

SubsMatch 2.0: Scanpath comparison and classification based on subsequence frequencies

Thomas C. Kübler $^1\cdot Colleen \ Rothe^2\cdot Ulrich \ Schiefer^2\cdot Wolfgang \ Rosenstiel ^1\cdot Enkelejda \ Kasneci^1$

Published online: 21 July 2016 © Psychonomic Society, Inc. 2016

Abstract Our eye movements are driven by a continuous trade-off between the need for detailed examination of objects of interest and the necessity to keep an overview of our surrounding. In consequence, behavioral patterns that are characteristic for our actions and their planning are typically manifested in the way we move our eyes to interact with our environment. Identifying such patterns from individual eye movement measurements is however highly challenging. In this work, we tackle the challenge of quantifying the influence of experimental factors on eye movement sequences. We introduce an algorithm for extracting sequence-sensitive features from eye movements and for the classification of eye movements based on the frequencies of small subsequences. Our approach is evaluated against the state-of-the art on a novel and a very rich collection of eye movements data derived from four experimental settings, from static viewing tasks to highly dynamic outdoor settings. Our results show that the proposed method is able to classify eye movement sequences over a variety of experimental designs. The choice of parameters is discussed in detail with special focus on highlighting different aspects of general scanpath shape. Algorithms and evaluation data are available at: www.ti.uni-tuebingen.de/scanpathcomparison.html.

Keywords Scan pattern · Eye tracking · String kernel · Comparison · Eye movements

Thomas C. Kübler thomas.kuebler@uni-tuebingen.de

Introduction

The world around us contains by far too much information to be processed all at once. Therefore, we have adapted a strategy of selective attention, i.e., visual processing is concentrated on currently relevant objects and withdrawn from less important ones. Further limitations arise due to the anatomy of the visual system. The human eye is a foveated system, where optimal visual perception is only possible at a small, central area of the retina, which is known as the fovea.

For this reason, already in 1931, Harry Moss Traquair describes our field of view as an island of vision in a sea of darkness (Traquair, 1931): Visual acuity drops rapidly as we move from the fovea towards the periphery. In fact, 5° from the fovea only 50 % of visual acuity is reached (Jacobs, 1979). Visual perception would therefore not be possible without eye movements. Although we are mostly unaware of them, when viewing a scene, our eyes are constantly moving (performing saccades) to enable the fovea to fixate different parts of the scene. This process of attentional shift by means of eye movements is called a change in overt attention. However, it is possible to attend to the low-resolution, more peripheral regions of the visual field. Such a covert attention shift occurs without performing eye movements, but is usually performed in preparation of a saccade and directed towards the saccade's landing location (Nobre & Kastner, 2013).

This work focuses on eye movement data, addressing therefore only overt attention. Covert attention is included in so far as that it is likely to manifest in a saccade shortly after the covert attentional shift.

Although limitations in processing capacity and the anatomy of our visual system force us to perform eye movements and to shift attention sequentially, we perceive a

¹ Universität of Tübingen, Sand 14, 72076 Tübingen, Germany

² Computer Engineering Department, University of Aalen, Anton-Huber-Strasse 23, 73430 Aalen, Germany

comprehensive impression of our world stitched together from different attentional foci in the brain like pieces of a jigsaw. Focusing on a relevant object in all its detail and maintaining a complete and up-to-date overview of our large visual world are both essential, yet opposing goals. This internal struggle drives our visual exploratory behavior.

The resulting spatiotemporal sequence of eye movements forms a pattern that is known as the visual scanpath (as coined by Noton and Stark (1971a,b) in their scanpath theory, without the implication of direct correlation to cognitive processes).

Although behavioral patterns are manifested in the scanpath, it is typically a challenging task to identify them from individual measurements. For example, a best practice approach for the construction of heatmaps is to average over at least 30 measurements in order to reach convergence (Pernice & Nielsen, 2009). For this reason, reproducing the results obtained by Yarbus is in his famous experiment (where subjects were instructed to look at a painting under different tasks and exhibited typical, task-dependent scanning patterns) has proven a lot more difficult than expected (DeAngelus & Pelz, 2009; Greene et al., 2012).

Therefore, automated methods to compare eye movements and to identify common parts as well as discriminative sequences between groups of scanpaths are required. Up to date, most eye movement analysis approaches are based on time integrated measures, such as the average fixation duration, or the number of fixations directed towards a specific region of interest (ROI). Thus, they ignore one of the most essential features of a scanpath, its sequential nature. While for some tasks the information of gaze density is sufficient, for other applications the exact fixation sequence is essential, e.g., does gaze follow composition lines when viewing fine art, while exploring driving scenarios for potential hazards (Tafaj et al., 2013), identifying characteristic visual exploration patterns of patients with impaired visual field (Sippel et al., 2014; Kasneci et al., 2014b), or in the context of activity recognition (Braunagel et al., 2015).

Scanpath comparison metrics are usually heading at sequences of fixations and saccades. Other movements, such as smooth pursuits, micro-saccades, ocular drifts, and micro-tremor are ignored, since it is difficult to extract them from the eye-tracking signal. Most methods for scanpath comparison disregard even saccades, since visual perception is suppressed during such fast movements. Smooth pursuits are often represented as a quick succession of short fixations and low amplitude saccades.

The first step towards scanpath construction is event detection. Indeed, distinguishing between fixations and saccades is an essential and non-trivial preprocessing step in data analysis, for which several algorithms are known, e.g. (Berger et al., 2012; Kasneci et al., 2015; Tafaj et al., 2012).

Scanpath comparison builds on fixation and saccade data, but can generally be considered a separate analysis step, although some algorithms for scanpath comparison come packed with their own fixation filters.

The aim of this paper is to (1) provide a review on sequence-sensitive scanpath comparison techniques, (2) introduce a new string kernel method for scanpath comparison, and (3) perform an extensive evaluation on eye movements data collected from a broad spectrum of experimental designs.

Furthermore, we show the applicability of the string kernel method to a wide variety of experiments. It is of utmost importance to be aware of the nature of scanpath differences that can be detected by as well as the restrictions of a scanpath comparison algorithm. By addressing several exemplary use-cases (reaching from simple image viewing to complex real-world driving tasks), we aim at exploring the level of generalization of our method to different experimental designs as well as identifying its limitations. Further potential of the proposed scanpath feature embedding and Support vector machine (SVM) classification step is explored by showing how the identification and extraction of discriminative features between groups of scanpaths can be performed.

Review on scanpath comparison methods

Scanpath comparison metrics are mostly designed with a specific application in mind - in fact they are usually published together with a data set to prove their applicability for that use case. However, their level of generalization to other applications remains mostly unexplored. The decision, which algorithm is suited for a specific experimental design is therefore challenging, since it requires detailed knowledge about the internal details of the algorithm. This is probably one of the major reasons why automated scanpath comparison methods have not found broad application yet. Instead, researchers employ simple time-integrating measures, such as the average fixation duration, or the number of fixations directed towards specific ROIs.

We collected a set of representative experiments that cover a wide range of typical eye tracking applications. Furthermore, we hypothesize that the degree of freedom that the subject is allowed to perform during the experiment is an important factor; while some experiments avoid large head movements by showing stimuli on a screen (e.g., in (Borji & Itti, 2014; Wang et al., 2012)), real-world scenarios (e.g., Kübler et al. (2015b) and Turano et al. (2002)) require free movement and interaction with a dynamic environment (Land et al., 1999; Land and Tatler, 2009). These factors also impose consequences for the eye-tracking signal quality and scanpath comparison. In the following, we will first give an overview on sequencesensitive scanpath comparison methods. A comprehensive summary of mostly non-sequential eye tracking metrics can be found in Holmqvist et al. (2011). A good scanpath comparison metric will produce small distances between scanpaths with a similar general shape and timing, and high distances between heterogeneous, spatially and temporally different scanpaths. How this heterogeneity is defined computationally depends mainly on the scanpath model representation chosen by the computational algorithm. Common scanpath representations are strings, probabilistic models (such as Markov chains), or geometrical vectors.

