Automated Visual Scanpath Analysis Reveals the Expertise Level of Micro-neurosurgeons

Thomas C. Kübler¹², Shahram Eivazi³, and Enkelejda Kasneci¹

¹ Wilhelm-Schickard-Institute of Computer Science, University of Tübingen, Tübingen, Germany thomas.kuebler@uni-tuebingen.de

² Competence Center Vision Research - Study Course Ophthalmic Optics/Audiology, University of Applied Sciences Aalen, Aalen, Germany

³ School of Computing, University of Eastern Finland, Joensuu, Finland

Abstract. Exploring the effects of expertise on eye movements and visual search behavior of surgeons may help to improve and speed up the training of novices. However, comparing scan patterns to each other is a non-trivial task. This work employs several state-of-the-art, automated scan pattern comparison methods to re-analyze eye-tracking data of neurosurgeons that was captured during the observation of a tumor removal. We evaluate whether these methods can reproduce the findings from the original manual analysis and compare their performance regarding different eye-tracking metrics. Among the considered methods, SubsMatch revealed significant differences between eye movement patterns of expert surgeons and novices for all stimulus images.

1 Introduction

Eye-tracking technology promises to provide insights into the current attentional focus and thereby into cognitive processes related to visual processing [8,18]. Despite a rich landscape of techniques that have been developed for the analysis of eye-tracking data, the comparison of visual search scanpaths, i.e., the identification of common or similar patterns among several visual scanpaths as well as the quantification of differences between them, is a non-trivial task and therefore still challenging. Insights that are gained from such analysis could be of enormous impact for example in therapy and rehabilitation of visually impaired patients [9,10,13,15,20]. Beyond issues of rehabilitation, scanpath analysis has found application in the assessment of expertise level in domain such as medicine and arts [5,12,16,17,19].

To date, most of the scanpath studies are mainly limited to the comparison of time-integrated features of eye movements, namely average fixation duration, average saccade length, or comparison of heatmaps derived from viewing tasks. However, due to the sequential nature of the scan pattern, such metrics are mostly not suitable to capture essential pattern characteristics of the viewing behavior. For example, Buswell [1] reported that during art viewing, fixation durations increase with observation time, suggesting an initial exploration phase of short fixations, followed by a series of longer fixations [1,2]. Similarly, an initial exploratory phase with long saccades may be followed by in-depth examination with lots of short saccades. Such an effect, however, cannot be found once an integrating and averaging over the time dimension has been conducted. Exploring such events requires the analysis of time series.

Although recent studies have employed automated sequence-based scanpath metrics, most of these analysis methods are yet limited to the simple string edit distance or matching of similar fixations. Furthermore, the output of most of automated scanpath comparison methods is in the form of a pairwise similarity matrix which is not intuitively comprehensible. Instead, further post-processing steps and visualization techniques are required to gain relevant semantics from such output.

This work focuses on the question whether state-of-the-art approaches can reveal the level of expertise of surgeons based on automated analysis of differences in their eye movement features. First, we provide an overview of existing methods on automated scanpath comparison and discuss then their strengths and weaknesses based on a data set of micro-neurosurgeons from [5] during the observation of a tumor removal surgery. We do not aim at the identification of new gaze characteristics associated with the specific task, but focus on the applicability and potential of automated scanpath comparison methods. In addition, we highlight ways of statistical testing and visualization of the results.

2 Review on Scan Pattern Comparison

With increasing distance from the fovea our visual acuity drops rapidly. Thus, we have to continuously perform eye movements in order to scan the environment and sequentially project different sections of the scene onto the fovea. During such a projection, i.e., *fixation*, the eye is kept stable on the object of interest. Rapid eye movements, i.e., *saccades*, enable an attention shift towards the next object or area of interest. The resulting spatio-temporal sequence of such eye movements is called a scanpath. The key to comparing scanpaths is to find an adequate computational representation of the data. Usually, the conversion to the model space is associated with massive simplification of the scan pattern. Scan patterns can be reduced to fixations and saccades and simplified further to so-called regions of interest (ROIs), discarding relevant information, such as the fixation duration or the temporal sequence of fixations. An overview on scanpath comparison algorithm is given in Table 1.

Fixation heatmaps, i.e., time integrated visualizations of the spatial gaze behavior, are used to compare scanpaths where areas that are often hit by the gaze gradually become hot in color. Although easy to comprehend, the construction and interpretation of heatmaps for dynamic scenarios is quite challenging. Heatmaps can be compared against each other statistically, however correcting for multiple testing for every pixel has to be considered [2].

