Master Thesis

Ophthalmic Optics and Psychophysics

Colleen Rothe

Matriculation Number: 26367

Evaluation of Scanpath Comparison Metrics for Static and Dynamic Tasks

Approved thesis of the course Ophthalmic Optics and Psychophysics to obtain the academic degree Master of Science

Day of submission: May 6, 2015

Auditor:	Prof. Dr. med. Ulrich Schiefer, Aalen University
Second Auditor:	Dr. Enkelejda Kasneci, Eberhard-Karls-University of Tübingen

Acknowledgement

Over the past eight months I have received support and encouragements from a great number of individuals. My very special thanks and appreciation go to Thomas Kübler for always being available for questions and issues concerning the thesis on hand and everything else. He is a great mentor who knows how to motivate and how to explain complex issues very clearly. He was the one giving me a push to get finished.

Furthermore, I would like to thank Prof. Dr. med. Ulrich Schiefer. He introduced me into the topic of eye tracking and knows how to enthuse students for optics and (neuro-) ophthalmology. He always has an open ear and gives useful feedback concerning all questions of life, even in his leisure time.

Thanks to Dr. Enkelejda Kasneci from the University of Tübingen for being second auditor and especially for the good cooperation.

Aditionally, I would like to thank Matthias Müller for helping me with LaTeX.

I owe particular thanks to Judith Ungewiß for her mental support and helpfuld advice whenever needed and to Brigitte Lanigan for proofreading.

Abstract

Purpose

Automated scanpath comparison metrics should deliver an objective method to evaluate the similarity of scanpaths. The aim of this thesis is an evaluation of seven existing scanpath comparison metrics in static and dynamic tasks in order to provide a guidline that helps to decide which algorithm has to be chosen for a special kind of task.

Methods

The applicability of the algorithms for a static, visual search task and a dynamic, interactive video game task as well as their constraints and limitations were tested. Therefore, binocular gaze data were recorded by using the eye tracking system *The Eye Tribe* (The Eye Tribe ApS, Copenhagen/ Denmark). Objective task performance measures from 21 subjects were used in order to create scanpath groupings for which a relevant effect of dissimilarity was to be expected. Objective task performance measures such as task performance time were statistically evaluated and compared to the results gained by the comparison metrics.

Results

Four of the algorithms being used successfully identified differences for static and dynamic tasks: *MultiMatch*, *iComp*, *SubsMatch* and the *Hidden Markov Model*. *ScanMatch* was very sensitive for the static task but not applicable to the dynamic task whereas *FuncSim* was suitable for dynamic but not for static tasks. *Eyenalysis* failed to detect any effect.

Conclusion

The applicability of scanpath comparison metrics depends on the state of the task, respectively on the kind of experimental set up. In future, the application area for eye tracking will expand and an improvement of automated scanpath comparison metrics is therefore required.

Keywords

Eye tracking, Scanpath, Comparison metrics, Visual search, Fixation, String alignment, Vector, Heat map, Pattern recognition

Contents

1 Introduction

Perception is a highly complex process which describes the incorporation of information, for instance visual, auditive and tactile sensations. Humans do not recognize their surrounding as a whole scene on one glance, it is rather sampled from single images by eye movements [10]. This is necessary due to the steep decline of spatial resolution from the foveola to the periphery caused by the decrease of cones. Images of the environment are projected onto the fovea centralis, encoded as electrical impulses and sent to the visual cortex where they are put together. This forms a reconstruction of the environment.

Furthermore, perception can be understood as a hypothesis due to diverse options of interpretation of retinal images. Visual attention and cognition is typically split into two models to describe and to demonstrate complex issues to get an acurate image of reality. The first one is called *top down*. From the abstract, general and superordinate, it goes out to the concrete, particular and subordinate. The second one is called *bottom up* which goes vice versa from the particular to the abstract. This model of top down and bottom up can not fully explain a real complex system but it helps to reduce its complexity by decomposition into its parts [36]. Due to the prediction of what can happen, humans are able to react adequatly and fast what formerly used to help them survive and also prevents humans from disadvantages today. [15]

Human gaze movements can be characterised as fixations, saccades and smooth pursuits (section 2.1). These parameters are important to understand human gaze behavior and how eye tracking works.

While creating huge masses of eye tracking data has become relatively easy, the analysis of such data is still a very time-consuming process. The manually evaluation is still the standard with severe drawbacks like subjectivity, time consumption and difficulties in reproduction. Automated scanpath comparison algorithms do exist, however, their applicability to experimental designs that differ only slightly from the one the algorithm was originally created for remains unclear.

The motivation of this study is to characterize the application areas as well

as strengths and weaknesses of the algorithms in order to provide a guideline that defines which algorithm is suited for which kind of experimental design. Properties of a good scanpath comparison metric include giving objective evidence with high sensitivity to scanpath differences. However, high sensitivity is hard to achieve with the overall high variability of eye movements. This automated exploratory comparison would be helpful in many areas of application since most eye tracking experiments boil down to a comparison of scanpaths, either between subject groups (e.g. patient versus control, expert versus novice) or experimental conditions (e.g. advertisement placement options). For example, the question whether subjects with visual deficites exhibit extraordinary scanpaths could be answered. Eventually existing compensatory eye and head movements could be proven. It is also interesting which differences have a positive or a negative influence on the performance at certain tasks and if a suboptimal scanpath could be influenced by gaze guidance in order to increase task performance.

In this thesis, an evaluation of different eye movement metrics of subjects performing a visual search task and viewing static stimuli is carried out as well as a comparison of eye movements of the same group of subjects while playing a video game simulating dynamic 3D scenes. The sensitivity of these metrics in detecting objective differences in task performance solely from the eye movement data is investigated. Compared to the tasks with static stimuli, interactive virtual environments are a huge step towards a more realistic environment and allow much more interaction and dynamics.

2 Theoretical background

In the following chapter, general aspects of gaze movements and behavior are explained. The technique of eye tracking and an overview of the term *scanpath similarity* is given. The state of the art in scanpath comparison algorithms and previous work are presented.

2.1 Eye movements

There are five different kinds of eye movements. (1) *Versions* (conjugated eye movements) are movements in the same direction (rectified). This means, both eyes are looking left or right, up or down at the same time. Versions can be splitted into rapid eye movements including (1a) *saccades* and slow eye movements including (1b) *smooth pursuits*. (3) *Vergences* (disconjugated eye movements) describe movements of both eyes in opposite directions (non-rectified) as in convergence when both eyes are adducted [16]. The (4) *optokinetic nystagmus* enables a stable picture on the retina caused by a change of moving targets as it is known by looking out of a moving train. The (5) *vestibulo-ocular reflex* compensates the (head-) movements done by the individual itself to stabilize gaze on a stationary target. The related neuronal structures are all located in different areas in the human brain where they are triggered and control oculomotor nerve (N III), trochlear nerve (N IV) and abducens nerve (N VI) [25].

In ophthalmological optics, the term *fixation* is used for target viewing in static outer space, respectively attention guided gaze movements to Area of Interests (AOIs). Objects are kept stable at the fovea centralis, the location on the retina with an extent of 20' (minutes of arc) where the sharpest vision is possible due to the high density of photoreceptors and neural encoding. Towards the periphery, the amount of cones decreases rapidly. Eye movements have to be performed in order to stabilize the image of the environment at the fovea centralis. The term central or foveal fixation is also used commonly. The condition for intact binocular vision is given when fixation lines cross at the gazed object point. The fovea represents the direction straight forward whereas peripheral regions on the retina have another

direction in space. Retinal correspondence exists if areas on both retinae have the same direction value. If bi-central fixation is given, the foveolae fulfill the retinal correspondence. This leads to a melting of the visual input of both eyes which is called fusion and allows binocular single visual impressions [10]. In general, minimal eye movements like microsaccades with an extension of 1-2' occur by fixating an object in order to avoid vision fading away by having stabilized a scene on the retina. When measuring fixations using an eye tracker, this effect adds up to the measurement noise. However, fixation defines the location where people attend to the best [12].

Saccades are fast eye movements. They are performed unintentionally as well as reflexively by fixating objects to repositioning the fovea. The duration of a saccade is about 10-100 ms [12] with an angular velocity of about 800 degrees per second [16]. The areas where they are triggerd are the superior colliculi, the paramedian pontine reticular formation (PPRF) and the rostral interstitial nucleus of medial longitudinal fasciculus (riMLF) [25]. When a saccade occurs, the processing of visual information is surpressed but cannot be recognized by the individual itself during that timespan. Saccades are used in order to switch between different fixation targets. Correcting saccades can occur when the saccade to fixate the next object is too short (undershoot) whereas an overshoot is often caused by a cerebellar lesion [16].

Smooth pursuits are eye movements caused by tracking a moving target. This can be voluntarily done by the individual if a moving target is present. The areas where they are triggered are the cortex, the pons and the vestibulocochlear nerve (N VIII) [25]. Smooth pursuits are performed much better along the horizontal direction than along the vertical meridian [10]. The angular speed is up to 30 degree per second. If the moving target is faster, corrective saccades are required to catch up.

2.2 Eye tracking

Eye tracking is a method of measuring eye movements. Insights into the cognitive processes at the bottom of the eye movements can often be concluded. The processes

of attention and pattern recognition can be indirectly detected with this method. There are several applications for eye tracking in medical diagnostics and academic research. Furthermore, it can be distinguished between *active applications* where the eye tracker is used for device control (e.g. control a computer cursor) and *passive application* where the eye movements are measured but have no active influence (e.g. improvement of web design) [29].

There are four different techniques commonly used for eye tracking: Electro-Oculo-Graphy (EOG), search coil, Video-Oculo-Graphy (VOG) and video-based pupil center detection [12]. In this thesis, only the latter is used and explained in detail.

Video-based eye tracking bases on corneal reflection. It is the least invasive method and offers high spatial and temporal accuracy. The devices can be wearable glasses on the subject's head or can be placed on the table in front of the subject. Today, computer-based methods delivering real-time images are commonly used. To separate head movements from eye rotations, it is often necessary to fix the head with a chin rest. With modern devices, it is possible to track the pupil center and the corneal reflection with infrared light (wavelength > 780 nm). A calibration is required in order to detect the pupil center. In most cases, subjects have to fixate predefined points whose images are represented on the cornea (Purkinje reflex). The first image being perpendicular planar is taken to approximate the viewers point of regard [12].

One of the first steps of data analysis is determining where the eye position changes from a stable fixation into a rapid saccade. This process is called *fixation and saccade filtering*. When subjects gaze at a specific area repeatedly, the region is called an Area of Interest (AOI). When a saccade moves the attentional focus from one AOI to the next, this is called a transition [4].

A *scanpath* describes the spatial and temporal sequences of fixations, saccades and smooth pursuits. Scanpaths can be visualized as shown in figure 1. According to Noton and Stark [26] [27], it is characteristic for each subject to make the same first eye movements by watching the same scene repeatedly. In general, the order

of eye movements differs between subjects and trials but they all have regions of interest (ROIs) in common.

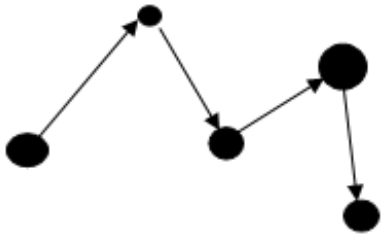

Figure 1: Visualization of a scanpath. Circles correspond to fixation positions with the circle radius representing the dwell time. The arrows connecting the fixations correspond to saccades.

The gold standard for comparing scanpaths in dynamic scenes is still an evaluation done manually by a human examiner. This leads to some severe drawbacks such as the subjectivity of the results, time consumption and difficulties in reproduction. Therefore it is necessary to automate the analysis of scanpaths. Automatic methods are commonly used for static scenes. Distances between sequent fixations and viewing durations can be evaluated. A differentiation of scanpaths has to be done with regard to their form, scaling, AOIs and fixation duration.

A *similarity measure* is a mathematical function that assigns a value of similarity to a pair of scanpaths. This value estimates the distance between the eye movement sequences and can be used for clustering and for the detection of differences [24]. Figure 2 visualizes that defining such a distance is non-trivial since differences can occur in many different properties of the scanpath:

1. *Random:* A random baseline can be used in order to quantify the statistical significance of a distance found between two scanpaths. Random pairs are not necessarily similar but essential to form a baseline against the other similarity values can be compared. Sampling fixation locations from a uniform distribution will most likely result in a different baseline than permuting the actual fixation locations of the scanpaths.

2. *Spatial offset:* Two sequences are translated with a spatial offset to each other. The effect can often be caused by inaccurracies of the eye tracker and a degrading calibration, resulting in a drift.

3. *Ordinal offset:* One sequence is shifted at the position p to $p+1$ in relation to the other sequence at every position. Such a misalignment occurs frequently since the variability of eye movements is high and the probability of a subject performing an additional fixation is high as well.

4. *Reversed:* Sequences contain the same position, however, the order is reversed. The relevance of this effect is highly task dependent: For an image viewing or reading task, this would speak for an entirely different pattern. For a search task, it would speak for an identical pattern but started from a different position.

5. *AOI border:* Sequences lie in adjacent AOIs and are interpreted differently. This is a general problem of equally sized spanned grid sectors.

6. *Local/ Global:* Along with a local cluster, local/ global sequences are constructed. Global shapes are always similar whereas the local clusters possibly differ.

7. *Scaled:* The degree of the covered stimulus space is different between the pairs of sequences. This can be caused by a deviation from the predefined observer distance.

8. *Duration:* The positions of the scanpaths are random, the durations of each (pair of) dot(s) is unmatched.

As it can be seen in figure 2, number two to eight represent a particular aspect of similarity. A requirement for a good scanpath similarity algorithm is to detect and weight each of these aspects of similiarity. Depending on which of the aspects they incorporate, how well they separate actual similarity from noise and how the individual factors are weighted in the final similarity measure, the algorithms are suited for a certain application area or not.

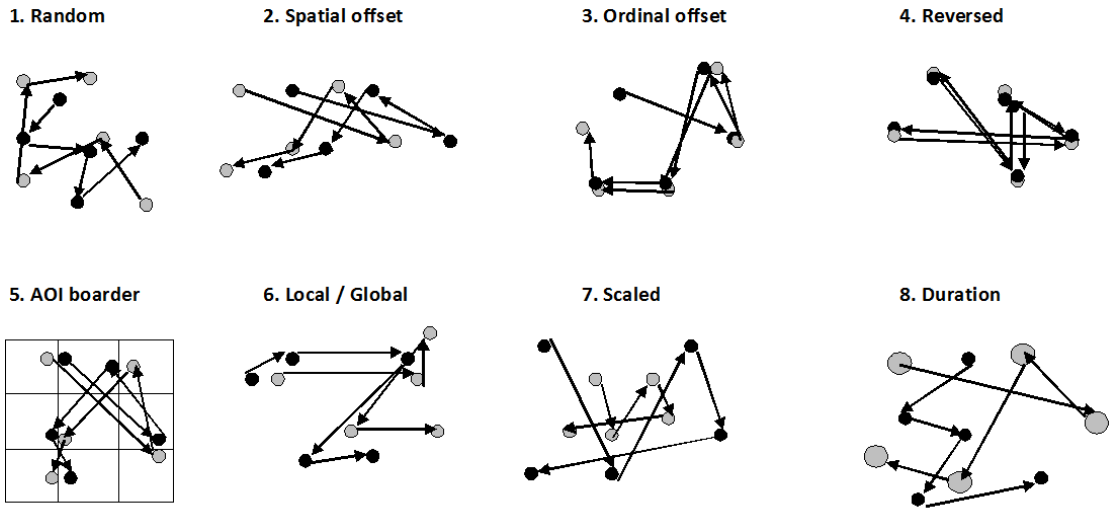

Figure 2: Similarity of two scanpaths (adapted from [9])

2.3 Scanpath Comparison Algorithms

The algorithms used in this study can be assigned to four categories: String alignment (section 2.3.1), Vector-based methods (section 2.3.2), Heat maps (section 2.3.3) and Pattern recognition (section 2.3.4). Individual implementations are assigned to these categories in table 1. Visualization techniques are helpful to evaluate eye tracking data and can be divided into point-based and AOI-based [4]. To evaluate fixations, there are several metrics like counting the number of fixations per minute, the fixation duration [ms] or the position. To evaluate sacccadic movements, the amplitude as well as the saccadic duration [ms] and the saccadic velocity [degrees per second] are frequently used metrics. Areas of Interest can be evaluated by transition count, by dwell time within the AOI [ms] or with the AOI hit which describes a fixation within or outside of an AOI.

2.3.1 String alignment

A string alignment algorithm is a transformation of spatial local information of sequences of fixations into sequences of letters. It was invented in 1965 by

Levenshtein. Using the so-called Levensthein distance [21] two scanpaths in letter-representation can be compared to each other. The similarity of two strings is thereby defined as the number of insertion, deletion and substitution operations required to make the strings equal. In a first step fixations are assigned to AOIs, either defined by stimulus properties or geometrical shapes. Then, the AOIs are divided geometrically or semantically. A letter is assigned to each fixation depending on which AOI the fixation falls into. This results in a representation of a scanpath as a word. Along with the string representation, two scanpaths can be compared within a Levensthein distance metric that delivers the minimal edit score between two strings. A double-letter representation was developed in order to enables the user to compare more than 26 AOIs. Technically, letters can also be numbered and the numbers can be compared to each other, solving the problem of limited AOIs. Disadvantages of this procedure occur if the areas are too large. Then, scanpaths seem to be more similar than they are and if fixations lie close to others but in different AOIs, a different result is gained. Furthermore, this method does not take the fixation duration into account. When choosing the AOIs geometrically, the grid overlaid on an image is defined independently of the image content. Especially in regions of interests, this may be too inaccurate whereas in other regions, it is too detailed. In order to avoid this effect, AOIs divided in half or multiple AOIs may be assigned to just one grid position. Figure 3 shows a scene from Mario Kart as an example for a string representation. The AOIs are shown as circles, fixations are marked as white and grey dots. Two scanpaths can be compared along their order of AOIs as described on the right side of the figure.