String-based comparison of scanpaths

Scanpaths are mostly represented as a string sequence, i.e., an encoding of the spatial location information of a sequence of fixations as a sequence of letters. A letter is assigned to each fixation based on the ROI that contains the fixation location. An example of such an encoding is shown in Fig. 1: The presented scanpath can be encoded by the first letter of the ROI label, e.g., in the order of the fixations as *MPMIOM* for the left scanpath, and *MPMPOM* for the right scanpath.

Based on such encoding, the scanpath similarity problem can be mapped to the string similarity problem - a well known problem in bioinformatics and spell-checking software. A basic and often employed string similarity metric is the Levensthein distance (Noton and Stark, 1971b), i.e., which represents the minimal number of letter insertions, deletions, and substitutions required to convert one string into the other. In the above example from Fig. 1, just one edit operation (substitution of I by P) is required to map one string to the other.

The more sophisticated Needleman Wunsch algorithm is employed by ScanMatch (Cristino et al., 2010). It can model relationships between ROIs (e.g., spatially close ROIs or ROIs with similar semantic context can be associated with smaller substitution scores). It is also possible to add gaps to the alignment. Up to here the string representation neglected fixation durations, but they can be represented by repeating a letter multiple times. In SubsMatch, Cristino et al. suggest an interval of 50 ms per letter to encode the temporal information (Cristino et al., 2010). To convert the scanpath into a letter sequence they recommend either a ROI-based approach or the employment of a regular grid.

Furthermore, when comparing sequences of different lengths, one has to be aware of the normalization problem, since the longer the sequences, the more edit operations are required to align them. Therefore, the number of edit operations per letter is a good normalization measure - if both sequences are roughly of the same length. For different sequence lengths, normalization is usually done by the length of the longer sequence. However, when comparing a very short sequence to a significantly longer one, there is a chance of achieving a very good alignment score just by placing the short sequence at the most similar segment of the longer one. The likelihood of a good match by chance increases with the difference in sequence lengths. Normalizing the number of edit operations by the longer sequence cannot compensate adequately for this effect. Since it is hard to normalize for this stochastic process, most scanpath comparison methods exhibit insufficient normalization when comparing sequences of different lengths.

An implementation of the Levensthein distance and the Needleman-Wunsch algorithm for the analysis of eyetracking data can be found in the eyePatterns software (West et al., 2006) together with some visualization options.

In another approach, Zangemeister and Oechsner (1996) employed a string comparison method, where saccadic direction are encoded by letters indicating to the compass directions (N, NE, E,...). This method was applied to find differences in visual scanning behavior of hemianopic patients (Zangemeister & Oechsner, 1996). The authors utilized different weights between ROIs and a special weighting of the first fixation. Significant differences in patients' scanning were identified using the Kolmogorov-Smirnov test on the distributions of distances between scanpaths.



Fig. 1 Example scanpaths from a Mario Kart driving game. White circles mark fixations, arrows indicate saccades. ROIs were annotated as colored box overlays for Mario, the position indicator, a power-up item, and the other players

The iComp method (Duchowski et al., 2010; Heminghous & Duchowski, 2006) circumvents the subjective ROI labeling step: fixations are mean-shift clustered, leading thus to a data-driven string conversion by assigning a letter to each fixation cluster in an automated way. The string alignment step requires identical regions of both scanpaths to be labeled by the same letter. Since the meanshift clustering produces slightly different clusters for each scanpath, corresponding clusters have to be identified by intersecting the clusters. A similar procedure was employed by Privitera and Stark (2000) to determine the overlap between automatically generated and human annotated ROIs.

In a similar approach (Over et al., 2006), Voronoi cells are constructed around fixation locations, leading to small cell sizes in densely fixated regions and larger sizes in homogeneously viewed regions.

Feusner and Lukoff (2008) propose a permutation test to determine the significance of differences between scanpath distance distributions. While the proposed permutation test is easy to apply (since it is parameter free and does not require any prior conditions), its sensitivity is limited.

Scanpath comparison based on fixation maps

iMap (Caldara & Miellet, 2011) is based on the comparison of 3D fixation maps (dimensions are x, y, *fixation density*). Random field theory is then applied to statistically test each pixel of two fixation maps against each other and to correct for the multiple testing. Since fixation maps are usually smoothed by a Gaussian filter, neighboring pixels are not independent of each other, which has further implications for the statistical analysis. Later versions of iMap (Lao et al., 2015) use therefore different statistics, such as pixel-wise linear mixed effect models and a bootstrapping approach. This leads to an increase in statistical power so that even more subtle effects can be detected.

The main advantage of heatmap over string-based methods is that no semantic a-priori expectations of the experimenter are introduced to the data analysis process, since no ROI annotation is required. Attentional constraints can be integrated into the fixation map, e.g., by weighting the Gaussian by fixation duration or modifying the spread of the Gaussian to the extend of the fovea region or the eye-tracker accuracy.

However, a fixation map is unable to represent the temporal dimension and order of a scanpath. Subsequent fixations to the same location will simply add up and give a similar impression as one continued, long fixation towards the region.

There are various fixation map comparison approaches and some even incorporate a time dimension to a limited degree (e.g., by considering several fixation maps from different time slices), but only few like (Leonards et al., 2007) provide a robust statistical testing.

Geometric scanpath comparison

Given the numerical output format of most eye trackers, one of the probably most straight-forward approaches to scanpath comparison is based on their geometrical representation. In this case, the problem of scanpath comparison can be defined as finding an optimal mapping between fixation locations in both scanpaths by matching the closest neighbor fixations, as done by the Mannan distance (Mannan et al., 1996). This approach requires no ROI annotation and is easy to implement. However, the devil is in the details: should we allow to match several fixations of one scanpath to just one fixation of the other scanpath? Eyenalysis (Mathôt et al., 2012) for example, performs a double mapping to circumvent this problem (see Fig. 2). How do we proceed with scanpaths of different lengths, where some fixations do not have a matching partner in the other scanpath? How to include the time dimension? How to choose the distance/neighborhood of a fixation? The Eyenalysis authors leave open, which features of fixations to use and how to weight them, and thereby, formulate a very general measure. As an extension to the Mannan distance, they normalize by the length of the longer sequence and suggest to include a time stamp in the feature vector in order to achieve sequence sensitivity.

The MultiMatch algorithm (Dewhurst et al., 2012; Jarodzka et al., 2010) is a more sophisticated vector-based method. It produces scanpath distances in various dimensions, such as location and duration. Therefore, fixations are converted to a vector representation by first simplifying



Fig. 2 Double mapping between two scanpaths S_1 and S_2 represented by their fixations (*white and gray circles*). In the first step, for every fixation of S_1 the spatially closest fixation of S_2 is determined (*solid arrows*). This procedure is then repeated the other way round - each fixation in S_2 is assigned to its closest neighbor in S_1 (dotted arrows). Adapted from Mathôt et al. (2012)

the scanpath shape (i.e. deleting small saccades and merging subsequent saccades towards the same direction), then choosing representative values such as the location and the fixation duration as vector dimensions. An optimal mapping of fixations is determined by the Dijkstra algorithm, finding the shortest path through a fixation vector similarity matrix. The method conserves fixation sequence. Scanpath difference distributions can be statistically tested by the Kolmogorov-Smirnov test.

FuncSim (Foerster and Schneider, 2013) splits a task into different functional units (subtasks) and performs comparisons only within the same functional unit. This way, the normalization problem for different scanpath lengths does not occur. A plausible way to segment a task into subtasks and a way to label the data (automatically or manually) is required. Saccade length, direction, fixation duration, and spatial characteristics of the scanpath can be modeled. The splitting into functional units represents a sequence conservation, but the algorithm can also include fixation durations in its alignment step. Additionally, FuncSim provides a similarity score baseline by comparing to the similarity of a scanpath to its permuted derivative, making the scores statistically meaningful.

