TOWARDS AUTOMATED COMPARISON OF EYE-TRACKING RECORDINGS IN DYNAMIC SCENES

Thomas C. Kübler 1,2, Dennis R. Bukenberger 1, Judith Ungewiss 2, Alexandra Wörner 1, Colleen Rothe 2, Ulrich Schiefer 2, Wolfgang Rosenstiel 1, Enkelejda Kasneci 1

1 University of Tübingen  
Department of Computer Engineering  
thomas.kuebler@uni-tuebingen.de

2 University of Applied Sciences Aalen  
Study Course Ophthalmic Optics/Audiology

ABSTRACT

Experiments involving eye-tracking usually require analysis of large data. While there is a rich landscape of tools to extract information about fixations and saccades from such data, the analysis at a higher level of abstraction (e.g., comparison of visual scanpaths between subjects) is still performed manually. Especially, the comparison of scanpaths derived from dynamic scenarios, where the observer is in permanent interaction with her environment, is highly challenging. In this work we (i) introduce a new work-flow for automated scanpath comparison in dynamic environments, which combines image processing, object tracking, and sequence comparison algorithms, and (ii) provide a new data set for performance evaluation of scanpath comparison methods that was extracted from eye-tracking data during an interactive tea-cooking task, referring to the experiments by Land et al. [1]. Furthermore, to showcase the applicability of our work-flow, we applied our method to the above data set to find differences in visual behavior between several runs for the tea-cooking task.

Index Terms— Scan pattern, scanpath comparison, area of interest annotation, eye tracking data analysis, image segmentation, automation

1. INTRODUCTION

The human eye is a foveated system, i.e., sharp vision is only possible within a small area of the retina called the fovea. Thus, in order to build and maintain a good representation of the environment, our eyes are constantly moving from one object to another, building a sequence of fixations (the eye is kept “stable” on an object of interest) and saccades (fast eye movements to change the focus of attention). Objects which have been fixated are considered as visually perceived. Such a sequence of fixations and saccades is called a scanpath. The cognitive processes that drive our eye movements are still not completely understood. However, recording of eye movements by means of eye tracking and analysis of eye-tracking data is helping to get more insights into these processes.

With the development of fast, high-end eye-trackers, the recording of eye movements has become affordable and easier, but experiments involving eye-tracking produce a large amount of data. The analysis of eye-tracking data, however, is still challenging and time-consuming, since some steps are still performed manually. Especially challenging is the analysis of data collected from dynamic and interactive scenarios, e.g., driving a car, walking through a supermarket, or cooking tea. Current measures describing eye-tracking recordings are usually time-integrated, e.g., the average fixation duration, the average saccade length, or the total number of fixations in a certain region of the scene. Indeed, there is a variety of tools that can be employed to determine the above measures [2, 3, 4, 5]. To date, there are few tools that support the analysis of eye movements at a higher level of abstraction, i.e., when visual scanpaths (the sequence in which certain areas of interest are traversed) of different subjects (inter-subject) or of repeated measurements of the same subject (intra-subject) have to be compared. Usually such analysis is performed manually and is mostly found in static viewing tasks (when a static image or a video sequence is being observed). The above tasks involve no or few motion of the observer. Stand-alone methods in this realm can be assigned to one of the following categories:

(i) String alignment algorithms, which are mostly based on the Levenshtein’s string metric [6]. The given stimulus (i.e., scene area) is first divided into geometric or semantic (e.g., in [7] eyes, mouth, or a specific object) AOs. Fixations are then mapped onto these AOs resulting in one word per scanpath. Two scanpaths can then be compared by comparison of their string-representation [8, 9]. However, the use of geometric AOs has two major disadvantages: if the areas are too large, scanpaths seem to be more similar than they in fact are. Moreover, fixations that are close to each other may lie in different AOs which also might influence the result. On the other hand, assigning labels to semantic objects in dynamic scenarios is non-trivial, unless the temporal order is discarded, such as in [10].

(ii) Other approaches are based on attention maps, which
are heat maps visualizing fixation data (the hotter a region, the higher the concentration of fixations on it). The scanpath comparison is then reduced to the comparison of attention maps, e.g., in [11, 12, 13]. Heat maps that combine spatial and temporal information have also been generated. However, due to the separation of the stimulus into several regions and independent analysis of those, important correlations between regions may be not recognized.

(iii) The most recent category in this realm contains vector-based algorithms, where fixations are represented as a set of vectors. Scanpath alignment is then done by finding the minimal distance between the simplified scanpaths using a shortest path algorithm, e.g., Dijkstra in [14].

The above methods have been evaluated mostly on artificial, short scanpaths extracted from static viewing tasks or on simulated data. In contrast, in interactive scenarios the stimuli changes with time, motion, and action. For example, when driving a car the scene changes for each driver with changing speed - even under standardized conditions (e.g., in a driving simulator).

