Enhanced Representation of Web Pages for Usability Analysis with Eye Tracking

Raphael Menges
University of Koblenz-Landau
Institute WeST
Koblenz, Germany
raphaelmenges@uni-koblenz.de

Hanadi Tamimi
University of Koblenz-Landau
Institute WeST
Koblenz, Germany
htamimi@uni-koblenz.de

Chandan Kumar
University of Koblenz-Landau
Institute WeST
Koblenz, Germany
kumar@uni-koblenz.de

Tina Walber
EYEVIDO GmbH
Koblenz, Germany
walber@eyevido.de

Christoph Schaefer
EYEVIDO GmbH
Koblenz, Germany
schaefer@eyevido.de

Steffen Staab*
University of Koblenz-Landau
Institute WeST
Koblenz, Germany
staab@uni-koblenz.de

ABSTRACT

Eye tracking as a tool to quantify user attention plays a major role in research and application design. For Web page usability, it has become a prominent measure to assess which sections of a Web page are read, glanced or skipped. Such assessments primarily depend on the mapping of gaze data to a Web page representation. However, current representation methods, a virtual screenshot of the Web page or a video recording of the complete interaction session, suffer either from accuracy or scalability issues. We present a method that identifies fixed elements on Web pages and combines user viewport screenshots in relation to fixed elements for an enhanced representation of the page. We conducted an experiment with 10 participants and the results signify that analysis with our method is more efficient than a video recording, which is an essential criterion for large scale Web studies.

CCS CONCEPTS

• **Human-centered computing → Systems and tools for interaction design**;

KEYWORDS

Eye tracking, Web page usability, viewport-relative elements.

ACM Reference Format:
Raphael Menges, Hanadi Tamimi, Chandan Kumar, Tina Walber, Christoph Schaefer, and Steffen Staab. 2018. Enhanced Representation of Web Pages for Usability Analysis with Eye Tracking. In *ETRA '18: 2018 Symposium on Eye Tracking Research and Applications, June 14–17, 2018, Warsaw, Poland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3204493.3204535

*Also with University of Southampton, Web and Internet Science Research Group.

ETRA '18, June 14–17, 2018, Warsaw, Poland

ACM ISBN 978-1-4503-5706-7/18/06…$15.00
https://doi.org/10.1145/3204493.3204535

1 INTRODUCTION

Eye tracking has emerged as a powerful tool to perform usability studies on graphical user interfaces. Estimation of gaze allows to measure the visual attention that users dedicate to regions of the interface in a viewport [Nielsen and Pernice 2009]. Thus, gaze data gives insight into user experiences by capturing both the sequence of fixations and the attention within certain areas of interest (AOI). Such information provides the usability analysts with implicit user feedback to evaluate a design and compare to different designs. The most commonly used method for eye tracking analysis maps the gaze data (e.g., scan paths or heatmaps) on a visual representation of the interface [Blascheck et al. 2017; Eraslan et al. 2015]. Hence, the analysts can assess and correlate which elements of the presented interface have drawn attention and which have been ignored.

The interface representation is critical for the accurate mapping of gaze data with respect to the elements inspected by a user. For static interfaces, the mapping is trivial as the stimuli can be simply represented with an image, e.g., captured as a screenshot of the interface that was shown to users, as it is common in many usability studies [Barreto 2013; Cutrell and Guan 2007; Walber et al. 2014]. In the Web environment, however, the users experience the Web page through a viewport and adapt what is visible by scrolling and by clicking on links to Web page anchors. Hence, the users actively influence the dynamic stimulus by their actions. This makes the synchronization of the data non-trivial, and has been identified as a challenging problem of gaze-based analysis [Blascheck et al. 2017].

One approach to deal with this issue is to create a virtual screenshot that comprises the complete Web page, an area that is possibly larger than what can be displayed to the user at once in the viewport, and to map all gaze data onto this screenshot. However, this naïve method suffers from *inaccurate* gaze-mapping on elements that are not affected by scrolling, on which gaze data is then misinterpreted by the usability analysts. For example, gaze data at an advertisement banner that is shown on a fixed viewport position is misplaced if the user scrolls the page. The banner is rendered once on the virtual screenshot at its initial position on the page. When the user scrolls the viewport down such that a lower region of the Web page becomes visible and the topmost region disappears, the

banner may stay in the same place relative to the viewport rendered on the screen. Though, the naïve method transforms *all* gaze data from screen to page space. The gaze data is then registered somewhere below the initial rendering of the banner. Therefore, the attention of the users cannot be appropriately associated with the banner and the usability analyst underestimates the effect of the banner on the user experience.

An alternative approach to capture the user experience of the page is to record a video of the interaction session and to overlay it with gaze data for analysis. A video recording of the user viewport is able to capture interaction sequences of each individual user encountering arbitrary dynamics on the Web page. However, every user usually reaches a different viewport position at a certain point of time, caused by scrolling and navigation on the Web page. Hence, gaze data of multiple users cannot be simply overlaid. Therefore, analysts must analyze the video of each user one by one to understand all the effects of the interface design, making the analysis time-consuming and tedious. In general, this is *not a scalable* solution, because usability analysis requires feedback from a reasonable number of users [Eraslan et al. 2016].

The literature [Blascheck et al. 2017; Eraslan et al. 2015; Špakov and Miniotas 2007] investigating gaze-based usability analysis, proposes several methods for clustering and visualizing gaze data. However, the accurate representation of the interface, which is critical to associate gaze data visualizations, has received less attention. In this work, we argue the need of better representation method of Web pages to facilitate relevant gaze data visualization and analysis, and to support large scale usability studies.

We propose an *enhanced representation* method for Web pages to tackle the above-mentioned challenges of accuracy and scalability. We make use of structural information from the Web page document to identify fixed elements, and to combine them coherently with scrollable content, with the goal to portray a representation of the Web page that is simplifying the actual user interaction yet is close enough to allow insights by the gaze data analyst. Our approach tightly integrates the eye tracking environment with the Web browser, as it extracts information about the Web page elements from the Web page document and it considers pixel patterns rendered in the viewport of the Web browser.