Algorithms of the category string alignment are *ScanMatch* [8], *iComp* [18] and *SubsMatch* [20] which are compared in present study.

2.3.2 Vector-based methods

At this juncture, scanpaths are represented as a sequence of mathematic vectors that describe single movements. To describe the scanpath alignment, the shortest path through a graph of the vector distances is searched to find the minimal distance between the simplified scanpaths. In this form, a sequence of vectors

Figure 3: Example for a string representation. AOIs are labeled by the black circles and fixations by dots (grey and white). The equal sign marks matching letters while the dotted line marks a mismatch that has to be resolved by a substitution. The example assigns a score of +1 to a match, -1 to a mismatch.

contains details about fixations and saccades, however, a presentation of semantic information is hard to do. An advantage of this method is that no AOIs are needed.

An example for a vector-based alorithm is called *MultiMatch* [9].

2.3.3 Heat maps

Attention or heat maps are time integrated visualization techniques that help to analyze eye tracking data based on a qualitative approach. Heat maps can be calculated as a superpositioning of Gaussian distribution density functions with their respective means at the fixation location. In general, areas that are fixated more often, are illustrated in warm colors (red) and areas that were less fixated, are presented in colder colors (blue). To compare several scanpaths, a comparison of attention maps can be done. A disadvantage of the illustration of heat maps is that important correlations between regions may not be recognized due to the separation of the stimuli into several regions and independent analysis of those. To get robust heat maps, many subjects are needed due to the high variability

among the subjects. For dynamic scenarios, heat maps are not suitable because they cannot take moving objects into account.

One algorithm of the category heat maps is called *iMap* [6].

2.3.4 Pattern recognition

Pattern recognition is the ability of clustering similarities, replications, regularities and principles of an amount of raw data. Today, there are three different approaches. The first one is called *syntactic pattern recognition*. Here, objects are described as a sequence of symbols to cluster objects of the same category with the same description. Probabilistic methods complete the syntactic method but they are not very common. The second one is called *statistical pattern recognition*. The aim of this approach is to determine the probability of an object belonging to one or another category. With the help of a feature vector, the values are clustered together. With a mathematically function, every feature vector is assigned to a special category. The third one is called *structural pattern recognition*. For the evaluation of eye tracking raw data, it is the most promising method and it combines the other methods mentioned above. Examples for algorithms of pattern recognition are *T-Patterns* [23] and the *Spatial Assembling Distance (SpADe)* [7]. SpADe is able to handle shifting and scaling in temporal and amplitude dimensions which makes it efficient for continuous pattern detection in streaming time series.

2.4 State of the art

In the following sections a short summary of previous work is given which is essential for this thesis. Studies were chosen due to their experimental set up and the implementation of algorithms that were used in the present study as well. An overview is given in table 1. The applicability of the algorithms according to their classification into evaluation tasks done by the authors of the algorithms is tested in present study.

Table 1: Overview of the scanpath comparison algorithms used in the present study. [Stimuli state s: static; d: dynamic; i: interactive]

Name	Method	AOI	State	Evaluation Task
MultiMatch	Vector + String	-	s, d	sequential looking
ScanMatch	String	+	s, d	sequential looking
FuncSim	Vector	-	i	real-world
iComp	String	+	s	visual search
SubsMatch	String	-	i	real-world driving
HMM	Probabilistic	+	s	free-viewing images
Eyenalysis	Vector + String	-	i	exploratory

With the help of post-processing techniques, results of the distance metrics can be made visible. The pairwise similarity of scanpaths is shown in a small distance between them. Vice versa, the more the distance between the scanpaths increases, the greater the dissimilarity between the scanpaths. The results can be visualized by multidimensional scaling. Therefore, scanpaths are plotted in 2D or 3D space or are visualized as dendograms.

2.4.1 Algorithm: MultiMatch

MultiMatch is a vector-based string-editing approach that allows a detailled comparison of scanpaths in multiple dimensions by saccades and fixations. The temporal sequence and spatial structure is dependent on the scanpaths under test. Saccades in the same direction and fixations that are close to others are simplified to just one eye movement in order to avoid large scanpath distances although the actual scanpath differences are small. Therefore, a threshold is needed and determined iteratively.

There are five dimensions. For each of them, differences can be calculated:

1. *vector*: Similarity in scanpath shape due to the vector difference of two linked saccade pairs.

2. *direction*: Angular distance between saccade vectors. A similarity of shape is measured when saccadic amplitudes are different.

3. *length*: Similarity in saccadic amplitude by taking the difference in length between the endpoints of saccade vectors into account.

4. *location*: Similarity in terms of the Euclidean distance by taking the difference in position between aligned fixations into account.

5. *duration*: Similarity in processing time and calculation through the difference in fixation duration between aligned fixations.

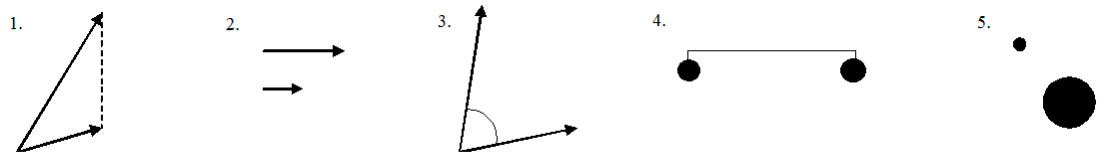

Figure 4: The five dimensions of MultiMatch (adapted from Dewhurst et al [9]).

The advantage of MultiMatch is the independence of AOIs which leads to a reduction of noise and a higher sensitivity to commonalities between scanpaths. Additionally, scanpaths can be compared even if they are of different length.

The MATLAB code for the MultiMatch algorithm can be directly aquired from the authors of the article. [9]

2.4.2 Algorithm: ScanMatch

ScanMatch is an algorithm based on the Needleman-Wunsch algorithm (string alignment) to compare saccadic eye movement sequences and describes how similar they are. Fixation location, time and the order of information are gained by binning the saccade sequence spatially and temporally and recoding these data into a sequence of letters.

To encode an eye tracking recording as a sequence of letters, eye movement data has to be filtered into saccades and fixations. Then, the stimulus is divided into AOIs. A letter is assigned to each area. Fixation duration is encoded by repeating

the letter linked to the current AOI proportionally to the fixation duration. The advantage of this step is that spatial, temporal and sequential information is contained in a string. Thereby, it is possible to classify identified AOIs and the spatial distance can be evaluated and compared. It is possible to compare more than 26 AOIs (length of the alphabet) with the help of the double-string procedure whereas the first letter is written lower-case, the second in upper-case and does not need to go up to Z. If the implementation would represent numbers instead of letters, an almost arbitrary number of AOIs could be encoded. According to Cristino et al [8], ScanMatch is a robust algorithm against noise in fixation sequences. It allows a more precise differentiation concerning the influence of eye movements on complex tasks. Furthermore, it allows gaps in the sequences and defines similarity scores between AOIs. In this way, two very similar AOIs can produce a high scanpath similarity while very dissimilar AOIs can be rated with a high distance.

ScanMatch is available at www.scanmatch.co.uk [8]

2.4.3 Algorithm: FuncSim

According to Foerster et al [14], FuncSim (functionally sequenced scanpath similarity method) is an algorithm suitable for the comparison of sequential tasks, especially in real-world tasks and everyday action. It allows the calculation of difference scores on multiple dimensions such as location and duration of fixations as well as length and direction of saccades. First of all, the location of every single fixation has to be standardized in the order of its delivering visual input during the task. Secondly, along with its functional unit, each fixation has to be labeled. Thereby, the mean fixation location can be calculated for each scanpath and the distance between the scanpaths can be compared. These distance values can be tested whether they are smaller across scanpaths than the distance values within a scanpath.

The precondition for using FuncSim is the standardization of the location of each fixation in so-called *world coordinates* according to the visual input. This step has to be done by manual frame-by-frame coding. Within five steps, the FuncSim

algorithm is completed. (1) The length and direction values are calculated, based on two successive fixations. (2) A random version of the scanpath is created as a baseline for statistical testing. (3) The scanpaths that should be compared, are aligned by FuncSim either in average concerning fixation location and duration as well as saccade length and direction or in relative duration which takes the fixation duration into account. (4) Difference values between the fixations are calculated among the Euclidean distance in pixel, cm or degree of visual angle. (5) Random baseline differences are calculated in all dimensions and are usable for statistical testing. If a task does not consist of an inherent sequence, functional units cannot be defined and therefore, fixations are aligned according to their temporal position in scanpaths which results in an underestimation of scanpath similarity. The applicability for FuncSim is dependent on the task structure and on the research question.

FuncSim is available at http://www.uni-bielefeld.de/psychologie/ ae/Ae01/Research/ FuncSim [14].

2.4.4 Algorithm: iComp

iComp is an objective automatic data-driven visualization tool for scanpath comparison in both sequence and location. It is based on string editing and defines AOIs automatically over fixations. With the two-step scanpath comparison method, fixations are clustered for spatial comparison of location and furthermore, the temporal sequences of fixations are ordered into strings prior to being compared. After that, parsing diagrams are gained from a matrix containing all similarity coefficients. Similar measurements seem to correlate. The current implementation of the present iComp algorithm (available at: http://andrewd.ces.clemson.edu/iComp/) uses a spatial Gaussian kernel. The advantages of iComp are the visualization of the whole recorded gaze data of all subjects for each image and the output into a parsing diagram. A color code illustrates the eye movements. Additionally, a velocity threshold can be used for fixation classification and saccade removal. [18]

2.4.5 Algorithm: SubsMatch

SubsMatch is an algorithm based on string representation, designed for dynamic, interactive real-world scenarios. To determine the similarity of scanpaths, the frequency of attention shifts is examined by searching for repeated patterns in visual scanpaths.

SubsMatch proceeds on three steps. (1) Scanpaths are transferred into string representation. The same amount of data is assigned to each letter, resulting in an optimal usage of the available encoding resolution. Spatial offsets caused by noise or incorrect calibration can be eliminated. (2) The string representation is divided into subsequences and listed in a hash table. (3) The difference in occurence frequency is calculated for each substring. Averaged over all patterns, this results in a distance value. To compare various groups of scanpaths, a weighting factor is used (longer scanpaths have higher weight). The advantage of SubsMatch as an algorithm usable for scanpath comparison in dynamic scenes is that there were no AOIs needed because the algorithm just examines the similarity of the process of exploration and not why scanpaths focus on the same objects at the same time.

The algorithm SubsMatch is freely available at http://www-ti.informatik.uni-tuebingen.de/~kueblert/SubsMatch1.0.zip [20].

2.4.6 Algorithm: Hidden Markov Model

The Hidden Markov Model (HMM) is a stochastic model and a special case of a dynamic Bayesian network. It can adapt to the individual viewing behavior of the subjects as well as to changes in scenes by traversing through several states. Each state has a different emission probability. The aim of the HMM is the evaluation of the hidden states on the basis of the emissions. In the context of the study, emissions are the eye movements, observable through eye tracking. States represent subject's mental states and unknown intentions that are hidden. E.g. during reading one could distinguish the states of proceeding in the text jumping from one line to the next line. Corresponding emissions of the states would be a small

saccade to the right side on the line and a large saccade to the bottom left in the following line. The probability of repeating the text proceeding step is quite high with only few transitions to the next line, while the probability that after jumping to the next line the state changes to text progression is quite low because jumping two lines is very unlikely. HMMs can be graphically represented as shown in figure 5.

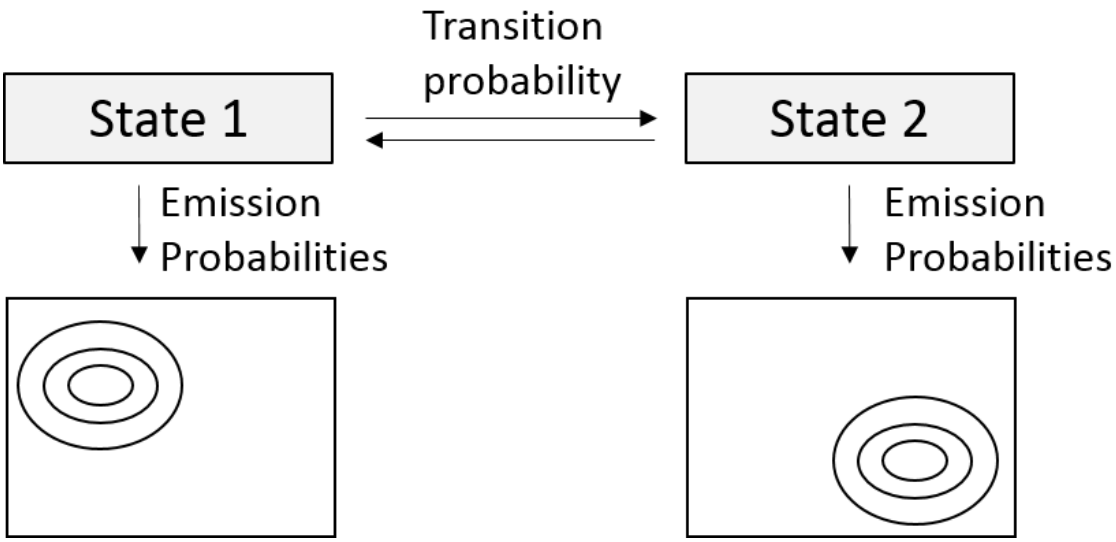

Figure 5: Simplification of the Hidden Markov Model.

The HMM has two characteristics (*Markov assumption*): Transitions from one state to another do only depend on the current state. Previous states are not considered. If a HMM has been trained on a certain scanpath, the probability of this model emitting another scanpath can easily be calculated as the product of transition and emission probabilities on its path through the graphical representation. This results in a similarity metric that is easy to use. For pretraining the HMM, the mean-shift clustering algorithm is used in order to find potentially interesting areas and thereby estimate the number of states required in order to represent the data. [35] [31]

In the application of this study, a HMM with Gaussian emissions is used, meaning that the probability of a fixation towards a certain location is represented by a gaussian probability density function over the expected location.

2.4.7 Algorithm: Eyenalysis

Eyenalysis represents eye movement sequences as a set of fixations that are defined through a number of dimensions such as location, duration or timestamp. Between two eye movement sequences, a mapping is constructed so that each point of a sequence is mapped onto at least one point from the other sequence. With a point-mapping, the Euclidean distance $d(p, q)$ between data points p and q of two different sequences can be calculated:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{1}$$

n: number of dimensions, p_i and q_i: i-th dimension of p and q

By collecting all point-mappings, a sequence-mapping is gained as well as results in a distance $D(S, T)$ from each sequence pair S, T.

$$D(S,T) = \frac{\sum_{i=1}^{n_S} d_S^i + \sum_{j=1}^{n_T} d_T^j}{max(n_S, n_T)} \tag{2}$$

n_S and n_T: length of the sequences S and T, d_s: distance between point i in S to the nearest neigbour in T and vice versa (double-mapping technique)

A limitation of distance ratings is that they are not absolutely but meaningful within a particular set of data. It is not possible to deteremine why two sequences are similar. By using artificial eye movement data, Mathôt et al [24] state that Eyenalysis is especially usable for exploratory analyses and less complex than other methods.

2.4.8 Conjunction Search Task

In the study *Visual search disorders beyond pure sensory failure in patients with acute homonymous visual field defects*, Machner et al [22] tested the hypothesis that

visual search patterns in subjects with homonymous visual field defects (HVFD) are caused by visual-sensory deficits. A special test was introduced to determine search behavior that is also used in the study on hand. Subjects had to solve three different search tasks: finding objects of same color [in CIE-coordinates: red (0,601/0,322); blue (0,218/0,579); green (0,147/0,070)], of same shape (triangle, square or circle) or conjuncted in color and shape. In each search task, an amount of zero, one, four or eight targets had to be detected among 40, 60 or 80 distractors which surrounded the targets. Nine subjects with HVFD, nine healthy subjects with technically simulated (virtual) HVFD and nine healthy persons without any visual defect were tested. In the following only the results relevant to this thesis are summarized. The relating scanpaths and strategies had circular, line-wise, column-wise, eight-shaped or chaotic patterns. There was no significant difference in the average rate of pictures with structured scanpaths between controls, patients or virtuals, neither in color nor in shape nor in conjunction search. Only in the control group, significant differences of structured scanpaths between color and shape were found. Concerning search duration, a significant influence could be observed for the categories *group* and *task*. In detail, patients with HVFD need longer for searching than the control and virtual HVFD group. Furthermore, subjects' search duration correlates significantly with the absolute number of fixations. In color search, less saccades were required than in shape search over all three subject groups. The authors concluded that objects with a defined color were significantly faster to find than the objects of a defined shape. The *factor of search duration (FSD)* was calculated for each subject as the relation between individual search duration and mean duration of all control subjects at the same stimulus.

This study was chosen because of the definition of the *Conjunction Search Task* that is used for the experimental evaluation of scanpath comparison metrics in the present study. With this standardized test, all parameters that could modulate scanpath shape and complexity are controllable and it provides different tasks with different difficulties due to the variety of distractors and targets. The focus lies on a regular search pattern shape due to the influence of the kind of task, number of targets and number of distractors.

2.4.9 Interactive Virtual Environments in Eye Tracking Research

In their study *Applying computational tools to predict gaze direction in interactive visual environments*, Peters and Itti [28] compared fully automated computational heuristics for gaze prediction in order to identify locations at which people look at in dynamic scenes. The aim of the study was to predict human gaze behavior with naturalistic dynamic stimuli in an interactive task. Subjects had to play five contemporary video games at the nintendo GameCube with simulated three-dimensional environments to provide the visual input and the interactive task. A comparison of the recorded gaze behavior with the predictions of nine different bottom-up computational heuristics based on low-level image features (color, intensity and motion) and saliency was done. All nine heuristics scored better at predicting observers gaze movements during exploration games than during racing games. There was less variability in heuristic performance across subjects than across games. Concerning both kinds of games, the best predictor was the heuristic motion, followed by flicker and full saliency. In racing games, the heuristic color scored the best. The authors explained this by the obstacle avoidance of players when they navigate through a predetermined course under time pressure.