This work presents a workflow for scanpath comparison (including sequence and duration for which objects are fixated) that not only works completely automated even for data extracted from dynamic scenes. We face the challenge of scanpath comparison with regard to various aspects, such as shape, scaling, spatial dimension, fixation sequence, AOI traversing, fixation duration per AOI, or similarity of parts of the scanpath. Many of these effects can be caused by small deviations in the experimental setup or lack of precision during eye-tracking measurement, others can be induced by the method of evaluation used.

Furthermore, we provide a new data set for evaluation of scanpath comparison methods. This data set was obtained from eye tracking during a interactive tea-cooking task, and was designed as described by Land et al. [1]. We evaluated our method on scanpaths derived from this data set, where scanpaths have a duration of several minutes and involve many degrees of freedom (head movement and free spatial movement).

2. METHODS

2.1. Task and evaluation data set

In order to compare results of different scanpath comparison algorithms, we established a new data set of scanpaths of different subjects during a tea-cooking task. 10 subjects (age range 20 - 56) recruited from students and staff of the Aalen University in Germany participated in the following study.

Subjects were instructed to make tea, referring to the experiment introduced by Land et al [1]. The standardized setting of the experiment (see Fig. 1) contained a sink, 5 different cups, a spoon, 5 kinds of tea, two kinds of sugar, honey, artificial sweetener, a bowl to dispose waste, and a water boiler. There were two different tasks conditions: first run (Run 1), each subject had to make tea for herself, whereas she could decide which ingredients to use. Second run (Run 2), each subject was instructed to make tea under more closely defined conditions (e.g., use the blue cup, make some herbal tea, add honey). The movements of the left eye of the subjects were recorded by means of a monocular Dikablis Eye-Tracker by Ergoneers GmbH at 25 frames per second. A 4-point calibration was performed using a calibration paper grid positioned on the table surface. All subjects gave their written consent for participation in the study.

Excerpts with a mean duration of 56(± 13) seconds were extracted from the eye movement recordings, excluding the eye-tracker calibration. For each subject, gaze accuracy (the visual field angle by which the measured gaze differed from the actual gaze due to calibration error e.g., by displacement of the eye tracker during the experiment) was checked at the end of the experiment. The data set is available online (http://www-ti.informatik.uni-tuebingen.de/~kuebler/euvip.zip). We encourage its use for evaluation and comparison of algorithms that support free moving subjects and scene manipulation.

Fig. 1. Setup of the tea cooking experiment with sink, water boiler, different sorts of tea in boxes and other items like honey, sugar, and a spoon.

2.2. Automated scanpath analysis

In focus of this work is an automated pre-processing step to a scanpath alignment method based on the Needleman-Wunsch-algorithm. More specifically, we introduce an automated procedure to assign AOI labels to specific objects in dynamic scenes. Scanpaths are then aligned and compared based on these labels. The automated workflow for scanpath comparison consists of two steps described by the following subsections.
2.2.1. Automated AOI annotation

In the context of dynamic, interactive scenarios, assigning AOI labels to specific objects is more challenging than for static tasks (e.g., image viewing) where the labeling has to be performed only once at the beginning of the experiment. In contrast, in interactive tasks, objects may change their position over time and most certainly will change their position relative to a moving observer.

In the first step of our work-flow, eye-tracking data is processed with a fixation identification filter [2, 3] in order to separate fixations from saccades. The extracted fixations were then projected into the video of the scene.

An Object Of Interest (OOI) was than established for each fixation cluster as following: starting with the fixation coordinates, a region growing method was initialized on the video frame at which the fixation occurred. SIFT features from that specific video frame and location were collected, while the region under consideration grows. In each iteration, the radius increases by 50%, until at least m SIFT features were detected in the considered region. m is a parameter that increases feature traceability when chosen to be large, but also decreases feature specificity. We chose m dependent on the resolution of the eye tracker image as follows:

\[ m = \left\lfloor \sqrt{\text{width}} + \sqrt{\text{height}} \right\rfloor / 2 \]

In our case, this gives us m = 25 feature points for a 576×768 video.

The descriptors of these new SIFT features are then compared using a FLANN-Matcher against all other collected descriptors. If the new descriptors pass the ratio test (as proposed by [15]) for one unique OOI, the features are assumed to originate from the same object. The descriptors of the original OOI are updated by the newly collected SIFT features (see Fig. 2).

With every fixation and SIFT features added to an OOI, its singularity grows and the ratio test thresholds dynamically assimilate to that growing reliability. The reason to establish a new OOI for a new fixation with several fitting OOIs is that the new SIFT features may not be unique enough and, consequently, adding them to the (at the moment) best fitting OOI would degenerate this OOI’s reliability. During the process most OOI reliabilities will grow.