2 RELATED WORK AND TOOLS

We describe how eye tracking is used in studies in general and in the Web specifically, and how page elements can be identified.

2.1 Eye Tracking in Usability Analysis

Eye tracking has been employed to analyze attention in several application domains, such as medical, sports, commerce and human-computer interaction studies [Duchowski 2002; Holmqvist et al. 2011; Nielsen and Pernice 2009; Poole and Ball 2005]. Gaze data provides implicit feedback on the interaction behavior of users, which is arguably more intuitive and natural than the conventional indicators [Schiessl et al. 2003]. There are various visualization techniques of the recorded gaze data in context or without the presented stimulus [Blascheck et al. 2017]. The visualization techniques can be classified into *point-based* and *AOI-based* methods. Both require an accurate registration of gaze data to the underlying stimulus.

In this work, we focus on supporting studies that aim to assess the usability within Web environments by quantifying user attention. For this purpose, eye tracking has been utilized to assess Web search efficiency [Cutrell and Guan 2007], online advertisements [Barreto 2013] and Web page navigation usability [Ehmke and Wilson 2007]. Apart from the scientific research studies, eye tracking analysis has also gained importance in commercial Web page usability analysis [Buscher et al. 2009].

2.2 Tools for Web Page Usability Analysis

There are various tools for conducting usability studies with eye tracking. Many tools utilize a webcam to estimate gaze and create a heatmap of user attention on a virtual screenshot of the Web page, like sticky.ai, eyezag.de or realeye.io. Other tools aim for a more precise, but also more costly, analysis. Cooltool.com utilizes dedicated remote eye tracking hardware with infra-red illumination and a video-based screen recording. Tobii Studio Pro [Tobii AB 2016] also offers gaze-mapping on a Web page representation, either by a virtual screenshot for each accessed URL, or through analyzing gaze data on a video recording of the interaction session. The experts can also manually merge videos from multiple users and define a certain video frame to be overlaid with the accumulated gaze data.

There are tools that are solely specialized in taking screenshots of Web pages without the intention to perform a usability analysis. FireShot[1] is a tool to capture the entire Web page in a single image. Fixed elements are transformed to an absolute position on the page. Then, the tools stitches viewport screenshots, while scrolling down the page automatically.

2.3 Web Page Element Identification

In the enhanced representation method, we assert specific handling of Web page elements. Hence, we need to identify the elements and retrieve their associated information. [Sano et al. 2013] have presented a tool to manually crop parts of a Web page. The pixel data and the hyperlinks, including their position and size, are stored as "clickable image map" of the cropped Web page block. [Lamberti et al. 2017] describe a method to accumulate attention on responsive Web pages with element-aware heatmaps. They inject JavaScript code into the Web page to retrieve information about elements of interest and store data accordingly. However, their heatmaps do not use eye tracking data but visualize the duration for which parts of the page had been visible on the screen. Both approaches would mean that the experts who analyze the attention of users had to manually label the fixed Web page elements before conducting the user study. Certainly, a more robust approach for automatic identification of relevant Web page elements can enhance the efficiency of the analysis workflow. In that regard, [Kumar et al. 2017] have presented a Chromium based Web browser that can be controlled by gaze-driven interaction. To adapt Web browsing for gaze input, they identify elements like hyperlinks or text input fields on Web pages. They utilize real-time observation via a DOM tree mutation observer to detect automatically structural changes in the Web page, and analogous to our work, they employ position style attributes to recognize certain elements, to improve various interactions such as automatic gaze-based scrolling.

[1] https://www.getfireshot.com

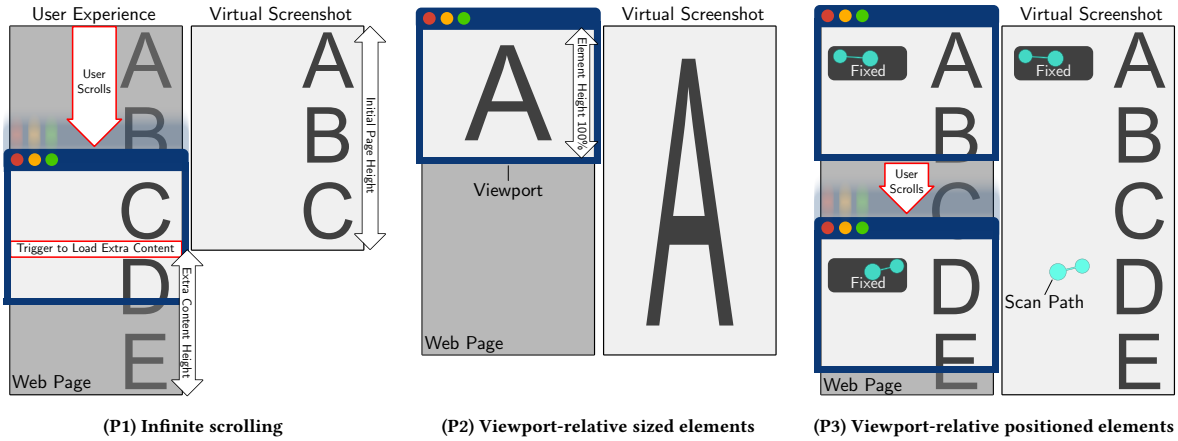


Figure 1: Problems P1 to P3 of the virtual screenshot as Web page representation. The actual experience of the users on the Web page is visualized on the left. The corresponding virtual screenshot is shown on the right.

