdn = 169.28 ms) regarding Average Fixation Duration, as shown in Figure 4.8.

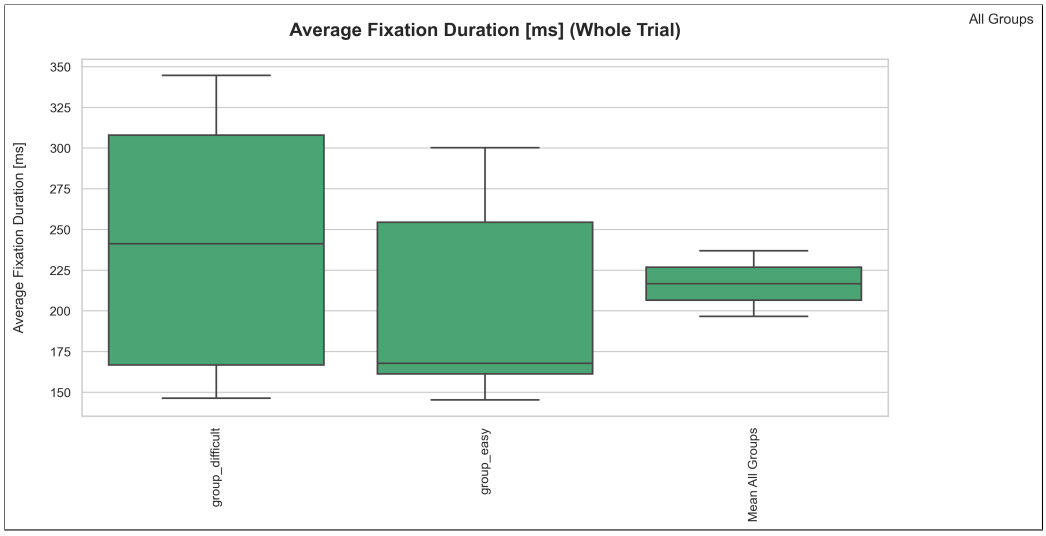


Figure 4.8: Box plot showing the Average Fixation Duration [ms] of *GD* (M = 238.20 ms) and *GE* (M = 202.31 ms).

In the box plot in Figure 4.9, the Average Fixation Duration of each trial per participant is represented. It is evident that there are more participants with a low average in *GE* compared to *GD*. Further, most participants show low variability over their three attempts. This could indicate that the cognitive effort was at a similar level among the three trials.

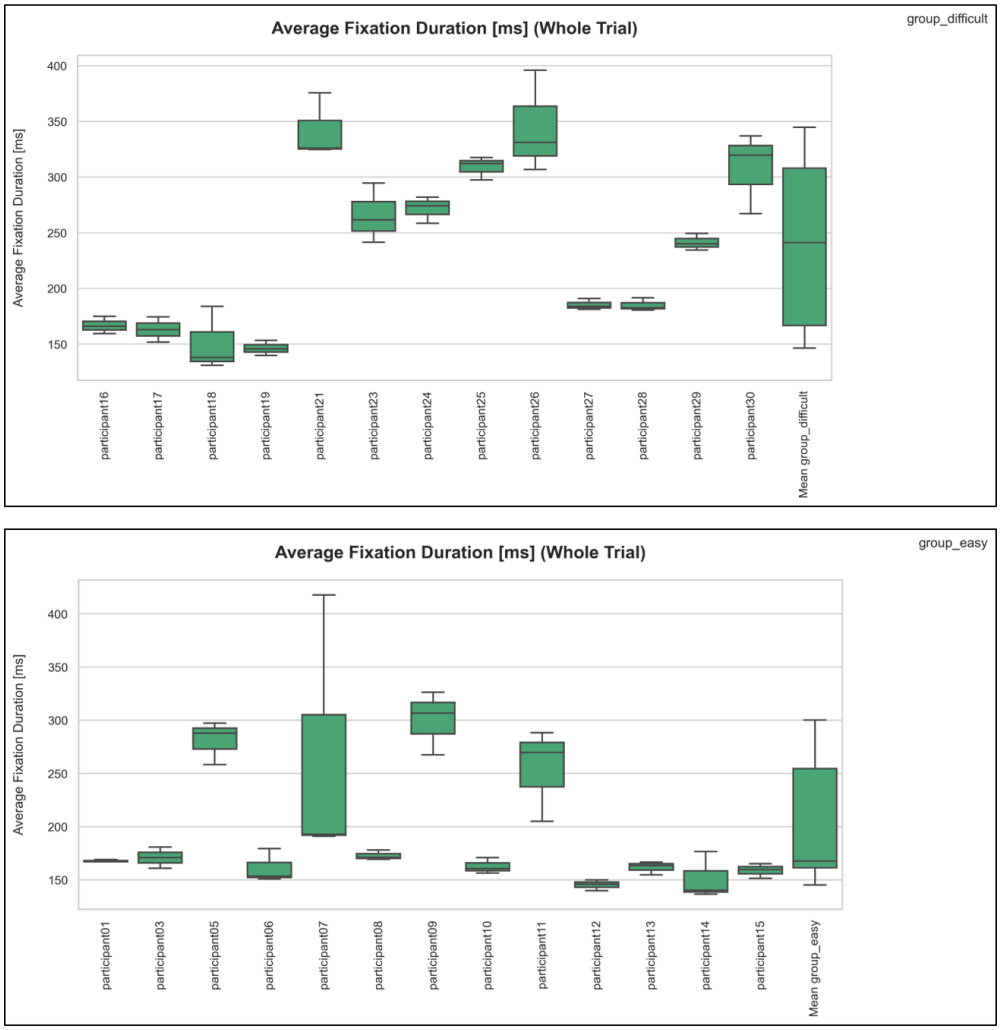


Figure 4.9: Box plots showing the Average Fixation Duration [ms] over all three trials of *GD* (top) and *GE* (bottom) for each participant and their mean. Note: The range of the y-axes differs from each other.

However, this seems to be in agreement with the literature, since fixation durations have also shown to be highly individual. For instance, Rayner et al. (2007) have reported high consistency in eye movement behaviour in terms of fixation duration and saccade amplitudes across multiple distinct tasks. Indeed, this gets evident when looking at individual participants, e.g. *participant01* and *participant23* in Figure 4.10. For both participants, the interquartile range (IQR) of the box plots is very consistent throughout all their three runs, even though the task is likely to get easier every time due to the learning process. Summed up, although fixation durations can provide valuable insight, they do not allow for straightforward conclusions on the performance due to being highly individual.

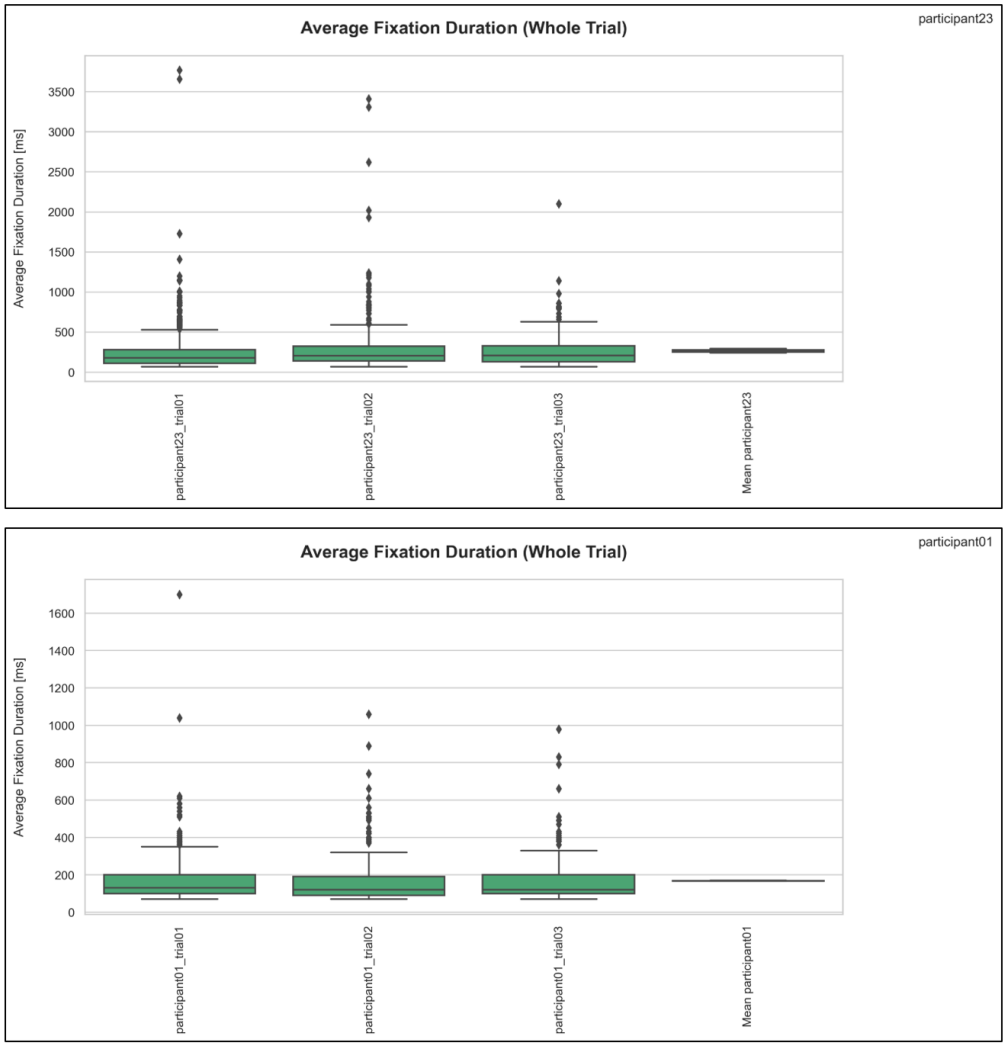


Figure 4.10: Box plots summarizing the Average Fixation Duration [ms] of each trial and their mean of *participant23* (top) and *participant01* (bottom). Note: The range of the y-axes differs from each other.

H6: A higher Fixation/Saccade Ratio in *GD* than in *GE*

To ensure maximal comprehensibility, the Saccade/Fixation Ratio was calculated as a percentage and rounded to zero decimal points for each trial. This measure makes statistical testing inaccurate, which is why only the means and standard deviations are reported in this case. *GD* had an average relative fixation time (of total fixation and saccade time) of M = 85.57% and a standard deviation of SD = 3.52%. *GE* showed an average of M = 85.14% relative fixation time and a standard deviation of SD = 2.51%. Since

the two means are very similar with a relatively high standard deviation, this hypothesis can be regarded as not supported by this outcome. This is illustrated in the pie charts Figure 4.11.

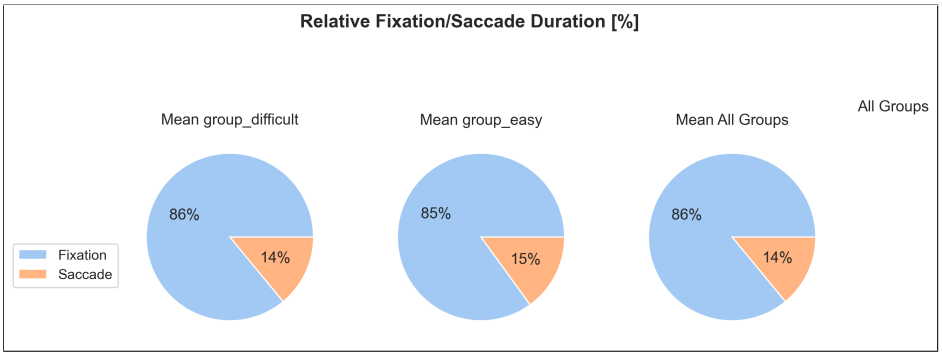


Figure 4.11: Pie charts illustrating the mean Fixation/Saccade Ratio of the two groups.