String-based representations encode the location information of a sequence of fixations as a sequence of letters. This way, scanpath comparison can be reduced to the problem of string-alignment (e.g. ScanMatch [3]). Sequential as well as spatial information is conserved. A similar approach is taken by the algorithm SubsMatch [11]. By determining transition probabilities for sequences of transitions instead of just for single transitions, SubsMatch can be applied to the comparison of complex search patterns.

Vector-based representation of scanpaths contain both information on the fixations as well as on the saccade characteristics. MultiMatch [4] and Func-Sim [6] are two methods in which pairs of vectors with a low distance to each other are matched together. In vector-based methods, usually several measures are computed, e.g., vector difference, saccade length difference or Euclidean distance between fixations. The vector representation is mathematically elegant and fast.

Probabilistic methods for scan pattern comparison can cope with the problem of high individual variability between repetitions of the same task and noise by learning the extend of individual noise and distinguishing effects that exceed this noise range, e.g., Hidden-Markov-Models (HMM).

Other methods for scanpath comparison are iComp [7] and Eyenalysis [14]. iComp automatically determines ROIs by mean-shift clustering of fixations. Eyenalysis [14] works on an unordered set of fixations of arbitrary dimensionality and performs a mapping of fixations with minimal distance towards each other.

Metric	Simplification	Temporal order	Repre-	Major
		perserved	sentation	characteristics
iMap	Simplified	No	Heatmap	Position, Duration
MultiMatch	Simplified	Yes	Vector	Shape, Position, Duration
ScanMatch	Grid	Yes	String	Sequence, Position
FuncSim	-	Weak	Vector sets	Shape, Position, Duration
iComp	AOI	No	String	Sequence, Position
SubsMatch	Binning	Weak	String / Prob.	Repeats, Position
HMM	Clustering	Weak	Probabilistic	Transition, Position
Eyenalysis	-	No	Vector	Position

Table 1: Overview of state-of-the-art methods for scan pattern comparison.

Scanpath similarity metrics as computed by the above algorithms represent a similarity score between two scanpaths. However, experimental designs usually involve the comparison of more than two scanpaths. Therefore, we need to handle a matrix of pairwise scanpath similarity scores such as the example depicted in Figure 3a. For n scanpaths this would result in a symmetric $n \times n$ matrix. Since the distances d(a, b) and d(b, c) do not necessarily predict d(a, c), we need to operate in a multidimensional space that is non-trivial to interpret. A simplification method is multidimensional scaling, where the high dimensional distance matrix is compressed into fewer dimensions. This process is however associated with an error that is subject to minimization. Scan pattern similarities can be reduced to one point in 2D or 3D space. Scan patterns with small distances between each other will be drawn closely together, while scanpaths with large distances are come to lie further apart (Figure 3b).

3 Methods

Our analysis is based on data from a study presented in [5]. Aim of that work was to study whether expert and novice micro-neurosurgeons differ in gaze behavior while viewing images of a surgery. The data acquisition is described in detail in [5], here the above authors thankfully provided us more participants data compared to their study. Seven expert surgeons and seven novices looked at four images of a tumor removal surgery (Figure 1) for 10 seconds. The subjects' gaze was recorded by means of a Tobii T120 eye tracker. Regions of interest were annotated for the tumor cavity, the instruments, and for the bleeding areas. The authors found differences in the viewing behavior regarding the amount of gaze directed towards the instruments and towards the areas highlighted by a fluorescence marker. Furthermore, they found longer fixation durations in the expert group as well as shorter saccades. The overall viewing behavior of experts was characterized as *more compact*, especially for the stimuli images 3 and 4 where a fluorescence marker was applied.



Fig.1: The sequence of stimuli employed in the study with the areas annotated [5]. The stimulus image number matches the descriptions in the text. Full scale versions at https://docs.google.com/presentation/d/ 1yHAGXdSDgjUkyKXs1_jAqM780ReLFQ_Kc-kyR_QIFRY

We calculated the scanpath similarities for the above data (14 subjects and 4 different stimuli) using ScanMatch [3], SubsMatch[11], MultiMatch [4], Func-Sim [6], iComp [7], Eyenalysis [14], and a HMM. Each comparison resulted in a similarity matrix of dimensions $14 \times 14 \times 4$. To achieve the best possible result for each algorithm, the parameter choice was optimized using a grid-search approach. It should be noted here that the parameter optimization step may have a significant influence on the performance of the algorithms, i.e., algorithms with many parameters may return better results due to our optimization criteria of group separability. However, they showed to be surprisingly robust to parameter choices. For algorithms that provide multiple distance output dimensions, all dimensions are treated as a separate, independent distance measure.