3 LIMITATIONS OF THE NAÏVE REPRESENTATION METHOD

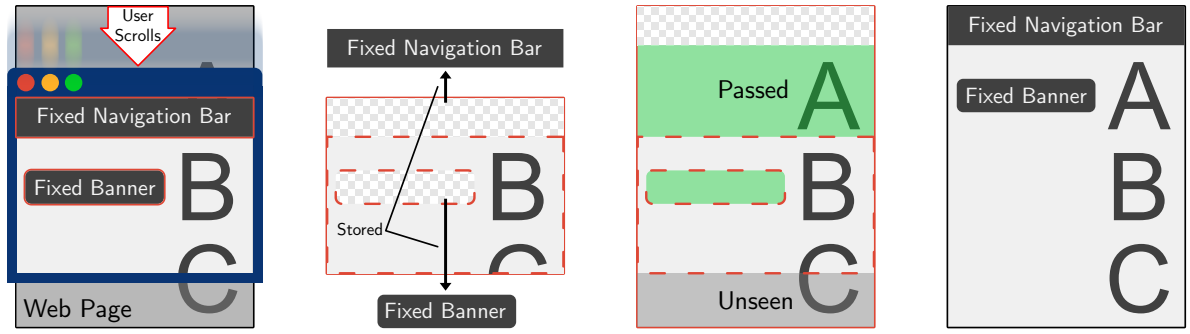
To analyze usability with eye tracking, gaze data is visualized as an overlay on the rendering of the interface design that the users experienced during their interaction with the system. Modern Web page usability analysis tools extend this approach for Web pages that are larger than what can be seen in a viewport by the user at a time. They assemble a screenshot by capturing pixels from a virtual viewport that is large enough to render the complete Web page at once. We call this method *virtual screenshot*. Gaze data coordinates, however, are received not relative to the virtual, but relative to the actual viewport of the user, which is usually too small to host the complete Web page at once. Therefore, the gaze data coordinates must be mapped between these two reference systems. There can be no mapping that is able to retain the experiences of the users, as it relates to their eye movements and the actual renderings in the observed viewports. This why we refer to this method as *naïve*. The problems are given in the following and illustrated in Figure 1:

- **(P1) Infinite scrolling pages.** The virtual screenshot method implies the assumption that at one point of time the complete Web page is loaded by the Web browser. However, this is not true or at least it is not trivial to determine the point of time when the complete Web page is loaded. As of today, there are "infinite" scrolling pages that use asynchronous server communication to request further Web page content when the user reaches a certain vertical scrolling position. The user scrolls down the Web page and may reach a trigger, which causes the Web browser to request for more content. Before the user reaches the bottom of the Web page, the retrieved content is appended to the bottom of the Web page and the user can continue with scrolling. The scrolling interaction dynamically expands the Web page height during an interaction session. *Example: social media news feeds.*
- **(P2) Viewport-relative sized elements.** Elements on Web pages can be sized in relation to the viewport. The height and width of elements are then provided in percentage values. When a virtual screenshot is taken, the viewport for capturing the visual content is set to the size of the complete Web page, including the content beyond the viewport of the user. Relatively scaled elements are scaled accordingly to the virtual viewport that covers the complete page, to capture the virtual screenshot. For example, an image with a width and height of 100% at the top position of the Web page would cover wrongly the complete Web page representation instead of the initial viewport of the user. *Example: A welcome message filling the entire screen with further details beneath, which the user can reach through scrolling.*
- **(P3) Viewport-relative positioned elements.** Elements on Web pages can be defined to remain at a certain screen position, so they are positioned viewport-relative and are not affected by scrolling. We refer to this property as position-*fixed*. In the virtual screenshot, the gaze data on fixed elements is wrongly stored when the user scrolls the page. The virtual screenshot method assumes the page content to move accordingly when the user scrolls the page and transforms all gaze data to page space. The gaze data on fixed elements is therefore wrongly transformed and cannot be associated with the actually viewed content in the later analysis. *Example: A fixed advertisement banner or a navigation bar at the top of the viewport.*

4 ENHANCED REPRESENTATION METHOD

In this work we propose the enhanced representation of a page, which is a composed image of user viewports from the interaction session, while keeping the position and size of page elements consistent as per the actual experience perceived by the user. To achieve this, we first identify the fixed elements in the viewport with the structure extraction approach described in in step 1. We crop the identified fixed elements from the viewport screenshots and combine the screenshots for a consistent Web page representation, as discussed in step 2. Finally, we show the composition of the screenshots with the fixed elements in step 3.

Step 1: Extraction of Fixed Elements. Fixed elements on a Web page are viewport-relative elements that stay visually in the same position within the viewport while the user scrolls the page. An



(a) Information about the fixed elements on the Web page are acquired.

(b) Cropping away of fixed elements pixels from a viewport screenshot.

(c) Integration of cropped viewport screenshot into the so-far stitched screenshots.

(d) Combination of stitched screenshots and fixed elements towards enhanced representation.

Figure 2: The enhanced representation of the visual Web page content for attention analysis.

example of a fixed banner and a fixed navigation bar within the viewport is shown in Figure 2a. For the identification of these viewport-relative positioned elements in the Web page document structure, we search for elements with a style position attribute value set to `fixed` [Atanassov and Eicholz 2016]. The style position attribute is the standard for creating a viewport-relative positioned element and is widely used in the modern Web. To acquire the information about fixed elements, we inject JavaScript code into the loaded Web page. The code contains a function to search recursively for fixed elements in the Web page document structure and returns their position, size and visibility via a callback to the Web browser. Furthermore, the XPath, the path from the root node to the fixed element in the Web page document structure, is used for explicit identification of the fixed elements. The identification can be used for dynamic updates of the fixed elements, e.g., if the elements move or change their visibility. When a formerly identified fixed element is no more contained in the returned list of identified fixed elements, its visibility is set to false. Gaze data that is registered within the boundaries of a visible fixed element is associated with the extracted element and stored alongside for further processing.