While the saccade durations were expected to be similar in both groups, *GD* was thought to have longer fixations owing to the higher cognitive workload. The fact the latter has been proven to be true implies that the Average Saccade Duration was longer in *GD* than in *GE*, which can be confirmed in the box plot in Figure 4.12. However, the difference is only small (*GD*: M = 32.80 ms, *GE*: M = 31.87 ms). Again, this non-significant difference can be due to the aforementioned individual gaze behaviour.

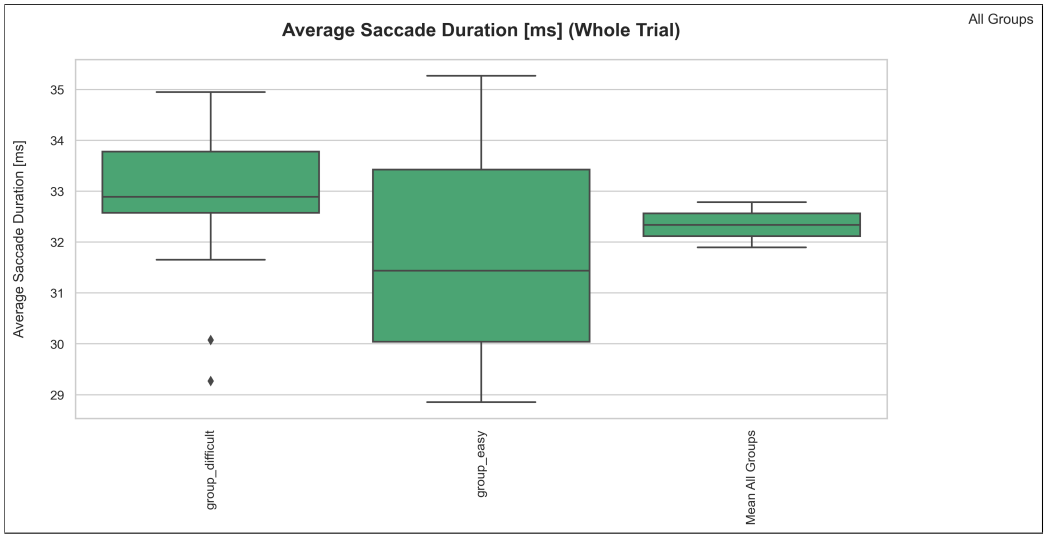


Figure 4.12: Box plot showing the Average Saccade Duration [ms] of *GD* and *GE*.

H7: A higher K-coefficient in *GD* than in *GE*

An independent t-test showed a non-significant difference ($t(26) = 0.768$, $p = .450$) in the K-coefficient in *GD* (M = 0.13) and *GE* (M = 0.07). Hence, the hypothesised higher focus needed in *GD* could not be confirmed in terms of the K-coefficient. Both means indicate a focal view but are close to zero. When looking at the individual participants in Figure 4.13, it gets evident that on average only very few individuals showed ambient vision.

Figure 4.13: Bar plots displaying the K-coefficient for each participant of *GD* (top) and *GE* (bottom).

Participant12 had the lowest mean K-coefficient of both groups. As depicted in the bar plot of Figure 4.14, two of their trials were outside two standard

deviations of the overall mean. However, when comparing these trials to recordings with a focal mean K-coefficient, we could not find any clear differences.

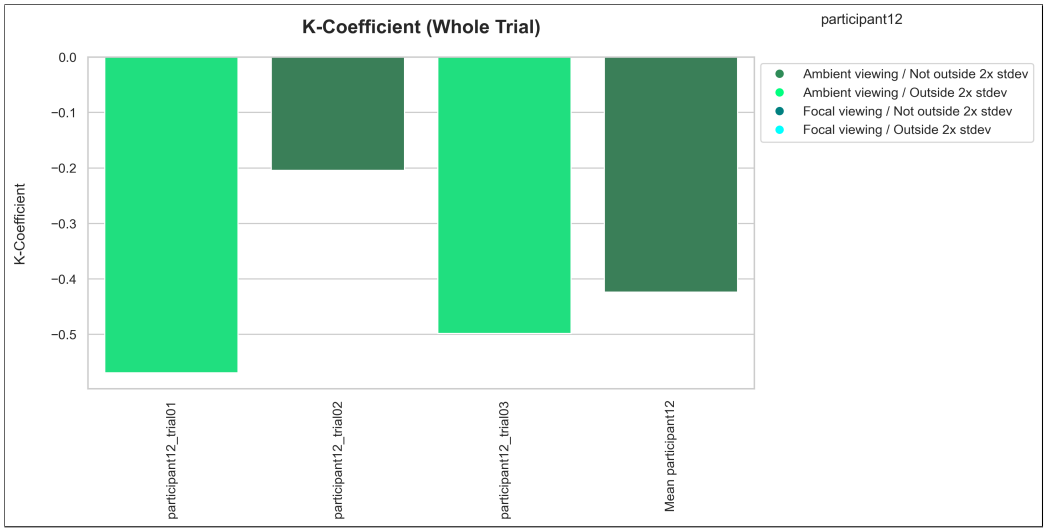


Figure 4.14: K-coefficient of all trials of *participant 12*. Trials 1 and 3 are marked as being outside two standard deviations, meaning they were highly ambient compared to the rest of the trials.

Indeed, these trial means have to be treated with caution. When examining the changes of coefficient K over the course of each of these trials in Figure 4.15, the reason becomes evident. In this participant, only 19.26 Ks per minute could be calculated on average. This means that only a small fraction of all the existing fixation-saccade pairs could be identified from the recorded gaze. Consequently, it is not possible to draw firm conclusions from the trial mean in this case, as it might be far off the real value. Thus, a minimum sample frequency should be set for the interpretation of the mean K-coefficient over a trial. This could be implemented in a future version of the program, e.g. in form of a warning on the generated graphs that display mean K-coefficients from small samples.

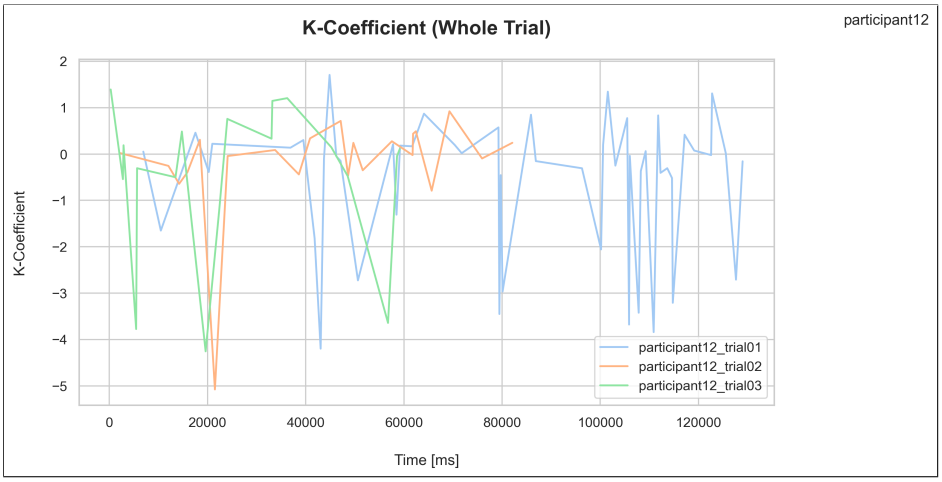


Figure 4.15: K-coefficient of all runs of *participant12* over the time course of the trial with an average of 19.26 sampled Ks/min.

Even in cases of an adequate number of Ks per minute, it makes sense to take a closer look. Figure 4.16 shows trial 2 of *participant20* with a sampling frequency of 90.26 Ks per minute. The mean K-coefficient of this trial is 0.39, implying overall focal vision. In the first and the last few seconds of the trial, the calculated Ks were rather high, while in the middle they were closer to zero. Thus, one could argue that it makes sense to calculate K not only over the entire trial but over certain time intervals. Hence, the K-coefficients of trial 2 of *participant20* were averaged over five and ten-second intervals and plotted again (Figure 4.17). Both graphs seem a lot easier to read and interpret. Again, this could be considered when further developing this framework. In the *Action-based analysis*, the K-coefficient is averaged over each action. This will be shown in the outcomes of the validation analysis in Section 4.2.

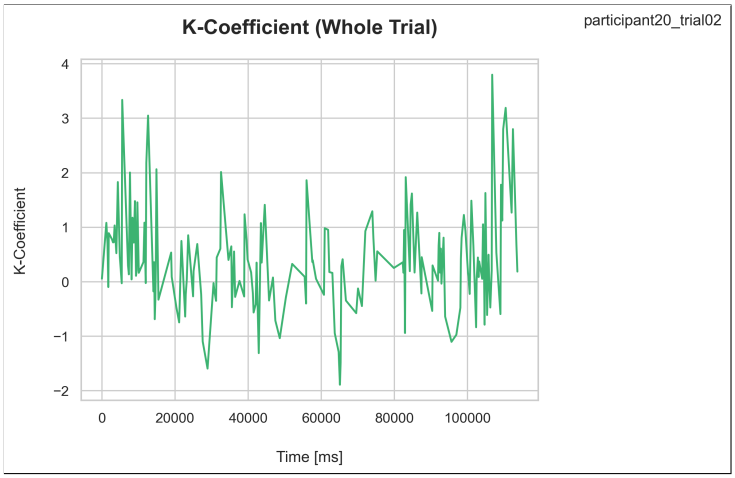


Figure 4.16: *K*-coefficient of *participant20* over a whole trial with a higher sampling frequency of 90.26 *K*s/min.

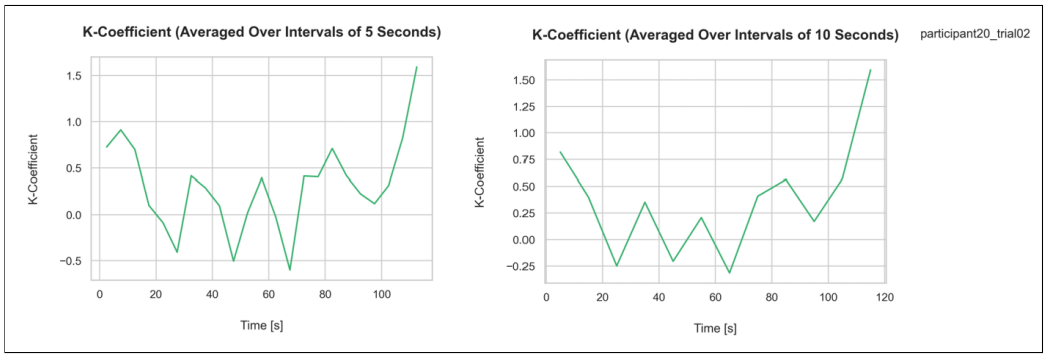


Figure 4.17: *K*-coefficient of *participant20* over the same trial as 4.16 averaged over five-second intervals (left) and ten-second intervals (right).

