Eye tracking scanpath analysis on web pages: how many users?

DOI:

[10.1145/2857491.2857519](https://doi.org/10.1145/2857491.2857519)

[Link to publication record in Manchester Research Explorer](#)

Citation for published version (APA):

Eraslan, S., Yeliz, Y., & Harper, S. (2016). Eye tracking scanpath analysis on web pages: how many users? In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications* (pp. 103-110). (ETRA '16). Association for Computing Machinery. <https://doi.org/10.1145/2857491.2857519>

Published in:

Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications

Citing this paper

Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights

Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy

If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [<http://man.ac.uk/04Y6Bo>] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Eye Tracking Scanpath Analysis on Web Pages: How Many Users?

Sukru Eraslan^{1,2}, Yeliz Yesilada², and Simon Harper¹

¹School of Computer Science, University of Manchester, Manchester, United Kingdom
sukru.eraslan@postgrad.manchester.ac.uk | simon.harper@manchester.ac.uk

²Middle East Technical University Northern Cyprus Campus, Kalkanlı, Guzelyurt, Mersin 10, Turkey
seraslan@metu.edu.tr | yyeliz@metu.edu.tr

Abstract

The number of users required for usability studies has been a controversial issue over 30 years. Some researchers suggest a certain number of users to be included in these studies. However, they do not focus on eye tracking studies for analysing eye movement sequences of users (i.e., scanpaths) on web pages. We investigate the effects of the number of users on scanpath analysis with our algorithm that was designed for identifying the most commonly followed path by multiple users. Our experimental results suggest that it is possible to approximate the same results with a smaller number of users. The results also suggest that more users are required when they serendipitously browse on web pages in comparison with when they search for specific information or items. We observed that we could achieve 75% similarity to the results of 65 users with 27 users for searching tasks and 34 users for browsing tasks. This study guides researchers to determine the ideal number of users for analysing scanpaths on web pages based on their budget and time.

Keywords: eye tracking, scanpath, web pages, sample size, users

Concepts: •Human-centered computing → Human computer interaction (HCI); Usability testing; Laboratory experiments;

1 Introduction

The number of users required for usability studies is being discussed over 30 years in the literature [Lewis 1982; Schmettow 2012]. Specific numbers, ranging from five users to 20 users (See Section 2), have already been suggested despite the fact that more users are typically used for statistical analysis (such as Jerger et al. [2014]). There is also a considerable number of views which argue against these suggested numbers [Bevan et al. 2003; Alroobaea and Mayhew 2014; Schmettow 2012]. They assert that each study has its own characteristics, and therefore the number of users should not be the same for all usability studies. In particular, it is typically suggested to include more users in more complicated studies [Alroobaea and Mayhew 2014; Schmettow 2012].

Eye tracking studies are increasingly used in usability research, especially for improving the usability of web pages [Ehmke and Wilson 2007]. In particular, the transition probabilities between web page elements can be a guide to evaluate the efficiency of the arrangements of the elements on the web page [Ehmke and Wilson

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. © 2016 ACM.
ETRA 2016, March 14 - 17, 2016, Charleston, SC, USA
ISBN: 978-1-4503-4125-7/16/03
DOI: <http://dx.doi.org/10.1145/2857491.2857519>

Figure 1: An example of a scanpath on the BBC web page which is segmented into its visual elements

2007]. Moreover, longer scanpaths can be associated with less efficient searching on web pages [Ehmke and Wilson 2007]. However, as researchers, we commonly experience difficulties in finding users for our eye tracking studies. A separate session should be allocated for each user in these studies. When there is only one eye tracker and/or one researcher, the study cannot be conducted in parallel. Therefore, researchers may need a considerable amount of time to complete their studies. To counteract with these problems, it is important to determine the ideal number of users. In this paper, we investigate the effects of the number of users in eye tracking studies that are conducted for analysing eye movement sequences of users (i.e., scanpaths) on web pages.

Whilst we are reading web pages, our eyes make fixations where they become relatively stationary. Our eyes also make quick movements between fixations called saccades. The sequences of fixations and saccades represent our scanpaths on web pages. Figure 1 illustrates an example of a scanpath on the BBC web page which is segmented into its visual elements, such as headers, search boxes and menu items [Akpınar and Yeşilada 2013]. These elements help users to complete their tasks. For example, on the BBC web page, users can click on the *Weather* menu item (the element D) to access weather estimations. As can be seen from this figure, fixations are represented by circles where longer fixations are illustrated by larger circles. Moreover, saccades are represented by straight lines.