The study of Peters and Itti is pioneering due to the use of interactive visual envirmonments in eye tracking research. For the purpose of this study, Mario Kart seems to be an adequate choice for gaze comparison since the environment conditions are relatively stable. Furthermore, it delivers a free and objective performance measure, the race completion time. With this fast racing game, it is possible to separate the eye movements for experienced players and novices that will very likely be different, just as in real-world driving. This experiment is a huge step towards a more realistic environment because it allows much more interaction and dynamics.

3 Methods

In this section, the aquisition of subjects (section 3.1), the used questionnaire (section 3.2), the experimental set up (section 3.3) and the procedure for the Conjunction Search Task (section 3.3.1) as well as for Mario Kart (section 3.3.2) are described in detail.

3.1 Subjects

The ophthalmological inclusion criteria for participants were a maximum of ±6.00 dpt spherical and a maximum of ±2.00 dpt cylindrical ametropia with a visual acuity of at least 20/20 and without media opacities. Ophthalmological exclusion criteria involved diabetic retinopathy, infections, amblyopia, defective color vision, strabism, nystagmus and macular deseases. General excluding criterias were epilepsy, neurological deseases and a known intake of medication affecting the visual field and a differential luminance sensitivity. All subjects were recruited within Aalen University and at least 18 years old. They were informed about the study and gave written consent.

3.2 Questionnaire

The questionnaire consisted of ten questions regarding gender, date of birth and refractive correction (spectacles, contact lenses or no correction worn during testing). Furthermore, the participants were asked to state their experience with smartphones, tablets, computer and specifically racing games. As a quantitative measure of gaming experience it was asked for the daily playing time. The questionnaire helps to differenciate between gaming experts and novices, more precisely, people who often play games on a smartphone, tablet or gaming console. A major focus will be on the presumption that subjects experienced in video gaming explore and perceive the game contents differently by distributing their fixations to different game elements. The difference in the duration and attention areas is distributed to the virtual driver and to the surroundings.

3.3 Experimental Set up

A 24" monitor (Fujitsu Display B24T-7 LED) with 1920 x 1080 pixel was used to show stimulus screens. Its border was surrounded with black paper to reduce reflexes from the white monitor border and to minimize distraction from the status LED and the surrounding. A chin rest, adjustable in height, was used to keep the head in position to measure eye movements with high accuracy. In both examinations, the eye tracker *The Eye Tribe* (The Eye Tribe ApS, Copenhagen/ Denmark) was placed in front of the subject within a distance of 45-55 cm. This eye tracker works with a high resolution sensor and infrared illumination. It is possible to record gaze location on the screen in real-time. Technical details are shown in table 2 [30]. With *The Eye Tribe* binocular gaze data was created. A DELL Precision M4700 notebook was used on which the eye tracking software (The Eye Tribe Server 0.9.36) was installed and started for calibration. MATLAB (version R2013a) was used to run the Conjunction Search Task. Mario Kart was run directly from a Nintendo Wii. The video signal was captured by the notebook and recorded, then forwarded to the stimulus screen.

Table 2: Technical details of The Eye Tribe.

Parameter	Technical details
sampling rate	30 Hz
accuracy	0.5° - 1.0°
calibration	9 points
operating range	45 - 75 cm
data output	binocular gaze data
dimensions (W/H/D)	20.0 x 1.9 x 1.9 cm

Screen-gaze coordinates were represented by a pair of (x, y) coordinates. A 9-point calibration was done. Calibration quality was defined by a maximum of five stars. For this study on hand, a minimum of four stars was required, otherwise the calibration was repeated. According to the manufacturer, contact lenses and spectacles can be worn without a huge influence on measurement quality. The illumination was 250 lux in the examination room with no influence of sunlight on

the device. The recorded data was preprocessed in MATLAB. Scanpath comparison scripts were used as provided by the respective authors. Where parameters were available, grid search was employed in order to find optimal settings. For statistical evaluation, Excel (Microsoft Office 2007) and GNU R (R version 3.1.1 (2014-07-10)) were used.

3.3.1 Procedure for the Conjunction Search Task

Visual attention is driven by bottom-up features of stimuli where an interesting feature such as bright red color pops out. The reaction to this input is highly dependent from individual experiences, emotions and current intentions of a person. This is called the top-down effect. The Conjunction Search Task according to the study of Machner et al [28] combines both models since subjects were shown a screen with either three geometrical symbols with same color or shape or one symbol combining color and shape. The latter is called conjunction. The symbols were triangles, squares and circles colored in CIE-coordinates red [0.601/0.322], blue [0.218/0.5790] and green [0.417/0.070] on a black background. The distance between the symbols was at least 1°. On the screen, there we either one, four or eight search objects (targets) between 40, 60 or 80 stimuli. Subjects were not informed about the amount of passive static targets that they have to count but were instructed to search for the predefined targets. The viewing distance was defined by 60 cm from chin rest to monitor with a visual field of 47° x 29°.

After a calibration with The Eye Tribe software, a second calibration written in MATLAB was done which was evaluated for each task. After performing ten tasks and after the last task a re-calibration was done in order to detect deteriorated calibration quality due to drift and head movements. 35 search tasks were presented in one block. There was no time limit per task. For counting, subjects had to click the left mouse button (wireless mobile mouse, Microsoft®, Model 1383). The right mouse button finished a trial and proceeded to the next one. The difficulty of the test can be modified by varying the *kind of tasks*, *number of targets* and *number of stimuli*. If the change in difficulty is strong enough, an associated change

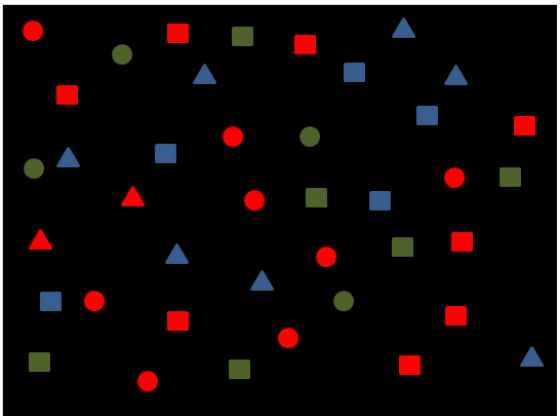

Figure 6: Example for a Conjunction Search Task screen with 40 stimuli.

in scan patterns is assumed. Therefore, performing time per task, number of fixations and errors were compared by using the Shapiro-Wilk test for normal distribution. In case of non-parametric independant values, a Kruskal-Wallis test was done that delivers only a global significance. In order to find statistically significant differences between the groups, a Wilcoxon rank-sum test has to be done. The 5% α-level for significance was corrected for multiple testing using the Bonferroni-Holm method. After the statistical evaluation, a comparison of seven different scanpath metrics was done according to table 1, described in section 2.4. On the gained similarity matrix, a Kolmogorov-Smirnov Test was used in order to compare distances between and within the groups. Due to the high amount of tests used, results had to be false discovery rate corrected with the method of Benjamini & Hochberg.

3.3.2 Procedure for Mario Kart

Mario Kart is a cartooned racing game invented in the 1990s by the Japanese company Nintendo. The game was introduced in 2008 for Nintendo Wii [34]. There are various other versions of the game for a variety of other consoles that share central elements. Mario Kart was chosen because it is very popular and many students are common with it from early childhood on. Furthermore, it is well suited since it delivers an objective performance measure (track completion time)

that can be used to form groups of expert and novice players. It can be expected that gaze behavior differs between these groups. Objects relevant to the game and distracting decorational elements as well the meta-views (minimap, position) are likely sources of diverging attention distribution. The speed with which relevant objects are scanned and orientation on the road may be different. Due to the relatively stable environmental conditions, the test delivers good data quality and is relatively low-cost.

During the interactive race, subjects watch their chosen driver from behind. On the track, there are several items (red boxes with questionmarks) which help the driver to gain advantages by having mushrooms to get faster, bananas on which other drivers slip and so on. On every track, three laps have to be completed, the fastest driver wins the race. In the present study, two tracks were chosen. In contrast to the search task mentioned in section 3.3.1, subjects actively influence their stimulus (Mario) with their interaction, so the stimulus becomes dynamic, meaning, few seconds after the race started, the screen shown to each player differs. In this study, the Nintendo Wii was used with a nun-chuck attached to the remote control to minimize head and body movements during testing while having the head rested on the chin rest. Viewing distance was set to 110 cm from chin rest to monitor. The visual field of the whole monitor was 27° x 16°. For technical reasons, the Mario Kart test screen did not extend to the whole monitor size, instead, the visual field was 18° x 12°. Every subject was introduced to chose the Grand Prix version in which 11 other computer-controlled racers compete against the subject. Mario was chosen as driver by 100 ccm (motorcycle) for the first cup called *mushroom*. The first track was *Luigis Cup* to get used to the controller. It was recorded after a 9-point calibration and stopped after finishing lap three. After a short break to store the data, the second track called *Moo moo* has to be done. Therefore, a re-calibration of the eye tracker was perfomed. Subjects were not allowed to speak or cheer during the race. Tricks like jumping over obstacles for speeding up was forbidden and also chewing gum was banned to minimize head movements.

Algorithms as described in section 2.4 and listed in table 1 were used for scanpath comparison.

3.4 Statistics

In this section, a short summary over all statistical tests used in the present study is given.

Shapiro-Wilk test
The Shapiro-Wilk test is the most popular procedure for testing normal distribution. If the null hypothesis (H_0) has to be rejected ($p \leq 0.05$), no normal distribution is given, meaning that the data does not come from a normally distributed population. For ($p > 0.05$), the assumption of a normal distribution cannot be rejected. It is necessary to display at least one figure (e.g. boxplot) to show wether normal distribution is given for further testing. [17]

Pearson's Chi-square test
The Pearson's Chi-square test is used for testing independent samples for nominal scaled values with different sample size. Within a contingency table, the squared distance between observed and expected cell frequencies is calculated in relation to the expected frequencies. The test value is X^2. If H_0 is accepted or has to be rejected for a predefined α-value, $X^2_{critical}$ can be read out of a table by considering the degree of freedom (df). H_0 is assumed if the calculated $X^2 < X^2_{critical}$, whereas H_0 is rejected when $X^2 > X^2_{critical}$. [32]

Kruskal-Wallis test
To compare more than two independant non-parametric ordinal-scaled samples, the Kruskal-Wallis test can be done. If each sample has the same distribution function, H_0 is assumed. With this global test, it is not possible to point out between which samples a statistical significance occurs. It can only be proven if a difference occurs between all samples under test. Due to the assumption of a multiple significance level (testing of several null hypothesis simultaneously), at least one of the H_0 has to be rejected. Even if each H_0 calculated is correct, a post-hoc test like the Wilcoxon rank-sum test has to be done. [17]

Wilcoxon rank-sum test

The Wilcoxon rank -sum test is the post-hoc test of the Kruskal-Wallis test. It is tested between which factor levels a significance occurs by comparing all possible pairs of groups with regard to the Bonferroni-Holm-correction for multiple testing. The α-level of 5% must be maintained. If $p \leq 0.05$ the null hypothesis has to be rejected. This leads to a significant difference between that pair of samples. [17]

Kolmogorov-Smirnov test

With the Kolmogorov-Smirnov test, the user is able to compare the distribution of two independent samples of unknown variance and unknown distribution. It is a non-parametric test that quantifies a distance between the empirical distribution function of two samples. With a given α-level of 0.05, the null hypothesis means that both samples originate from the same population and there is no statistical significance ($p \geq 0.05$). The alternative hypothesis (H_1) means that the samples originate from different distributed populations which leads to significant differences between the samples ($p < 0.05$). [32]

FDR correction method of Benjamini & Hochberg

The false discovery rate (FDR) correction is a post-hoc test that is necessary if a high amount of tests is used simultaneously. The FDR allows to tolerate a certain number of tests to be incorrectly discovered. In detail, it describes the proportion of incorrect (false positive) rejections among all discoveries of H_0. The FDR method of Benjamini & Hochberg allows a gradual correction of the significance level. This is necessary due to the increase of Type I error (alpha-error-cumulation). The given level of significance ($\alpha = 0.05$) has to be divided by the amount of all tests (m_i) used. Then, the first and lowest P-value results. The last P-value is calculated by 0.05 and has to be divided by $(m_i - i - 1)$. With the assumption that all P-values are significant, the FDR correction is calculated from the first to the last value and evaluates the significance of them having regard to the significance of each single P-value. By exceeding the last significance barrier, the FDR correction stops and a statistical significance is below limit of detection. [37]

4 Results

In the following section, results of the Conjunction Search Task (section 4.1) and Mario Kart (section 4.2) are shown with respect to task performance and scanpath similarity. These measures were compared to each other statistically in order to establish a statement of applicability of the scanpath similarity algorithm with respect to the task.

According to the questionnaire, the subject's age range was from 22 to 43 years (average 26.5 ± 4.05). There were 10 female and 11 male participants in this study. Nine of 21 subjects (43%) wore glasses. Additionally, six persons (29%) alternated between glasses and contact lenses whereas six persons (29%) did not need any refractive correction. One person performed the examination without any correction even though he normally alternates between contact lenses and glasses.

4.1 Conjunction Search Task

In a first step, the measurement quality of the eye tracker had to be assessed for all trials ($21 subjects \times 35 tasks = 735$). This guaranteed results for fixation number and task completion time as accurate as possible. The boxplot of measurement quality during the experiment is shown in figure 7. Due to blinks, illumination conditions and other measurement errors, the detection of the pupil may have failed and resulted in an invalid data point.

In the box, the median is shown as a thick black line. The edges of the box mark the 25^{th} and 75^{th} percentiles. The whiskers above and below the box extend to the most extreme point that is not considered as an outlier and is plotted individually. An outlier is a value that is larger than $q_3 + w^{(q_3 - q_1)}$ or smaller than $q_1 - w^{(q_3 - q_1)}$, where q_1 and q_3 are the 25^{th} and 75^{th} percentiles and w is 1.5, corresponding to ± 2.7 sigma (Euler's number) and 99.3 % coverage if the data are normally distributed.

As it can be seen in figure 7, the asymmetric quality distribution has its median at 98% valid data points, the 25^{th} percentile is at 93%. The whiskers extend down to 83%. This value was chosen as a quality measurement threshold, identifying 51 out of 735 measurements as minor quality runs and removing them from the evaluation for number of fixation and task completion time.

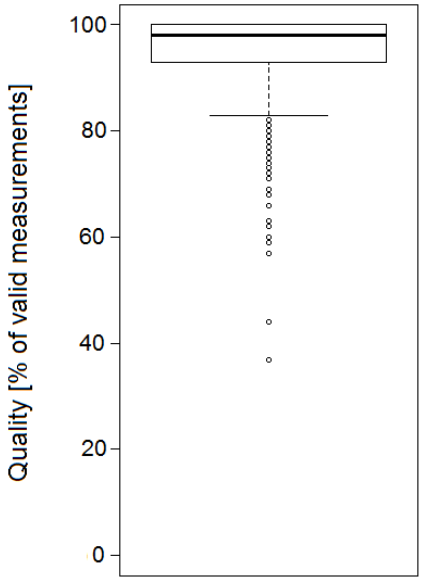

Figure 7: Boxplot of the measurement quality for the Conjunction Search Task (percentage of valid gaze measurements) in all 735 trials.

The Conjunction Search Task trials can be subdivided into three kinds of categories: *Kind of Task, Number of Targets* and *Number of Stimuli*. Within these categories, it was evaluated if there are statistical differences concerning the objective task performance measures of number of errors, task completion time as well as a very simple scanpath measure, the number of fixations. It can be interpreted as the complexity of visual search. These three categories are described in detail in the section 3.3.1 and their results are reported in the following.

4.1.1 Kind of Task

Within the category *Kind of Task*, the Conjunction Search Task was divided into the factors color (red, blue and green), shape (triangles, squares and circles) and conjunction (combining shape and color).

Evaluation of the number of errors

For the evaluation of error frequency meaning the report of a an incorrect number of search objects, the quality of measurement had no influence. It was just relevant if an error occurred and not by how much the reported number was off. Therefore, all 735 values were considered. The Pearson's Chi-square test was used for testing independent samples for nominal scaled values with different sample size. The null hypothesis stated no significant difference in the failure frequency between the tasks conjunction, color and shape.

Table 3: Number of tasks and associated error counts for the category *Kind of Task*.

	error	no error	Total
Conjunction	10 (3.7%)	263 (96.3%)	273
Color	8 (3.8%)	202 (96.2%)	210
Shape	30 (11.9%)	222 (88.1%)	252
Total	48 (6.5%)	687 (95.5%)	735

A comparison of all three tasks showed that the shape task yielded an increased error rate. A statistical significance occurred at the 5% level ($X^2 = 18.1483, df = 2, p = 0.0001$. When comparing conjunction and color, the null hypothesis could not be rejected ($X^2 = 0.0071, df = 1, p = 0.9328$). The alternative hypothesis was valid for the tasks conjunction and shape ($X^2 = 12.6462, df = 1, p = 0.0004$) and color and shape ($X^2 = 9.9442, df = 1, p = 0.0016$). This means that the tasks color and conjunction are more similar in their error frequency and should therefore be more difficult to distinguish from the eye tracking measures than combined groups with the factor shape that has a higher error frequency.

Evaluation of the number of fixations

To evaluate whether the number of fixations was normally distributed, the Shapiro-Wilk test was done. Normal distribution had to be rejected for each task (color: $W = 0.7044, p < 2.2e-16$; conjunction: $= 0.745, p < 2.2e-16$; shape: $W = 0.7641, p < 2.2e-16$). The distribution of the number of fixations is graphically shown in the violin plots in figure 8.