For statistical testing the distances within the expert surgeon and within the novice groups were compared to the distances between the classes. The two distributions were tested against each other by the Kolmogorov-Smirnov test. The resulting p-values were corrected for multiple testing by the Benjamini-Hochberg false discovery rate procedure (14 techniques $\times 4$ stimuli). A significant result (with alpha-level 0.05) would therefore mean that distances within groups and distances between groups do not follow the same distribution.

4 Results

Figure 2 shows the statistical evaluation of scan pattern distances, where significant results (p < 0.05) regarding distances within and between the surgeon groups are shown in black. For the stimulus images 3 and 4 (fluorescence marker has been applied), significant distances between the expertise groups are found by 7 of the evaluated algorithms. While ScanMatch and FuncSim (Direction) detected differences for the stimulus images 3 and 4, SubsMatch revealed scanpath differences between experts and novices for all four stimulus images. Algorithms working mainly on fixation location information (ScanMatch, SubsMatch) and respecting the temporal order of fixations (ScanMatch, FuncSim, SubsMatch) are performing well. For the presented images the local information also corresponds to tool locations. Results will probably change in favor of other methods (such as FuncSim direction) once a video or real operation procedure would be used as stimulus.



Fig. 2: Result of automated scan pattern comparisons on four stimulus images. Reported are the false discovery rate corrected p-values of a Kolmogorov-Smirnov test applied to the distances within and between expertise groups.

To visualize the results, we chose the similarity matrix produced by the SubsMatch algorithm for stimulus image 4 as an example. However, any of the other algorithms with significant results should produce similar results, although the group separability may vary. Figure 3b was created by a multidimensional scaling of the distance matrix of Figure 3a: three-dimensional positions are assigned to each scanpath in a way that the distance to all other scanpaths is as similar to the distances of the matrix as possible. The result in Figure 3b suggest that there is a group of three novices showing very similar gaze behavior, while experts spread wide around them. This indicates a strong focus on certain image regions for the novice group and a more heterogeneous viewing behavior for the expert group. The novice group can probably be split into two subgroups - one with very homogeneous gaze behavior and one more similar to the experts gaze behavior. Figure 3c reveals that experts repeatedly focus on certain image regions with the first few fixations, then they begin a broader exploration phase. Novices show an overall more repetitive viewing behavior.

A separation of the expertise groups could be achieved by examining the distance to the cluster center (the mean of all the scanpath positions). This indicates that the level of expertise is a major causes of systematic variance within the eye-tracking data and that, consequently, differences in viewing behavior between experts and novices do exist.



Fig. 3: (a) Distance matrix of the pairwise scan pattern comparisons with SubsMatch for stimulus image 4. (b) Scatterplot of the multidimensional scaling of the scanpath distance matrix. Novices show an overall smaller distance towards each other, resulting in a denser MDS plot while expert scanpaths exhibit larger distances. (c) Repeated patterns as found by SubsMatch. They reveal a more repetitive viewing behavior for the novices. Experts focus their repetitive scanning to the first few fixations, then transition to broader exploration.

In summary, most of the evaluated algorithms found significant differences in viewing behavior between expert surgeons and novices for the stimulus images 3 and 4, where a fluorescence marker was applied. Among the state-of-the-art algorithms SubsMatch revealed scanpath differences for all four images.

The above findings from the automated analysis correspond to findings from a previous study by Eivazi et al. [5], where the viewing behavior between the surgeon groups was analyzed manually based on aggregated features. Eivazi et al. [5] found a significant effect of expertise on the number of fixations performed. Furthermore, the authors reported that the average fixation duration was found to be significantly longer (for all stimuli) and average saccade length larger (for stimuli 3 and 4) for the expert group. Neither MultiMatch nor FuncSim were able to show this effect consistently, Figure 2. This is probably due to the fact that the above algorithms respect the temporal order instead of averaging for the viewing time. In the light of the automated scanpath analysis we can now conclude that longer saccade lengths probably resulted in the more heterogeneous scanpaths for the expert group. Most algorithms performed better on stimulus images 3 and 4, for which Eivazi et al. were also able to find more significant effects than for the other images. Novice gaze was probably captured by image saliency at the fluorescent image parts stronger than expert gaze.