Scanpaths can allow to recognise how users traverse web pages. By analysing user scanpaths, we can understand which visual elements were mostly fixated and in which order. This can then be used for different purposes, specifically for re-engineering (i.e., transcoding)

web pages to make them more accessible in constrained environments such as on small screen devices [Akpınar and Yeşilada 2015] and in audio representation [Yesilada et al. 2013]. For example, visual elements can be re-ordered to give a direct access to firstly and commonly fixated visual elements, and therefore users can directly focus on the desired content. This kind of transcoding is known as *Experiential Transcoding* [Yesilada et al. 2013]. Scanpaths can also analysed for guiding web designers to improve their designs or testing whether users follow the expected path for a particular goal on a web page [Ehmke and Wilson 2007].

The aim of this paper is to investigate whether a smaller group of users can also reveal the approximate results for scanpath analysis. Hence, our research question here is: *Is it possible to approximate the same results in scanpath analysis with a smaller group of users?*

In order to achieve the aim, we conducted experiments with our algorithm called *Scanpath Trend Analysis (STA)* [Eraslan et al. 2016b]. This algorithm analyses individual scanpaths of multiple users on a web page and then identifies their most commonly followed path in terms of the visual elements of the web page. We call this path *Trending Scanpath*. Trending scanpath differs from an absolute path (i.e., fully shared path). Existing algorithms typically try to find an absolute path, and therefore they are likely to produce unacceptably short results because of the variations caused by individual differences. In the STA, if a particular visual element is not shared by all individual scanpaths but it gets at least the same attention as the shared visual elements, the visual element is included in the resulting scanpath. In comparison with other existing algorithms, this algorithm provides the most representative scanpath for multiple users [Eraslan et al. 2016b].

In our previous paper, we described the STA algorithm with its evaluation by using an eye tracking study in depth [Eraslan et al. 2016b] (Sections 3 and 4). In this paper, we present a new study to investigate whether we can approximate the same results in scanpath analysis with a smaller group of users (Sections 5-8).

2 Related Work

Researchers typically come across difficulties in finding users for their usability studies [Renkema-Padmos et al. 2014]. There are researchers who indicate that there is no need for a large volume of users for usability studies. These researchers commonly analyse some previous studies and model their findings to estimate the ideal number of users [Nielsen and Landauer 1993; Law and Hvannberg 2004; Hwang and Salvendy 2010]. These numbers are known as *Magic Numbers* in the literature [Bevan et al. 2003]. This section firstly discusses these magic numbers and then associate them with scanpath analysis on web pages.

The most popular magic number is five users which is suggested by Nielsen and Landauer [1993]. In order to estimate this number, they analysed their own previous studies and then fit their findings with a Poisson mathematical model. The equation of this model is shown in Equation 1 where i is the number of users, N is the total number of problems, and λ is the probability of finding a usability problem when a single, average subject, or a single, average heuristic evaluator, is used [Nielsen and Landauer 1993]. Based on their own previous studies, they suggest that the typical value of λ is equal to 31%.

$$Found(i) = N * (1 - (1 - \lambda)^i) \quad (1)$$

Based on the Poisson model, they suggest that it is possible to discover 85% of usability problems of a particular product with only five users. However, their study has got some considerable limitations. Although they calculated the λ value by using their own

Figure 2: An example of a heat map on the BBC web page

previous studies, there are views which argue that each study has its own features, and therefore this value cannot be unique. For example, it can be changed according to the complexity of products and/or the characteristics of users [Spool and Schroeder 2001; Caulton 2001]. In order to address this problem, Lewis [2001] suggests to conduct some initial sessions and then estimate this value based on the sessions. However, Nielsen [2012] suggests that when a web page consists of many different features, several different usability tests are needed instead of increasing the number of users for each test.

Further studies were conducted by Faulkner [2003] to test whether five users are sufficient to find 85% of usability problems. She noticed that this is not the case for all five users. Based on her experiments, five users found only 55% of the usability problems in some cases. She then states that 15 users should be included in a study to find 85% and above of usability problems. Moreover, Hwang and Salvendy [2010] conducted a meta-analysis by analysing many published research papers since 1990 and suggest that at least 10 ± 2 users are required to discover 80% of usability problems. Furthermore, Alroobaea and Mayhew [2014] recommend to consider 20 users as a maximum number of users and 12 users as a minimum number of users for usability studies along with the complexity of the study and the criticality of its context.