The output of Figures 4.15 and 4.16 allows to compare the development of the *K*-coefficient over time to the scene video with the overlaid gaze point. In the first twelve seconds, *participant20* finds and inserts the first shape without problems. When selecting and tossing in the second shape, the subject seems insecure and rapidly looks back and forth between the instructions and the shape. This is nicely reflected in the first negative peak. However, the next negative peak appears to be caused by a few large saccades within a short time period, caused by looking from the grid to the block to the instructions in a straightforward manner. In this case, the negative value is misleading, since the subject seems very focused and clear during this time interval. This is caused by the dynamic setting in this study, as opposed to a static experiment, e.g. where participants simply have to explore a painting

in front of them. Thus, even more caution is needed when interpreting K in dynamic tasks which provides a challenge to most medical device usability studies and surgical workflow assessments.

H8: A higher Total Duration in *GD* than in *GE*

Two outliers were removed from the sample of *GD* since these durations were a lot higher in these participants (*participant16* and *participant23*). An independent t-test revealed that participants of *GE* (M = 94473 ms) were significantly faster in completing the task than *GD* (M = 145119 ms); ($t(24)$ = 7.873, $p > .001$). This finding supports the hypothesis and is illustrated in Figure 4.18. This suggests that *GD* was less efficient in tossing the same amount of shapes into the block. As found in H1, *GD* likely spent more time identifying and searching for the correct shape than *GE*, while both groups were similarly efficient in properly aligning the shape and finding the corresponding opening.

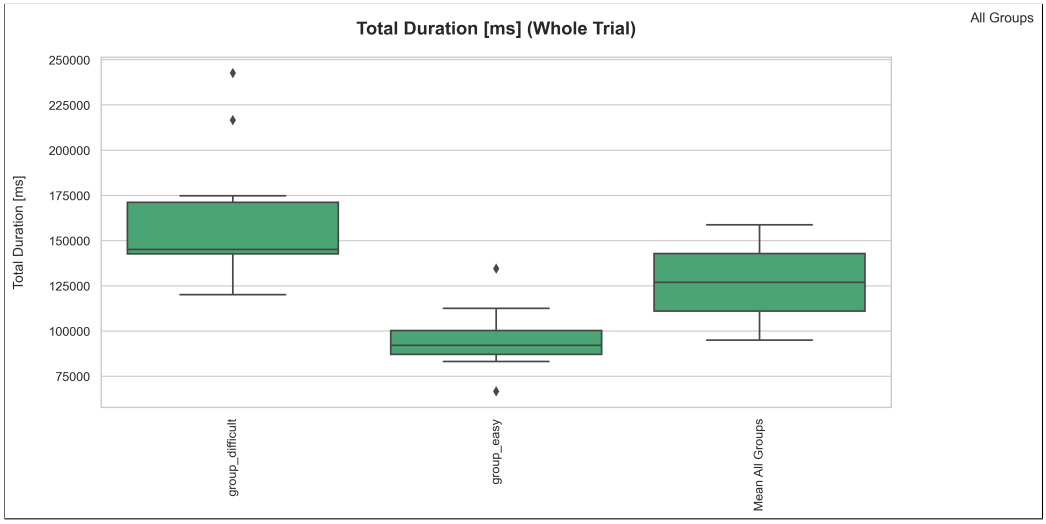


Figure 4.18: Box plot showing the average Total Duration [ms] for the two groups as well as their mean.

Participant16 of *GD* was slower than the rest of the group. In Figure 4.19 it can be appreciated how they got more efficient over time. Hence, this metric is a good indicator of efficiency, but it does not require an ET device to measure the time of a trial.

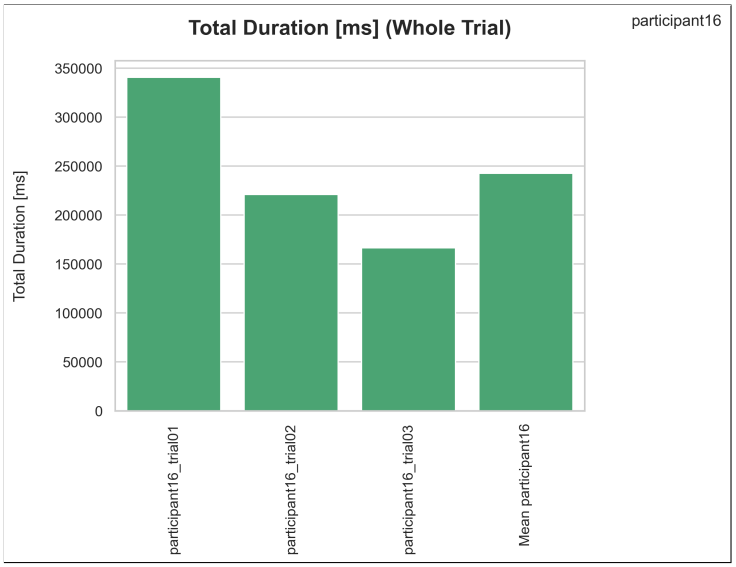


Figure 4.19: Bar plot showing Total Duration [ms] of *participant16* for each trial and the mean.

H9: Higher Number of Fixations and Saccades in *GD* than in *GE*

One outlier of *GD* (*participant16*) was removed in the statistical test regarding the total Number of Fixations. An independent t-test showed a significant difference ($t(25) = 3.966$, $p > .001$) in the total number of fixations in *GD* (M = 446.49) and *GE* (M = 317.33).

Two outliers were removed from each group (*participant16*, *participant23* from *GD* and *participant03*, *participant11* from *GE*), since a lot more saccades were detected in these two participants compared to the rest of their group. An independent t-test showed a significant difference ($t(22) = 7.9$, $p > .001$) in the total Number of Saccades in *GD* (M = 308.50) and *GE* (M = 469.89).

In other words, *GD* exhibited significantly more fixations and saccades than *GE* (see Figure 4.20). Naturally, both of these metrics are strongly correlated to the duration. However, calculating the number of fixations or saccades per minute makes the metric better comparable. On average, *GD* counted 181.85 fixations and 204.52 saccades per minute and *GE* 201.66 and 211.99, respectively. Again, it is suggested to include this calculation in the future in order to simplify comparing different conditions. However, while all metrics are slightly biased by the gaze sample rate, fixations or saccades per minute are more affected by this than others.

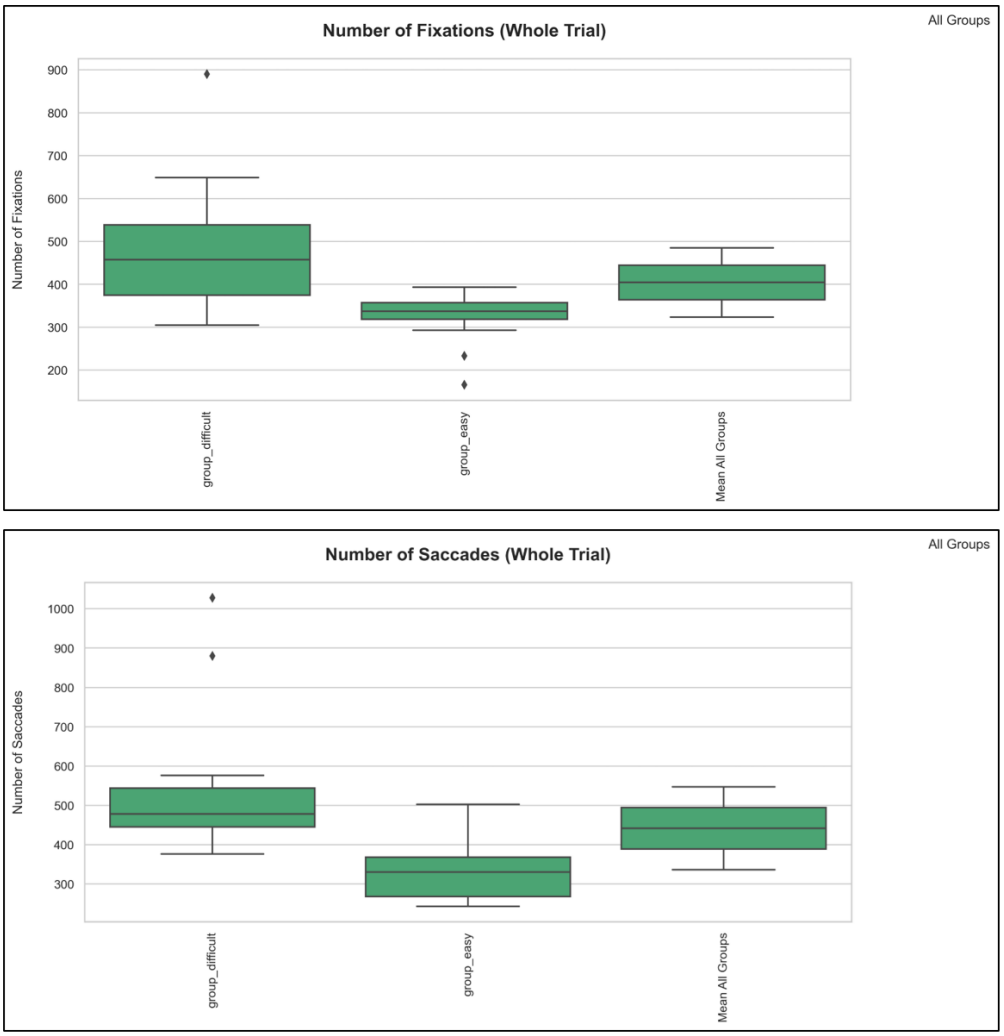


Figure 4.20: Box plots showing the average Number of Fixations (top) and Saccades (bottom) of the two groups as well as their mean.

4.1.2 Summary of Findings from the Study

Metrics

The goal of this proof-of-concept study was to assess whether the metrics implemented in this framework allow to find meaningful differences between two tasks of distinct difficulty. Indeed, many significant differences between *GD* and *GE* could be identified while testing the formulated hypotheses. However, in total, only four of the nine hypotheses were supported by a statistically significant result. Further, apart from the metrics that reflect

attention on OOIs (Total Hits, Total Fixation Time, Revisits, and Total Dwell Time per OOI) and efficiency (Total Duration), none of the outcomes is able to provide a qualitative statement. Moreover, the underlying reason for increased attention towards an OOI cannot be determined directly, since it depends on the given task. Instead, each metric has to be considered within the given context and together with other outcomes as a whole. However, for certain tasks, correlations between single metrics and a variable, e.g. expertise, have been manifested in the literature. In that case, they can be interpreted individually and compared to other studies. Moreover, caution is needed when interpreting certain averages over an entire trial. In particular, this became evident with coefficient K. Thus, in a dynamic setting that features multiple different subtasks, as is the case in a medical device usability study or in a surgical skill assessment, the analysis will always be more informative if it is separated into actions.

Nevertheless, many different conclusions could be drawn from the outcome of this study. Firstly, the fixation and saccade-based metrics of the general analysis (Average and Total Fixation Duration, Total Duration, Fixation/Saccade Ratio) revealed valuable information on efficiency, focus, and cognitive workload. Moreover, the K-Coefficient could reveal some insights into the ambient or focal vision of the subject when averaged over time intervals in certain subjects. In the eyes of the author, the OOI-based metrics revealed the most prominent differences between the two groups, in particular the attention towards the instructions and the grid. This clearly indicated that *GD* spent a lot more time trying to identify the shape on the instructions and finding it in the grid. This is further underlined by the higher SGE in *GD* as a consequence of a more even hit distribution. Overall, the addition of an *OOI-based analysis* clearly enhances the informative value of an ET study.