Step 2: Viewport Screenshot Stitching. A simple stitching approach of appending viewport screenshots would replicate the fixed elements on every captured scroll position. Therefore, we use the extracted information about the fixed elements to crop the pixel data within the boundaries of the fixed elements from each viewport screenshot, see Figure 2b. The cropped screenshot is transformed accordingly to the page scrolling and stitched together with the previously stitched screenshot of the Web page, as shown in Figure 2c. Areas that have been cropped away are left transparent, so they can be filled with information from other viewport screenshots at different scrolling positions.

Step 3: Composition of Stitched Viewport Screenshots and Fixed Elements. The procedure described in step 1 and step 2 is triggered by `window.onload` and `window.onscroll` events, and is additionally executed every 200ms, in case the fixed elements are changed in their properties by scripts while the user is not scrolling the page. A lower sampling frequency has produced visual gaps in a dry run, whereas higher sampling frequencies do not add more value but increase computational load. Once the interaction session is ended, for example through an URL change, a composition of the stitched screenshot and the collected information about the fixed elements is created. The challenge is to place the fixed elements on the stitched screenshot as close as possible to the user experience. We employ the heuristics to align a fixed element to either the top or the bottom of the stitched screenshot, depending on which vertical half-space the element was displayed in the viewport of the browser. See Figure 2d for the composition of the stitched screenshot and the fixed elements in the example. Both fixed elements in the figure (navigation bar and banner) are originally placed within the upper half of the viewport and therefore aligned at the top in the final composition. Any gaze data on these fixed elements would be displayed in the analysis accordingly to the final position of these elements. Notably for saccade analysis, we must ensure the correct fixation sequence with respect to the fixed and scrollable elements. The associated gaze data provides this information, and we propose the analysis tool to show this information by appropriate means of visualization. This method is able to solve the problems of the naïve method as described in the following:

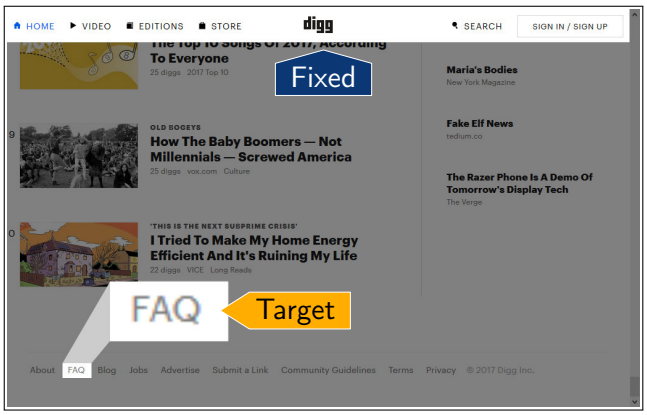
- The frequent stitching of viewport screenshots from the user experience does include the dynamic additions on infinite scrolling Web pages, solving (P1).
- The screenshots of the user viewport guarantee that Web page elements are displayed in the same scale as they have been presented to the user, solving (P2).
- The extraction and cropping of fixed elements and their association with registered gaze data allows for accurate visualization and metrics calculation, solving (P3).

An open-source prototype based on the Qt WebEngine example using mouse data is available on GitHub.[2]

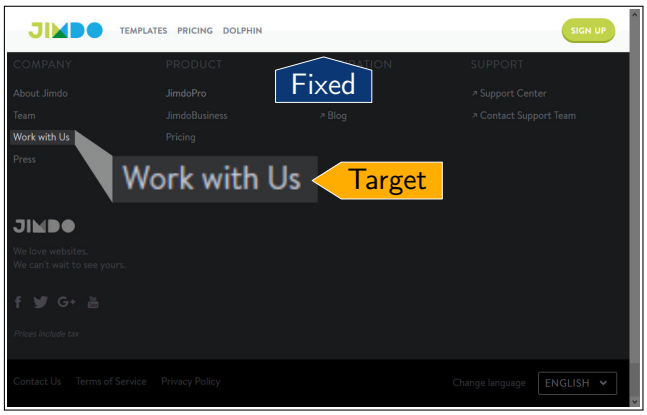
5 EVALUATION

In the proposed enhanced representation method, we overcome the limitations of the naïve representation approach, described in

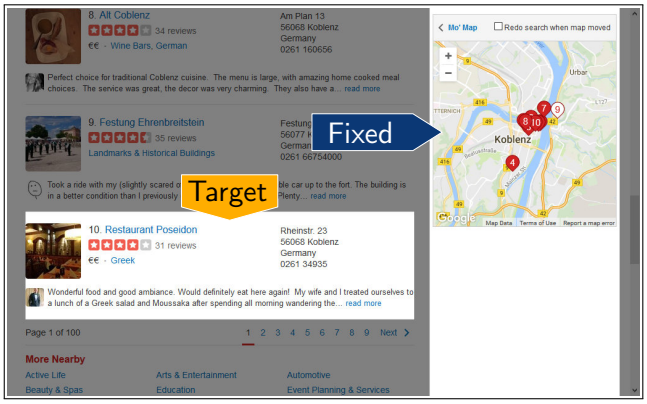
[2]https://www.github.com/Institute-Web-Science-and-Technologies/MTB



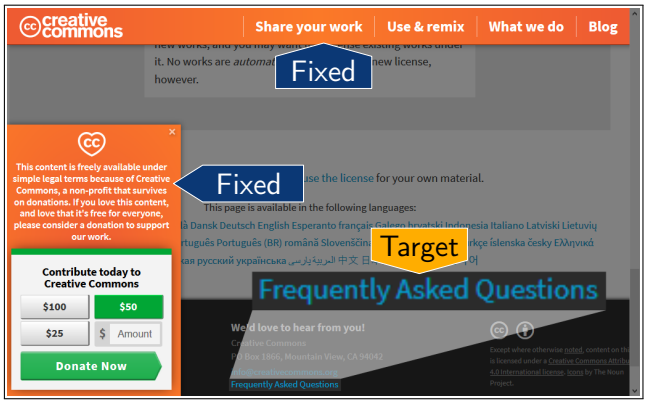
(a) Digg: The Web page contains a fixed navigation bar on the top. The users were advised to find and click the Frequently Asked Questions (FAQ) link.