As can be seen from the summary above, existing research typically tries to find the minimum number of users for usability studies which attempt to find usability problems but do not focus on eye tracking studies. However, eye tracking studies have been widely conducted to investigate user interactions with web pages for different purposes in usability research, particularly for improving the usability of web pages [Ehmke and Wilson 2007].

Pernice and Nielsen [2009] have certain suggestions for eye tracking studies. They suggest researchers to include 39 users if they want to have a stable heat map. Heat maps represent mostly fixated areas on visual stimuli [Tobii Technology 2010]. As can be seen from an example of a heat map in Figure 2, they consist of different colours where red colour usually represents the highest number of fixations or the longest time, in contrast, green colour represents the least [Tobii Technology 2010]. Although these maps are useful to directly recognise the most commonly fixated areas, they do not take a sequence into account. Therefore it should not be interpreted

Figure 3: An example of 40 user scanpaths on the BBC web page

that 39 users can also be sufficient for analysing user scanpaths. As illustrated in Figure 3, user scanpaths are much more complicated than heat maps for analysis.

Pernice and Nielsen [2009] also state that if a qualitative study will be conducted by watching eye movements of users, six users are sufficient. Watching eye movements of users on a particular web page may provide a general idea about the user behaviours on the web page. However, as eye trackers collect a large amount of data, some important details can be eluded if researchers only watch the recordings. Similar to the views against the magic numbers for usability studies, there are also views which claim that there is no unique number for eye tracking studies because of their objectives and characteristics [Rösler 2012].

In this paper, we investigate whether a smaller group of users can also reveal the approximate results in scanpath analysis by using our STA algorithm. Similar to the most of other existing scanpath analysis algorithms (such as West et al. [2006], Goldberg and Helfman [2010] and Eraslan et al. [2014]), our algorithm conducts a scanpath analysis based on the areas of visual stimuli that we call them visual elements of web pages.

Other existing scanpath analysis algorithms can mainly be used for determining similarities or dissimilarities between scanpaths, calculating transition probabilities between visual elements, detecting sequential patterns within given scanpaths, and identifying a representative scanpath for multiple users [Eraslan et al. 2016a]. These algorithms typically do not take the durations of fixations into consideration, but the STA algorithm considers the durations to identify a trending scanpath [Eraslan et al. 2016b]. Furthermore, the algorithms are typically affected a lot because of the variations caused by individual differences while they are detecting patterns and identifying a representative scanpath for multiple users [Eraslan et al. 2016a]. Hence, their results might not be sufficiently representative for the users. In comparison with other algorithms, the STA algorithm provides the most representative scanpath [Eraslan et al. 2016b].

3 STA: Scanpath Trend Analysis

The STA algorithm is designed for analysing multiple scanpaths on a web page to identify their trending scanpath in terms of the visual

elements of the web page [Eraslan et al. 2016b].

We already evaluated the STA algorithm by using an eye tracking study with 40 participants and six web pages for browsing and searching tasks [Eraslan et al. 2016b]. In our evaluation, we compared the resulting scanpaths of the STA algorithm and other existing algorithms with the individual scanpaths. We then applied the paired T-Test or its non-parametric alternative Wilcoxon Signed Rank Test to investigate whether the resulting scanpaths of the STA algorithm were more similar to the individual scanpaths in comparison with the resulting scanpaths of other algorithms. Based on our evaluation, the STA algorithm performs significantly better than other existing algorithms in terms of providing a similar scanpath to individual scanpaths, and this is not an occurrence by chance ($p < 0.05$).

In order to identify a trending scanpath, the STA algorithm consists of three main stages which are briefly explained below.

3.1 Preliminary Stage: Individual Scanpaths

In the preliminary stage, the eye movement sequences of users are represented in terms of the visual elements of the web page. For example, the individual scanpath in Figure 1 is represented as LLLTTP.

To segment web pages into their visual elements, we use the extended and improved version of the Vision-Based Page Segmentation (VIPS) algorithm [Akpınar and Yeşilada 2013]. This segmentation algorithm automatically segments a web page into its visual elements by using both its source code and visual representation (See Figure 1). As it also correlates the visual elements with the underlying source code, user scanpaths can be used for further processing of web pages, especially for the purpose of re-engineering. The VIPS algorithm segments web pages based on a granularity level which is inversely proportional to the sizes of segments. Specifically, when the granularity level increases, more and smaller segments are provided. A user study conducted by Akpınar and Yeşilada [2013] suggests the fifth segmentation granularity level as the most successful level with approximately 74% user satisfaction. Hence, we use this level to segment web pages.