Probabilistic scanpath comparison

Normal variability in scanpaths may be high and mask the subtle but existing differences between scanpaths. Probabilistic representations can handle this by learning the level of variability in scanpaths and comparing against the level of variability found between two specific scanpaths. An early mentioning of the transition matrix approach based on fixation location can be found in Ellis and Smith (1985). A detailed description will be given in Section 2. One advantage of this approach is that transition frequencies can be normalized for each scanpath separately, resulting thus in a good normalization for different sequence lengths. To better cover general scanning patterns, changes in saccadic directions (N, S, E, W, NE,...) instead of local features were later proposed (Ponsoda et al., 1995). The authors mention the necessity of dimension reduction (i.e., deleting seldom occurring directions from the transition matrix), especially when employing statistical testing (e.g., χ^2 test) on the transition matrices.

A similar approach to transition matrices is SubsMatch (Kübler et al., 2014b). The algorithm builds a string representation of the scanpath but splits it into smaller subsequences. The frequencies of these subsequences are then compared between scanpaths. This is basically an extension of the transition matrix approach to more than just one transition. SubsMatch constructs the scanpath string by binning the data by percentiles. Therefore, the number of occurrences of each letter in the scanpath representation is

the same, resulting thus in an efficient spatial resolution usage.

Another representative of this group is the Markov chain/model (e.g., Engbert and Kliegl (2001)). States of the model are typically represented by ROIs and state transitions are calculated as the transition probability between the different ROIs (Kanan et al., 2014).

Mast and Burmeister (Mast & Burmester, 2011) employ t-pattern detection (Magnusson, 2000) in order to find repetitive scanning patterns. Based on a statistical critical interval test, repetitive sequences can be found, even if there is a constant time-delay contained in the pattern (similar to an alignment with a sequence of gaps/substitutions inside). Short t-patterns can be then combined to longer, more complex ones.

Most of the aforementioned algorithms interpret scanpath comparison as the problem of finding one similarity value between the scanpaths. But different experimental factors may have a huge influence on this measure, masking thus the effect of other factors. To evaluating the goodness of such a measure, it is necessary to test for significant differences in the similarity within and between scanpath groups. However, this measure does not imply anything about actual group separability. Therefore, comparing the performance of these algorithms with the classification method proposed here is not adequate and we will restrict the comparison to other scanpath classification approaches.

String Kernel construction

In this section, we will describe the construction of a new string kernel approach for scanpath comparison and its application to scanpath classification. First, the scanpath is encoded as a string and spliced into short subsequences. The frequency of specific subsequences that resemble typical, repeatedly occurring behavioral patterns is then used as a feature of similarity.

The general idea and concept can be derived from the transition matrix approach: a transition matrix (Ponsoda et al., 1995) contains the number of transitions from one ROI to another. An example transition matrix for the two scanpaths visualized in Fig. 1 is shown in Table 1. The transition matrix for the left scanpath was constructed by starting at the first fixation, directed towards Mario and noting the transition from Mario to the position indicator by adding one to the corresponding entry (row 1, column 2) in the transition matrix. We continue to add the remaining transitions for the whole scanpath.

In the next step, transition count matrices are normalized (Table 2), since longer scanpaths would naturally lead to a higher number of transition counts for all transitions. In this work, however, we are only interested in the general

 Table 1
 Transition count matrices for the two scanpaths MPMIOM (left) and MPMPOM (right) from Fig. 1

From \To	Mario	Position	Item	Others
Mario	0	1	1	0
Position	1	0	0	0
Item	0	0	0	1
Others	1	0	0	0
Mario	0	2	0	0
Position	1	0	0	1
Item	0	0	0	0
Others	1	0	0	0

exploratory patterns. Therefore, we need to compare scanpaths of different lengths with regard to their shape. L2 normalization is achieved by making each transition matrix sum up to 1, with each entry corresponding to a transition frequency. Such transition frequencies can easily be embedded as a Markov chain.

n-gram features

Note that exploratory gaze patterns often consist of gaze sequences of more than just two subsequent fixations. A typical example is the glance to the wing-mirrors and over the shoulder while driving or changing the lane, e.g., in the order windshield-side, mirror-side, and window-windshield. In order to capture such patterns, we need to look at more than just two subsequent transitions. In the scenario presented in Fig. 1, there is a pattern of looking towards Mario, the position indicator, and back towards Mario in both of the exemplary scanpaths.

While there is no theoretical limitation regarding the number of subsequent transitions that can be considered, there are practical implications, since the possible number of unique transition patterns increases dramatically with the pattern length. A scanpath of length l encoded with an

Table 2 Normed transition frequency matrices for the abovescanpaths

From \To	Mario	Position	Item	Others
Mario	0	0.2	0.2	0
Position	0.2	0	0	0
Item	0	0	0	0.2
Others	0.2	0	0	0
Mario	0	0.4	0	0
Position	0.2	0	0	0.2
Item	0	0	0	0
Others	0.2	0	0	0

alphabet of a unique letters can contain a theoretical number a^l combinations of letters. However, actually occurring patterns are limited to a fraction of the theoretically possible patterns, thus making them computationally tractable. Furthermore, we are specifically interested in patterns that occur more than once (representing a typical behavioral pattern), and will therefore choose the parameters in a way to reduce the number of occurring patterns to produce a certain overlap.

An efficient method for the calculation of such scanpath subsequences and their frequencies was suggested in SubsMatch (Kübler et al., 2014b). With the following classification step in mind, we call this subsequence frequency calculation a conversion to the *n*-gram feature embedding of the scanpath, with *n* being the length of the subsequences. Such a feature embedding can be performed in linear time. An easy to use implementation can be found in the Sally tool (Rieck et al., 2012). The tool associates each letter with one dimension in the vector space of the embedding. Each feature is indexed by a hash function to create the usually large and sparse matrices of feature counts efficiently. These feature counts are then normalized to represent feature occurrence frequencies.

The scanpath comparison step proposed by Kübler et al. in SubsMatch (Kübler et al., 2014b) is based on a subsequence histogram comparison and suffers from the following limitations: subsequences that are frequent in two groups of scanpaths, but not discriminative between the groups, may have a major influence on the scanpath similarity score. While the absolute difference in transition frequencies for these patterns may be large for individual scanpaths, the difference relative to their likelihood can be quite small. Furthermore, there is no notion of similar subsequences. Two subsequences that differ only in one fixation are treated as just as different as completely different subsequences. Therefore, typical histogram comparison methods such as Earth Movers Distance (Rubner et al., 1998) cannot easily be employed. Finally, the SubsMatch approach can only produce a difference score that sums over all effects present in the scanpath. When comparing scanpaths with multiple factors that may have an effect on scanpath similarity, the strongest effect is likely to overwhelm all others.

String conversion

For many applications, labeling ROIs is not an option, either because there is no unequivocal region (e.g., when viewing abstract art) or in case of an interactive, dynamic scenario. While the annotation of a video frame-by-frame is already an annoyingly and time-consuming task, participants viewing the video will all see the same stimulus image at a given point in time. For scenarios that interact with the user, such as the Mario Kart game, each participant will be confronted with an individual stimulus. ROIs would therefore require separate labeling for each participant.

There have been approaches to semi-automatically determine interesting objects and to track them Kübler et al. (2014a), however they are not generically applicable. We already mentioned approaches to cluster fixation locations in order to determine ROIs from the data (Heminghous and Duchowski, 2006).