Summed up, all the metrics have the potential to be meaningfully interpreted in the given context. However, at this point, none of them can provide the user with ready-to-use feedback on the performance without requiring interpretation by the user.

Workflow

The implemented GUI made running the two data sets *DS1* and *DS2* through the framework quick and convenient. While the generation of the output graphs and tables from the different modules was relatively fast, it took a few hours to export all the summary reports. Since they are only created once all the other modules have been executed, the user can already look at the exported graphs and tables while the reports are still created.

For each hypothesis, we could find and discuss possible explanations by either reviewing related metrics or examining the recordings. The interpretation and exploration of the outcome are made easy for the user in different ways. Firstly, all information is visually available in ready-to-use graphs. Additionally, for the entropy metrics and the K-coefficient, particularly high or low values are marked within the graphs. Also, all metrics are summarised on each level; per trial, per participant, per group, and for all groups. In addition, for the user who desires to take a closer look, the .csv outputs make it easy to run statistical analyses on the exported metrics, similar to how it has been done in this study. Lastly, various different aspects of the performance can be analysed, i.e. cognitive workload, focus, entropy, attention, and efficiency.

4.2 Action Recognition Validation

As indicated before, this analysis is based on fictional participants and groups. Therefore, its outcome will not be interpreted regarding the usability of the UnoPen or the SmartPilot. Instead, the outputs will be presented and commented on with respect to their utility in an automated ET analysis. In particular, the two modules *Action-based analysis* and *Sequence comparison* will be addressed, since they were not part of the proof-of-concept study.

4.2.1 Action-Based Analysis

When running the *Action-based analysis* module, the metrics of the previously executed subanalyses are computed again, but per action instead of over the entire trial [1] Consequently, all graphs and tables are now additionally exported for each action. This is particularly valuable in studies that consist of diverse subtasks, as is the case in most medical device usability tests and surgical workflow assessments. It does not only allow comparing the same action across different conditions, e.g. to assess a specific feature that differs in two medical device prototypes. It also enables to compare different actions within the same participant or group, e.g. to identify what surgical steps a subject struggles with. Figure 4.21 shows the Average Fixation duration of fictional *participant03* in two different actions.

[1] In particular, these are the *General analysis*, the *K-Coefficient analysis*, and the *OOI-based analysis*. Please refer to Figure 3.6 again for an overview of the modules within the pipeline.

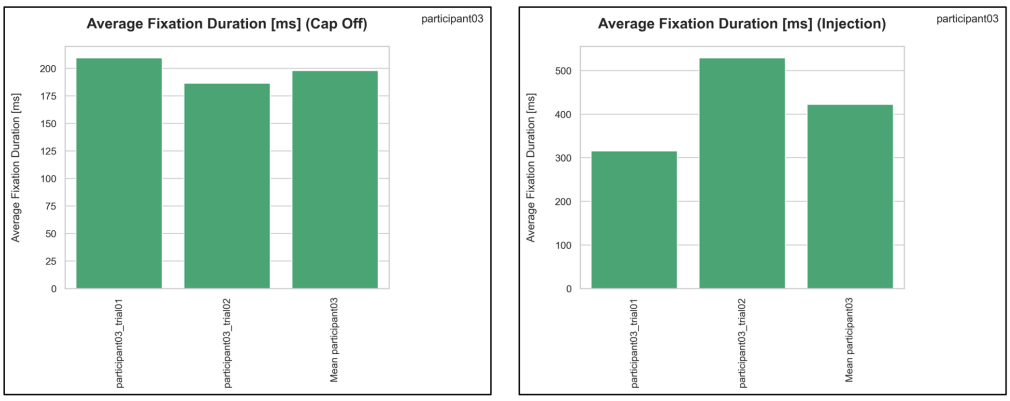


Figure 4.21: Average Fixation Duration [ms] of two different actions for the same participant.

Further, the Total Duration of each action is determined. An example is illustrated in Figure 4.22. Next to the Total Duration, also the Average Duration per Action is determined. This provides additional information in case at least one action occurs more than once within a trial, which is quite probable within the intended use of the program. Although this output is very straightforward in its calculation, it might be one of the most useful action-based metrics.

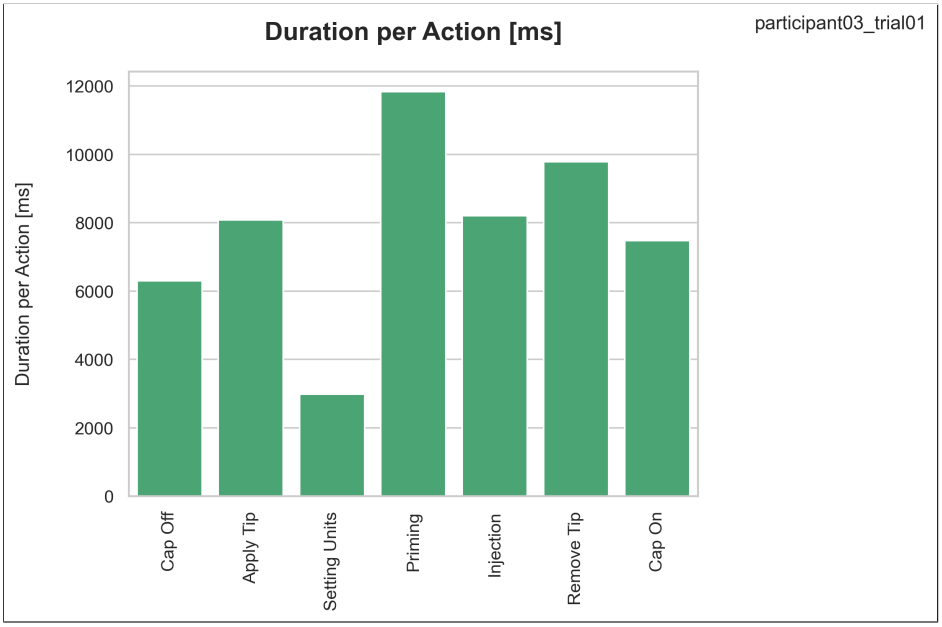


Figure 4.22: Average Duration [ms] of all actions of $participant03_trial01$.

As concluded in Section 4.1.1, the calculation of the coefficient K makes more sense when calculated over small time intervals or actions. The latter is presented in Figure 4.23 and provides the user with much more in-depth information on the focal and ambient vision of the participants.

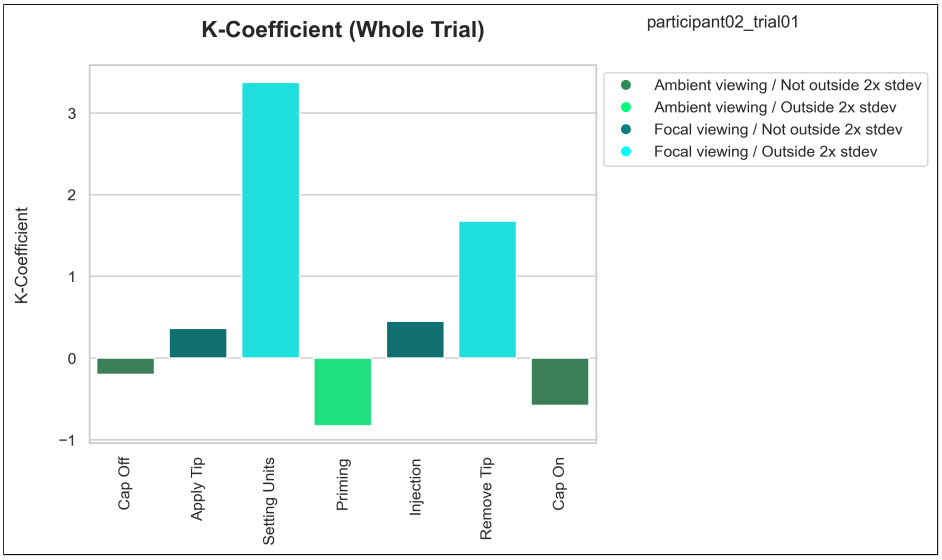


Figure 4.23: Coefficient K calculated over each action. In this case, three average Ks (for Setting Units, Priming and Remove Cap) lie outside two standard deviations from the mean.

Moreover, the sequence of the actions, along with the duration, are exported in a table, as seen in Figure 4.24. This information will be further used in the *Sequence comparison* module.

	A	B	C	D
1		action	start_time_action	end_time_action
2	0	Cap Off	0	6296
3	1	Apply Tip	6296	14382
4	2	Setting Units	14382	17050
5	3	Priming	17050	28883
6	4	Setting Units	28883	32711
7	5	Injection	32711	37958
8	6	Setting Units	37958	40417
9	7	Injection	40417	51580
10	8	Remove Tip	51580	61364
11	9	Cap On	61364	68840
12				

Figure 4.24: Duration per step [ms] of an example trial where each step corresponds to one of the predefined actions.

Summed up, similarly to the inclusion of OOIs seen in Section 4.1.2, action separation adds enormous value to the analysis. However, it involves not only more effort but also some expertise to set up the PVHMM algorithm needed to generate the *OGD/Action file*. Moreover, the results can only be as accurate as the underlying action classifier. This will be further discussed in Section 4.3.

4.2.2 Sequence Comparison

For each trial, the Levenshtein distance between the sequence of actions calculated in the *Action-based analysis* (see Figure 4.24) and the template sequence provided via the user input is calculated. An example is provided in Figure 4.25.

	A	B	C
1		participant03	participant04
2	trial01	2	4
3	trial02	4	2

Figure 4.25: Screenshot of the exported .csv file with the Levenshtein distances of all trials of *group_2*.

Consequently, for each trial, one can estimate how similar the performance was to the given template. The latter could represent e.g. the instructions for use of a medical device or the gold standard of a surgical technique, making this module particularly beneficial.

Besides the Levenshtein distance, many other edit-distance algorithms exist to quantify how similar sequences are, e.g. the Hamming distance (Hamming (1950)) or the Jaro-Winkler distance (Winkler (1999)). These could additionally be implemented in a future version of the program. Moreover, this part of the framework could be extended to a more in-depth sequence analysis, e.g. to determine expertise development according to Wang et al. (2022).

4.2.3 Report

A .pdf summary is automatically generated for each participant, for each group, as well as for all groups. The auto-generated report of *group_1* in the validation analysis can be found in the Appendix A.6. Although it is mainly intended for people with little experience in analysing ET data, it can be useful to quickly get a first overview of the outcomes of the evaluation. The short introduction to each metric provides the reader with a

brief explanation of what was measured, as well as suggestions on how to interpret the values. Moreover, the fact that metrics are separated into three categories is additionally helpful for an ET novice. However, as stated above, it can be problematic to interpret means according to the suggestions without taking any closer look at the data, as wrong conclusions could be drawn from it. Thus, summing up, the summary report might be useful as a quick first impression on the outcomes or for somebody with only limited experience with ET, but it has to be taken with caution.

4.3 Pre-Processing

The main goal of the project was to create an automated evaluation framework by making use of the previous work from $pd|\mathbf{z}$, in particular, the OGD algorithm and the PVHMM algorithm. To this end, the output of these methods, the *OGD file* and the *OGD/Action file*, are employed as input for the proposed framework. This enables an OOI-based and action-based analysis with minimal effort for the user compared to conventional methods. In case this data input is not present, the framework's modular structure still enables the user to run the evaluation only with the *Fixation/saccade file* exported from Tobii Pro Lab.