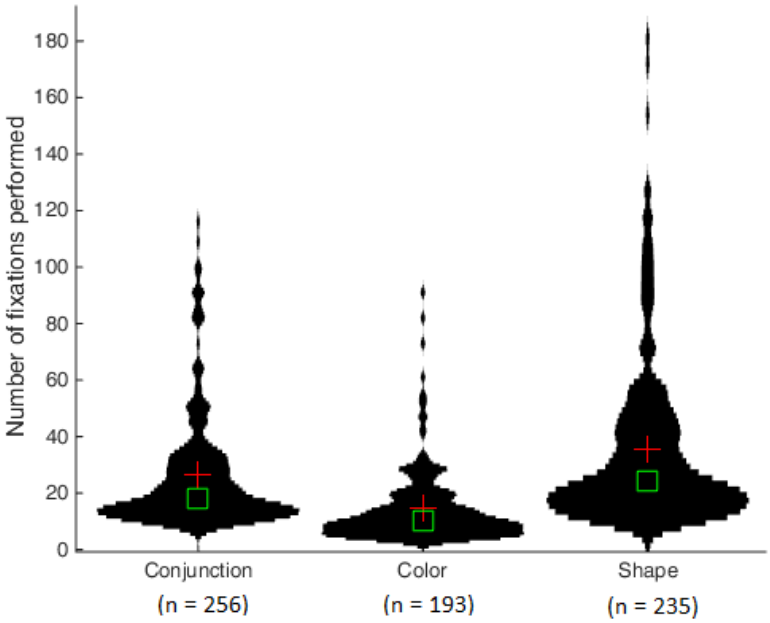

Figure 8: Violin plots of the number of fixations performed during the tasks separated by the category *Kind of Task*. A violin plot uses the density function due to the combination of a box plot and a kernel density plot. The geometrical mean is shown as a red cross and the median lies within the green square.

In figure 8, the median for color lay at 10, for conjunction at 18 and for shape at 24 fixations in each task. The number of fixations varied the most for the task shape. The geometrical mean could be neglected due to the asymmetric distribution of the values.

To compare more than two independant non-parametric samples, the Kruskal-Wallis test was done at the 5% level. For the comparison between all three categories,

the null hypothesis had to be rejected (Kruskal-Wallis $X^2 = 110.9461, df = 81, p = 0.0152$). A Wilcoxon rank-sum test was performed in order to test the effect of these groups statistically. The number of fixations was found to be significantly different between all groups (all $p < 0.00005$, corrected for multiple testing by Bonferroni-Holm method). This means, acoording to the number of fixations made within a task, it can be differenciated between the groups conjunction, color and shape. The number of fixations in group color is lower than in the other groups due to the considerable *pop-out-effect* [33] for the factor color.

Evaluation of the task completion time

To evaluate whether the task completion time was normally distributed, a Shapiro-Wilk test was done and had to be rejected (color: $W = 0.6755, p < 2.2e-16$; conjunction: $W = 0.8023, p < 2.2e-16$; shape: $W = 0.6852, p < 2.2e-16$). Violin plots in figure 9 show the asymmetric distribution graphically.

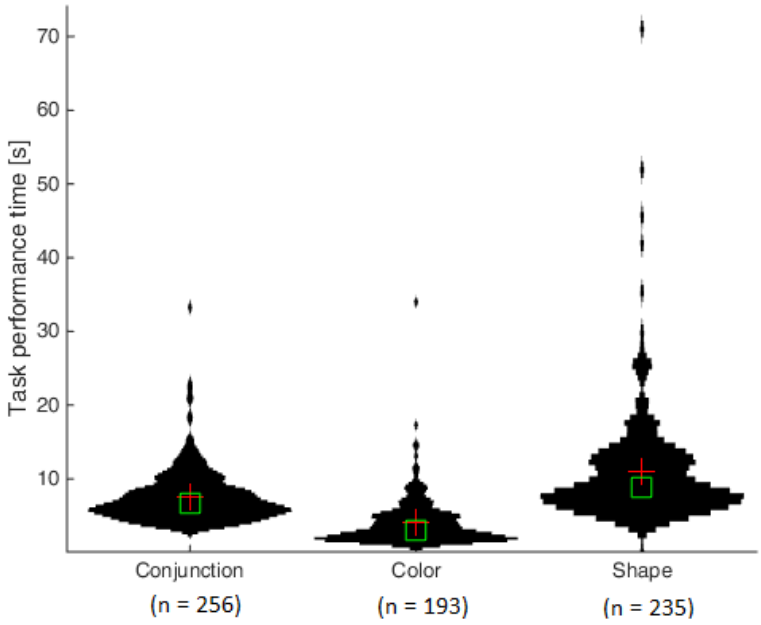

Figure 9: Violin plots of the task performance time [s] performed during the trials separated by the the category *Kind of Task*.

The task completion time varied most for task shape and thresholds ranged from

0 (no answer was given) up to 20.5 seconds. The median was higher for the task shape with 8.8 s than for color with 3.1 s and conjunction with a value of 6.6 s per task.

With the Wilcoxon rank-sum test for independent non-parametric values regarding the Bonferroni-Holm correction coefficient, the null hypothesis had to be rejected for all three tasks (conjunction vs. color: $p < 2e-16$; conjunction vs. shape: $p < 7e-13$; color vs. shape: $p < 2e-16$). Statistically, there was a significant difference in task completion time between the groups.

4.1.2 Number of Targets

According to the statistical evaluation done above, the section *Number of Targets* of the Conjunction Search Task was divided into the number of one, four or eight targets.

Evaluation of the number of errors

For the evaluation of error frequency meaning the report of an incorrect number of search objects, all 735 measured values were taken into account. It was only important whether an error occurred and not how many errors were made within a trial. The Pearson's Chi-square test was used for testing independent samples for nominal scaled values with different sample size. When comparing all three tasks, a statistical significance occurred at the 5% level ($X^2 = 21.8224, df = 2, p = 1.825e-05$). The null hypothesis could not be rejected for the comparison of the groups of one target with four targets ($X^2 = 0.093, df = 1, p = 0.7604$). For the groups of one and eight targets, the alternative hypothesis was valid ($X^2 = 16.7658, df = 1, p = 4.229e-05$) as well as for the groups of four and eight targets ($X^2 = 11.7894, df = 1, p = 0.0006$). This means that the groups of one and four targets are more similar in their error frequency and should therefore be more difficult to distinguish than the other groups with more than four targets. Table 4 shows the number of errors being made during the tasks.

Table 4: Number of tasks and associated error counts for the category *Number of Targets*.

	error	no error	Total
1 Target	12 (3.8%)	303 (96.2%)	315
4 Targets	10 (4.3%)	221 (95.7%)	231
8 Targets	26 (13.8%)	163 (6.2%)	189
Total	48 (6.5%)	687 (93.5%)	735

Evaluation of the number of fixations

Along with the Shapiro-Wilk test, the assumption of a normally distributed population of the number of fixations had to be rejected for all three groups (1 target: $W = 0.7483, p < 2.2e-16$; 4 targets: $W = 0.738, p < 2.2e-16$; 8 targets: $W = 0.7124, p = 3.617e-16$). This asymmetric function can be seen in figure 10.

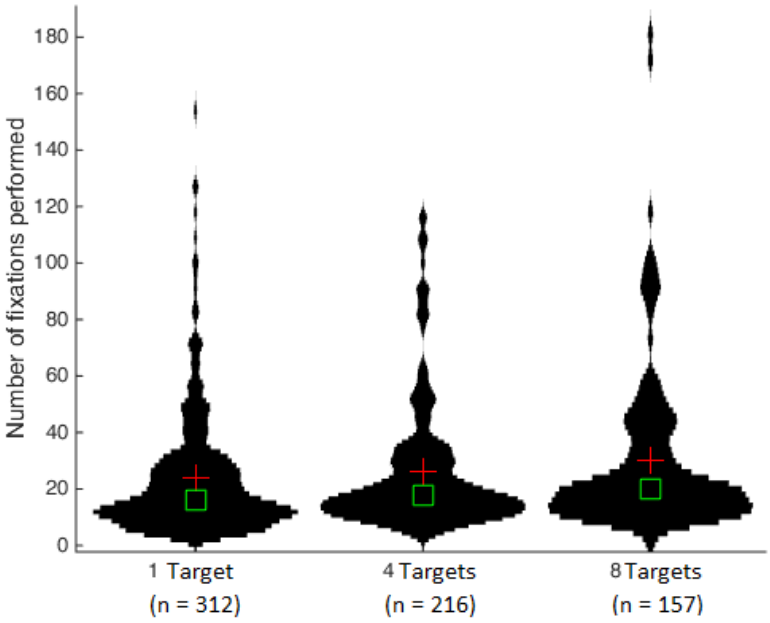

Figure 10: Violin plots of the number of fixations performed during the trials separated by the category *Number of Targets*.

The median for the group of one target is 16 errors, 18 errors were made in the

group of four targets and 20 errors in the eight targets group. It seems that the distribution for four targets is similar to the distribution of one target as well as to the distribution of eight targets and lie in between.

With the Kruskal-Wallis test on the 5% level, a statistical significance occurred by comparing all three groups (Kruskal-Wallis $X^2 = 126.8065, df = 81, p = 0.0009$). A Wilcoxon rank-sum test, corrected for multiple testing by the Bonferroni-Holm method, was performed in order to test the effect of these groups statistically. The number of fixations was found to be significantly different between the groups of one and eight targets ($p = 0.011$) but neither between one and four targets ($p = 0.147$) nor between four and eight targets ($p = 0.151$). This leads to the assumption that the amount of fixations is more similar, the more similar the tasks are in relation to their set up.

Evaluation of the task completion time

With the Shapiro-Wilk test, the assumption of a normally distributed population had to be rejected for all three tasks (1 target: $W = 0.7406, p < 2.2e-16$; 4 targets: $W = 0.7523, p < 2.2e-16$; 8 targets: $W = 0.639, p < 2.2e-16$). This asymmetric function, shown as violin plots, can be seen in figure 11. The median for the task completion time of one target was 5.9 *s* , for four targets 6.7 *s* and for eight targets 8.7 *s*. The geometrical mean was very similar to the respective median although there was no normal distribution given. Furthermore, the Wilcoxon rank-sum test delivered statistical differences between all three groups (1 vs. 4 targets: $p = 0.0003$; 1 vs. 8 targets: $p = 1.4e-11$; 4 vs. 8 targets: $p = 6.8e-5$, Bonferroni-Holm corrected). Depending on the number of targets, the performing time varies significantly.

4.1.3 Number of Stimuli

Within the section *Number of Stimuli*, the Conjunction Search Task was classified into 40, 60 and 80 stimuli.

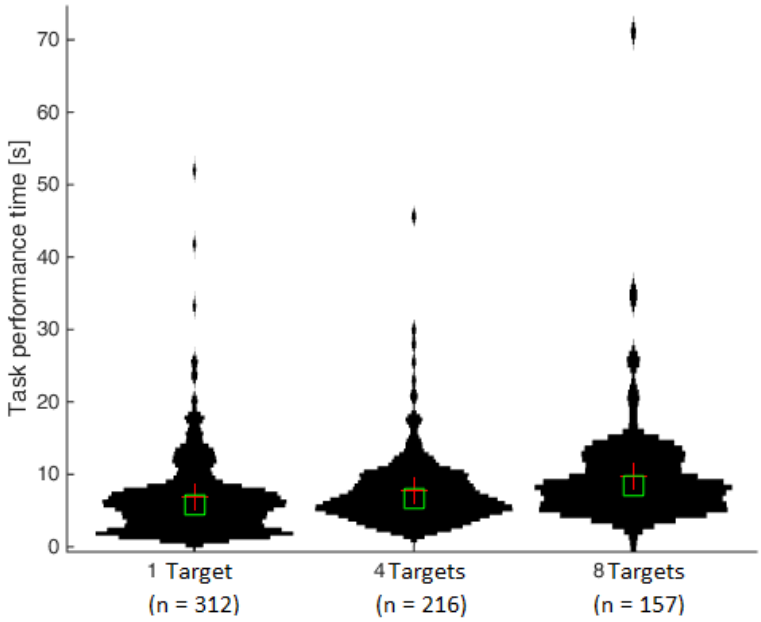

Figure 11: Violin plot of the task performance time [s] performed during the trials separated by the category *Number of Targets*.

Evaluation of the number of errors

For the evaluation of error frequency, meaning the report of a wrong number of search objects, all 735 values were taken into account. Table 5 shows the number of errors for the category number of stimuli. The Pearson's Chi-square test was used for testing normal distribution between the tasks. On the 5% level, the null hypothesis could not be rejected between all three tasks ($X^2 = 1.8772, df = 2, p = 0.3912$). Regarding the number of errors, there was no statistically significance given between 40, 60 and 80 stimuli and the number of errors made was independent from the number of presented stimuli.

Evaluation of the number of fixations

The Shapiro-Wilk test was used to evaluate if the number of fixations was normally distributed in the tasks of 40, 60 and 80 stimuli. There was no normal distribution given (40 stimuli: $W = 0.6824, p < 2.2e-16$; 60 stimuli: $W = 0.7547, p < 2.2e-16$; 80 stimuli: $W = 0.7587, p < 2.2e-16$). This asymmetric distribution can be seen in figure 12, shown as violin plots.

Table 5: Number of tasks and associated error counts for the category *Number of Stimuli*.

	error	no error	Total
40 Stimuli	16 (6.3%)	236 (93.7%)	252
60 Stimuli	13 (5.2%)	239 (94.8%)	252
80 Stimuli	19 (8.2%)	212 (91.8%)	231
Total	48 (6.5%)	687 (93.5%)	735

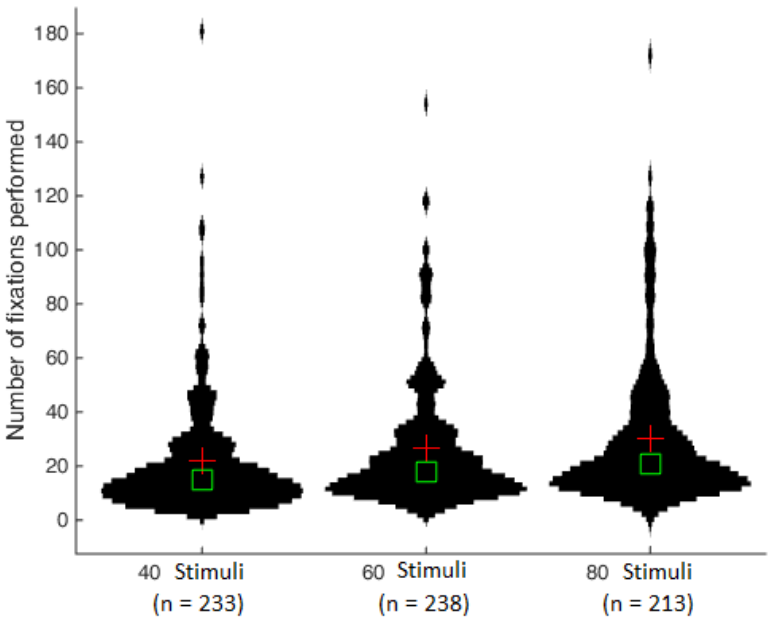

Figure 12: Violin plots of the number of fixations performed during the trials separated by the category *Number of Stimuli*.

The median lay at 15 fixations for 40 stimuli, 18 fixations for 60 stimuli and 21 fixations for 80 stimuli.

With the Kruskal-Wallis test on the 5% level, the null hypothesis had to be rejected for all tasks (Kruskal-Wallis $X^2 = 107.7964, df = 81, p = 0.025$). To specify, a Wilcoxon rank-sum test regarding Bonferroni-Holm correction was done. In all three cases, the null hypothesis had to be rejected (Stimuli 40 vs. 60: $p = 0.0022$; Stimuli 40 vs. 80: $p = 1.5e-06$; Stimuli 60 vs. 80: $p = 0.0474$). That means, there is a significant difference in the amount of fixations depending on the amount of stimuli shown. The less stimuli are presented, the less fixations are necessary to fulfill the search task.

Evaluation of the task completion time

The Shapiro-Wilk test was used to evaluate if the the task completion time [s] was normally distributed in the tasks of 40, 60 and 80 stimuli.

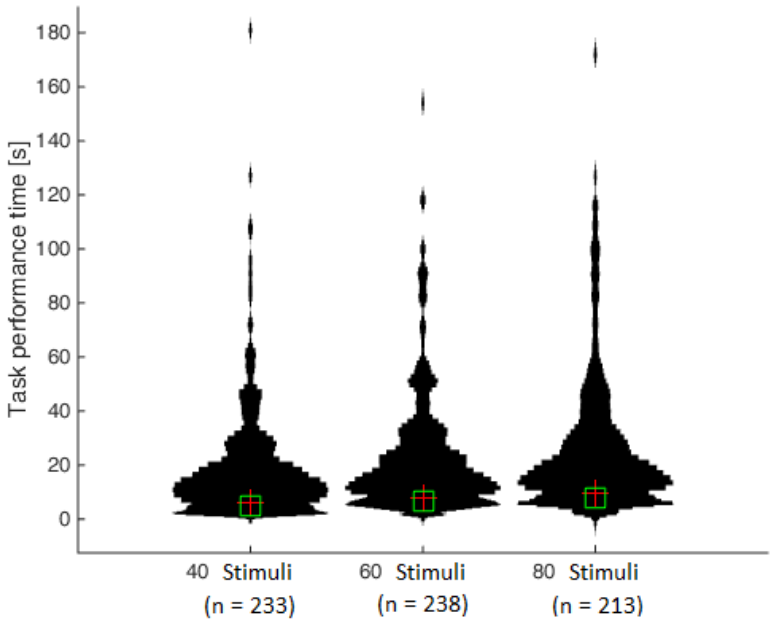

Figure 13: Violin plots of task performance time performed during the trials separated by the category *Number of Stimuli*.

The null hypothesis had to be rejected in all three cases (40 stimuli: $W = 0.8049, p =$

2.246e−16; 60 stimuli: $W = 0.7957, p < 2.2e{-}16$; 80 stimuli: $W = 0.6767, p < 2.2e{-}16$). The asymmetric function can be seen in the violin plot in figure 13. The median for task completion time was 5.1 s for 40 stimuli, 6.7 s for 60 stimuli and 7.9 s for 80 stimuli.