In this work we employ two different approaches: a regular grid over the stimulus as used by Cristino et al. in ScanMatch (Cristino et al., 2010) and the percentile mapping utilized by Kübler et al. in SubsMatch (Kübler et al., 2014b). The differences between these approaches are visualized in Fig. 3. The percentile mapping approach should be able to compensate for calibration offsets, drift and small differences between experimental runs, while losing some sensitivity versus the regular grid approach in very controlled experimental conditions.

Discriminative features

There are many scenarios where we want to infer more than just a scanpath similarity score from the data, e.g., Foerster et al. (2011) compared between-session, between-subject and random baseline similarity by applying a statistical test on the similarity scores. Alternatively to these posthoc tests, the challenge of identifying scanning patterns associated with specific experimental factors can be tackled by applying machine learning techniques. In this work, we employ a SVM with a linear kernel (using liblinear (Chang and Lin, 2011)). SMVs can learn differences in



Fig. 3 Forming regular, gridded bins (*bottom*) over the whole data range versus determining data percentiles (*top*) and assigning fixations to the percentile bins. Bins are marked by the background intensity and assigned fixations by their fill intensity

features of two different classes, e.g. differences in the scanpath features between two different experimental factors. For the linear kernel the SVM learns feature weights that represent the discriminative power between two classes. When applying these weights to a new scanpath feature vector, the SVM assigns the scanpath into one of the learned classes.

We are also able to extract feature weights from the SVM after learning. This way, a ranking of subsequence frequencies in terms of their discriminative power can be performed.

The method was evaluated by computing the classification accuracy of the SVM, the percentage of correctly assigned class labels. We performed a 10-fold crossvalidation, during which the scanpaths used for training and testing were permuted. The 2-regularized L2-loss support vector classification solver of the dual problem was used and the SVM cost parameter was not optimized, except for adjustments for unequal group sizes (unless stated otherwise). Compensation for unequal group sizes was done by adjusting the misclassification penalties relative to the inverse proportion of the group label in the data set. A misclassification of a less frequent label would therefore be penalized more than a misclassification of frequent label. A cost parameter optimization was not done since we are already optimizing over a range of n-gram lengths (1-10) as well as alphabet sizes (2-26) and five different string encodings. Optimizing over more parameters would require a validation data set to avoid over-fitting. An adequate sample size for the validation is not given for most of the data we analyzed.

Since the scanpath groups in our data were not necessarily of equal size, chance level for the classification accuracy can be determined by always guessing the majority class label.

Experimental settings and data

As revised above, there is a rich body of scanpath similarity measures that have been published so far. However, when to apply which measure remains unclear. In fact, little is known about the sensitivity of individual algorithms to specific experimental factors, nor is their sensitivity to different sources of noise explored. Some of the methods were evaluated on hours of dynamic real-world tasks, others on static laboratory conditions with only some seconds of data per trial, some even exclusively on simulated data. Furthermore, temporal resolution of the eye-tracking data varies from 25 to 1000 Hz across applications. Considering the enormous implications that these factors impose on data quality (pupil detection, calibration accuracy, number of samples per saccade) and content (less than ten fixations

Table 3	Three replications of the	Yarbus experiment that	were used	for evaluation	01
---------	---------------------------	------------------------	-----------	----------------	----

	subjects	different tasks	stimuli	viewing time [s]	eye-tracker [Hz]
OWN	20	2	2 paintings	30-120	EyeTribe (30 Hz)
RY (Greene et al., 2012)	17	4	20 grayscale images	60	SR Research
DY (Borji & Itti, 2014)	21	7	15 natural scenes	30	Eyelink (1000 Hz)

versus thousands per recording), no method can currently claim to cover the whole range of applications.

Short trial durations and repeated stimulus display in a laboratory setting is likely to lead to highly similar scanpaths and few noise. On the other hand, real-world experiments are always associated with a high level of pupil detection failures (a review can be found in Fuhl et al. (2016)) and identical experimental conditions are hard to reproduce (e.g., the same amount traffic while driving). Scanpaths of such experiments are likely to be more dissimilar and noisy.

The aim of this evaluation section is to give an idea of the applicability of the proposed measure and to quantify the influence of experimental parameters on the results. To approach this, we chose a set of experiments with representatives for certain aspects of typical eye-tracking experiments. To show the level of generalization of our proposed measure over different settings as well as its limitations, we evaluate the classification power and the discriminative quality of the feature vectors.

In the following, we first introduce the eye-tracking experiments in the order of their complexity.

Yarbus: image viewing with different tasks

We performed an evaluation on three replications of the classic Yarbus experiment (Table 3).

Our replication (further called OWN) contains recordings of 20 subjects (age range 20-57, 6 male and 14 female). All subjects viewed Ilya Repin's painting *The unexpected visitor* in two different settings: free-viewing (without a specific task requirement) and estimating the age of the people in the painting. We further presented a similar image by Ilya Repin with the free-viewing task and varied the viewing time on the original image (Fig. 4). The complete experiment protocol is shown in Fig. 5. Eye-tracking data of 19 out of 20 subjects was of sufficient quality to be included in the evaluation. Ten subjects wore glasses, one contact lenses during the measurement.

Greene et al. (2012) previously published eye-tracking data (further called RY) of 17 observers performing 4 tasks (memorize, determine the decade when the picture was taken, estimate how well the people know each other, estimate the wealth of the people) on 20 different grayscale images for 60 s viewing time. Four to five observers viewed

the same image with the same task. Greene et al. reported that replicating Yarbus' findings is harder than expected. In response to the Green et al. work, Borji and Itti (2014) showed that it is in fact possible to determine the observer task from the scanpath above chance level. In addition, Borji et al. provided another data set (further called DY) containing recordings of 21 subjects performing 7 different tasks on 15 images of natural scenes (free-viewing, estimate wealth, age, what was the family doing before the arrival, memorize the clothes, memorize positions of people and objects, estimate how long the visitor had been away). The block design resulted in 3 observers viewing the same image with the same task assignment.

There are further studies on this topic, but their data is not publicly available (DeAngelus & Pelz, 2009; Haji Abolhassani & Clark, 2014). Therefore, a direct comparison of our results to their performance is not possible.

Conjunction search task

By performing a visual search task according to Machner et al. (2005), subjects have to count all stimuli of a specific *color* (red, blue, green), all stimuli of a specific *shape* (square, triangle, circle), and the conjunction of both features, namely all stimuli of a specific *shape and color*. For these three different tasks, the number of total stimuli (targets plus distractors) and the number of target objects can be varied (see Figs. 6 and 8), resulting thus in different difficulty levels.

This study is of particular interest for scanpath comparison, since it involves abstract and quite simple stimuli. As for the Yarbus experiment, different tasks can be



Fig. 4 The painting *Unexpected* by Ilja Repin (*left*) and a similar, earlier version (*right*) were used as stimuli



Fig. 5 Stimulus presentation and task order in our own experiment replicating parts of Yarbus' original experiment

distinguished from the viewing pattern. Additionally, a measure of task difficulty is available by considering the number of counting errors that occur.

In this work, we looked at the association between task, task difficulty, and scanpath shape.

The study was performed on a 24 display of 1920×1080 pixel resolution. 21 subjects (age range 22 to 43 years, average 26.5 ± 4.05 , 10 female and 11 male) participated and completed 35 different stimuli, resulting in 735 individual trials. Nine subjects wore glasses, four wore contact lenses. Subjects counted target objects by clicking the left mouse button and could move on to the next stimulus display with the right mouse button. Data was recorded by means of an EyeTribe (Dalmaijer, 2014) tracker at 30 Hz using a chin rest to minimize calibration errors due to head movements. 51 out of 735 trials were discarded due to insufficient eye tracking (conj. 256, color 193, shape 235). Data included in the analysis has a minimal tracking rate of 83 % (median 98 %).