However, it could be shown that the most prominent value of this automated ET analysis clearly comes from the integration of the OOIs and the actions. If the appropriate input data is provided, no manual annotation is required by the user, which is a great advantage compared to conventional approaches. However, it is important to note that the generation of these input files through the OGD algorithm and the PVHMM for HAR algorithm, i.e. building the models, come with a substantial amount of time and effort, as seen in this study. Also, they afford a certain level of expertise in machine learning. For the instance segmentation algorithm, appropriate images have to be extracted from the video frames and the OOIs have to be labelled. Then, a good augmentation algorithm and the right training parameters have to be found through parameter optimization. Further, the results are highly dependent on the accuracy of the underlying instance segmentation of the trained model. Thus, the latter has to be thoroughly tested before using it to generate the *OGD file* which is either fed to the framework or used as a basis for HAR in the PVHMM.

Nevertheless, the manual labelling of OOIs as the current state-of-the-art arguably comes with higher time expenditure, depending on the number of OOIs and recordings. Hence, in each case, one has to weigh up the effort and return anew. Naturally, if the deep learning-based instance segmentation

and action recognition can be reused for multiple studies, it is more likely to be worth the effort. This is presumably the case in surgical skill assessment, where the setting, the objects, and the task stay the same and the pipeline could be reused multiple times. However, when it comes to usability testing, examining the same medical device with the same actions multiple times is less likely.

4.4 Limitations

Although the implemented metrics provide valuable and diverse information on a performed task, the study has shown that only a few of them are able to deliver a plain suggestion for interpretation. Moreover, these suggestions are limited to efficiency and attention towards OOIs. Thus, while the automation is provided in the pre-processing step of the pipeline, it is not fulfilled yet in terms of delivering an automated performance assessment as an output nor have suitable metrics been found within this project.

The findings in this work are solely supported by a proof-of-concept study conducted with a shape sorter toy. In order to more accurately evaluate the implemented metrics, the framework should be tested with a task that is closer to the program's intended use. Unfortunately, the analysis of the action-based input with PVHMM could only be examined regarding its functionality and not its qualitative value. Also, no data from surgical workflow assessments have been fed to the framework.

Certain metrics have the potential to be extended, as suggested in the discussion of the individual hypotheses in Section 4.1. For instance, the SGE and the GTE could additionally be computed using a grid instead of OOIs, or the coefficient K over time could be exported after reducing the noise. Also, there is the possibility to introduce new metrics that can fully exploit the *OGD file*.

Regarding the output, some of the created graphs are hard to read if there are too many participants per group or trials per participant. For these graphs, a limit of trials or participants could be defined for their generation and export. Also, is not yet possible to easily compare trial numbers, e.g. all first trials to all last trials. As a workaround at this point, one could assign each trial number to a different group in order to compare them. The next major step includes fully integrating the pre-processing, i.e. generation of the *OGD file* or the *OGD/Action file*, since they are two consecutive steps at the moment.

5 Conclusion

With the ultimate goal of advancing the automation of medical device usability testing and surgical skill assessment, we determined a range of suitable ET metrics. We then generated a framework that with different data inputs calculates and summarises these metrics on different levels, and we tested it in a proof-of-concept study and with already existing data from a medical device usability testing.

The conducted study and the action recognition validation run proved that the selected metrics allow making implications on various aspects of a performance, in particular cognitive workload, focus, visual entropy, attention, and efficiency. Thereby, also more complex metrics like the K-coefficient, the SGE, and the GTE have been successfully implemented. While almost all the metrics have the potential to provide an interpretive approach within the given context and when combined together, none of them can deliver a ready-to-use statement for a performance assessment on their own.

The major added value of the proposed pipeline in terms of automation is the integration of the object-gaze mapping and the action recognition that spares the user with any manual annotation tasks. Even though the generation of the deep learning-based models is time-intensive too, it arguably outperforms the conventional method very quickly, especially if it can be reused multiple times.

Besides that, the framework comes with a diverse set of attributes. The extensive output with the auto-generated graphs and summaries on different levels grants the user with various possibilities for the analysis. On top of that, the modular structure of the framework allows to conduct the evaluation with different data inputs and research focuses. In addition to integrating the calculation of all the metrics per action, a sequence comparison module was added to compute the Levenshtein distance, which can be very useful regarding the program's intended use. Further, with the intention of setting the basis for many different features, a few elements have been integrated into the framework even though their functionality could not be fully exhausted

within this work. Examples of this are the *Statistics* and the *Sequence comparison* modules. These parts are built in a way so they can easily be extended in the future. The extensively equipped GUI makes the program quick and easy to use. Together with the auto-generated summary report, the program can even be operated by people with very limited experience in ET data analysis.

In conclusion, the proposed pipeline clearly has the potential to facilitate the analysis of ET data. With the integration of the OGD and the PVHMM algorithm, the pipeline is largely automated on the input side. However, regarding the output, the proof-of-concept study showed that a fully automated assessment that does not require any additional interpretation by the user is not feasible at this time point with the selected metrics.

To this end, an alternative approach to generate fully automated feedback could be to provide information on the context and the task in the input of the pipeline. As an example, the user could choose for each OOI, if increased attention means higher complexity, engagement or interest. Instead of exporting e.g. the Total Dwell Time per OOI, the output could make a qualitative statement, e.g. that the subject had a high interest in a specific OOI. Moreover, if certain correlations have been found in a given context, they could be entered as input as well, e.g. if for a specific task, a higher GTE has been associated with confusion. With the goal of a fully automated performance assessment from the input all the way to the output, this represents a promising approach that could be pursued in the future.

A Appendix

A.1 Source Code

The source code developed and used during this thesis can be found on `https://github.com/reneemuriel/Automated_ET_Analysis_Framework_MT.git`

A.2 Framework Directory Structure

```
Framework/
├─ app.py
├─ config/
|  ├─ example.yaml
├─ data/
|  ├─ input/
|  ├─ output/
|  ├─ report/
├─ scripts/
|  ├─ append_input_type.py
|  ├─ split_tobii_output.py
|  ├─ tobii_to_fixations_script.py
├─ src/
|  ├─ analysis/
|  |  ├─ action_analysis.py
|  |  ├─ general_analysis.py
|  |  ├─ kcoeff_analysis.py
|  |  ├─ ooi_analysis.py
|  |  ├─ report.py
|  |  ├─ statistics.py
|  ├─ gui/
|  ├─ util/
|  |  ├─ action_separation.py
|  |  ├─ general_metrics.py
|  |  ├─ ...
```

Figure A.1: Directory tree of the framework.

The directory structure of the framework is depicted in Figure A.1 and explained in the following.

- To start the analysis, *app.py* has to be executed. First, it collects the configurations either via the specified .yaml file stored in the *config* folder or the GUI. Then, the analysis is run by calling the chosen modules in the *src/analysis/* directory.
- *src/utils/* contains the functions used in the different modules.
- The *src/gui/* directory holds all the files needed to run the GUI.
- The sub-directory *data* is meant to store the input data and the output folder. However, these can be placed anywhere specified by the user. The summary reports get exported into *data/report*.

- In the *scripts* folder, tools that may be needed for data preprocessing are stored.
  - *tobii_to_fixations_script.py* filters all fixations from the *Fixation/saccade file* exported from Tobii Pro Lab and converts it to a .txt file, e.g. to be used as input for the OGD program.
  - *append_input_type.py* appends a desired string to filenames without changing the extension, e.g. to append '_tobii' to exported .tsv files from Tobii Pro Lab.
  - *split_tobii_output.py* splits a single "Metrics Export" file (exported from Tobii Pro Lab) that contains multiple recordings into single trials to get the *Fixation/saccade files*.

A.3 Data Input Format

The following input format should be followed (see Figure A.2):

- One folder per group should be created in a selected input directory. All types of input files from all trials belonging to one group have to be dumped into the respective group folder.
- The *Fixation/saccade file*'s name must follow this structure: *participantXX_trialYY_tobii.tsv*, whereas *XX* denotes the participant number and *YY* denotes the trial number.
- The *OGD file* or *OGD/Action file*, depending on availability, must be named *participantXX_trialYY_tobii.tsv*, whereas *XX* denotes the participant number and *YY* denotes the trial number.

```
input/
├─ group1/
├─ group2/
├─ group3/
|   ├─ participant05_trial01_tobii.tsv
|   ├─ participant05_trial01_ogd.txt
|   ├─ participant05_trial02_tobii.tsv
|   ├─ participant05_trial02_ogd.txt
|   ├─ participant06_trial01_tobii.tsv
|   ├─ participant06_trial01_ogd.txt
|   ├─ participant06_trial02_tobii.tsv
|   ├─ participant06_trial02_ogd.txt
```

Figure A.2: Directory tree of an example input. The study is divided into three groups, whereas group3 is made up of two participants with two trials each.

A.4 DTL Tree Structure

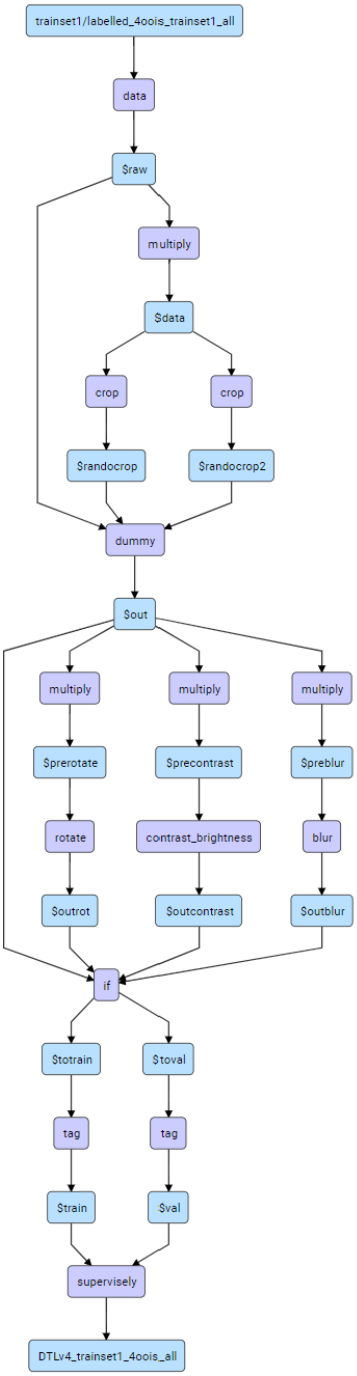


Figure A.3: DTL Tree Structure extracted from supervise.ly.

A.5 OOIs and Actions of the Validation Study

The following OOIs were chosen in the OGD algorithm:

- App
- Cap
- Gauge
- Pad
- Pen
- Safety
- Tip

The following actions were chosen for the HAR by PVHMM:

- Cap Off
- Apply Tip
- Setting Units
- Priming
- Injection
- Remove Tip
- Cap On

The correct template sequence was defined as: Cap Off, Apply Tip, Setting Units, Priming, Setting Units, Injection, Remove Tip, Cap On.