(b) Jimdo: A fixed header is shown after a certain amount of scrolling. The users had to search for the "work with us" link and click on it.



(c) Yelp: The search page makes use of a fixed column on the right to display a map with location information about the results. Users were asked to select the 10th entry of the search results in the list.



(d) Creative Commons: The Web page can be considered as a complex layout in regards to fixed elements. The navigation bar at the top is only shown when the user scrolls up and the donation pop-up on the left is displayed after a certain timeout. For the evaluation, only the donation pop-up was considered. Analogously to Digg.com, users were asked to click on the FAQ.

Figure 3: Annotated viewport screenshots of the Web pages selected for evaluation.

Section 3. It allows the accurate mapping of gaze data on the Web page content for "infinite" scrolling pages and viewport-relative scaled elements, because these features are reproduced identically to the actual user experience. However, we need to assess if the mapping on the enhanced representation is perceived correctly for viewport-relative positioned elements. Users have inspected the fixed elements on different scrolling states of the Web page, which alter the page content around the fixed elements. The composition in the enhanced representation, however, gathers fixed elements and maps associated gaze data on the top or bottom of the page. Hence, we compare its effectiveness against the method of video recording, which provides individual user experiences and is supported in commercial tools for the gaze data analysis on dynamic Web pages. One major limitation of the video-based approach is the effort required by the analysts to analyze individual videos, hence we aim to evaluate if our approach would reduce the workload on analysts, while being equally effective. In summary, we came up with the following two hypotheses to assess the analysis on fixed elements with the enhanced representation method:

(H1) Accuracy: The enhanced representation method supports the analysis of gaze data on fixed elements as accurate as the video recording.

(H2) Scalability: For analyzing gaze data from multiple users, the enhanced representation method would be more efficient than a video recording.

Procedure. The goal was to quantify the effectiveness and efficiency of the two methods in supporting analysts in assessing user attention for Web page usability. For the usability test, the representation methods were used in the EYEVIDO analytical tool[3] environment, where interactions are presented and analyzed in the same manner for both methods (straight line as visualization for saccades, enumerated circles for fixations, a timeline to play and skip gaze data and the video recording, zooming and scrolling

[3] https://www.eyevido.de/cloud-eye-tracking/software-eyevido-lab

Table 1: Dataset distribution among analysts. For each Web page analysis, the analysts worked either on the enhanced representation (E) or on the baseline video recording (V).

	Digg.com	*Jimdo.com*	*Yelp.com*	*CC*
Group A	V	E	V	E
Group B	E	V	E	V

tools for the enhanced representation, AOI utilities). The gaze data of one user was displayed for both the enhanced representation and the video recording. Furthermore, the enhanced representation allowed to overlay data of multiple users at once. The timeline was used to play and skip through the gaze data and the video.

We recruited ten participants (two females and eight males, with a mean age of 25.4 years, standard deviation 1.77), who have been trained in the analysis tool and referred as analysts. The analysts were randomly divided into two static groups called A and B and the groups performed an analysis either on basis of the enhanced representation of the Web page or on the video recording. The utilization of the enhanced representation or the video recording was the independent variable of the experiment, see Table 1 for the counter-balanced dataset distribution among the groups. The analysts were asked to draw AOIs on the fixed elements of the Web page. The analysis tool calculated standard measures of eye tracking [Bylinskii et al. 2017; Poole and Ball 2005] in the AOIs, of which the following were reported by the analysts: Time To First Fixation (TTFF) and Total Fixations (TF).

The dependent variables of the experiment were the AOI measurements of the metrics, for which the evaluator created a ground truth, the completion times, the outcome of a NASA-TLX questionnaire with scores from one to seven and a subjective estimation by the analyst about the difficulty for the user to fulfill the task. It was formulated as the following question: "It was challenging for the participant to solve the task." The analyst answered on a 5-step scale from *absolute disagreement* to *absolute agreement*.

Dataset. The dataset for the evaluation was created by two users on four Web pages with a viewport of 1920x984 pixels. During their interaction session, both a video recording and an enhanced representation were created along with gaze data. We have chosen the following pages because of their visiting ranks noted on moz.com/top500 and their utilization of fixed elements: (a) Digg.com, (b) Jimdo.com, (c) Yelp.com[4] and (d) Creativecommons.org[5] (CC), as depicted in Figure 3. The tasks have been designed to let the users explore the Web pages in their complete height. The targets were page-relative links that are placed on the bottom of the Web pages. These Web pages include the limitation for fixed elements of the naïve representation, and analysts would fail to analyze them correctly on a virtual screenshot. Furthermore, the analysts were asked to assess the attention on fixed elements, which is a straightforward task for the video recording. If the analysts had to assess attention on page-relative content, they would have to adapt the position and size of the AOIs in the viewport manually to the individual user experience. Please refer for the corresponding figure captions for more details about the tasks for the users and analysts.

[4]https://www.yelp.com/search?find_loc=koblenz

[5]https://www.creativecommons.org/licenses/by/2.0

The outcome has been a dataset with a total of four enhanced representations, eight videos and eight interaction sessions, and the recorded gaze data. Additionally, we acquired subjective feedback about the perceived difficulty from the users, analogously to the question asking about the difficulty estimation by the analysts. The users have been asked to answer "How challenging was the task for you?" after viewing each Web page on a 5-step scale from *very easy* to *very difficult*.

Apparatus. We used a Visual Interaction myGaze-n 30Hz remote eye tracking device for recording the dataset, which was then analyzed by the analysts. Furthermore, we implemented the enhanced representation method in the EYEVIDO recording tool to collect the dataset. For the baseline, an additional video of the Web page interaction of each user has been recorded with the Open Broadcaster Software.[6] The EYEVIDO analysis tool has been expanded to allow either the use of the enhanced representation or the video recording, which were used by the analysts to perform the analysis.