Mario Kart video game

The video game is interesting in terms of eye movements, since it allows to study the effect of dynamic, interactive stimuli under laboratory conditions. Some seconds after the game starts, no subject encounters the same stimulus screen. Even when replicating the same experiment, the racing game will produce very different settings depending on user interaction with the game and some random effects. More specifically, we examine:

- 1. performance in terms of average lap time.
- 2. two different routes and how their different attentional requirements influence scanpath shape.

Data of 21 subjects (identical to those of the conjunction search task) was recorded using an EyeTribe (Dalmaijer, 2014) tracker at 30 Hz and a chin rest. Median tracking rate was 91 %. Subjects drove three laps per route on two different routes. A Nintendo Wii console and controller were employed.

Driving with visual field defects

20 patients with blind areas in their visual field (related to binocular glaucoma or homonymous visual field defects) and a healthy sighted control group wore a mobile 30 Hz Dikablis eye-tracker by Ergoneers GmbH during a \sim 40 minutes on-road driving session in real traffic. Details on data collection can be found in Kasneci et al. (2014a), Kübler et al. (2015a), and Kübler et al. (2015b). Aim of these studies was to investigate the relationship between binocular visual field defects, eye movements, and driving performance.

In fact, previous work on this topic often mentions the hypothesis of compensatory eye/gaze movements that allow for the compensation of the visual field defect and lead to a successful driving test outcome despite the visual field loss (Kasneci et al., 2014a; Kübler et al., 2015b). By moving their eyes or head, subjects affected by a visual



Fig. 6 Conjunction search task with different conditions: Counting all green circles requires different effort depending of the number of green circles actually present as well as on the number of distractor objects

field defect are able to shift the blind area of their visual field. In this work, we look at the scanpaths of the subjects and search for evidence of this hypothesis by identifying the above mentioned exploratory patterns.

More specifically, we examine:

- 1. driving fitness of each subject as judged by a driving instructor who was blind to the specific health status of the subject.
- 2. the effect of a visual field loss on exploratory gaze behavior.

Preprocessing of eye-tracking data

Eye-tracking data preprocessing is essential to ensure good data quality, which in this work is measured in terms of accuracy and the percentage of tracking failures. Such failures mostly occur due to algorithmic flaw in detecting the pupil center in the eye tracker's image. Main error sources include changing lighting conditions, reflections on the eyeglasses worn by the subject and dark make-up. If a tracking failure occurs, no reliable statement about the current gaze direction can be made. Tracking failures can be corrected if the eye image is recorded by manually marking the pupil center in the image. However, this process is very time consuming. Tracking accuracy on the other hand is associated with the calibration step, during which a mapping of pupil center coordinates towards 3D gaze directions is performed. Over time, calibration accuracy degrades due to positional changes between the tracker device and the eye (for example by displacement of an head-mounted device). Therefore, long recordings without recalibration will usually exhibit a larger error margin in the determined 3D gaze direction than short recordings. Correcting for accuracy problems post-experimentally is not trivial, since one has to find objects that are fixated for sure at a specific point in time. Is such information available, a recalibration is possible.

It should be noted that for the very different experimental designs in this work data quality varies: while the conjunction search task can be performed with constant lighting conditions and regular recalibration was done between the short trials, on-road driving involves regular tracking losses and no opportunity for recalibration.

Data from the costly on-road driving experiments was manually annotated in order to reach a tracking failure rate of less than 20 %. Accuracy was corrected whenever the data analyst found accuracy to be worse than 5°, usually after around 20 minutes of driving. Data from screen-based laboratory experiments (Yarbus, Conjunction search task and Mario Kart) with a tracking loss of more than 20 % was discarded. Fixations and saccades were identified using a Mixture of Gaussians algorithm (Kasneci et al., 2015) with a maximum likelihood fit for the Gaussian distributions. String encoding was performed at 40 ms fixation duration intervals, resembling the 25 Hz recording rate of the Dikablis device and close to the 50 ms per letter suggested by Cristino et al. (2010).

Results

The overall scanpath classification results shown in Table 4, suggest that our method is applicable to most of the experiments and experimental factors. Our results exceed by far the guessing chance baselines. As stated in the Section Methods, some trials were excluded from data analysis for quality reasons, leading to a non-balanced design in some cases. Therefore, the resulting guessing chance, e.g., for a four-class classification may exceed 25 %. In the following subsections, we will report the results for each experiment in detail.

Yarbus: image viewing with different tasks

On our own Yarbus data (OWN) reported in Table 4 we varied task, image, and viewing time in order to demonstrate the influence of each of these conditions on the scanpath similarity measure. A four-class SVM was trained in order to distinguish between the four conditions. The classification results are presented as a confusion matrix in Fig. 7. The diagonal of the matrix shows the percentage of scanpaths with a certain label (e.g., free-viewing), which are classified correctly. The off-diagonal elements represent a misclassification, e.g., a free-viewing task scanpath is wrongly classified as an age task by the SVM in 29.41 % of the cases. The confusion matrix reveals that the extremely long viewing time of three minutes has the strongest influence on the scanpath measure, i.e., scanpaths of category free-viewing (3 min) can be assigned correctly to their category in 93 % of the cases. We examined this effect by looking at heatmap visualizations of the data for each subject and found that there were two ways of dealing with the very long viewing time: subjects get bored after some time and start to explore more subtle details of the painting or they just stare at one point. This effects causes the scanpaths to be easily distinguishable from the other categories.

We can also find an effect for the task and the displayed image: scanpaths of the free-viewing task and the age estimation task can be classified to the correct group above chance level (53 % and 59 %, Fig. 7), but 29 % of freeviewing scanpaths are misclassified as age estimation tasks. Task classification is working, but seems to be a hard task,

Experiment	Research Question						Baseline	Best
	OWN: condition	58.5	50.8	66.2	58.5	43.1	26.2	66.2
	RY : task	30.3	32.1	30.9	28.8	17.4	25.0	32.1
	RY : image	19.1	17.1	8.5	10.6	3.2	5.0	19.1
	DY : task	21.6	17.8	18.7	19.0	18.7	14.3	21.6
	DY : image	16.8	14.0	49.8	22.2	16.5	6.7	49.8
<pre></pre>	Search task	61.6	61.7	57.5	54.2	54.5	37.2	61.7
•	# distractors	27.2	27.1	27.1	26.7	27.9	33.3	27.9
	# targets	45.8	45.5	51.5	46.6	48.2	45.8	51.5
	Stimulus	8.2	7.2	17.3	14.3	8.7	2.9	17.3
	Lap time route1	76.2	81.0	76.2	76.2	85.7	61.9	85.7
	Lap time route2	76.2	81.0	71.4	76.2	85.7	61.9	85.7
0	Route1/Route2	92.9	71.4	88.1	73.8	76.2	50.0	92.9
	Driving fitness	77.8	69.4	75.0	63.9	69.4	69.4	77.8
	Patient/Control	77.8	72.2	69.4	58.3	75.0	58.3	77.8
	Fitness & Glaucoma/Hemi	25.0	33.3	22.2	22.2	25.0	22.2	33.3
Statistically significan horizontal and vertical	it results ($p < 0.05$) of a binomial test (axis percentile binning, horizontal and w	without correction f ertical axis regular b	or multiple testing inning. The differe) are reported in build that the two setting on the setting the two setting th	old. Different string ants are characterize	g encoding pattern od in Table 3	s are displayed: circul	ar encoding

 Table 4
 Classification accuracies (10-fold cross-validation) for all experiments and research questions

 $\underline{\textcircled{O}}$ Springer



Fig. 7 Confusion matrix of the Yarbus (OWN) experiment with percentile binning in vertical direction. The largest influence originates from the long viewing times. However, also task and image exhibit an effect that allows to classify many of the scanpaths correctly

considering the relatively high error rate. When we compare this to the presentation of an alternative image, the confusion with the original image is smaller, suggesting a strong effect of the image on scanpath shape.