With the Wilcoxon rank-sum test regarding Bonferroni-Holm correction, a statistical significance was given between all kinds of stimuli (40 vs. 60 stimuli: $p = 2.1e-6$; 40 vs. 80 stimuli: $p = 8.8e-14$; 60 vs. 80 stimuli: $p = 0,0002$). This leads to a significant difference in the task performance time depending on the amount of stimuli.

To summarize the findings of the previous chapter, figure 14 displays the significant factors in experimental design that had an influence on task performance. The symbole * stands for a statistical significance between the two classes linked together. It is shown for the categories *Kind of Task*, *Number of Targets* and *Number of Stimuli*.

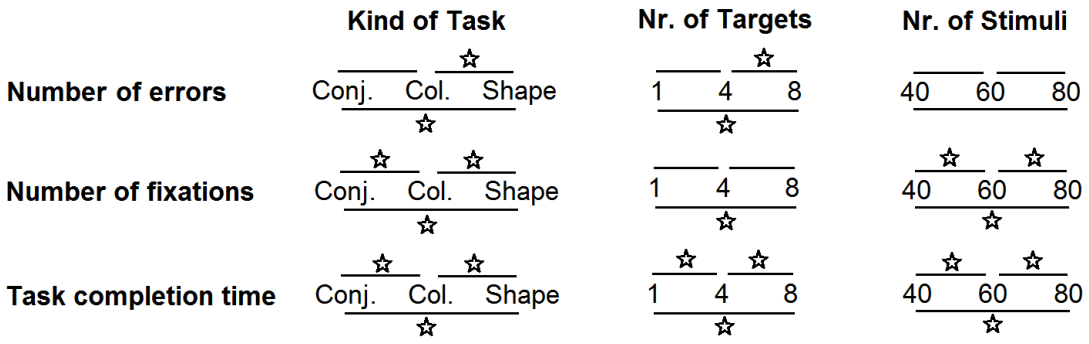

Figure 14: Overview of experiment design factors and their influence on task performance measures. Significant differences (p<0.05) are marked by *.

As can be seen in table 6, the strongst effect was shown for the category kind of task, with color being a lot easier than conjunction and shape being the most difficult task. In detail, most errors were done for the factor *shape*, followed by the factor *eight targets* and *80 stimuli*. The median of the evaluated values for the number of fixations was higher than for the other factors of the category. This was also reflected in the task completion time. It seemed that the factors shape, eight targets and 80 stimuli were harder to answer than the other tasks. E.g. it could be

Table 6: Overview of the number of errors and the median for the number of fixations and task performance time [s] for the factors *Kind of Task*, *Number of Targets* and *Number of Stimuli*.

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

assumed that the difference between tasks with 40 and tasks with 80 stimuli were bigger than the distance of both of them towards 60 stimuli. Both, the *Kind of Task* and the *Number of Targets* and the *Number of Stimuli* have a statistical influence on task performance time as well as on the number of fixations.

4.1.4 Evaluation of Scanpath Comparison Metrics

The following section shows which scanpath comparison algorithms correspond well to the effects and effect sizes determined by the task performance measures. Figures 16, 18 and 19 display the result of the statistical test between scanpath distance groups: If distances between scanpaths within one experiment condition distribute differently than distances between scanpaths of different experiment conditions, the effect is considered as detectable by the algorithm. In the figures, the algorithms used are presented on the y-axis and the categories *Kind of Task*, *Number of Targets* and *Number of Stimuli* are presented on the x-axis. Each field of the performance matrix was constructed by a statistical test performed on a distance matrix as following: All scanpaths originating from the task on the x-axis were collected and the average distance to each other was compared to the average distance between the remaining scanpaths. Figure 15 visualizes this process. The distance between scanpaths of group x is supposed to be quite small while the inter-group distance between scanpaths of group x and of group y is supposed to be quite large. It was tested statistically for equal distribution.

A distance matrix is a symmetrical quadratic matrix with the dimension $n \times n$

containing the distance of n-scanpath towards every other scanpath. n corresponds to the number of scanpaths involved in the comparison and in the case of this study, n is equal to the total number of scanpaths recorded. Values on the diagonal of the distance matrix are zero because they contain the distance of the scanpath to itself and need to be excluded from the statistical analysis.

The distance between the scanpaths was tested by using the Kolmogorov-Smirnov test which compared the distances between the groups with the distances within the groups. Due to the high amount of tests used (number of algorithms × number of categories = 45), the results had to be false discovery rate (FDR)-corrected with the Benjamini-Hochberg method. In figure 16, white fields show a p-value < 0.05 that means there is a statistical significance and the method is able to find a predefined effect. Black fields show a p-value ≥ 0.05 meaning there is no statistcal significance and the method cannot be differentiated from the others and is therefore less sensitive.

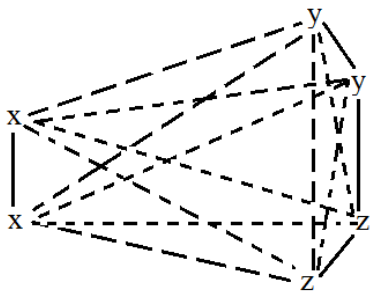

Figure 15: Demonstration of the distances within (solid lines) and between (dashed lines) the groups x, y and z.

Kind of Task

In figure 16, each field of the performance matrix was constructed by a statistical test performed on the 735×735 distance matrix which was split into a 273×273 conjunction to conjunction comparison matrix (13 trials x 21 subjects), a 210×210 color to color comparison matrix (10 trials x 21 subjects), a 252×252 shape to shape comparison matrix (12 trials x 21 subjects) and a $273 \times 210 \times 252$ conjunction to color to shape comparison matrix.

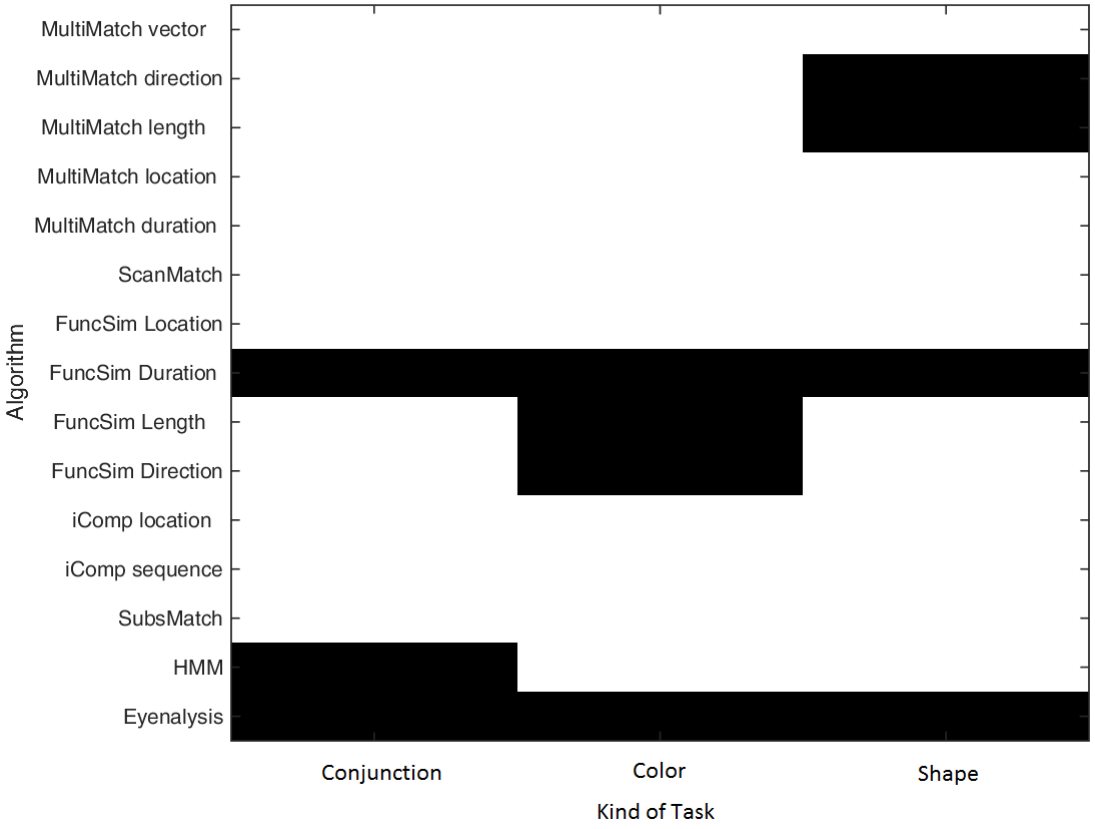

Figure 16: Results of the effect detection strength of the scanpath comparison algorithms with respect to the task performed. Significant values ($p < 0.05$) are shown in white.

Figure 16 shows statistical effects ($p < 0.05$) for the algorithms MultiMatch vector, location and duration and FuncSim location according to the Kolmogorov-Smirnov-Test. SubsMatch, iComp and ScanMatch are successful in all three groups. Two algorithms are not able to detect differences in all three groups, namely FuncSim Duration and Eyenalysis. Furthermore, five of the algorithms are not able to detect the differences in all the categories. Of these, two originate from the MultiMatch and two from the FuncSim metric. It is notable that for MultiMatch and FuncSim the same parameters (length and direction) fail, but at different categories. This finding suggests that the respective parameters are instable and perhaps barely able to detect the differences. The Hidden Markov Model does not detect differences in

the group conjunction.

Concerning the *frequency of errors*, the evaluated statistic (see overview figure 14) shows a statistical significance, both, for conjunction and shape as well as for color and shape, whereas the groups conjunction and color are difficult to differentiate. These algorithms being able to detect differences deliver a result that is similar to the statistically determined results. FuncSim duration and Eyenalysis do not show a statistical difference at all ($p \geq 0.05$) and seem to be imprecise with regard to the error frequency. The algorithms MultiMatch length and MultiMatch direction deliver a significance for conjunction and color but not for the factor shape which is in contrast to the statistically determined results in section 4.1.1. From these results the shape and color tasks should be most unique with the conjunction task somewhere in-between.

Regarding the *number of fixations*, a statistical significance occurs between all three groups in the evaluated statistics. It can be differentiated what kind of task the subject had looked at. With the exception of Eyenalysis and FuncSim duration, all algorithms are able to detect statistical differences for the number of fixations in dependency of the kind of task. In detail, MultiMatch length and direction are less sensitive for the group shape, FuncSim length and direction are less sensitive for color and the HMM is less sensitive for conjunction. Summarized, all algorithms with the exception of FuncSim and Eyenalysis deliver results that are consistent with the statistically determined results.

Concerning the *task performance time*, the evaluated statistic show a statistical significance between the groups conjunction, color and shape. This leads to the assumption that the kind of task has an influence on the task performance time. Eyenalysis and FuncSim are not sufficiently precise for the detection of the task performance time in dependency of the kind of task. The remaining algorithms are more sensitive and show significant differences.

Using multidimensional scaling as a post-processing step, threedimensional data points can be approximated from the pairwise distances of the distance matrix. This process introduces a small error since the distance matrix is high dimensional

and dimensions have to be dramatically reduced in order to be able to visualize them in 3D space. This technique is used for the algorithm ScanMatch and is shown in figure 17.

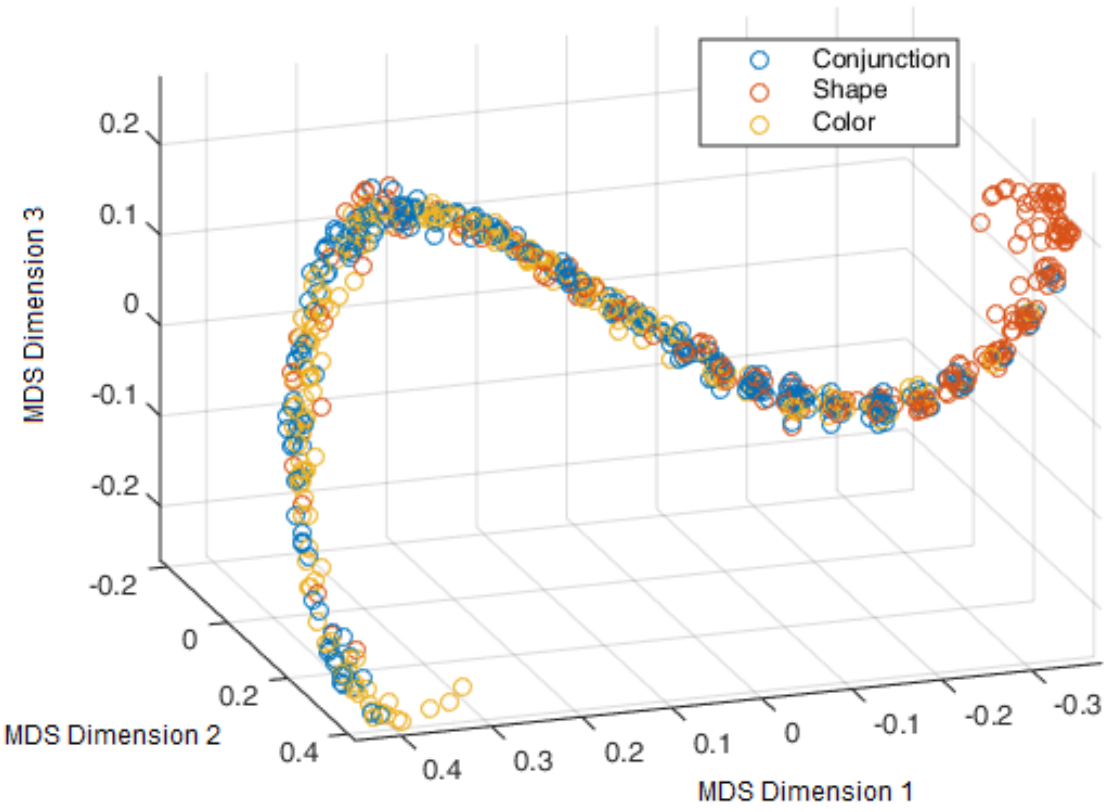

Figure 17: Multidimensional scaling (MDS) with ScanMatch on the pairwise scanpath distances for the category *Kind of Task*. Each point represents one scanpath and is placed close to the scanpath which is determined to be very similar. The axes represent virtual space coordinates (1., 2. and 3. Dimension) derived from the distances.

All distances between the groups conjunction, color and shape are represented. It shows that within the groups, distances are smaller than between the groups, especially for the group *shape* where most scanpaths cluster very close together. The conjunction and color task cannot be separated that clearly. This is consistent with the results gained in the statistical evaluation in section 4.1.1. Such a visualization

could be created for each of the scanpath comparison algorithms and all of them showing statistically significant results should yield in a similar group separability effect.

Number of Targets

In figure 18, each field of the performance matrix was constructed by a statistical test performed on the $735x735$ distance matrix which was split into a 336×336 one to one target comparison matrix, a 231×231 four to four targets comparison matrix, a 168×168 eight to eight targets comparison matrix and a $336 \times 231 \times 168$ one to four to eight targets comparison matrix. FuncSim duration is the only algorithm unable to find any significant differences. Furthermore, the algorithms MultiMatch direction, SubsMatch and Eyenalysis do not show a significance for the group of four targets and the last named for eight targets, too.

Concerning the *frequency of errors*, statistical significances (see overview figure 14) occur between one and eight targets as well as between four and eight targets whereas the variation between one and four targets did not yield much of a difference. Therefore, a very good separability of the eight targets group and worst separability for the group of four targets is expected. This expectation also makes intuitive sense, since the one and eight target cases should be most different from each other, while the four targets stimulus lies somewhere in-between. Compared to the results of the Kolomogorov-Smirnov test, all algorithms except FuncSim duration and Eyenalysis deliver statistical differences between the tasks, too. Furthermore, the results for MultiMatch direction and SubsMatch lie in between with regard to their precision. They do not show a significant difference for the group of four targets but recognize differences between one and eight targets as expected. The extreme sensitivity of some algorithms like ScanMatch and the HMM appears surprising.

The assumption made above also applies to the *number of fixations*. The statistical evaluation shows significant differences between the amount of one and eight targets. There is no statistical significance between the groups of one and four targets as well as between four and eight targets. According to the number of fixations, it

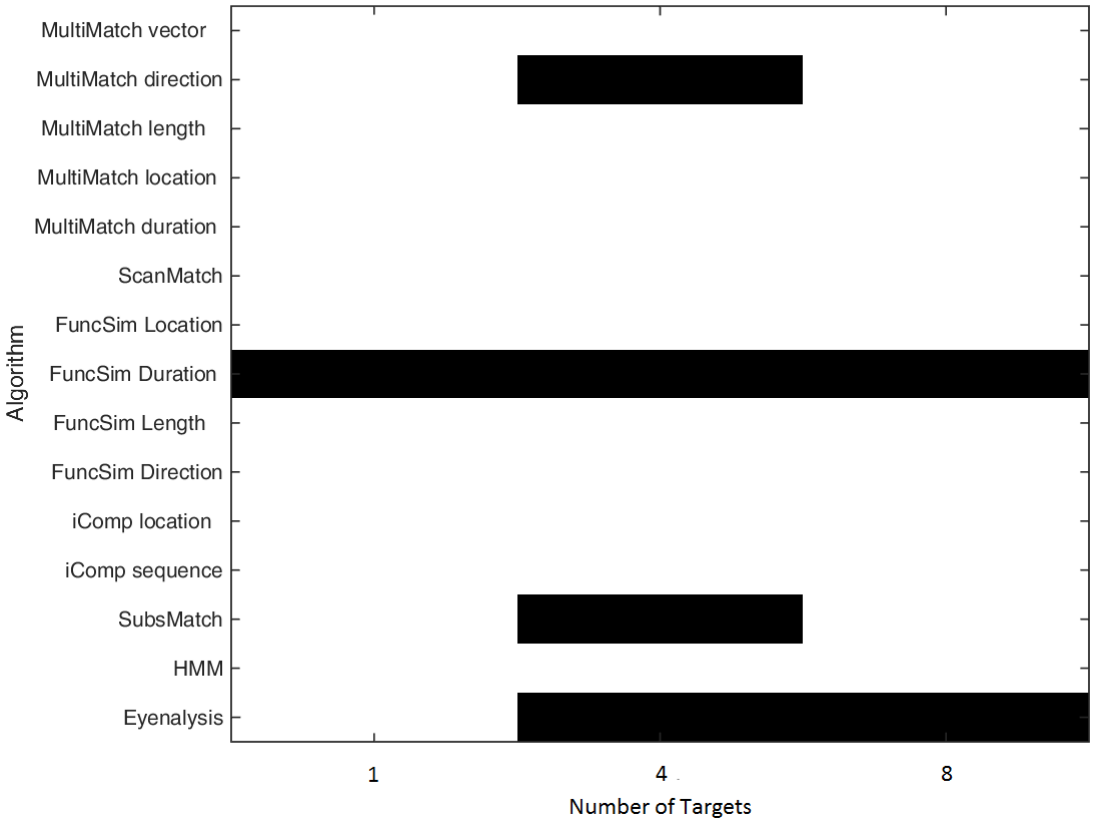

Figure 18: Results of the effect detection strength of the scanpath comparison algorithms with respect to the *Number of Targets*. Significant values ($p < 0.05$) are shown in white.

cannot be differentiated if subjects look at one or four and four or eight targets. When comparing these results to the results of the scanpath comparison metrics shown in figure 18, it seems that all algorithms except FuncSim duration and Eyenalysis are more exact to determine statistical differences between the amount of targets.