In a second step, when all fixations are grouped into OOIs, the resulting objects can be merged with each other. An OOI that was isolated during the process because of none or several unreliable matches, has a good chance to find a good match during this merge process. At this point, we allow for user intervention, i.e., manual feature matching is also possible using an optional GUI as depicted in Figure 3. Thus, the above

2.2.2. Data analysis and statistics

The comparison of scanpaths is based on the Needleman-Wunsch-algorithm [16] (match score 1, mismatch score -1, gap penalty 3), including a normalization by the length of the scanpath sequences. Furthermore, the distance matrix was converted to a dissimilarity matrix \( D_{\text{is}} \) of value range [0, 1]
by processing each entry as follows:

\[
A = \max_{i,j} |D(i, j)|
\]

\[
Dis(x, y) = \left| \frac{D(x, y) + A}{\max_{i,j} D(i, j) + A} - 1 \right|
\]

where \( D \) is the pairwise distance matrix of dimensions \( i,j \). The resulting dissimilarity matrix was used as input of a multidimensional scaling \([17]\). The pairwise comparison distances do not necessarily fit into 2D space and therefore a dimensionality reduction is required in order to visualize the results. The multidimensional scaling performs a dimensionality reduction while minimizing the resulting stress.

3. RESULTS

We applied the presented workflow to automatically compare scanpaths of subjects who participated in the tea-cooking task. In order to analyze the results statistically, a Wilcoxon rank sum test for equal medians was performed for distances between:

- Person A cooking tea for the first time and Person A cooking tea for the second time.
- Person A cooking tea versus Person N cooking tea.

Our labeling algorithm assigned 110±50 fixations per scanpath to the objects on the scene. The distances of recorded scanpaths from each other were approximated by multidimensional scaling and visualized in Fig. 4.

The Wilcoxon rank sum test revealed significant differences between scanpaths of the same subject cooking tea for the first and second time \((p < 0.01)\). In contrast, inter-subject differences (person A cooking tea versus person N cooking tea) were not significant. However, Fig. 4 shows an outlier with ID 8.

The distance matrix of pairwise scanpath comparisons as calculated by the Needleman Wunsch algorithm is shown in Fig. 5.

The eye tracking accuracy achieved after the experiment is shown for each subject in Table 1.

<table>
<thead>
<tr>
<th>ID</th>
<th>Accuracy run 1 [°]</th>
<th>Accuracy run 2 [°]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.5</td>
<td>3.5</td>
</tr>
<tr>
<td>2</td>
<td>3.5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4.5</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>5.5</td>
</tr>
<tr>
<td>9</td>
<td>&gt;7</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1. Eye tracking accuracy achieved after the experiment. Accuracy was measured by the subject looking at four points surrounded by circles of different diameters. Depending on the diameter of the smallest circle that the mapped gaze point comes to lie in, the accuracy can be estimated.

4. DISCUSSION

We proposed a method to analyze visual scanpaths extracted from eye-tracking recordings in dynamic scenes based on feature extraction and feature matching algorithms for AOI annotation. While almost all available eye-tracking data sets consist of short scanpaths from viewing static images or video viewing tasks (mainly for evaluation of saliency map algorithms), the data used in this work was collected during a real-world task, i.e., cooking tea, where the observer is in interaction with the scene.
tion targets, while the background is likely to be deprived of features. Furthermore, (iii) SIFT features are scale invariant and quite robust to looking at an object from different distances or slightly different angles.

Our analysis of the tea-cooking data set indicate that scanpaths derived from the second run, which was more controlled than the first run, show less variability. This may originate from the standardization or the knowledge about the exact position of task-relevant objects. We could observe that some individuals show only a small change in visual scanpath between the first and the second run, while others seem the show a learning effect. In the latter case the visual scanpath of the second run is closer to the denser region of the mds-plot (Fig. 4) and therefore closer to the probably optimal scanpath for the tea-cooking task (e.g., subject 8, 7 or 1).

The dissimilarity score matrix in Fig. 5 suggests that there are individual scanpaths that tend to be dissimilar from most other scanpaths (e.g., 8, 9.2). Having a closer look at such outliers may improve data quality of eye-tracking experiments. In our experiment, subject 8 was the only person who tried to measure the required amount of water for boiling with the cup, spilled some water and had to clean that up. This resulted in a scanpath that differs from most others.

**Limitations**

In-depth analysis of the alignment matrices would be possible and could give us some insight in when the similarities and differences in scanpaths occur. At the current stage, the pure scores do not tell us anything about why two sequences are similar or dissimilar to each other. We are currently working on improving the annotation quality and user interface for semi-automated AOI annotation. Our plans include the creation of algorithms for identification of repeated behavioral patterns and deviating patterns between different subject groups.

**5. CONCLUSION**

We have presented a novel method to analyze visual scanpaths extracted from eye-tracking recordings in dynamic scenes based on feature extraction and feature matching for AOI annotation. We think that our algorithms can accelerate the analysis of eye-tracking studies in naturalistic environments. Furthermore, we provided an accompanying data set from eye-tracking recordings during a tea-cooking task involving several runs at different degrees of freedom. Our evaluation on the above data set indicate that scanpaths derived from experiment in more standardized conditions (less degrees of freedom) show less variability in visual scanning. As future work, we will enhance the AOI sequence comparison as well as the scanpath clustering step. While the proposed methods provide an overview over the similarity of scanpaths, their power to distinguish between experiment groups still needs to be improved.

**6. REFERENCES**