In order to verify these effects, we applied our string kernel method to data provided by Greene et al. (2012) (RY) and Borji and Itti (2014) (DY). Both data sets allow us to compare the string kernel method with methods from related work. The authors of these data sets used them for task-from-scanpath classification, so we are able to compare directly to their results.

On the (RY) data, chance level is at 25 %. Borji et al. reach a classification performance of 34.1 % using a Boosting classifier, whereas (Kanan et al., 2014) reach 33 % using a SVM on the parameters of a Hidden Markov Model. Our approach performs similarly, reaching a classification accuracy of 34.4 % with SVM parameter optimization and 32.1 % without optimization (Table 4).

On the (DY) data provided by Borji et al., our approach achieves a classification accuracy of 24.2 % with optimization, 21.6 % without optimization, being thus far above the chance level of 14.3 %. Our results support the findings of these studies that scanpaths contain information that allows the prediction of the observer's task. It should however be noted that this holds for the overall classification rate; reliably predicting the task from one individual scanpath still remains a challenging problem.

In a second run, we tried to predict the stimulus image that caused a specific scanpath. The high correct classification rates of up to 50 % are caused by the good spatial resolution of the string kernel method: for large alphabet sizes and a percentile binning approach the method basically performs fixation clustering and thereby learns the positions of relevant objects in each image. The structure of the images, i.e., the axis that separates the relevant objects in the image best, obviously has a major influence on the success of the classification step.

We can thus conclude that the string kernel method is sensitive to the performed task at a similar classification rate as other approaches, but the effect of the stimulus image displayed is much stronger than that of the task performed.

Conjunction search task

The conjunction search task can be performed under standardized conditions and has a well defined setup. Therefore, it is well suited to study the effect of specific experimental factors on the results of the string kernel method. Further we can correlate the scanpath results with a performance measure.

As shown in Fig. 8, the number of fixations performed differs significantly between the three tasks conditions and the three stimuli counts employed in this experiment. Overall the performance measures indicate that it is very easy to perform the pop-out color task, but difficult to distinguish targets by shape. The distribution of task completion times (median conj. 6.6 s, color 3.1 s, shape 8.8 s) as well as the number of counting errors (conj. 3.7 %, color 3.8 %, shape 11.9 %) suggest that the conjunction task of filtering by color and shape together is more similar to the color pop-out task than to the shape task - probably because a large number of stimuli can easily be discarded by their color feature and shape has to be considered only for this reduced stimuli set. These performance results are consistent with the findings of Machner et al. (2009) in their control group.

This effect is also resembled in the scanpath confusion matrix (Fig. 9), showing that the shape task can be classified at a high accuracy of 72 %. The confusion between color and conjunction task is much higher (31 %/33 %) than towards the shape task (11 %/15 %). The tasks that are more

	Kind of Task	# Targets	# Stimuli
Number of Errors	Co <u>nj. Col. Sh</u> ape	1 <u>4</u> 8	40 60 80
Number of Fixations	Conj. Col. Shape	1 <u>4</u> 8 *	4 <u>0 60 8</u> 0
Task completion time	Conj. Col. Shape	<u>*</u> * 1 <u>4</u> 8	4 <u>0 60 8</u> 0

Fig. 8 Kruskall-Wallis tests with false discovery rate corrected pvalues on task performance and scanpath length for the different experimental settings of the conjunction search task. p-values < 0.05are marked by a *



Fig. 9 Confusion matrix for the task classification in the conjunction search task with percentile binning in horizontal direction. The shape task is relatively easy to classify, while color and conjunction get mixed up more often. Averaged over all target types, target counts and distractor counts

similar in terms of task performance also result in more similar scanpaths. The more demanding shape task results in a more unique, distinguishable scanpath. The overall classification accuracy of 62 % supports the finding of the Yarbus task that observer task classification from scanpath is possible, also for the conjunction search task with its much simpler stimulus complexity.

Surprisingly, the string kernel method did not reliably succeed in separating trials of different stimulus count or different target object count from each other. This might be related to the normalizes the transition frequencies in order to account for different scanpath length. Contrary to the Yarbus example with the very long viewing time and subjects getting bored, the duration of the conjunction search experiment was determined by the subject herself. Thus, we can assume that visual search behavior occurred for the whole trial duration. In this case, a long search may appear very similar to a short search process due to normalization. Especially remarkable is the high accuracy (17 % versus 3 % chance level) of assigning gaze recordings to a specific stimulus. Since the same stimuli were presented to all subjects, exact position information is very sensitive and a high spatial resolution (represented by a large alphabet size), can capture this effect, just as for the Yarbus image.

We can conclude that it is possible to detect high level effects such as the task given to a subject, but also to maintain high spatial sensitivity for image classification. Viewing time does not seem to have a relevant influence on scanpath similarity, as long as viewing behavior is similar for the whole duration. This implies scanpath length normalization is working well in eliminating the duration factor from the scanpath.

Mario Kart

60% of the study participants were familiar with the Mario Kart game. In consequence, we found a high correlation of experience and average lap times. In this work, racing game average lap times are therefore split at the 60% marker as shown in Fig. 10 in order to separate *slow* from *fast* drivers.

Separating scanpaths from players with a fast average lap time from scanpaths with a slow average lap time was possible significantly above chance level (86 %). This video game experiment is especially interesting since it involves a dynamic, interactive stimulus: every interaction of the payer with the game (and some random events) will change the game. In consequence, a different image is displayed on the screen. Contrary to the Yarbus and conjunction search experiment, local fixation positions will map to very different in-game objects and ROIs. However, this does not seem to have a negative impact on the separability of the scanpaths.

As for the Yarbus and conjunction search data, we found that different stimuli (here different routes) invoke very different scanning patterns, leading thus to a very high classification accuracy of 93 % as shown in Table 4.



Fig. 10 Mario Kart lap times for both routes split by participants who played the game before and novices

Driving experiments

For the driving experiment, we would expect that patients with visual field defects employ so-called compensatory gaze in order to shift their intact visual field towards potentially relevant areas of the driving scenario. This compensatory mechanism should be reflected in the resulting scanpath, and can therefore be used to distinguish between patients and the control group. However, not all patients are able to utilize these patterns and some of them exhibit a gaze behavior indistinguishable from the control group.

Our results suggest that compensatory gaze patterns can be identified using the proposed method: separating the patient and the control group is possible above chance level at 78 % (Table 4). Using the linear SVM weights, we are able to identify these discriminative gaze patterns that distinguish patients from the control group. These patterns suggest that the frequency of a switch between left and right hemifield is altered between the patient and the control group.

However, predicting driving fitness from the scanpath is not possible above chance level (78 % at 69 % chance level). The correlation between employed gaze patterns and driving fitness seems to be weak and not what we would expect from the compensatory gaze theory.

This finding rather suggests that there might be no such thing as an alteration of gaze behavior that facilitates driving fitness, but rather that there is a break-down of normal gaze behavior in patients that are not fit to drive. This is supported by the confusion matrix shown in Fig. 11. Distinguishing



Fig. 11 Confusion matrix of the on-road driving experiment. The group of patients failing the driving test can be separated from the other groups best. Patients fit to drive exhibit a gaze behavior similar to that of the control group

patients that passed the driving test from the control group seems to be impossible - 67 % of the control group get misclassified as patients that passed the test and 38 % of fit-to-drive patients get misclassified as control group members. However, only few of the fit-to-drive subjects are misclassified as not fit-to-drive patients (13 %/0 %).

Maintaining a normal gaze behavior may be a difficult task for patients with visual field defects and can as such of course be considered a compensatory pattern. However, the often mentioned increase in saccadic length and distribution of attention towards the blind side in hemianopic patients rather suggested an alteration of the viewing behavior of fit-to-drive patients. Of course it is possible that there are compensatory movements that we were unable to capture with our approach.

It should also be noted that we mixed various kinds of visual field defects (glaucoma and hemianopic patients, with both left- and right-sided defects) for this analysis and that a typical compensation routine for a specific defect type may exist, however we were unable to identify it.