The evaluated statistic show statistical significance between the tasks of one, four and eight targets in relation to the *task completion time*. This leads to the assumption that the number of targets have a great influence on the task performance time. This is not consistent with FuncSim duration and partly with Eyenalysis (four and eight targets). All other algorithms are able to detect statistical differences as can

be seen in the statistical evaluation as well. This fact makes them applicable for the evaluation of the task completion time.

Number of Stimuli

In figure 19, each field of the performance matrix was constructed by a statistical test performed on the 735×735 distance matrix which was split into a 252×252 40 to 40 stimuli comparison matrix, a 231×231 60 to 60 stimuli comparison matrix, a 231×231 80 to 80 stimuli comparison matrix and a $252 \times 231 \times 231$ 40 to 60 to 80 stimuli comparison matrix. The performance matrix shows significant differences completely for the algorithm iComp and partly for the algorithm Eyenalysis with 80 stimuli. The other algorithms do not show scanpath differences and are not able to differentiate between 40, 60 or 80 stimuli.

Concerning the *frequency of errors*, the evaluated statistic (see overview figure 14) does not show a significance for the number of stimuli at all. It is hardly possible to differentiate if subjects look at 40, 60 or 80 stimuli. These results are consistent with the results shown in figure 19. iComp seems to be an algorithm that is more sensitve concerning the number of stimuli because it detects scanpath differences.

By regarding the *number of fixations*, a statistical significance occurs between all groups for the number of stimuli. It is possible to differentiate whether subjects look at 40, 60, 80 stimuli. In this case, iComp is more sensitve than all the other algorithms under test that do not find statistically effects and thus gains best results.

The evaluated statistic for the *task completion time* shows a statistical significance between the groups 40, 60 and 80 stimuli. This leads to the assumption that the number of stimuli has an influence on the task completion time. These results are consistent with the algorithm iComp and with Eyenalysis in the case of 80 stimuli. All the other algorithms seem to be less sensitve concerning the task completion time in the category number of stimuli.

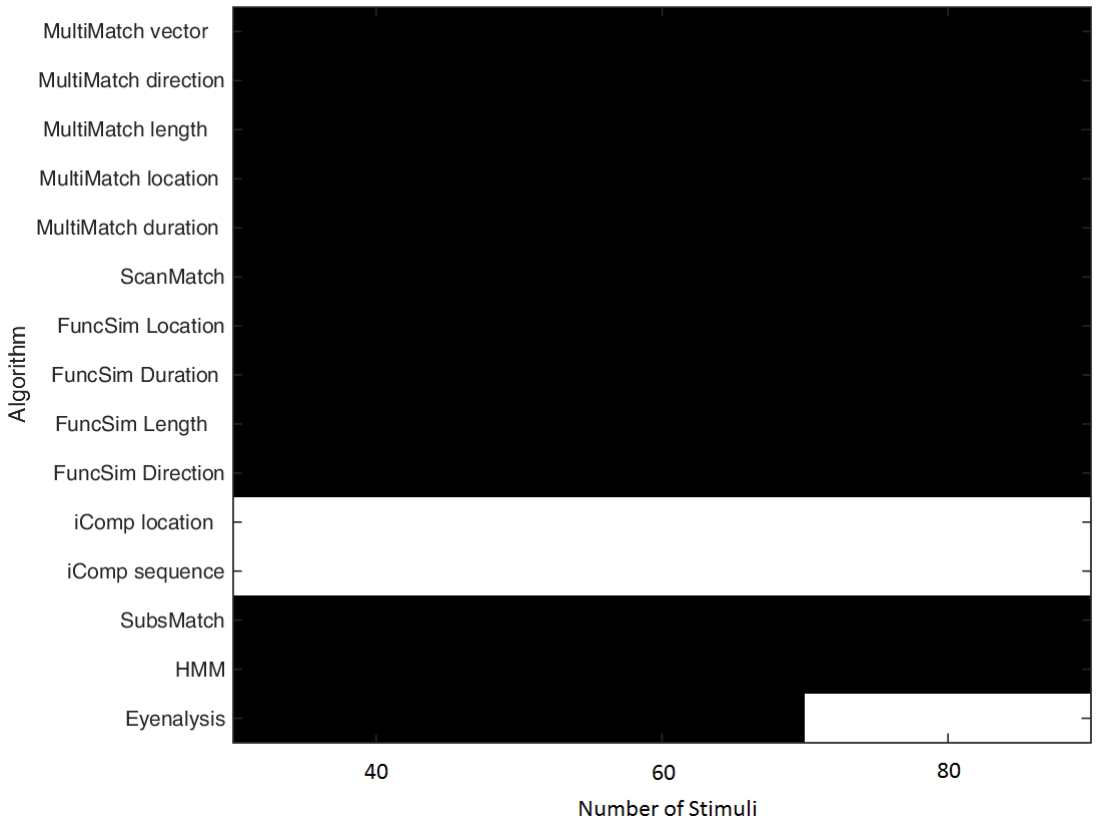

Figure 19: Results of the effect detection strength of the scanpath comparison algorithms with respect to the *Number of Stimuli*. Significant values ($p < 0.05$) are shown in white.

To summarize, the evaluated scanpath comparison metrics are shown in table 7. Values of ± and 0 are listed. Algorithms which mainly have a value of zero scored as expected and are defined by the accordance of the statistics with the respective algorithm. A value of + indicates that the distance metric shows a statistical difference but the statistic does not. The algorithm is therefore more sensitive for dissimilarities. A value of - means that the distance metric does not show significances but the statistic does. That fact makes the algorithm less sensitive to existing scanpath differences and in this case of inaccuracy, the algorithm is unusable.

In detail, the algorithms MultiMatch, ScanMatch, SubsMatch, the Hidden Markov

Table 7: Rating of the algorithms used for static visual search tasks. 0: distance metric gains the same results as the statistical evaluation. +: distance metric shows significant differences, the statistical evaluation does not. -: distance metric does not show significant differences, the statistical evaluation does. A: number of errors, B: number of fixations, C: task completion time

	Task			Targets			Stimuli		
Algorithm	A	B	C	A	B	C	A	B	C
MultiMatch	0	0	0	0	0	0	0	-	-
ScanMatch	0	0	0	0	0	0	0	-	-
FuncSim	0	0	0	0	0	0	0	-	-
FuncSim duration	-	-	-	-	-	-	0	-	-
iComp	0	0	0	0	0	0	+	0	0
SubsMatch	0	0	0	0	0	0	0	-	-
HMM	0	0	0	0	0	0	0	-	-
Eyenalysis	-	-	-	-	-	-	0	-	-

Model and FuncSim (location and length) scored identically in all three categories. For the categories kind of task and number of targets, these algorithms were consistent in their results with the statistic. For the category number of stimuli, they were rarely sensitive for the number of fixations and the task completion time. In the categories kind of task and number of targets, the results for the algorithm iComp were the same as for the last named. For the number of stimuli, iComp was more sensitive in detecting scanpath differences in the error frequency. For the number of fixations and the task completion time, the results were consistent with the performance measures. This algorithm scored best in static visual search tasks. The algorithms being totally different from the other algorithms were FuncSim duration and Eyenalysis. Although FuncSim duration was a part of the whole algorithm FuncSim, it was regarded separately at this place because it delivered completely different results than the FuncSim location and length. Eyenalysis and FuncSim duration were both identical in scoring. With the exception of the error frequency in the category number of stimuli, they were not able to detect significant differences whereas the evaluated statistic did. Therefore, these two algorithms were

not precise enough to answer questions of scanpath similarity.

4.2 Mario Kart

In the following section, the results of the Mario Kart experiment in a virtual interactive environment are shown. This racing game offers an objective performance measure, the lap or track completion time and suggests high differences between the subjects in driving behavior. According to questions five to seven of the questionnaire, 90% of the subjects used smartphones and tablets regularly. Only 10% owned neither a smartphone nor a tablet. 57% of the subjects stated playing games regularly. 52% of them played less than one hour a day and one subject played one to two hours daily. 43% did not play at all. Almost all subjects had experience with handling modern devices but only about the half of them had regular playing experience. According to questions eight to ten, eight of 21 subjects (38%) played racing games regularly. 28% played less than one hour and two persons (10%) played one to three hours a week whereas 62% of the subjects did not play racing games regularly. 67% stated that they had already played Mario Kart whereas 33% never had played Mario Kart on any console before.

In a first step, measurement quality of the eye tracker was evaluated for both tracks of Mario Kart. The percentage of valid gaze measurements during the experiment is shown in figure 20. The asymmetric distribution of measurement quality for track 1 has its median at 92% valid data points, the 25^{th} percentile is at 77%. The whiskers extend down to 69%. In track 2, the median is at 91%, the 25^{th} percentile is at 77% and the whiskers extend down to 54%. For the further analysis all data was used including the low quality recordings.

Based on question eight of the questionnaire, subjects were classified into either regular players or non-players depending on their statement of playing video racing games regularly or not at all. The time in seconds for completing track 1 and 2 is graphically shown in figure 21.

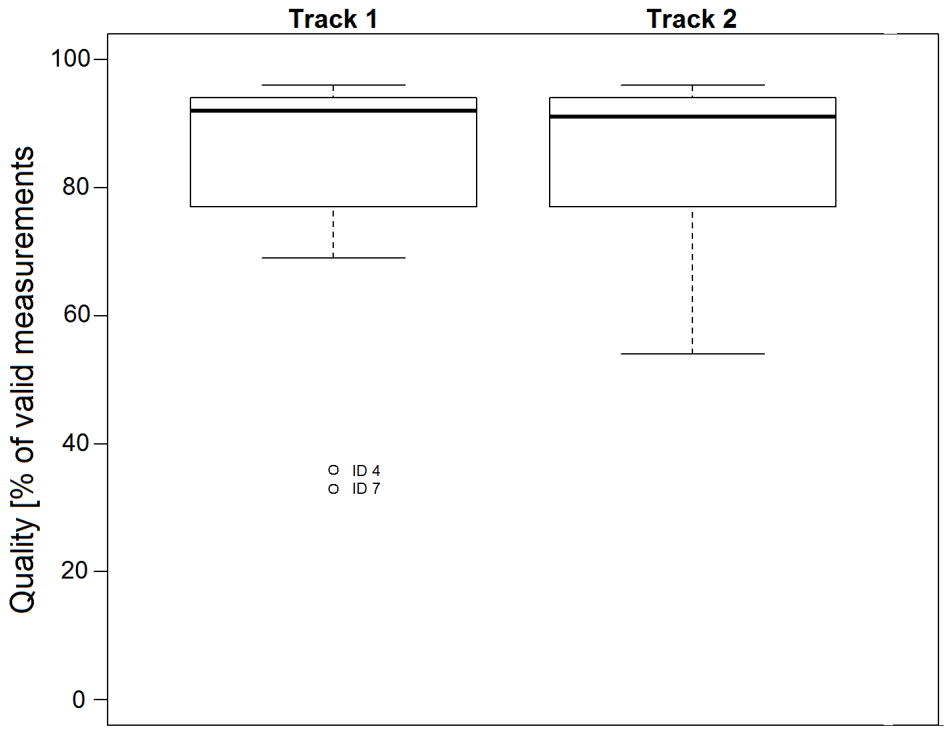

Figure 20: Measurement Quality for Mario Kart with $n_1 = n_2 = 21$.

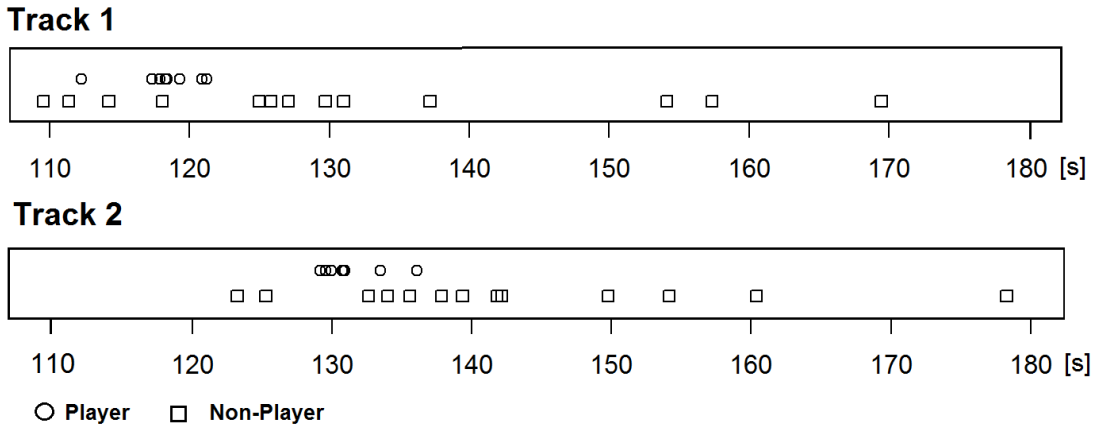

Figure 21: Track completion time [s] of regular players and non-players for Mario Kart.

Figure 21 reveals that subjects who rated themselves as regular players showed an overall good performance and were able to judge their skills adequately. Subjects who rated themselves as non-players however varied a lot in their track completion time, ranging from the shortest to the longest of all participants. Furthermore, two subjects of the group of non-players completed both tracks faster than the rest of the regular playing group. This led to the conclusion that the classification with the questionnaire was too subjective and therefore inaccurate.

In a next step, the more specific question regarding the knowledge of the video game Mario Kart was used for subject classification. Subjects, either being familiar with Mario Kart or not being familiar with Mario Kart, were classified. The results can be seen in figure 22. Being familiar with Mario Kart clearly had a strong influence on task performance time. The huge impact of specific Mario Kart experiences suggests that there are game elements that players being familiar with the game attend to and possibly interact with in order to get their advantage.

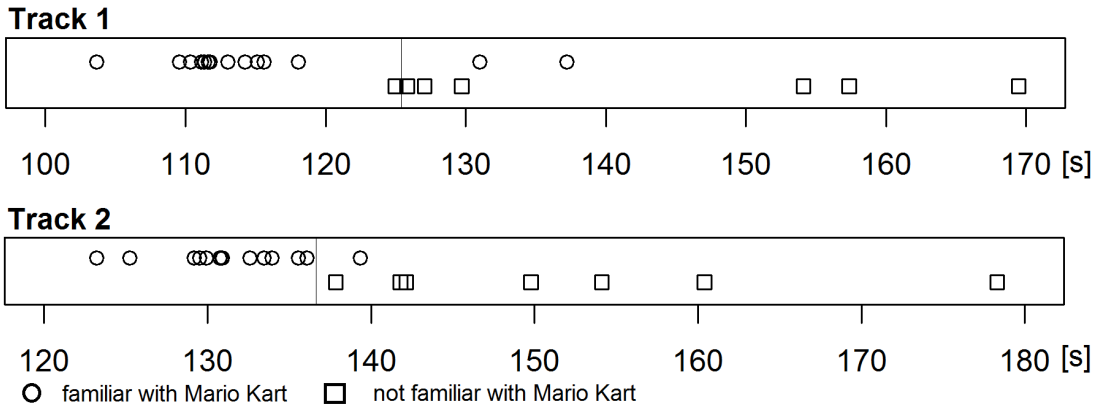

Figure 22: Track completion time [s] of subjects being familiar with Mario Kart and subjects not being familiar with Mario Kart.

In the following, the results were examined with respect to track completion time. This objective performance parameter corresponded quite well with Mario Kart experience but could also capture the effect of novices understanding the game fast.

Based on the placement when finishing the tracks, a separation of subjects could be

observed: 13 subjects finished at place one to eight in the first track and at place one to six in the second track whereas eight subjects finished at place 11 to 12 in the first track and 10 to 12 in the second track. Along with these findings, the fastest 60% (<125s for track 1 and <136s for track 2) were classified in the group of fast drivers whereas the remaining 40% of the subjects form the group of slow drivers. The separation is shown by the black line in figure 22. In the following, the number of fixations as well as the track completion time were compared according to the categories of track 1 and 2 for fast (n = 13) and slow drivers (n = 8) as well as for the whole subject group (n = 21).