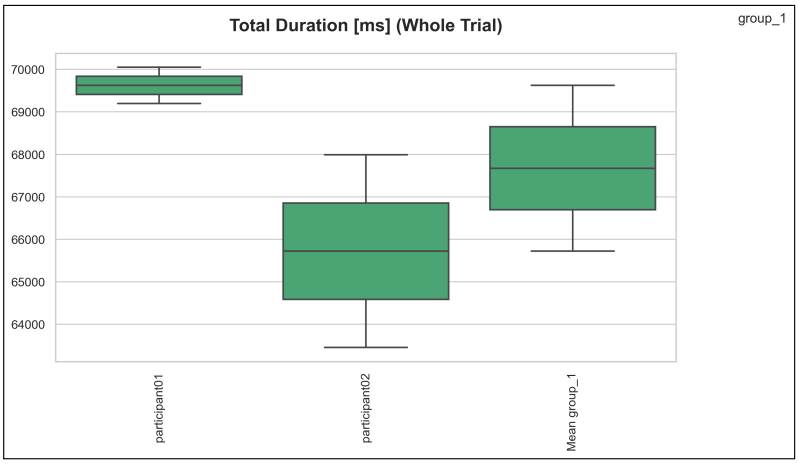
A.6 Summary Report Example

Summary of Gaze Analysis - group_1

1) Efficiency

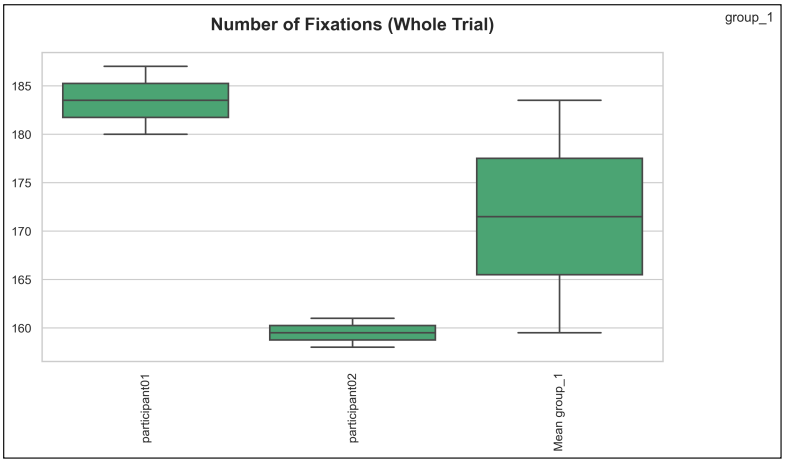
Total Duration

The average time [ms] it has taken the participant to complete the task. Generally, the shorter the duration, the more efficient the execution, and the higher the expertise.



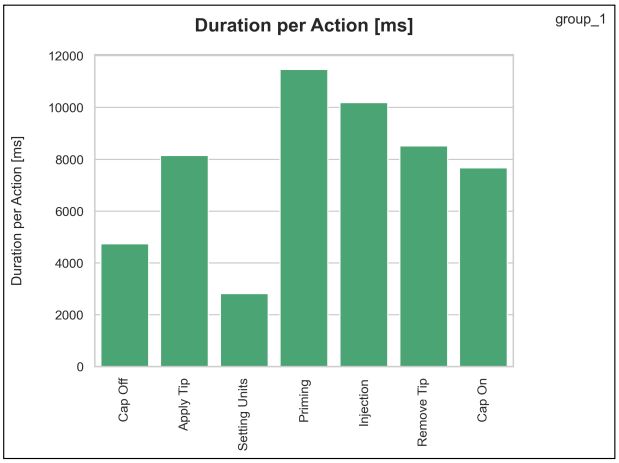
Number of Fixations

The average number of fixations per trial. Naturally, this correlates with the trial duration. Therefore, it is only useful if it is either normalised or if all recordings of a study have the same length. Generally, the number of fixations decreases with increasing expertise and in turn, increasing efficiency.



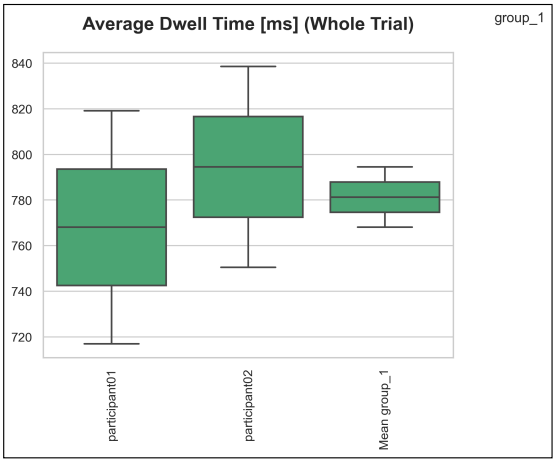
Average Duration per Action

The average time [ms] it has taken the participants to complete each of the identified actions. Similar to "Total Duration", the faster they were, the more efficient the task was performed.



Average Dwell Time

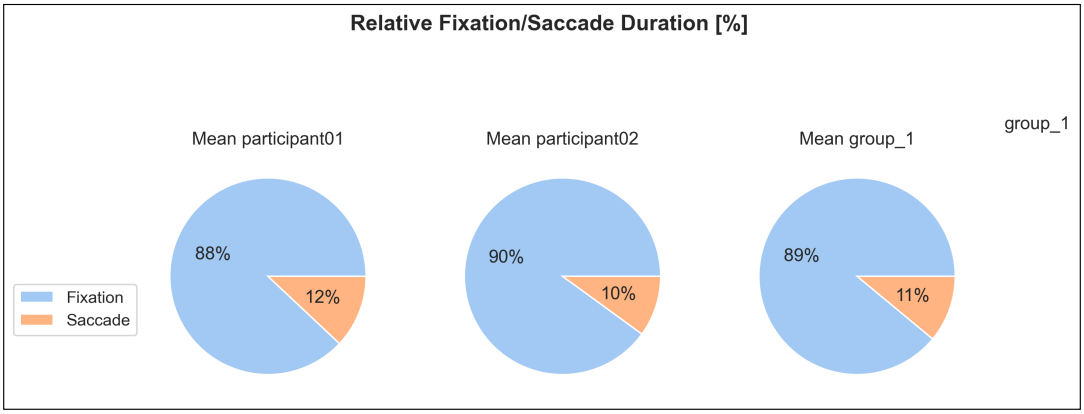
The average duration per dwell [ms] on the defined OOIs. Longer dwell times can mean that the chosen OOIs keep up the attention for longer individual time periods.



2) Focus

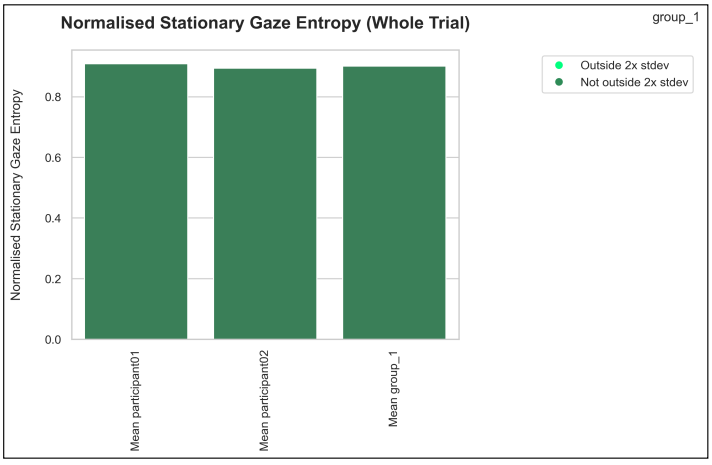
Relative Fixation/Saccade Duration

Relative percentage of fixation and saccade durations, i.e. the total time the person has spent fixating compared to travelling from one fixation to the next. The higher the ratio, the more time is spent processing compared to searching.



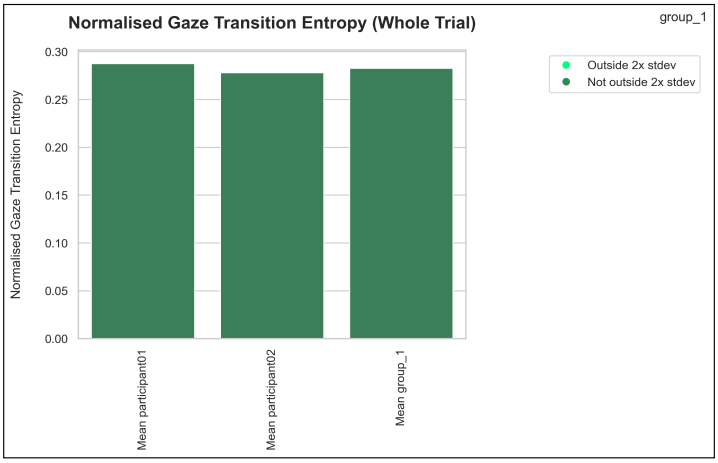
Normalised Stationary Gaze Entropy

A higher Stationary Gaze Entropy implies a more equal distribution of the visual attention between the OOIs. A lower value reflects when fixations tend to be concentrated on specific OOIs, either because they are more complex or more interesting to the subject.



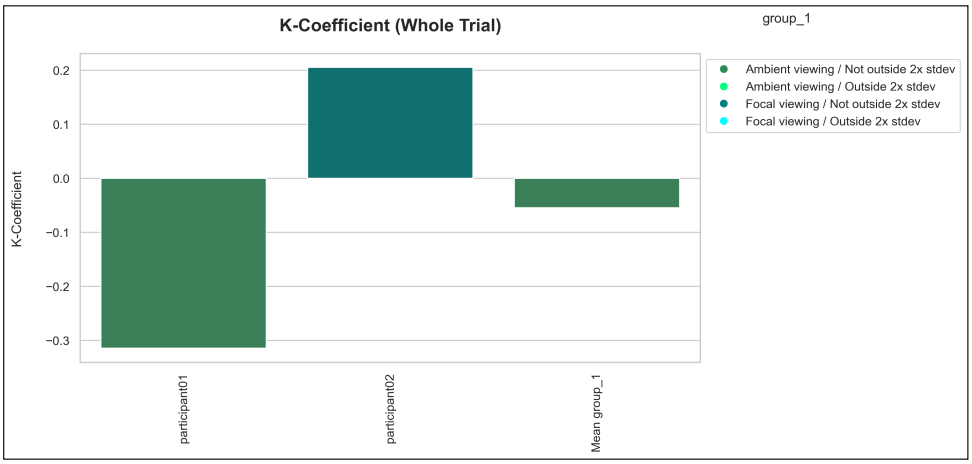
Normalised Gaze Transition Entropy

In short, higher entropy can imply more randomness in the visual scanning pattern and in turn, less focus and efficiency.



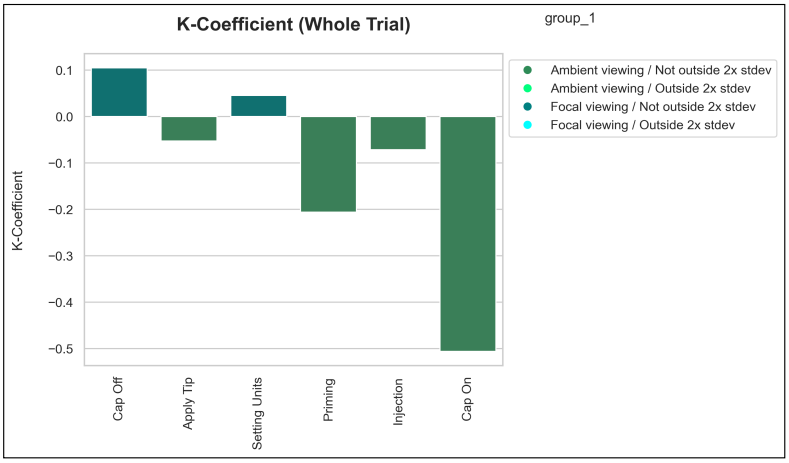
Average K-Coefficients

K > 0 indicates relatively long fixations succeeded by short saccades, implying focal vision. K < 0 indicates relatively short fixations succeeded by long saccades, implying ambient vision.