The influence of n-gram length and alphabet size

There are certain factors to consider when choosing the n-gram length: The method was designed to examine the frequency differences in n-grams. Therefore, n-grams have to occur multiple times in the scanpath. Choosing a large n will produce many unique n-grams and requires a long scanpath to allow subsequence counts to pile up. On the other hand, a low n value might not be able to capture characteristic patterns. For example, while n = 1 corresponds to simple ROI frequencies, n = 2 represents ROI transition frequencies as in a Markov chain. The higher the n, the more specific the gaze patterns will get. The influence on classification accuracy can be observed in Fig. 12 for the Yarbus experiment.

The choice of the alphabet size is highly stimulus dependent. We observed that for static stimuli and an expected high scanpath similarity a large alphabet is advantageous,



Fig. 12 Classification accuracies for different combinations of alphabet size and n-gram length. In general, larger alphabets require a smaller n, since subsequences get more and more unique

for more abstract tasks and dynamic stimuli with an expected low scanpath similarity a small alphabet size is preferable. For large alphabet sizes the method basically learns image properties, while low alphabet sizes seem to facilitate a better generalization behavior for more abstract factors.

The influence of string encoding

As a general rule of thumb we can conclude that binning, either horizontally or vertically, works well for stimuli presented on a computer screen. The real-world task adds additional complexity, i.e., relatively high measurement inaccuracies, calibration drift, and measurement errors call for a more robust string encoding. Using the data percentile seems to work well, however, one has to be aware of the severe limitations induced by this step. For example, a rightwards bias of one scanpath compared to another one would not be visible in this encoding.

Limitations and considerations

While the proposed method is able to separate scanpath groups for most experiments and experimental factors, results for the static experimental settings (Yarbus and Conjunction search task) are generally better than for the dynamic tasks (Mario Kart and driving experiment).

Our results suggest that a classification of both the viewed material (OWN: screen, RY: image, DY: image, search screen, Mario Kart route) and the performed task (OWN: screen, RY: task, DY: task, search task) is relatively easy. Classification of participants performance, however, is more challenging (patients vs. controls, lap times, driving fitness) either due to the amount of actual change in visual scanning, or the sensitivity of the algorithm to the nature of the change.

Yarbus data and intuition suggests that best separability between classes can be achieved by choosing an encoding that represents the stimulus ROIs in a meaningful way. Manual ROI annotation would be the gold standard and computer vision methods are currently lacking the potential of segmenting ROIs in a meaningful way for all possible experiment designs. An improvement over the current string mapping approach could be achieved by applying a principal component analysis to the gaze data and performing string conversion along the axis of highest variance. However, it is a common misconception that ROIs are necessarily required to represent semantic objects in the viewed scenery. While this is certainly advantageous, we were able to show that a simple grid- or percentile-based letter assignment is often sufficient. An approach similar to the string encoding is the use of shapelets (Rakthanmanon and Keogh, 2013). Shapelets are a one-dimensional representation of a 2d contour. The features constructed by shapelets will be very similar to our substring features, however there is the possibility to construct masked shapelets that are able to tolerate a limited number of mismatches, finally leading to a similar effect as the mismatch kernel. However, this comes at a computational cost.

A similar effect could be achieved by means of string mismatch kernels (Leslie et al., 2004) that allow for a certain degree of variation in the subsequence pattern and can therefore represent a similarity between features. This would however increase the complexity, i.e., for longer patterns not only the actually occurring patterns but also the ones similar to them have to be considered.

On the other hand the possibility to consider all patterns remedies the very sparse feature vectors for short experiments analyzed with a high alphabet and n-gram size. Classification accuracy without considering mismatches might drop to chance level. Allowing for mismatches in the subsequences would make the feature vector less sparse. As an alternative feature selection could be performed to identify the discriminative power of individual features relevant for the classification problem at hand.

By accident a programming error in an earlier version of the algorithm resulted in a unique labeling for the very first *n*-gram in the sequence (a space separating the class label from the fixation sequence was considered as part of the sequence). We found that for short scanning sequences marking the first fixation n-gram uniquely can improve the results. This underlines the importance of the first few fixations in a static context (Zangemeister & Oechsner, 1996).

This effect should also be observable in a global alignment of the sequences as a higher degree of conservation. It was not included in the final version of the algorithm and the evaluation at hand. For long lasting experiments and for studying general scanning patterns, the effect becomes negligible.

Conclusion

We introduced a new scanpath comparison and classification method. In an extensive evaluation the generalization over a vast field of eye-tracking applications, from laboratory conditions to real-world scenarios, was demonstrated. SubsMatch 2.0 is a measure for finding discriminating patterns between groups of scanpaths and for calculating scanpath similarity.

References

- Berger, C., Winkels, M., Lischke, A., & Höppner, J. (2012). Gazealyze: a MATLAB toolbox for the analysis of eye movement data. *Behavior research methods*, 44(2), 404–419.
- Borji, A., & Itti, L. (2014). Defending Yarbus: Eye movements reveal observers task. *Journal of Vision*, 14(3), 29. doi:10.1167/14.3.29.doi
- Braunagel, C., Kasneci, E., Stolzmann, W., & Rosenstiel, W. (2015). Driver-activity recognition in the context of conditionally autonomous driving. In 2015 IEEE 18th International Conference on Intelligent Transportation Systems, (pp. 1652–1657).
- Caldara, R., & Miellet, S. (2011). iMap: A novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods*, 43(3), 864–78. doi:10.3758/s13428-011-0092-x
- Chang, C.C., & Lin, C.J. (2011). LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 27, 1–27.
- Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I.D. (2010). Scanmatch: a novel method for comparing fixation sequences. *Behavior Research Methods*, 42(3), 692–700. doi:10.3758/BRM.42.3.692
- Dalmaijer, E. (2014). Is the low-cost eyetribe eye tracker any good for research? Tech. rep., PeerJ PrePrints.
- DeAngelus, M., & Pelz, J.B. (2009). Top-down control of eye movements: Yarbus revisited. *Visual Cognition*, 17(6-7), 790–811.
- Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). 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. doi:10.3758/s13428-012-0212-2
- Duchowski, A.T., Driver, J., Jolaoso, S., Tan, W., Ramey, B.N., & Robbins, A. (2010). Scanpath comparison revisited. Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA '10 p 219. doi:10.1145/1743666.1743719
- Ellis, S.R., & Smith, J.D. (1985). Patterns of statistical dependency in visual scanning, 9, 221–238.
- Engbert, R., & Kliegl, R. (2001). Mathematical models of eye movements in reading: A possible role for autonomous saccades. *Biological Cybernetics*, 85(2), 77–87.
- Feusner, M., & Lukoff, B. (2008). Testing for statistically significant differences between groups of scan patterns. Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA '08 p 43. doi:10.1145/1344471.1344481
- Foerster, R.M., & Schneider, W.X. (2013). Functionally sequenced scanpath similarity method (FuncSim): Comparing and evaluating scanpath similarity based on a tasks inherent sequence of functional (action) units. *Journal of Eye Movement Research*, 6(5), 1–22.
- Foerster, R.M., Carbone, E., Koesling, H., & Schneider, W.X. (2011). Saccadic eye movements in a high-speed bimanual stacking task: changes of attentional control during learning and automatization. *Journal of Vision*, 11(7), 9–9.
- Fuhl, W., Tonsen, M., Bulling, A., & Kasneci, E. (2016). Pupil detection for head-mounted eye tracking in the wild: an evaluation of the state of the art. Machine Vision and Applications 1–14. doi:10.1007/s00138-016-0776-4
- Greene, M.R., Liu, T., & Wolfe, J.M. (2012). Reconsidering Yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, 62, 1–8. doi:10.1016/j.visres.2012.03.019
- Haji-Abolhassani, A., & Clark, J.J. (2014). An inverse yarbus process: Predicting observers task from eye movement patterns. *Vision research*, 103, 127–142.
- Heminghous, J., & Duchowski, A.T. (2006). iComp: A tool for scanpath visualization and comparison. ACM SIGGRAPH 2006 Research posters p 186.

- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford: Oxford University Press.
- Jacobs, R. (1979). Visual resolution and contour interaction in the fovea and periphery. *Vision Research*, 19(11), 1187–1195.
- Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. In *Proceedings of* the 2010 symposium on eye-tracking research & applications, (pp. 211–218).
- Kanan, C., Ray, N.A., Bseiso, D.N., Hsiao, J.H., & Cottrell, G.W. (2014). Predicting an observer's task using multi-fixation pattern analysis. In *Proceedings of the symposium on eye tracking research and applications*, (pp. 287–290).
- Kasneci, E., Sippel, K., Aehling, K., Heister, M., Rosenstiel, W., Schiefer, U., & Papageorgiou, E. (2014a). Driving with binocular visual field loss? A study on a supervised on-road parcours with simultaneous eye and head tracking 9(2):e87,470. doi:10.1371/journal.pone.0087470
- Kasneci, E., Sippel, K., Heister, M., Aehling, K., Rosenstiel, W., Schiefer, U., & Papageorgiou, E. (2014b). Homonymous visual field loss and its impact on visual exploration: A supermarket study. TVST 3(6).
- Kasneci, E., Kasneci, G., Kübler, T.C., & Rosenstiel, W. (2015). Online recognition of fixations, saccades, and smooth pursuits for automated analysis of traffic hazard perception. In Koprinkova-Hristova, P., Mladenov, V., & Kasabov, N. K. (Eds.) Artificial neural networks, springer series in bio-/neuroinformatics, vol 4, springer international publishing, (pp. 411–434).
- Kübler, T.C., Bukenberger, D.R., Ungewiss, J., Wörner, A., Rothe, C., Schiefer, U., Rosenstiel, W., & Kasneci, E. (2014a). Towards automated comparison of eye-tracking recordings in dynamic scenes. In Visual Information Processing (EUVIP), 2014 5th European Workshop on, (pp. 1–6).
- Kübler, T.C., Kasneci, E., & Rosenstiel, W. (2014b). Subsmatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, (pp. 319–322).
- Kübler, T.C., Kasneci, E., Rosenstiel, W., Aehling, K., Heister, M., Nagel, K., Schiefer, U., & Papageorgiou, E. (2015a). Driving with homonymous visual field defects: Driving performance and compensatory gaze movements. *Journal of Eye Movement Research*, 8(5), 1–11.
- Kübler, T.C., Kasneci, E., Rosenstiel, W., Heister, M., Aehling, K., Nagel, K., Schiefer, U., & Papageorgiou, E. (2015b). Driving with glaucoma: Task performance and gaze movements. *Optometry & Vision Science*, 92(11), 1037–1046.
- Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11), 1311–1328.
- Land, M.F., & Tatler, B.W. (2009). Looking and acting.
- Lao, J., Miellet, S., Pernet, C., Sokhn, N., & Caldara, R. (2015). imap 4: An open source toolbox for the statistical fixation mapping of eye movement data with linear mixed modeling. *Journal of vision*, 15(12), 793–793.
- Leonards, U., Baddeley, R., Gilchrist, I.D., Troscianko, T., Ledda, P., & Williamson, B. (2007). Mediaeval artists: Masters in directing the observers' gaze. *Current Biology*, 17(1), R8–R9.
- Leslie, C.S., Eskin, E., Cohen, A., Weston, J., & Noble, W.S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4), 467–476.
- Machner, B., Sprenger, A., Kömpf, D., & Heide, W. (2005). Cerebellar infarction affects visual search. *Neuroreport*, 16(13), 1507–1511.
- Machner, B., Sprenger, A., Kömpf, D., Sander, T., Heide, W., Kimmig, H., & Helmchen, C. (2009). Visual search disorders beyond pure sensory failure in patients with acute homonymous visual field defects. *Neuropsychologia*, 47(13), 2704–2711.

- Magnusson, M.S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods*, *Instruments*, & Computers, 32(1), 93–110.
- Mannan, S.K., Ruddock, K.H., & Wooding, D.S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10(3), 165–188.
- Mast, M., & Burmester, M. (2011). Exposing repetitive scanning in eye movement sequences with t-pattern detection. Proceedings IADIS International conference IHCI 137–145.
- Mathôt, S., Cristino, F., Gilchrist, I., & Theeuwes, J. (2012). A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research*, 5(1), 1–15.
- Nobre, K., & Kastner, S. (2013). *The Oxford handbook of attention*. Oxford: Oxford University Press.
- Noton, D., & Stark, L. (1971a). Eye movements and visual perception. Scientific American.
- Noton, D., & Stark, L. (1971b). Scanpaths in eye movements during pattern perception. *Science*, 171(3968), 308–311.
- Over, E.A., Hooge, I.T., & Erkelens, C.J. (2006). A quantitative measure for the uniformity of fixation density: The voronoi method. *Behavior Research Methods*, 38(2), 251–261.
- Pernice, K., & Nielsen, J. (2009). How to conduct eyetracking studies. Nielsen Norman Group.
- Ponsoda, V., Scott, D., & Findlay, J.M. (1995). A probability vector and transition matrix analysis of eye movements during visual search. *Acta Psychologica*, 88(2), 167–185.
- Privitera, C.M., & Stark, L.W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9), 970– 982.
- Rakthanmanon, T., & Keogh, E. (2013). Fast shapelets: A scalable algorithm for discovering time series shapelets. In *Proceedings of* the thirteenth SIAM conference on data mining (SDM).

- Rieck, K., Wressnegger, C., & Bikadorov, A. (2012). Sally: A tool for embedding strings in vector spaces. *The Journal of Machine Learning Research*, 13(1), 3247–3251.
- Rubner, Y., Tomasi, C., & Guibas, L.J. (1998). A metric for distributions with applications to image databases. In *Computer Vision*, 1998. Sixth International Conference on, (pp. 59–66).
- Sippel, K., Kasneci, E., Aehling, K., Heister, M., Rosenstiel, W., Schiefer, U., & Papageorgiou, E. (2014). Binocular glaucomatous visual field loss and its impact on visual exploration - a supermarket study. *PLoS ONE*, 9(8), e106,089. doi:10.1371/journal.pone.0106089
- Tafaj, E., Kasneci, G., Rosenstiel, W., & Bogdan, M. (2012). Bayesian online clustering of eye movement data. In *Proceedings of* the Symposium on Eye Tracking Research and Applications, ACM, ETRA '12. doi:10.1145/2168556.2168617, (pp. 285– 288).
- Tafaj, E., Kübler, T.C., Kasneci, G., Rosenstiel, W., & Bogdan, M. (2013). Online classification of eye tracking data for automated analysis of traffic hazard perception. In *Artificial Neural Networks and Machine Learning–ICANN 2013*, (pp. 442–450): Springer.

Traquair, H.M. (1931). Perimetry in the study of glaucoma 51:585.

- Turano, K.A., Geruschat, D.R., & Baker, F.H. (2002). Fixation behavior while walking: persons with central visual field loss. *Vision Research*, 42(23), 2635–44.
- Wang, H., Freeman, J., & Merriam, E. (2012). Temporal eye movement strategies during naturalistic viewing. *Journal of Vision*, 12(1), 16.
- West, J.M., Haake, A.R., Rozanski, E.P., & Karn, K.S. (2006). eyepatterns: software for identifying patterns and similarities across fixation sequences. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, (pp. 149–154).
- Zangemeister, W.H., & Oechsner, U. (1996). Evidence for scanpaths in hemianopic patients shown through string editing methods. *Advances in Psychology*, 116, 197–221.