5.1 Results

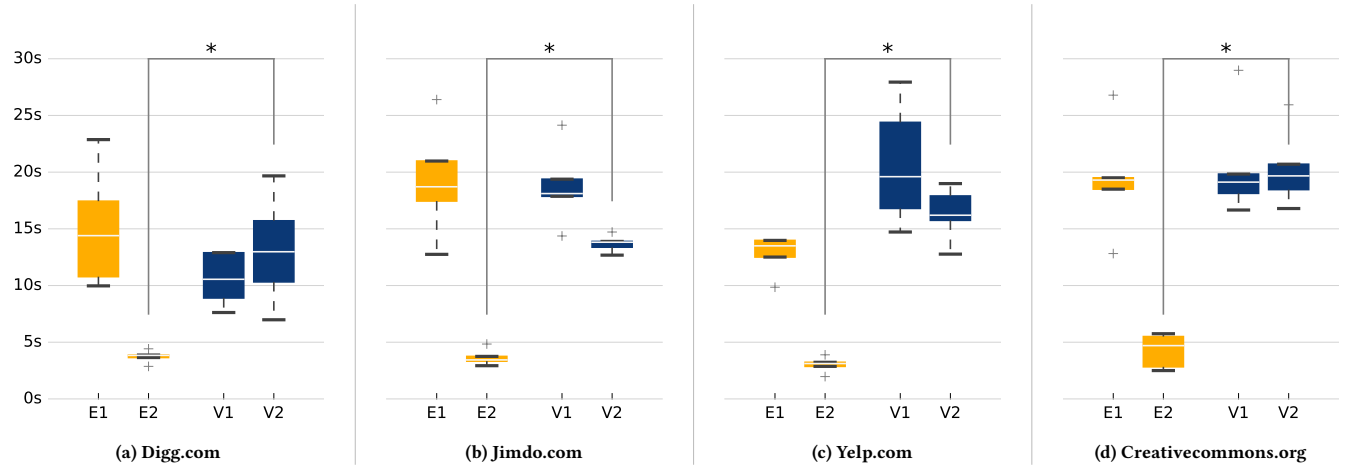
First, we present the results of the accuracy for both methods. Then, we provide details about the task completion times and the feedback from the NASA-TLX questionnaire.

Accuracy. The accuracy is measured by the average absolute percentage errors of the two metrics TTFF and TF, within the AOIs marked by the analysts. The error is defined as difference between the outcomes of the analysts and the ground truth by the experimenter. A lower value indicates a higher accuracy. The errors appear to be similarly low for all Web pages, as presented in Table 2. However, the average percentage error value of 31.3% ± 44.3% for the TF estimation on Jimdo.com, when using the enhanced representation of the Web page, appears to be high. But a Mann-Whitney significance test indicates that the TF estimation for the enhanced representation (Mdn = 0.0%) is not significantly different from the video recording (Mdn = 0.0%), U = 30, with p = .14, for the Jimdo.com Web page. A look into the data reveals that the high error was mainly caused by one analyst, who reported correct values for TTFF but 100% deviating answers for the TF values, in analysis of both users. There has been no systematic error, as three out of five analysts have reported values with zero percent error. These results validate our first hypothesis (H1), i.e., the analysis of gaze on viewport-relative positioned elements with the enhanced representation can be as accurate as with the video recording.

Task completion time. We measured the time that the analysts required to fulfill the tasks. This includes the marking of the AOIs and the output of the statistical values. See Figure 4 for a box plot per Web page. A Mann-Whitney test shows no significant difference between the analysis of the *first user* with the enhanced representation and with the video recording for each Web page.

In contrast, the Mann-Whitney test shows a significant difference in timings for analyzing the *second user* per Web page for both representations. In average, the analysts only need 24.8% of the time for analysis of the second user when working on the enhanced representation in comparison to the colleagues, who worked on the

[6]https://www.obsproject.com



(a) Digg.com (b) Jimdo.com (c) Yelp.com (d) Creativecommons.org

Figure 4: Box plot showing the times required by the analysts to solve the analysis tasks per Web page. The time for analyzing the first user with the enhanced representation is labeled with E1 and the second user with E2. Similarly, the timings of the analysts using the video recording are labeled as V1 for the first user and V2 for the second user.

video recording. The lower time requirement with the enhanced representation method supports our second hypothesis (H2).

NASA-TLX. The questionnaire is designed to measure the workload of the analysts. The analysts have been asked to answer the questionnaire after reporting the measurements for the user attention of each video recording, but only once for the enhanced representation of each Web page. This procedure has been defined due to the very low timings for the analysis of the second user when the enhanced representation of the Web page was utilized. If we had asked the participants to fill in the same questionnaire within a few seconds, the result would have been biased through the effort of filling questionnaire itself. A bar plot of the outcome is shown in Figure 5. Most raw values are similar for both methods, though the enhanced representation receives slightly better average ratings than the analysis of the first user via video recording. Especially, the temporal demand appears to be significantly lower. A Mann-Whitney test supports the impression that the temporal demand is in general higher for the video recording of the first user ($Mdn = 2.5$) than for analysis with enhanced representation of both users ($Mdn = 1.0$), $U = 111$, $p = .017$. The Pearson's correlation effect size is with 0.38 considered to be of medium to high value, according to [Cohen 1988]. This emphasizes that the analysts felt more rushed and hurried when using the video recording as Web page representation than with the enhanced representation. The positive outcome for both timing and temporal demand supports our second hypothesis (H2). For multiple users, the enhanced representation method requires less time and temporal demand than the representation as video recording.

Table 2: Average absolute percentage errors of the metrics estimations by the analysts for enhanced representation (E) and video recording (V), compared to the ground truth.