Average K-Coefficients per Action

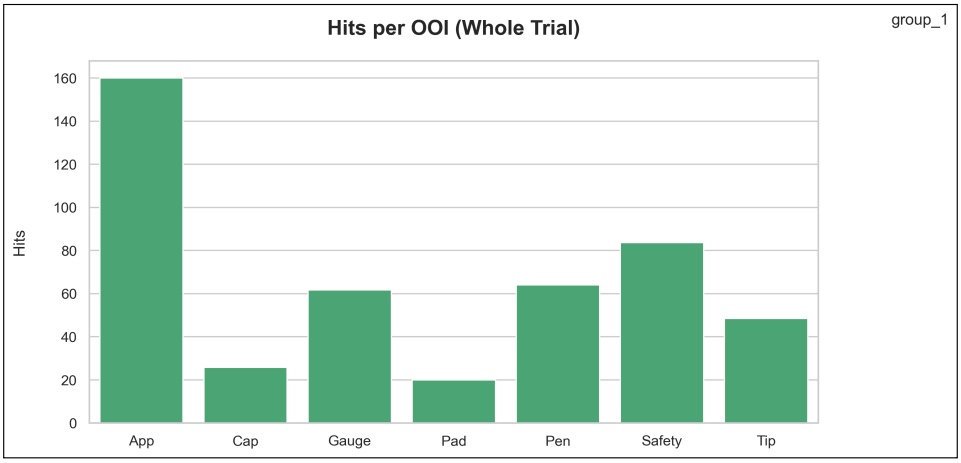
K > 0 indicates relatively long fixations succeeded by short saccades, implying focal vision. K < 0 indicates relatively short fixations succeeded by long saccades, implying ambient vision.



3) Attention / Object of Interest-Based Analysis

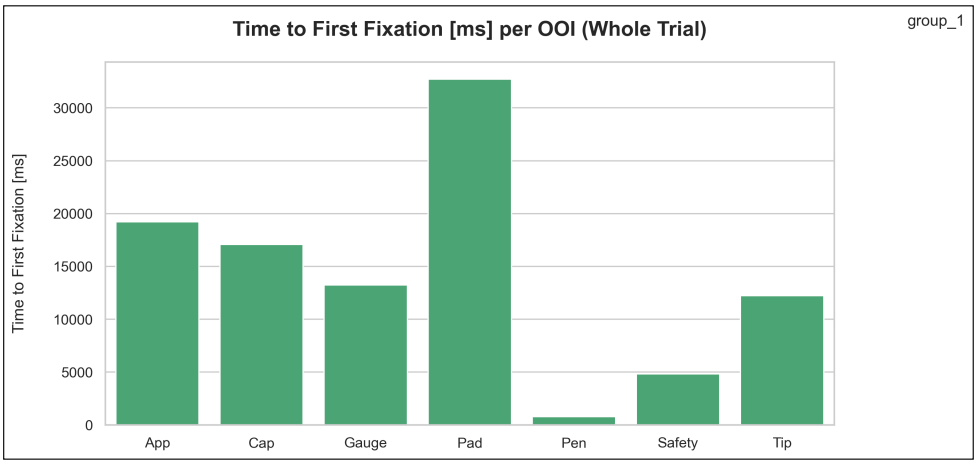
Hits per OOI

The amount of fixations that were identified on the respective object of interest. In general, the more hits an object has, the higher its importance.



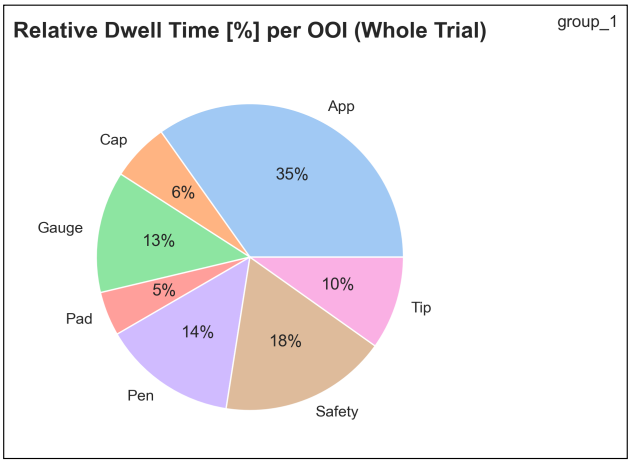
Time to First Fixation per OOI

The average time [ms] until the first fixation on a specific object took place. In general, the less time passes until the object is noticed, the higher its importance or the more noticeable it is.



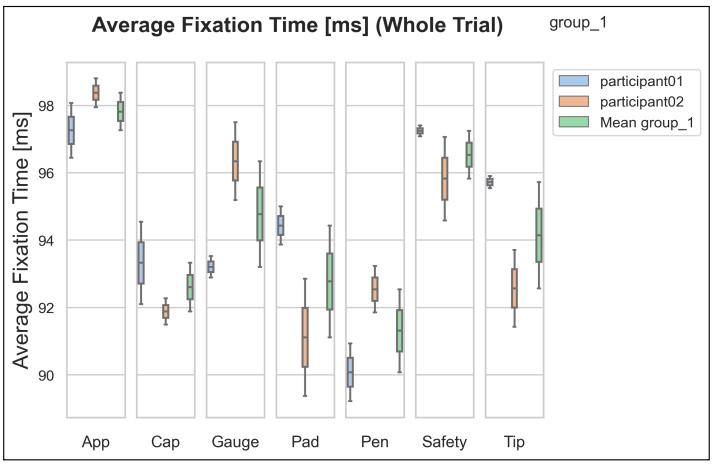
Relative Dwell Time per OOI

The relative amount of time the participants' gaze was focused on each OOI. In general, the higher the percentage of dwell time, the higher the objects' importance.



Average Fixation Time per OOI

The average fixation time per OOI is the mean duration of all fixations placed onto a particular OOI. Generally, higher fixation durations can be associated with more focus, concentration or interest.



Bibliography

Ahmidi, N., Hager, G. D., Ishii, L., Fichtinger, G., Gallia, G. L. & Ishii, M. (2010), 'Surgical task and skill classification from eye tracking and tool motion in minimally invasive surgery', *Med Image Comput Comput Assist Interv* **13**(Pt 3), 295--302.

Baloh, R. W., Sills, A. W., Kumley, W. E. & Honrubia, V. (1975), 'Quantitative measurement of saccade amplitude, duration, and velocity', *Neurology* **25**(11), 1065--1070.

Bastien, J. M. C. (2010), 'Usability testing: a review of some methodological and technical aspects of the method', *International Journal of Medical Informatics* **79**(4), E18--E23.

Bergstrom, J. R., Schall, A. J., Dunkerley, M., Scherer, H., Studholme, A. & ProQuest (2014), 'Eye tracking in user experience design'.

Branaghan, R. J., O'Brian, J. S., Hildebrand, E. A. & Foster, L. B. (2021), *Humanizing Healthcare - Human Factors for Medical Device Design*, 1st 2021. edn, Springer International Publishing : Imprint: Springer, Cham.

Bylinskii, Z., Borkin, M. A., Kim, N. W., Pfister, H. & Oliva, A. (2017), 'Eye fixation metrics for large scale evaluation and comparison of information visualizations', *Eye Tracking and Visualization: Foundations, Techniques, and Applications, Etvis 2015* pp. 235--255.

Carter, B. T. & Luke, S. G. (2020), 'Best practices in eye tracking research', *International Journal of Psychophysiology* **155**, 49--62.

Clancy, C. M. (2009), 'Ten years after to err is human', *Am J Med Qual* **24**(6), 525--8.

Dalveren, G. G. M. & Cagiltay, N. E. (2018), 'Insights from surgeons' eye-movement data in a virtual simulation surgical training environment: effect of experience level and hand conditions', *Behaviour Information Technology* **37**(5), 517--537.

Davson, H. (1990), *Physiology of the eye*, 5th edn, Macmillan, Basingstoke.

Deane, O., Toth, E. & Yeo, S.-H. (2022), 'Deep-saga: a deep-learning-based system for automatic gaze annotation from eye-tracking data', *Behavior Research Methods* pp. 1--20.

Devillez, H., Guyader, N., Curran, T. & O'Reilly, R. C. (2020), 'The bimodality of saccade duration during the exploration of visual scenes', *Visual Cognition* **28**(9), 484--512.

Doherty, S., O'Brien, S. & Carl, M. (2010), 'Eye tracking as an mt evaluation technique', *Machine Translation* **24**(1), 1--13.

Donaldson, M. S., Corrigan, J. M. & Kohn, L. T. (2000), 'To err is human: building a safer health system'.

Duchowski, A. T. (2017), *Eye tracking methodology : theory and practice*, Springer, Cham, Switzerland.

Duchowski, A. T. & Krejtz, K. (2017), 'Visualizing dynamic ambient/focal attention with coefficient k', *Eye Tracking and Visualization: Foundations, Techniques, and Applications, Etvis 2015* pp. 217--233.

Dzau, V. J. & Shine, K. I. (2020), 'Two decades since to err is human: Progress, but still a "chasm"', *Jama-Journal of the American Medical Association* **324**(24), 2489--2490.

Ellis, S. R. & Stark, L. (1986), 'Statistical dependency in visual scanning', *Human Factors* **28**(4), 421--438.

Estrada, S., O'Malley, M. K., Duran, C., Schulz, D. & Bismuth, J. (2014), 'On the development of objective metrics for surgical skills evaluation based on tool motion', *2014 Ieee International Conference on Systems, Man and Cybernetics (Smc)* pp. 3144--3149.

EU (2017), 'Regulation (eu) 2017/745 of the european parliament and of the council of 5 april 2017 on medical devices, amending directive 2001/83/ec, regulation (ec) no 178/2002 and regulation (ec) no 1223/2009 and repealing council directives 90/385/eec and 93/42/eec (text with eea relevance) text with eea relevance: European union, 2017', *J. Eur. Union* **5**(2017), 1--175.

Evans, K. M., Jacobs, R. A., Tarduno, J. A. & Pelz, J. B. (2012), 'Collecting and analyzing eye-tracking data in outdoor environments', *Journal of Eye Movement Research* **5**(2).

FDA (2011), Applying human factors and usability engineering to medical devices - guidance for industry and food and drug administration staff, Guidance, U.S. Department of Health and Human Services, Food and Drug Administration.

Funke, I., Mees, S. T., Weitz, J. & Speidel, S. (2019), 'Video-based surgical skill assessment using 3D convolutional neural networks', *International Journal of Computer Assisted Radiology and Surgery* **14**(7), 1217--1225.

Gawande, A. A., Thomas, E. J., Zinner, M. J. & Brennan, T. A. (1999), 'The incidence and nature of surgical adverse events in colorado and utah in 1992', *Surgery* **126**(1), 66--75.

Goh, J. O., Tan, J. C. & Park, D. C. (2009), 'Culture modulates eye-movements to visual novelty', *Plos One* **4**(12).

Goldberg, J. H. & Kotval, X. P. (1999), 'Computer interface evaluation using eye movements: methods and constructs', *International Journal of Industrial Ergonomics* **24**(6), 631--645.

Guo, F., Ding, Y., Liu, W. L., Liu, C. & Zhang, X. F. (2016), 'Can eye-tracking data be measured to assess product design?: Visual attention mechanism should be considered', *International Journal of Industrial Ergonomics* **53**, 229--235.