3.2 First Pass: Trending Visual Elements

In the first pass, the STA algorithm analyses the individual scanpaths to identify the trending visual elements by using their occurrence frequencies and the total durations of fixations on the visual elements (i.e., dwell time).

A user can visit the same visual element more than once which can be consecutive (such as LLTP) and/or non-consecutive (such as LTPL). Each non-consecutive visit is referred to as a visual element instance. For example, LLTPL has two visual element instances of L. These instances are differentiated by giving them different numbers based on their durations as the durations are associated with information processing [Velichkovsky et al. 2002; Follet et al. 2011]. In particular, the longest instance of a particular element will get the first number in an individual scanpath (such as L1 T1 P1 L2). The visual element instance can be trending if it meets with both of the following two conditions:

1. Its total occurrence frequency $\geq$ the minimum total frequency of the visual element instances shared by all of the individual scanpaths.
2. Its total duration $\geq$ the minimum total duration of the visual element instances shared by all of the individual scanpaths.

3.3 Second Pass: Trending Scanpath

In the second pass, the STA algorithm locates the trending visual element instances into a trending scanpath according to their overall positions in the individual scanpaths. The individual scanpaths are firstly abstracted (e.g.: L1 [200 ms] L1 [300 ms] T1 [100 ms] → L1 [500 ms] T1 [100 ms]). The algorithm then determines a priority value (ψ) for each instance in each individual scanpath by using Equation 2 where P is the instance position in the individual scanpath (starting with zero) and L is the length of the individual scanpath. The maximum value and the minimum value are taken as 1 and 0.1 respectively.

$$\psi = 1 - P \cdot \frac{\max\text{Value} - \min\text{Value}}{L - 1} \quad \Psi = \sum_{i=1}^n \psi_i \quad (3)$$

The total priority value (Ψ) for each instance in all the individual scanpaths is then determined by using Equation 3 where n is the number of individual scanpaths. The instances are then located in the resulting scanpath according to their total priority values in descending order. In case of the same total priority values, the total duration and occurrence frequencies are considered.

Finally, the numbers of the visual element instances are removed (e.g. L1 → L) and then the consecutive repetitions are abstracted (e.g. LLTP → LTP) so that the trending scanpath is represented in terms of the visual elements.

4 An Eye Tracking Study

In order to investigate the effects of the number of users on scanpath analysis, we used an eye tracking study conducted at both the University of Manchester and Middle East Technical University Northern Cyprus Campus (METU NCC) in 2013 with 40 users. This study was previously used to evaluate the STA algorithm [Eraslan et al. 2016b]. To increase the number of users, we conducted the same eye tracking study at METU NCC in 2014. Therefore, we now have 81 users in total to investigate whether or not it is possible to approximate the same results with a smaller number of users in eye tracking studies conducted for analysing user scanpaths.

4.1 Equipment and Materials

To record eye movements of the users, a 17" monitor was used with a built-in Tobii T60 eye tracker. The screen resolution was adjusted as 1280 x 1024. These users were asked to complete specific tasks on the saved versions of the Apple, Babylon, AVG, Yahoo, Godaddy and BBC web pages (See Open Data Section). This ensures that the pages were not changed or updated between the sessions.

These web pages were randomly chosen from a user study that aimed to evaluate the extended and improved version of the VIPS algorithm [Akpınar and Yeşilada 2013]. In that study, the top 100 web pages listed by ALEXA.com in 2013 were investigated with the Visual Complexity Rankings and Accessibility Metrics (ViCRAM) framework [Michailidou 2010]. The ViCRAM framework discovers visual complexities of web pages that are correlated with users' perception [Michailidou 2010]. Once the complexities of the web pages were determined, ten web pages were randomly selected for each of the low, medium and high complexity groups. To consider web pages from different visual complexity groups, two web pages were randomly selected from each complexity group for the eye tracking study. The Apple and Babylon pages have a low complexity, the AVG and Yahoo pages have a medium complexity, and the Godaddy and BBC pages have a high complexity.

4.2 User Tasks

The tasks used were categorised into two groups which were browsing and searching tasks. In the browsing tasks, the users serendipitously browsed on the web pages. In contrast, they were requested to find specific information or items on the web pages for the searching tasks. These tasks can fit into G. Marchionini's Search Activities Model which is one of the most popular models in categorising user tasks on web pages [Marchionini 2006]. In this model, user tasks are divided into the lookup, learn and investigation groups. The browsing tasks belong to the investigation group, and the searching tasks belong to the lookup group. The searching tasks are provided below.