		Digg.com	*Jimdo.com*	*Yelp.com*	*CC*
E	*TTFF*	3.0 ± 6.2	0.0 ± 0.0	1.2 ± 3.8	0.0 ± 0.0
	TF	9.4 ± 17.8	31.3 ± 44.3	2.5 ± 7.9	1.0 ± 3.2
V	*TTFF*	15.3 ± 31.8	0.0 ± 0.0	1.2 ± 3.8	0.0 ± 0.0
	TF	15.3 ± 17.1	0.0 ± 0.0	2.5 ± 7.9	0.0 ± 0.0

Difficulty Estimation. We report an average overestimation in difficulty perception on the value scale from one to five of 1.34 for the enhanced representation and 0.90 for the video recording.

5.2 Discussion

The results confirm our hypotheses and indicate that the proposed enhanced representation method allows a less time consuming analysis than the video recording, while enabling the analysts to reach a similar level of accuracy.

The time required for the analysis let us conclude that the analysts had to work on both users independently when presented with the video recording of the Web page interaction. This is because dynamic fixed elements like the pop-up on Creativecommons.org are not visible through the entire video. The analysts had to define at which point of time the pop-up was visible and at which it was not. However, the times required for the analysis of the first user are similar for both methods. We observed that while the time taken in video recording was invested to find out when the elements of interest are visible, for the enhanced representation the analysts

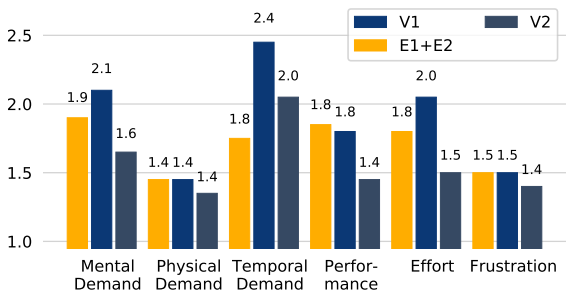


Figure 5: Raw NASA-TLX values

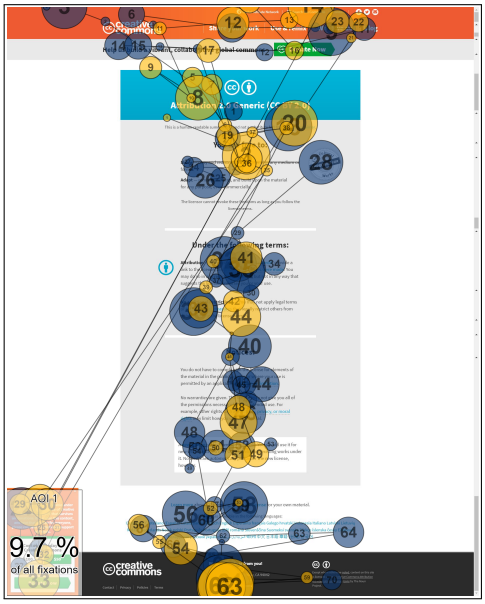


Figure 6: Enhanced representation of the Creativecommons.org Web page with scan paths from evaluation.

invested time in scrolling and zooming, to find the desired parts of the Web page before marking the AOIs.

The NASA-TLX results indicate that the temporal demand to analyze on basis of the first user's video is higher than doing the analysis of both users with the enhanced representation. Though, the mental demand, physical demand, effort and frustration metrics of the average NASA-TLX scores are lower for the second video than for the enhanced representation. One could argue that the analysts get into a "flow" and the task load might converge to a certain minimum, which is below the task load for the enhanced representation. However, when an analyst just has started to analyze the video of the second user, another analyst using the enhanced representation of the same task is already done with the analysis, because of the low time demand on successive users. The visualization of the gaze data of multiple users at once accelerates the analysis process significantly. After the first user, the analyst can reuse existing AOIs, regardless whether they lay upon page-relative or viewport-relative elements on the Web page.

We report a slightly higher estimation of the difficulty by analysts with the enhanced representation. This might be introduced through the visualization of saccades from page-relative into viewport-relative content and vice versa. For example, the saccades shown in Figure 6, relating to the fixed element in the lower left corner. These *virtual saccades* are visualized longer than their actual length, hence analysts might have been unaccustomed in interpretation. We argue, that this is only a matter of training or visual hints for the analysts in the enhanced representation. In contrast to recent visualizations by [Kurzhals et al. 2016a,b], which present cropped pixel data from the foveated area as thumbnails in a timeline per user, our method can retain the overall Web page appearance.

The presented evaluation covers the analysis of gaze data from two users on each Web page, and already shows a clear trend in favor of the enhanced representation. It is obvious that the video recording would suffer significantly once the dataset becomes larger, i.e., for large scale usability study with many users. Some of the recent studies have shown that Web usability studies would require at least 27 users for searching tasks and 34 users for browsing tasks [Eraslan et al. 2016], which emphasizes the potential of the enhanced representation method in real-world applications. Furthermore, with the evolution of cheap eye trackers, the user-base is expanding, and there is an increasing interest for Web page usability analysis using crowd-sourcing [Lebreton et al. 2015].

The composition of the page- and the viewport-relative content might be displayed more interactively, for example allowing the analyst to inspect only certain fixed elements or to move them on the locations as they have been perceived on the page by the users. Moreover, the visualization of accumulated attention might be combined with video recordings of individual attention and interaction. This enables experts to first get an impression about the overall attention pattern and then watch specific outlying behaviors in detail in the video recording. This has the potential to merge the benefits of both methods, the scalability of the enhanced representation method and the intuitive interpretation with the video recording.

6 CONCLUSION

In this paper, we provide an overview of the Web pages usability analysis with eye tracking. We argue that the representation of Web pages is an imperative aspect to support the analysis for pertinent gaze data mapping and visualization. However, current methods are limited in terms of accuracy and scalability. We propose an enhanced representation method that stitches screenshots of the user viewport and extracts fixed elements from the Web page document structure, to tackle both issues. The evaluation results signify the applicability of our method as an accurate and scalable approach. We envision the applicability of presented method in supporting large scale quantitative studies using eye tracking measures.