Hamming, R. W. (1950), 'Error detecting and error correcting codes', *The Bell system technical journal* **29**(2), 147--160.

He, K. M., Gkioxari, G., Dollar, P. & Girshick, R. (2017), 'Mask R-CNN', *2017 Ieee International Conference on Computer Vision (Iccv)* pp. 2980--2988.

Hegde, V. (2013), 'Role of human factors / usability engineering in medical device design', *59th Annual Reliability and Maintainability Symposium (Rams)* .

Holmqvist, K., Holmqvist, K. & Ebook, L. (2011), *Eye tracking : a comprehensive guide to methods and measures*, first edition. edn, Oxford University Press, Oxford.

IEC (2015), Medical devices - Part 1: Application of usability engineering to medical devices, Standard, International Organization for Standardization, Geneva, CH.

ISO (2018), Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts, Standard, International Organization for Standardization, Geneva, CH.

Jain, S. (2019), 'Applied physiology of eye movements', *Simplifying Strabismus* pp. 15--22.

James, J. T. (2013), 'A new, evidence-based estimate of patient harms associated with hospital care', *J Patient Saf* **9**(3), 122--8.

Just, M. A. & Carpenter, P. A. (1976), 'Eye fixations and cognitive-processes', *Cognitive Psychology* **8**(4), 441--480.

Kable, A. K., Gibberd, R. W. & Spigelman, A. D. (2002), 'Adverse events in surgical patients in australia', *International Journal for Quality in Health Care* **14**(4), 269--276.

Koester, T., Brøsted, J. E., Jakobsen, J. J., Malmros, H. P. & Andreasen, N. K. (2017), 'The use of eye-tracking in usability testing of medical devices', *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* **6**(1), 192--199.

Kreiner, T. (2021), Development of a peripheral vision based approach to human task classification, Master's thesis, ETH Zurich.

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A. & Kopacz, A. (2016), 'Discerning ambient/focal attention with coefficient k', *ACM Transactions on Applied Perception (TAP)* **13**(3), 1--20.

Krejtz, K., Szmidt, T., Duchowski, A. T. & Krejtz, I. (2014), Entropy-based statistical analysis of eye movement transitions, *in* 'Proceedings of the Symposium on Eye Tracking Research and Applications', pp. 159--166.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., Graham, A. R., Descour, M. R., Davis, J. R. & Weinstein, R. S. (2006), 'Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience', *Human Pathology* **37**(12), 1543--1556.

Levenshtein, V. I. et al. (1966), Binary codes capable of correcting deletions, insertions, and reversals, *in* 'Soviet physics doklady', Vol. 10, Soviet Union, pp. 707--710.

Levin, M., McKechnie, T., Khalid, S., Grantcharov, T. P. & Goldenberg, M. (2019), 'Automated methods of technical skill assessment in surgery: A systematic review', *Journal of Surgical Education* **76**(6), 1629--1639.

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P. & Zitnick, C. L. (2014), 'Microsoft coco: Common objects in context', *Computer Vision - Eccv 2014, Pt V* **8693**, 740--755.

Lohmeyer, Q., Schneider, A., Jordi, C., Lange, J. & Meboldt, M. (2019), 'Toward a new age of patient centricity? the application of eye-tracking to the development of connected self-injection systems', *Expert Opinion on Drug Delivery* **16**(2), 163--175.

Makary, M. A. & Daniel, M. (2016), 'Medical error-the third leading cause of death in the us', *BMJ* **353**, i2139.

Merali, N., Veeramootoo, D. & Singh, S. (2019), 'Eye-tracking technology in surgical training', *Journal of Investigative Surgery* **32**(7), 587--593.

Michael, P. W., Kendler, J. & Strochlic, A. Y. (2015), *Usability testing of medical devices*, CRC press.

Mussgnug, M., Singer, D., Lohmeyer, Q. & Meboldt, M. (2017), 'Automated interpretation of eye-hand coordination in mobile eye tracking recordings identifying demanding phases in human-machine interactions', *Kunstliche Intelligenz* **31**(4), 331--337.

Ooms, K., Coltekin, A., De Maeyer, P., Dupont, L., Fabrikant, S., Incoul, A., Kuhn, M., Slabbinck, H., Vansteenkiste, P. & Van der Haegen, L. (2015), 'Combining user logging with eye tracking for interactive and dynamic applications', *Behavior Research Methods* **47**(4), 977--993.

Polisena, J., Williams, D. & Ciani, O. (2020), *Health technology assessment of medical devices*, Elsevier, pp. 795--798.

Punde, P. A., Jadhav, M. E. & Manza, R. R. (2017), 'A study of eye tracking technology and its applications', *2017 1st International Conference on Intelligent Systems and Information Management (Icisim)* pp. 86--90.

Rahal, R. M. & Fiedler, S. (2019), 'Understanding cognitive and affective mechanisms in social psychology through eye-tracking', *Journal of Experimental Social Psychology* **85**.

Ravizza, A., Lantada, A. D., Sanchez, L. I. B., Sternini, F. & Bignardi, C. (2019), 'Techniques for usability risk assessment during medical device design', *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies, Vol 1 (Biodevices)* pp. 207--214.

Rayner, K., Li, X., Williams, C. C., Cave, K. R. & Well, A. D. (2007), 'Eye movements during information processing tasks: Individual differences and cultural effects', *Vision research* **47**(21), 2714--2726.

Reingold, E. M. & Sheridan, H. (2011), 'Eye movements and visual expertise in chess and medicine.'.

Rodziewicz, T. L., Houseman, B. & Hipskind, J. E. (2022), *Medical Error Reduction and Prevention*, Treasure Island (FL).

Sakurai, M. (2016), *Parafovea*, Springer New York, New York, NY, pp. 997--999.
URL: *https://doi.org/10.1007/978-1-4419-8071-7*$_2$15

Schieber, F. & Gilland, J. (2008), Visual entropy metric reveals differences in drivers' eye gaze complexity across variations in age and subsidiary task load, *in* 'Proceedings of the Human Factors and Ergonomics Society Annual Meeting', Vol. 52, SAGE Publications Sage CA: Los Angeles, CA, pp. 1883--1887.

Shannon, C. E. (1948), *A mathematical theory of communication*, Bell telephone system technical publications Monograph B-1598, American Telephone and Telegraph Company, New York.

Sheridan, H. & Reingold, E. M. (2017), 'Chess players' eye movements reveal rapid recognition of complex visual patterns: Evidence from a chess-related visual search task', *Journal of Vision* **17**(3).

Shiferaw, B., Downey, L. & Crewther, D. (2019), 'A review of gaze entropy as a measure of visual scanning efficiency', *Neuroscience and Biobehavioral Reviews* **96**, 353--366.

Skaramagkas, V., Giannakakis, G., Ktistakis, E., Manousos, D., Karatzanis, I., Tachos, N., Tripoliti, E. E., Marias, K., Fotiadis, D. I. & Tsiknakis, M. (2021), 'Review of eye tracking metrics involved in emotional and cognitive processes', *IEEE Rev Biomed Eng* **PP**.

Slawomirski, L., Auraaen, A. & Klazinga, N. S. (2017), 'The economics of patient safety: strengthening a value-based approach to reducing patient harm at national level'.

Stuart, S. (2022), *Eye Tracking : Background, Methods, and Applications*, Neuromethods 183, 1st 2022. edn, Springer US : Imprint: Humana, New York, NY.

supervisely (2022), `https://supervise.ly/`.

Tao, L., Wang, Q., Liu, D., Wang, J., Zhu, Z. Q. & Feng, L. (2020), 'Eye tracking metrics to screen and assess cognitive impairment in patients with neurological disorders', *Neurological Sciences* **41**(7), 1697--1704.

Tobii AB (2017), Product Description Tobii Pro Glasses 2, Technical report.

Tobii AB (2022), *Tobii Pro Lab User Manual*.

Unema, P. J. A., Pannasch, S., Joos, M. & Velichkovsky, B. M. (2005), 'Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration', *Visual Cognition* **12**(3), 473--494.

Uppal, K., Kim, J. & Singh, S. (2022), 'Decoding attention from gaze: A benchmark dataset and end-to-end models', *arXiv preprint arXiv:2211.10966* .

van Hove, P. D., Tuijthof, G. J. M., Verdaasdonk, E. G. G., Stassen, L. P. S. & Dankelman, J. (2010), 'Objective assessment of technical surgical skills', *British Journal of Surgery* **97**(7), 972--987.

van Kasteren, A. (2019), 'The contribution of eye tracking to quality of experience assessment of 360-degree video'.

Vansteenkiste, P., Cardon, G. & Lenoir, M. (2013), Dealing with head-mounted eye-tracking data: Comparison of a frame-by-frame and a fixation-based analysis, *in* 'Proceedings of the 2013 conference on eye tracking South Africa', pp. 55--57.

Vincent, C. J., Li, Y. & Blandford, A. (2014), 'Integration of human factors and ergonomics during medical device design and development: it's all about communication', *Appl Ergon* **45**(3), 413--9.

Wachter, R. (2012), *Understanding Patient Safety, Second Edition*, 2nd edn, McGraw-Hill Publishing, New York.

Wang, F., Kreiner, T., Lutz, A., Lohmeyer, Q. & Meboldt, M. (2022a), 'What we see is what we do: A peripheral vision based HMM framework for gaze guided human action recognition'.

Wang, F. L. S., Wolf, J., Farshad, M., Meboldt, M. & Lohmeyer, Q. (2021), 'Object-gaze distance: Quantifying near- peripheral gaze behavior in real-world applications', *Journal of Eye Movement Research* **14**(1).

Wang, F. S., Gianduzzo, C., Meboldt, M. & Lohmeyer, Q. (2022), 'An algorithmic approach to determine expertise development using object-related gaze pattern sequences', *Behavior Research Methods* **54**(1), 493--507.

Wang, F. S., Gianduzzo, C., Meboldt, M. & Lohmeyer, Q. (2022b), 'An algorithmic approach to determine expertise development using object-related gaze pattern sequences', *Behavior Research Methods* **54**(1), 493--507.

Watson, R. A. (2013), 'Quantification of surgical technique using an inertial measurement unit', *Simulation in Healthcare-Journal of the Society for Simulation in Healthcare* **8**(3), 162--165.

Wedel, M. (2018), 'Improving ad interfaces with eye tracking', *The Wiley handbook of human computer interaction* **2**, 889--907.

Wegner, S. (2021), Automating quantitative mobile eye tracking analysis for human factors testing of medical devices, PhD thesis, ETH Zurich.

Wegner, S., Lohmeyer, Q., Wahlen, D., Neumann, S., Groebli, J. C. & Meboldt, M. (2020), 'Value of eye-tracking data for classification of information processing-intensive handling tasks: Quasi-experimental study on cognition and user interface design', *JMIR Hum Factors* **7**(2), e15581.

Winkler, W. E. (1999), The state of record linkage and current research problems, *in* 'Statistical Research Division, US Census Bureau', Citeseer.

Wolf, J., Hess, S., Bachmann, D., Lohmeyer, Q. & Meboldt, M. (2018), 'Automating areas of interest analysis in mobile eye tracking experiments based on machine learning', *Journal of Eye Movement Research* **11**(6).

Wu, C. H., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J. & Yu, D. (2020), 'Eye-tracking metrics predict perceived workload in robotic surgical skills training', *Human Factors* **62**(8), 1365--1386.