It should be noted that significant differences derived by this analysis do not necessarily imply that groups are clearly separable. Although higher distances between than within groups indicate this fact, the extend of this effect is not analyzed thoroughly so far. Provided enough data, an effect even for largely overlapping but not entirely identical groups could be found. Therefore, the visualization step is required in order to reveal the actual extend of the effect.

5 Conclusions

We analyzed and compared eye movements of expert and novice microneurosurgeons when viewing images of a tumor removal surgery by means of several state-of-the-art algorithms for automated scanpath comparison. The automated analysis revealed that there are significant differences between expert and novice surgeons regarding their viewing behavior, confirming thus the assumption that the level of expertise is manifested in the viewing behavior. As future work, we will focus on methods that direct the surgeon's gaze into areas of interest, mimicking thus the viewing strategy of experts. When employed for training purposes, such method may not only increase the learning rate of novice surgeons but also contribute to an overall better surgery quality.

References

- 1. G.T. Buswell. How people look at pictures: a study of the psychology and perception in art. 1935.
- 2. R. Caldara and S. Miellet. iMap: a novel method for statistical fixation mapping of eye movement data. *Behavior research methods*, 43(3):864–78, September 2011.
- 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, August 2010.
- 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, December 2012.

- S. Eivazi, R. Bednarik, M. Tukiainen, M. von und zu Fraunberg, V. Leinonen, and J. E. Jääskeläinen. Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. In *Proceedings of the Symposium* on Eye Tracking Research and Applications, pages 377–380. ACM, 2012.
- R. M. Foerster and W. X. Schneider. 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), 2013.
- J. Heminghous and A. T. Duchowski. iComp: a tool for scanpath visualization and comparison. ACM SIGGRAPH 2006 Research, 1(212):59593, 2006.
- 8. Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures.* Oxford University Press, 2011.
- E. Kasneci, K. Sippel, K. Aehling, M. Heister, W. Rosenstiel, U. Schiefer, and E. Papageorgiou. Driving with binocular visual field loss? a study on a supervised on-road parcours with simultaneous eye and head tracking. *PloS one*, 9(2):e87470, 2014.
- E. Kasneci, K. Sippel, K. Aehling, M. Heister, W. Rosenstiel, U. Schiefer, and E. Papageorgiou. Homonymous Visual Field Loss and Its Impact on Visual Exploration: A Supermarket Study. *Translational Vision Science and Technology*, 3(6), 2014.
- T. C. Kübler, E. Kasneci, and W. Rosenstiel. Subsmatch: scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the Symposium* on Eye Tracking Research and Applications, pages 319–322. ACM, 2014.
- B. Law, M. S. Atkins, A. E. Kirkpatrick, and A.J. Lomax. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, pages 41–48. ACM, 2004.
- P.J. Marsh and L.M. Williams. ADHD and schizophrenia phenomenology: Visual scanpaths to emotional faces as a potential psychophysiological marker? *Neuro*science & Biobehavioral Reviews, 30(5):651–665, 2006.
- S. Mathôt, F. Cristino, I. D. Gilchrist, and J. Theeuwes. A simple way to estimate similarity between pairs of eye movement sequences. J. Eye Mov. Res, 5:1–15, 2012.
- D.P. Munoz, I.T. Armstrong, K.A. Hampton, and K.D. Moore. Altered control of visual fixation and saccadic eye movements in attention-deficit hyperactivity disorder. *Journal of Neurophysiology*, 90(1):503–514, 2003.
- 16. C.F. Nodine and H.L. Kundel. Using eye movements to study visual search and to improve tumor detection. *Radiographics*, 7(6):1241–1250, 1987.
- M. W. Pietrzyk, M. F. McEntee, M.E. Evanoff, P.C. Brennan, and C.R. Mello-Thoms. Direction of an initial saccade depends on radiological expertise. In *SPIE Medical Imaging*, pages 90371A–90371A. International Society for Optics and Photonics, 2014.
- Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.
- E. M. Reingold, N. Charness, M. Pomplun, and D. M. Stampe. Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, 12(1):48–55, 2001.
- K. Sippel, E. Kasneci, K. Aehling, M. Heister, W. Rosenstiel, U. Schiefer, and E. Papageorgiou. Binocular glaucomatous visual field loss and its impact on visual exploration-a supermarket study. *PloS one*, 9(8):e106089, 2014.

8