In future, we aim to extend the extraction approach to improve the precision of fixed element cropping. For example, fixed elements may cast shadows that stay on the page after cropping of the fixed elements. Additionally, fixed element might have overflowing children elements, which expand the bounding box or even define a shape different from a rectangular box. These challenges might be tackled by computer vision approaches, which can be augmented with structural information from the Web page document.

Furthermore, to support large scale studies with crowd-sourcing, we need to ensure that the Web page appears identically for every user. However, many modern pages make use of various interactive elements like image carousels or drop-down menus. Some pages like social networks even present each user a unique Web page or the same information in a different layout per user. This brings a challenging scenario of detecting dynamic elements, and adapting the enhanced representation for gaze data mapping and analysis.

ACKNOWLEDGMENTS

This work is part of project MAMEM that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement number 644780. We also acknowledge the financial support by the Federal Ministry of Education and Research of Germany under the project number 01IS17095B.

REFERENCES

Rossen Atanassov and Arron Eicholz. 2016. *CSS Positioned Layout Module Level 3*. W3C Working Draft. W3C. http://www.w3.org/TR/2016/WD-css-position-3-20160517.

Ana Margarida Barreto. 2013. Do users look at banner ads on Facebook? *Journal of Research in Interactive Marketing* 7, 2 (2013), 119–139. https://doi.org/10.1108/JRIM-Mar-2012-0013

Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomes Ertl. 2017. Visualization of Eye Tracking Data: A Taxonomy and Survey. *Computer Graphics Forum* 36, 8 (2017), 260–284. https://doi.org/10.1111/cgf.13079

Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. 2009. What Do You See when You'Re Surfing?: Using Eye Tracking to Predict Salient Regions of Web Pages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/1518701.1518705

Zoya Bylinskii, Michelle A. Borkin, Nam Wook Kim, Hanspeter Pfister, and Aude Oliva. 2017. *Eye Fixation Metrics for Large Scale Evaluation and Comparison of Information Visualizations*. Springer International Publishing, Cham, 235–255. https://doi.org/10.1007/978-3-319-47024-5_14

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Edward Cutrell and Zhiwei Guan. 2007. What Are You Looking for?: An Eye-tracking Study of Information Usage in Web Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 407–416. https://doi.org/10.1145/1240624.1240690

Andrew T. Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers* 34, 4 (01 Nov 2002), 455–470. https://doi.org/10.3758/BF03195475

Claudia Ehmke and Stephanie Wilson. 2007. Identifying Web Usability Problems from Eye-tracking Data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1 (BCS-HCI '07)*. British Computer Society, Swinton, UK, UK, 119–128. http://dl.acm.org/citation.cfm?id=1531294.1531311

Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2015. Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison. *Journal of Eye Movement Research* 9, 1 (2015). http://dx.doi.org/10.16910/jemr.9.1.2

Sukru Eraslan, Yeliz Yesilada, and Simon Harper. 2016. Eye Tracking Scanpath Analysis on Web Pages: How Many Users?. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 103–110. https://doi.org/10.1145/2857491.2857519

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Chandan Kumar, Raphael Menges, Daniel Müller, and Steffen Staab. 2017. Chromium Based Framework to Include Gaze Interaction in Web Browser. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 219–223. https://doi.org/10.1145/3041021.3054730

Kuno Kurzhals, Marcel Hlawatsch, Michael Burch, and Daniel Weiskopf. 2016a. Fixation-image Charts. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (ETRA '16)*. ACM, New York, NY, USA, 11–18. https://doi.org/10.1145/2857491.2857507

Kuno Kurzhals, Marcel Hlawatsch, Florian Heimerl, Michael Burch, Thomas Ertl, and Daniel Weiskopf. 2016b. Gaze Stripes: Image-Based Visualization of Eye Tracking Data. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016). http://dx.doi.org/10.1109/TVCG.2015.2468091

Fabrizio Lamberti, Gianluca Paravati, Valentina Gatteschi, and Alberto Cannavo. 2017. Supporting Web Analytics by Aggregating User Interaction Data From Heterogeneous Devices Using Viewport-DOM-Based Heat Maps. *IEEE Transactions on Industrial Informatics* 13, 4 (2017), 1989–1999. https://doi.org/10.1109/TII.2017.2658663

Pierre R Lebreton, Toni Mäki, Evangelos Skodras, Isabelle Hupont, and Matthias Hirth. 2015. Bridging the gap between eye tracking and crowdsourcing., In Human Vision and Electronic Imaging. *Proc.SPIE*, 93940W. https://doi.org/10.1117/12.2076745

Jakob Nielsen and Kara Pernice. 2009. *Eyetracking Web Usability* (1st ed.). New Riders Publishing, Thousand Oaks, CA, USA.

Alex Poole and Linden J. Ball. 2005. Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future. In *Prospects, Chapter in C. Ghaoui (Ed.): Encyclopedia of Human-Computer Interaction. Pennsylvania: Idea Group, Inc.*

Hiroyuki Sano, Shun Shiramatsu, Tadachika Ozono, and Toramatsu Shintani. 2013. Web Block Extraction System Based on Client-Side Imaging for Clickable Image Map. *Journal of Communication and Computer 10* (2013), 1–8.

Michael Schiessl, Sabrina Duda, Andreas Thölke, and Rico Fischer. 2003. Eye tracking and its application in usability and media research. *MMI-interaktiv Journal* 1, 06 (mar 2003), 41–50.

O Špakov and Darius Miniotas. 2007. Visualization of eye gaze data using heat maps. *Elektronika ir elektrotechnika* (2007), 55–58.

Tobii AB. 2016. *Tobii Studio User's Manual*. Version 3.4.5.

Tina Walber, Ansgar Scherp, and Steffen Staab. 2014. Smart Photo Selection: Interpret Gaze As Personal Interest. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2065–2074. https://doi.org/10.1145/2556288.2557025