4.2.1 Number of fixations

Track 1

The Shapiro-Wilk normal distribution test for the number of fixations could not be rejected for the categories of fast and slow drivers as well as for the whole subject group (fast: W = 0.9133, p = 0.203, slow: W = 0.8345, p = 0.06613, all subjects: W = 0.9478, p = 0.3093). Even if normal distribution is given, a Wilcoxon rank-sum test was performed instead of an unpaired t-test due to the small sample size (n_{fast} = 13, n_{slow} = 8). With the Wilcoxon rank-sum test, no significant differences could be obsereved between fast and slow drivers (p=0.94, Bonferroni-Holm corrected) as well as by comparing all three groups (p = 1 between each group due to the appearance of ties within the number of fixations). It seems that the number of fixations is not a reliable performance measure in dynamic tasks. The effect that the number of fixations is dependent on the track completion time (slow drivers fulfil a larger amout of fixations) could not be observed.

Track 2

The Shapiro-Wilk normal distribution test for the number of fixations could not be rejected for the categories of fast and slow drivers (fast:W = 0.8991, p = 0.13, slow: W = 0.8854, p = 0.2119, all subjects: W = 0.9515, p = 0.3629). With the Wilcoxon rank-sum test, no significant differences could be found between fast and slow drivers (p=0.12, Bonferroni-Holm corrected). By comparing all three groups, no significant differences could be observed, either (fast vs. slow: p = 0.36,

fast vs. all: $p = 0.59$, slow vs. all: $p = 0.59$). This is equal to the findings in track 1.

4.2.2 Track completion time

Track 1

With the Shapiro-Wilk test, the track completion time was tested for normal distribution within the categories of fast and slow drivers and could not be rejected (fast: $W = 0.9247, p = 0.29$, slow: $W = 0.8614, p = 0.1241$) but had to be rejected for the group of all subjects ($W = 0.834, p = 0.0023$). As designed by the group splitting, a significant group effect in terms of track completion time for fast and slow drivers (Wilcoxon rank-sum test: $p = 9.8e-06$, Bonferroni-Holm corrected) was observed. As expected, no statistical significances occurred between fast and all of the drivers ($p = 0.068$) but between the categories of slow and all drivers ($p = 0.024$).

Track 2

The assumption of a normal distribution for the track completion time within the categories of fast and slow drivers could not be rejected (fast: $W = 0.9444, p = 0.5165$, slow: $W = 0.8656, p = 0.1364$) but has to be rejected for the whole subject group ($W = 0.8285, p = 0.0019$, Shapiro-Wilk test). As already seen in the evaluation of track 1, the Wilcoxon rank-sum test showed a statistical significance between the categories of fast and slow drivers ($p = 9.8e-06$, Bonferroni-Holm corrected). By comparing all three groups, there were no statistical significances between fast and all drivers ($p = 0.068$) whereas a difference occurred between the categories of slow and all drivers ($p = 0.024$).

To summarize the findings of the previous chapter, figure 23 displays the significant factors in experimental design that have an influence on track performance.

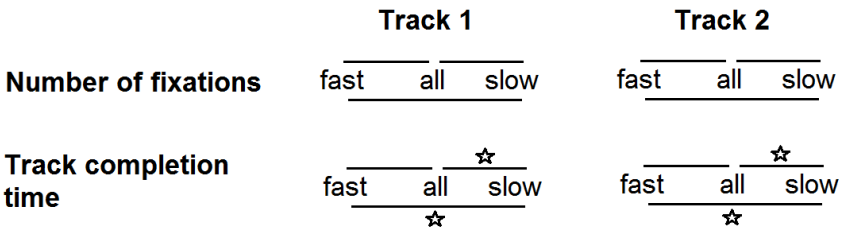

Figure 23: Overview of experiment design factors and their influence on track performance measures. Singificant values ($p<0.05$) between the categories of fast, slow and all drivers are marked by *.

4.2.3 Evaluation of Scanpath Comparison Metrics

In the following, an evaluation of scanpath comparison metrics was done for Mario Kart, Track 1 and 2 separately. The algorithms under test should be able to detect differences between the groups of fast and slow drivers. Figure 24 shows the performance of the comparison metrics with respect to track 1 and 2. Each field of the performance matrix was constructed by a statistical test performed on a distance matrix as follows: All scanpaths ($n = 21$) originating from the track currently under consideration were collected and the pairwise distance between scanpaths from the same group and scanpaths from different groups were compared to each other. The 21×21 distance matrix was therefore split into a 13×13 fast driver to fast driver comparison matrix, a 8×8 slow driver to slow driver comparison matrix and a 13×8 fast driver to slow driver comparison matrix. The distributions of scanpath distances between the performance groups are tested for equality by using the Kolmogorov-Smirnov test. Due to the high amount of tests used (number of algorithms $\times$ number of tracks $= 15 \times 2$), the results have to be false discovery rate (FDR)-corrected with the Benjamini-Hochberg method. In figure 24, white fields show a value of $p < 0.05$ meaning that the method was able to find the desired effect whereas black fields show a p-value ≥ 0.05 meaning that the method lacked the sensitivity to differentiate the two subject groups.

Track 1

Figure 24 shows significant differences for the algorithms MultiMatch and the

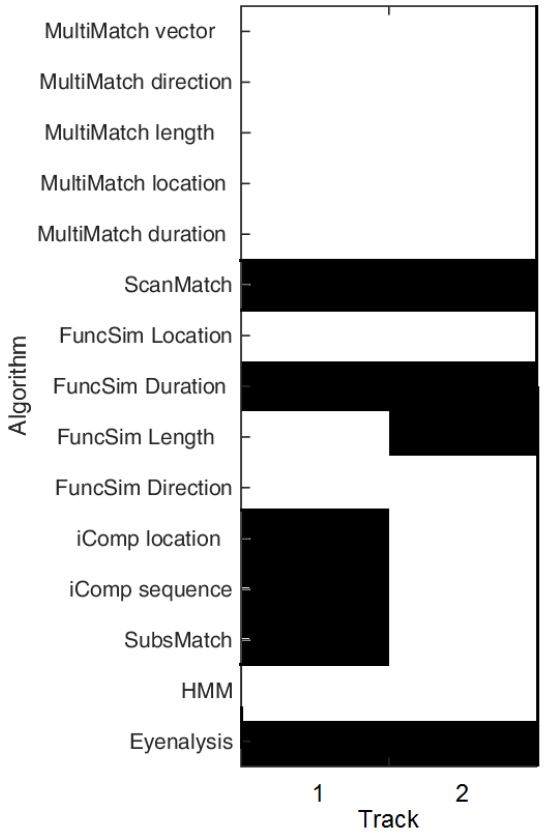

Figure 24: Results of the effect detection strength of the scanpath comparison algorithms with respect to the track performed. Significant values ($p < 0.05$) are shown in white.

Hidden Markov Model for Track 1. Partly, differences occurred for FuncSim (except duration) and no differences were detectable with ScanMatch, iComp, SubsMatch and Eyenalysis. Due to the separation of subjects into fast and slow drivers, statistcal differences occurred in track completion time between these groups which makes the number of fixations an unsuitable performance measure. Only the algorithms MultiMatch, HMM and FuncSim were able to detect differences.

Track 2

Figure 24 shows significant differences for the algorithms MultiMatch, iComp and the Hidden Markov Model for Track 2. Partly, differences occurred for FuncSim

(location and direction) whereas no differences were detectable with ScanMatch, FuncSim (duration and length) and Eyenalysis. It is noticeable that iComp and SubsMatch show a value of $p < 0.5$ for track 2 but were not able to detect differences in track 1.

Figure 25 shows the separation of fast and slow drivers graphically for track 2 of Mario Kart after multidimensional scaling with the algorithm ScanMatch.

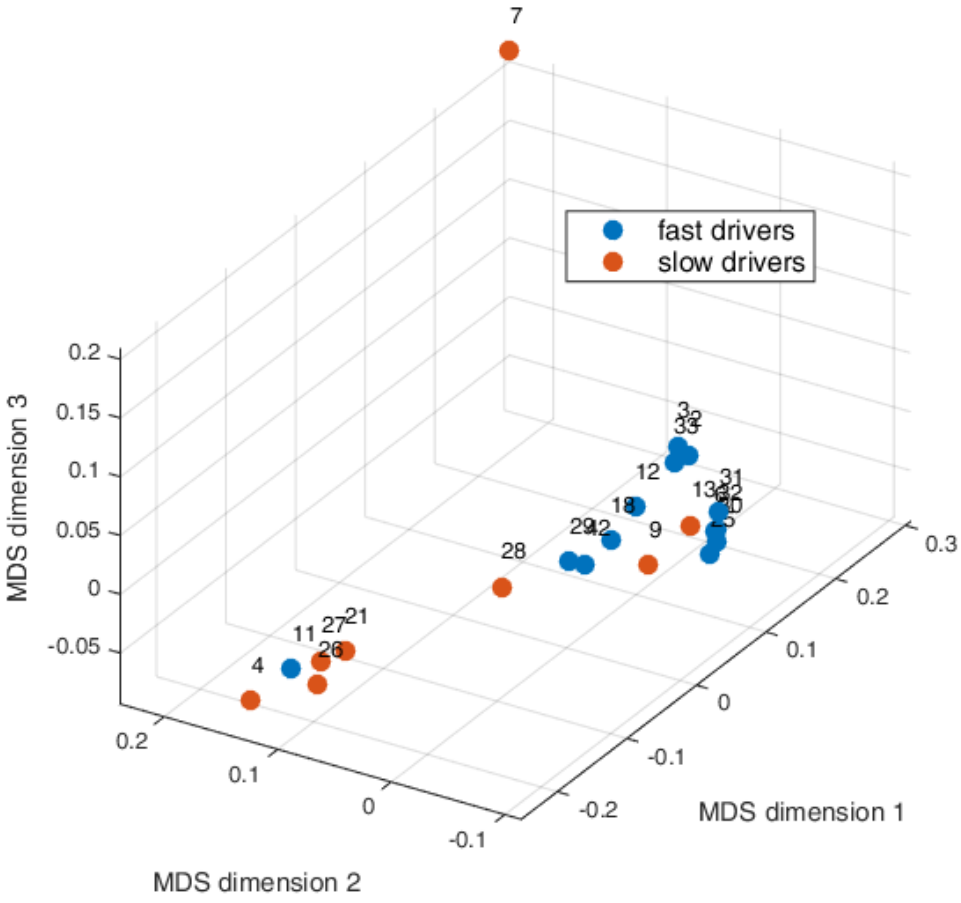

Figure 25: Multidimensional scaling for track 2 for Mario Kart with the algorithm ScanMatch.

Fast drivers (blue) lie close together as well as the group of slow drivers (red). A good separation of the two classes is possible in the first dimension suggesting that the task completion time effect is indeed the most relevant factor for scanpath

similarity in this dataset. Subject 7 was the only factor defining the third MDS dimension and could safely be considered an outlier. Measurement quality was 54% and even worse in track 1 (36%). The subject finished track 1 at last (169s) but improved in track 2 (138s) and thereby nearly missed the border for being considered a fast driver. Due to the high rate of data loss, the number of fixations performed by the subject was small. It is interesting that subject 7 neither having played Mario Kart before nor having a smartphone for playing games. Furthermore, for the recording of subject 7, the eye tracker had to be placed on books to heighten up in order to enable the detection of the pupil. This could have caused noise and could have let to inaccuracies in measurement quality.

To summarize, the evaluated scanpath comparison metrics for Track 1 and 2 are shown in table 8.

Table 8: Rating of the algorithms used for dynamic tasks. 0: Algorithm gains the same results as the performance measures. +: Algorithm shows significant differences, the performance measures do not. -: Algorithm does not show significant differences, the performance measures do.

	Track 1		Track 2	
Algorithm	Fixations	Time	Fixations	Time
MultiMatch	+	0	+	0
ScanMatch	0	-	0	-
FuncSim	+	0	0	0
iComp	0	-	+	0
SubsMatch	0	-	+	0
HMM	+	0	+	0
Eyenalysis	0	-	0	-

Algorithms which mainly have a value of zero scored as expected and are defined by the accordance of the performance measures with the respective algorithm. A value of + indicates that the distance metric shows a statistical difference but the performance measures do not. The algorithm is therefore sensitive for scanpath

similarities in contrast the performance measures. In general, this is considered as a good effect, but it may also mean that the algorithm is overly sensitive even to small differences. Therefore, other potentially stronger and more important effects could be covered. A value of - means that the distance metric does not show significances but the performance measures do. That fact makes the algorithm insensitive to existing scanpath differences and in this case of inaccuracy, the algorithm is unusable.

MultiMatch and the Hidden Markov Model seem to be the best choice for automated scanpath comparison metrics in dynamic tasks, directly followed by FuncSim. It is surprising that SubsMatch and iComp showed different results for the first and the second track. ScanMatch and Eyenalysis were not able to detect differences in scanpaths and should not be considered for the evaluation in dynamic tasks.

5 Discussion

In the following, the results for the Scanpath Comparison Metrics for the *Conjunction Search Task* were compared with the results gained in dynamic tasks by the video game *Mario Kart* and with the original data set of each algorithm.

The algorithm *MultiMatch*, implemented by Dewhurst et al [9], is vector-based on a string alignment and provides multiple dimensions by saccades and fixations without defining AOIs for the evaluation. In contrast to Jarodzka et al [19] where the scanpath comparison is done with eight pairs of fictious scanpaths, the scanpaths in this approach were recorded from 21 real subjects performing a search task and playing a video racing game interactively. In this study, the dimensions (vector, direction, length, location and duration) were usable independently from each other. A weighting factor would be helpful to deteremine which dimensions are more important and should play a major role in evaluation. Dewhurst et al [9] compared MultiMatch to ScanMatch with synthetic data and real eye movements from 20 subjects viewing sequences of dots. They found that MultiMatch outperformed ScanMatch since MultiMatch is able to show how similar two scanpaths are along with the providing of five different dimensions. In the Conjunction Search Task, ScanMatch delivered statistical significances more reliable than MultiMatch. The latter one did not show significant differences in some dimensions for a specific group, respectively. However, MultiMatch scored better in scanpath comparison in dynamic tasks than ScanMatch. In summary, MultiMatch is able to detect differences and commonalities between scanpaths in static visual search tasks as well as in dynamic tasks.

ScanMatch, implemented by Cristino et al [8], bases on the string alignment and contains spatial, temporal and sequential information within a substitution matrix. In comparison to MultiMatch, ScanMatch is not divided into multiple dimensions and the use of AOIs is necessary. Nevertheless, ScanMatch scored more reliably regarding the Conjunction Search Task than MultiMatch in some dimensions. Cristino et al [8] tested ScanMatch with synthetic data as well as with one subject in a sequential looking task and with eight participants performing a visual search

task. For static stimuli, the algorithm is robust as seen in the present study. The authors of ScanMatch do not expect a change over time in the substitution value and consider the algorithm as reliable. In this thesis, ScanMatch scored poorly for dynamic tasks. The algorithm was not able to detect differences for track completion time between fast and slow drivers for track 1 and 2 separately. Instead of AOI labeling, a grid was used for evaluation as the only practical way. In summary, ScanMatch is a reliable algorithm for scanpath comparison in static scenes but it is not suitable for dynamic tasks due to the way of analysis.

FuncSim, implemented by Foerster et al [14], is vector-based and provides the calculation of different scores on multiple dimensions. Foerster et al compared FuncSim with the algorithms MultiMatch and ScanMatch with eight pairs of artificial scanpaths. FuncSim does not provide advantages in tasks with no inherent sequences. In this case, the algorithm just aligns the fixations in relation to their temporal position in the scanpath instead of defining functional units. The Conjunction Search Task in the present study showed that FuncSim duration is less sensitive whereas FuncSim location, length and direction scored well. Altogether, FuncSim leads to an underestimation of scanpath similiarity in static visual search tasks. Nevertheless, FuncSim is perfectly suitable to compare scanpath similarity in dynamic tasks as proven with the second expermimental design in the present study.

iComp bases on a string alignement in the dimensions sequence and location. The inventors of iComp, Heminghous and Duchowski [18], tested the algorithm with six subjects viewing synthetic images for 500 ms. Local similarity (different subjects looking at the same image) was found to be significantly higher than random similarity (both for location and sequence). In the present evaluation of scanpath comparison metrics, iComp scored best in static scenes for location as well as for sequence. The algorithm delivered results that were consistent with the evaluated statistic and even more sensitive regarding scanpath similarities than the other algorithms under test. Concerning the video game, iComp scored poorly in the track 1 but fulfilled the expectations in track 2.

SubsMatch bases on the frequency of attention shifts and exploratory eye movements

in dynamic scenes. The author of this algorithm, Kübler et al [20], compared SubsMatch with MultiMatch and ScanMatch by performing a simulated driving task. Thereby, SubsMatch showed the best performance, followed directly by ScanMatch. Compared to MultiMatch, SubsMatch scored better for the inner-group distances and the distances between the groups. The advantages of SubsMatch are the search for repeated patterns in visual scanpaths instead of computing a general similarity score. In the subject groups, a characteristical pattern that occurs more often could be observed. For static scenes e.g. images and visual search tasks as performed in the present study, SubsMatch was not yet tested but gains good results comparable to MultiMatch and ScanMatch. For interactive virtual environments, SubsMatch scored poorly in the track 1 but fulfilled the expectations in track 2.

Hidden Markov Model is a stochastic model that keeps the transition probability constant over time and evaluates the hidden states on the basis of the emissions. In the static visual search task, the HMM scored identically with the algorithms MultiMatch and ScanMatch. In dynamic tasks like the interactive environment achieved with Mario Kart, the HMM scored well like MultiMatch. The results are consistent with the evaluated statistic and makes the HMM usable for the evaluation of static and dynamic tasks.

Eyenalysis bases on vectors and string alignment and combines the dimensions location, duration and timestamp which can be weighted. The inventors of Eyenalysis, Mathôt et al [24], compared their algorithm to ScanMatch by using artificial but realistic eye movement data gained from two images and a 26×26 grid for the evaluation. There, Eyenalysis performed much better than ScanMatch. This is in contrast to the results gained in the present study where Eyenalysis scored very poorly in each category for the Conjunction Search Task. The authors of Eyenalysis expect a great usability in real-world tasks what can currently not be fulfilled as seen in the experimental design for Mario Kart. In general, each scanpath comparison algorithm fulfils a step where the produced similarity score is corrected for the length of sequences compared to each other. Probably, this step was not performed correctly by the algorithm Eyenalysis and resulted in high sensitivity for scanpath length differences but low sensitivity for other interesting differences. The

difficulties in compensating different scanpath lengths could be caused by tasks being longer than a few seconds. For the evaluation in dynamic scenes, a correct normalization is much more important than for static scenes.

As shown by Machner et al [22], less saccades and fixations were required for color search than for shape search. This is consistent with the results of the present study where the number of fixations correlates with the task completion time (table 6). The fact that objects with a predefined color were faster to detect than objects with a predefined shape confirmed the assumption of the *pop-out effect*, described by Treisman and Gelade [33]. Baluch and Itti [3] found out that color was the dominant feature followed by size and orientation when Gabor patches, different in color, spatial frequency and orientation, were used. In contrast to the scanpath comparison presented by Anderson et al [1] where images were shown twice and a comparison of scanpaths between subjects and images as well as within subjects and images was done by using contemporary algorithms like ScanMatch and MultiMatch, this approach went one step further and evaluated the usability of algorithms in dynamic tasks. The paper of Anderson et al is currently the only published work dealing with scanpath comparison metrics using free-viewing tasks. Free-viewing tasks showed a higher variability in scanpaths between subjects than task-driven, visual search tasks as used in present study. Both, ScanMatch and MultiMatch, showed robust results in the free-viewing, top-down driven tasks as proven by Anderson et al as well as in this study, although the visual search task based on the bottom-up effect.

As shown in the study of Peters and Itti [28], there is less variability in heuristic performance across subjects than across games. It must be noted that only five subjects participated in their study. An evidence for gaze behavior of good and bad players can not be made. In the present study, 21 subjects splitted into fast and slow drivers participated. It was expected that the eye movements and gaze behavior between the two groups differ. Resulting from the statistical analysis in section 4.2.1, the number of fixations in dynamic tasks can not be considered to be a reliable performance measure. Dorr et all [11] claim that fixations are at least partially determined by the visual input. In the interactive racing game, a more frequent use of the item boxes (placed in the left upper corner on the screen) from

subjects being familiar with Mario Kart could be observed. This is likely to have caused much of the viewing behavior differences between expert and novice players. In general, when subjects perform a task for the first time, more fixations are done and the completion time is longer than after having acquired more expertise as shown by Epelboim et al [13]. In the racing game context, this effect may have been countered by faster explorative scanning of the experts, explaining why no differences in the number of fixations performed were found. However, since participants did not perform the same track twice, the size of the training effect cannot be determined. Due to the different track lengths, predictions about a training effect cannot be made. An increased frequency of using the special item boxes that gives advantages to drivers and thereby a relevant increase in task performance time could be observed for some players. Due to this fact, these subjects showed a large learning effect within the two tracks. Track completion time varies more among subjects never having played Mario Kart before or play racing games regularly than between experienced players. The position at the finish line is not taken into account for the evaluation. Since computer players are non-deterministic in their behavior, the same track completion times may lead to slightly different positioning. This means that a fast driver finishing at position 6 can be faster in completing than a slow driver finishing at position 2. Therefore, track completion time is considered to be a more stable factor. In real-world experiments as described by Foerster et al [14], look-ahead fixations on the so-called target locations have been found, for example when placing cups. It is expected that all subjects pay attention in front of the virtual driver at the target location instead of the driver itself.

Sources of errors could have been head movements, blinking during the measurement or a lack of the tear film. The latter one can occur due to non-blinking by starring concentrated on the screen. This issue could not be avoided in total but reduced by making a break between the trials, respectively tracks. The spectacle frame and the illumination condition could also have influenced the measurement. With the binocular eye tracking device *The Eye Tribe*, this should be inhibited [30]. Furthermore, the limitation of eye tracking systems could have caused errors. As described by Duchowski [12], the accuracy typically decreases the more the subject

looks to the periphery. Therefore, the screen was positioned 60 cm, respectively 110 cm in front of the subjects to keep the visual field as small as possible and especially to avoid distraction from the surroundings. Small deviations in the experimental set up could have caused a lack of precision during eye tracking measurement. In each experiment, the eye tracker was positioned manually in front of the subject in order to detect at least four of five stars permanentely. Calibration measurements being poorer were excluded from the evaluation. In order to avoid the influence of an interruption during recording, a threshold was defined for measurement quality. Values of lower quality were excluded from the evaluation. Errors could also be induced by the method of evaluation. Therefore, the statistical tests were false discovery rate corrected. Measurement errors for the *Conjunction Search Task* could be caused by the length of the test. In the study of Machner et al [22], subjects had to perform as many tasks as possible (max. 84) within 30 minutes and were under time pressure. Even though subjects in this experiment had as much time as they needed to complete all 35 tasks, a decrease of attention is likely and could have led to an inaccuracy concerning the recorded values and evaluation. For the video game *Mario Kart*, a calibration was done before the race started but not repeated at the end of the course. Therefore, an occurring deviation in measurement quality could not be proven. Despite briefing, some subjects screamed or just moved their chin a little bit while driving. This could have led to inaccuracies in measurement quality as well. In order to avoid the influence of head and body movement, subjects had their chin on a chin rest. Subjects reported on a time delay (<1s) in the transmitted video game. This comes from the fact that Mario Kart was first transferred from the Wii to the laptop for recording and then transmitted to the test screen. Unfortunately, the delay in data-transmitting could not be rectified and was only recognized by subjects being familiar with Mario Kart. Subjects playing Mario Kart for the first time reported that the controlling was very sensible and it was hard to use the nun-chuck and the controller at the same time. Regular players missed the music of the video game. It was off in order to avoid distraction. In the game, a warning signal normally rings if another racer or a shell comes from behind. This sensitizes drivers for dangers and catches the drivers' attention. It would be interesting whether a difference in scanpath behavior could be observed by hearing and non-hearing of the

sounds from the surrounding. For further experiments e.g. in driving simulators, a comparison of scanpaths recorded with and without sounds of the surroundings could be done.

6 Conclusion

In the present study, the applicability of seven algorithms being based on different scanpath representations was evaluated for static and dynamic visual tasks. For the evaluation of static tasks, a visual *Conjunction Search Task* was used. This is a relatively simple test where all parameters that could modulate scanpath shape and complexity are controllable. In interactive tasks like playing the video game *Mario Kart*, the intentions and emotions of the subjects have more influence on task performance which makes it much harder to control. Four of the scanpath comparison metrics under test scored well in static and dynamic tasks: *MultiMach*, *iComp*, *SubsMatch* and *The Hidden Markov Model*. *ScanMatch* was not able to detect differences in scanpath similarity in dynamic tasks. *Eyenalysis* did not fulfil the expectations, both, in the visual search task and in the interactive virtual environmental set up. An enhancement of this algorithm would be desirable.

The analysis of eye movement data which were composed of the objects positioned in space and time as well as space and time as described by Andrienko et al [2], need to be simplified. In this field, further research is requiered to enhance algorithms for the evaluation of dynamic tasks. Researchers should be aware of the influence of the task and stimulus set as well as of the subjects' understanding on decoding accuracy by planning an experimental set up as already demanded by Borji and Itti [5].

In future, technical progress will make eye tracking devices faster and more precise in eye movement detection. The handling will be easier and the devices will be affordable and comfortably portable like *google glasses*. The usability of eye tracking reaches from medical and academic research to active applications (e.g. for device control) and passive applications (e.g. improvement of web designs). Persons with visual field defects will profit from the medical field of eye tracking: Compensatory head and eye movements can be detected and scanpaths can actively be influenced by gaze guiding. Therefore, further algorithms as presented in the study on hand are necessary to validate gaze movements accurately and need to be automated to ensure objectivity.

References

[1] N. C. Anderson, F. Anderson, A. Kingstone, and W. F. Bischof.
A comparison of scanpath comparison methods.
Behav Res, (2008), 2014.

[2] G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel.
A conceptual framework and taxonomy of techniques for analyzing movement.
Journal of Visual Languages & Computing, 22(3):213–232, June 2011.

[3] F. Baluch and L. Itti.
Training top-down attention improves performance on a triple-conjunction search task.
PloS one, 5(2):e9127, Jan. 2010.

[4] T. Blascheck and K. Kurzhals.
State-of-the-art of visualization for eye tracking data.
Proceedings of the Eurographics Conference on Visualization, page 20, 2014.

[5] A. Borji and L. Itti.
Defending Yarbus: eye movements reveal observers' task.
Journal of vision, 14(3):29, Jan. 2014.

[6] R. Caldara and S. Miellet.
iMap: a novel method for statistical fixation mapping of eye movement data.
Behavior research methods, 43(3):864–78, Sept. 2011.

[7] Y. Chen, M. Nascimento, B. C. Ooi, and A. Tung.
SpADe: on Shape-based Pattern Detection in Streaming Time Series.
IEEE 23rd International Conference on Data Engineering, pages 786–795, 2007.

[8] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist.
ScanMatch: a novel method for comparing fixation sequences.
Behavior Research Methods, 42(3):692–700, Aug. 2010.

[9] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist.
It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach.
Behavior research methods, 44(4):1079–100, Dec. 2012.

[10] H. Diepes.
Refraktionsbestimmung.
DOZ-Verlag Heidelberg, 3. edition, 2004.

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

[12] A. T. Duchwoski.
Eye Tracking Methodology - Theory and Practice.
Springer London, 2007.

[13] J. Epelboim, R. Steinman, E. Kowler, M. Edwards, Z. Pizlo, and C. Erkelens.
The function of visual search and memory in sequential looking tasks.
Vision Research, 35(23-24):3401–22, 1995.

[14] R. M. Foerster and W. X. Schneider.
Functionally sequenced scanpath similarity method (FuncSim): Comparing and evaluating scanpath similarity based on a task 's inherent sequence of functional (action) units.
Journal of Eye Movement Research, 6(5):1–22, 2012.

[15] R. L. Gregory.
Auge und Gehirn.
Rowohlt Taschenbuch Verlag GmbH, Reinbek bei Hamburg, deutsche e edition, 2001.

[16] F. Grehn.
Augenheilkunde.
Springer Medizin Verlag Heidelberg, 29. edition, 2006.

[17] J. Hain.
Statistik mit R.
RRZN, 3. edition, 2014.

[18] J. Heminghous and A. Duchowski.
iComp: a tool for scanpath visualization and comparison.
In *Proceedings of the Symposium on Applied Perception in Graphics and Visualization (APGV)*, 2006.

[19] H. Jarodzka, K. Holmqvist, and M. Nyström.
A Vector-based, Multidimensional Scanpath Similarity Measure.

Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, 1(212):211–218, 2010.

[20] T. Kübler, E. Kasneci, and W. Rosenstiel.
SubsMatch: scanpath similarity in dynamic scenes based on subsequence frequencies.
ETRA '14 Proceedings of the Symposium on Eye Tracking Research and Applications, pages 319–322, 2014.

[21] V. I. Levenshtein.
Binary Codes capable of correcting deletions, insertions, and reversals.
Soviet Physics-Doklady, Vol. 10(No. 8):707–710, 1966.

[22] B. Machner, A. Sprenger, D. Kömpf, T. Sander, W. Heide, H. Kimmig, and C. Helmchen.
Visual search disorders beyond pure sensory failure in patients with acute homonymous visual field defects.
Neuropsychologia, 47(13):2704–2711, Nov. 2009.

[23] M. S. Magnusson.
Discovering hidden time patterns in behavior: T-patterns and their detection.
Behavior Research Methods, Instruments, & Computers, 32(1):93–110, Mar. 2000.

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

[25] F. H. Netter, A. Brass, and R. V. Dingle.
Nervensystem I - Neuroanatomie und Physiologie.
Georg Thieme Verlag Stuttgart, 5. edition, 1987.

[26] D. Noton and L. Stark.
Scanpath in saccadic eye movements.
Vision Research and Science, 11(9):929–942, 1971.

[27] D. Noton and L. Stark.
Scanpaths in Eye Movementes during Pattern Perception.
Science, 171(3968):308–311, 1971.

[28] R. J. Peters and L. Itti.
Applying computational tools to predict gaze direction in interactive visual

environments.
ACM Transactions on Applied Perception, 5(2):1–19, 2008.

[29] The Eye Tribe ApS.
What is eye tracking - basics.
Copenhagen. Online: http://dev.theeyetribe.com/general/, date of access: 2014-08-03.

[30] The Eye Tribe ApS.
The eye tribe - products - the tech specs.
Copenhagen. Online: https://theeyetribe.com/products, date of access: 2014-10-02.

[31] S. Theodoridis and K. Koutroumbas.
Pattern recognition.
Elsevier/Academic Press, 3. edition, 2006.

[32] H. Toutenburg, M. Schomaker, M. Wissmann, and C. Heumann.
Arbeitsbuch zur deskriptiven und induktiven Statistik.
Springer, 2. edition, 2009.

[33] A. M. Treisman and G. Gelade.
A Feature-Integration Theory of Attention.
Cognitive Psychology, 12:97–136, 1980.

[34] Wikipedia.
Mario Kart.
Online: http://de.wikipedia.org/wiki/Mario_Kart, date of access: 2014-10-03.

[35] Wikipedia.
Hidden Markov Model (HMM).
Online: http://de.wikipedia.org/wiki/Hidden_Markov_Model, date of access: 2014-10-16.

[36] Wikipedia.
Top-down und Bottom-up.
Online: http://de.wikipedia.org/wiki/Top-down_und_Bottom-up, date of access: 2014-10-29.

[37] Wikipedia.
False discovery rate.
Online: http://en.wikipedia.org/wiki/False_discovery_rate, date of access:

2015-01-09.

Acronyms

acronym	meaning
AOI	Area of Interst
df	Degree of Freedom
FDR	False Discovery Rate
FSD	Factor of Search Duration
FuncSim	functionally sequemced scanpath similarity method
H_0	Null hypothesis
H_1	Alternative hypothesis
HHM	Hidden Markov Model
MDS	Multidimensional Scaling
N III	Oculomotor Nerve
N IV	Trochlear Nerve
N VI	Abducens Nerve
OOI	Object of Interest
PPRF	Paramedian pontine reticular formation
riMLF	Rostral interstitial nucleus of medial longitudinal fasciculus
RMS	Root Mean Square
ROI	Region of Interest
s	Seconds
sc	Sine correction
SIFT	Scale-invariant feature transform
vs	Versus

List of Formal Signs

unit	sign	meaning
—	∘	Degree
—	’	Minutes of arc
ccm	—	Cubic centimeter
cm	—	Centimeter
D	—	Depth
g	—	Gram
H	—	Height
Hz	—	Hertz
nm	λ	Wavelength
m	—	Meter
W	—	Width

List of Figures

List of Tables

Studien-ID: _____

Fragebogen
Machner Test und Mario Kart

Der Fragebogen hilft bei der Auswertung beider Studien und ist freiwillig auszufüllen.

1. Geschlecht
   ☐ männlich ☐ weiblich ☐ keine Angabe
2. Geburtsdatum
   _____
3. Tragen Sie Brille oder Kontaktlinsen?
   ☐ Brille ☐ Kontaktlinsen ☐ keine Korrektion
4. Mit welcher Korrektion werden beide Tests durchgeführt?
   ☐ Brille ☐ Kontaktlinsen ☐ ohne
5. Besitzen Sie ein Smartphone / Tablet?
   ☐ ja ☐ nein ☐ keine Angabe
6. Spielen Sie Spiele wie Jewels oder 2048 oder ähnliche?
   ☐ ja ☐ nein ☐ keine Angabe
7. Wenn Sie Frage 6 mit ja beantwortet haben: Wie lange spielen Sie im Schnitt pro Tag?
   ☐ <1h ☐ 1-2h ☐ >2h
8. Spielen Sie Rennspiele am PC oder an einer beliebigen Konsole?
   ☐ ja ☐ nein ☐ keine Angabe
9. Wenn Frage 8 mit ja beantwortet wurde: Wie lange spielen Sie im Schnitt pro Woche?
   ☐ <1h ☐ 1-3h ☐ >3h
10. Haben Sie schon einmal Mario Kart gespielt?
    ☐ ja ☐ nein ☐ keine Angabe

Table 9: Results of the Questionnaire.

Nr.	Answer		
1	male: 11 (52.4%)	female: 10 (47.6%)	no answer: 0
2	average age: 26.5 ± 4.05		
3	lenses: 9 (43%)	lenses + CL: 6 (29%)	sc: 6 (29%)
4	lenses: 9 (43%)	CL: 4 (19%)	sc: 8 (38%)
5	yes: 19 (90%)	no: 2 (10%)	no answer: 0
6	yes: 12 (57%)	no: 9 (43%)	no answer: 0
7	$<$ 1h: 11 (52%)	1-2h: 1 (5%)	$>$ 2h: 0
8	yes: 8 (38%)	no: 13 (62%)	no answer: 0
9	$<$ 1h: 6 (29%)	1-3h: 2 (10%)	$>$ 3h: 0
10	yes: 14 (67%)	no: 7 (33%)	no answer: 0

Table 10: Order of the Conjunction Search Task used in present study.

Task Nr.	Conjunction									Color			Shape			Targets			Stimuli		
	red			bue			green														
	Δ	□	o	Δ	□	o	Δ	□	o	red	blue	green	Δ	□	o	1	4	8	40	60	80
1										x								x	x		
2															x	x			x		
3								x									x		x		
4						x											x				x
5															x		x				x
6														x		x					x
7								x										x			x
8	x																x			x	
9								x								x				x	
10															x			x		x	
11						x										x				x	
12							x									x				x	
13			x														x			x	
14					x											x					x
15												x						x			x
16													x			x					x
17		x																x			x
18										x						x			x		
19														x			x		x		
20												x					x		x		
21												x				x					x
22														x				x			x
23									x							x					x
24			x														x		x		
25											x					x			x		
26													x			x			x		
27											x							x		x	
28													x				x			x	
29															x	x				x	
30											x						x		x		
31												x				x			x		
32													x					x	x		
33											x					x				x	
34	x																x			x	
35														x		x				x	

Index

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

place, date	signature

J