- **Apple:** (a) Can you locate a link which allows watching the TV ads relating to iPad mini? (b) Can you locate a link labelled iPad on the main menu?
- **Babylon:** (a) Can you locate a link that you can download the free version of Babylon? (b) Can you find and read the names of other products of Babylon?
- **AVG:** (a) Can you locate a link which you can download a free trial of AVG Internet Security 2013? (b) Can you locate a link which allows you to download AVG Anti-virus Free 2013?
- **Yahoo:** (a) Can you read the titles of the main headlines which have smaller images? (b) Can you read the first item under the News title?
- **Godaddy:** (a) Can you find a telephone number for technical support and read it? (b) Can you locate a text box where you can search for a new domain?
- **BBC:** (a) Can you read the first item of the Sport News? (b) Can you locate the table that shows market data under the Business title?

4.3 Procedure

The users firstly read an information sheet about the study and then signed the consent form to accept that they were aware of their rights and their participations were voluntary. After that, they were requested to fill in a brief questionnaire that required their gender, age group and educational level. Following this, they sat in front of the monitor and viewed the web pages two times for browsing (30 seconds) and searching (maximum 120 seconds) in a random order. To deal with the memory effects, the half of the participants were asked to do the browsing tasks at their first visits to the web pages and do the searching tasks at their second visits. The task order was reversed for another half of the participants.

4.4 Participants

This study was conducted in 2013 with 20 female and 20 male users, and in 2014 with 20 female and 21 male users. In total, 40 females and 41 males participated in the study. Most of these users were students, along with some academic and administrative staff at the two universities. All participants used the web on a daily basis, except for one male who claimed that he used the web weekly. The majority of these users (55 users) were aged between 18 and 24, then 25-34 (20 users) and then 35-54 (6 users). Furthermore, 55 users completed their high schools, seven users held a bachelor degree, ten users completed their master degrees and nine users held a PhD degree.

5 Methodology

Before we started our experiments, we noticed that some of these users could not be successful in completing their tasks on the web

pages and/or had eye calibration problems. We defined these users as unsuccessful users and then had to exclude them from the experiments. Subsequently, there were different numbers of users on the web pages for the browsing and searching tasks. For instance, there were 67 successful users on the Apple page whereas 77 successful users on the BBC page for the searching tasks. To have an equal number of users, we also randomly excluded some of the successful users. We then had 65 successful users on each web page for the browsing and searching tasks.

In order to investigate whether or not it is possible to approximate the same results in scanpath analysis with a smaller group of users, we firstly segmented the web pages with the extended and improved version of the VIPS algorithm with respect to the granularity level five [Akpınar and Yeşilada 2013]. We then applied the STA algorithm to the users for identifying their trending scanpaths. After that, we randomly selected smaller groups of users with the size of one user to 64 users from the successful users and then applied the STA algorithm to identify their trending scanpaths. These trending scanpaths were then compared with the trending scanpaths of the entire group. In order to eliminate any possible effects of the selected users, we generated 100 different combinations for each group size and calculated their median similarity to the entire group as a similarity to the entirety of the group [Faulkner 2003]. The reason for using the median similarity instead of the average similarity was to deal with any possible extreme cases.

The Levenshtein Distance (String-edit) algorithm was used to compare two scanpaths in our experiments as it has commonly been used to calculate the distance between two scanpaths represented as strings [Josephson and Holmes 2002; Takeuchi and Habuchi 2007]. Since our scanpaths were represented in terms of the visual elements of the web pages, they were also in a string format. This enabled us to use this algorithm for the comparison. The String-edit algorithm uses a minimum number of operations (insertion, deletion and substitution) to transform one string to another [Josephson and Holmes 2002]. For example, the distance between ABCDE and ABCDF is calculated as one because the transformation can be done by only substituting E with F or vice versa. However, the substitution costs between all pairs of visual elements potentially are not the same because the sizes and the distances between the visual elements may be different [Josephson and Holmes 2002]. In order to take this into consideration, it has been suggested to use a substitution matrix which holds substitution costs between all pairs of visual elements [Takeuchi and Habuchi 2007]. As suggested by Takeuchi and Habuchi [2007], we used an Euclidean distance between the visual elements to create a substitution matrix. Equation 4 shows how to calculate a substitution cost between two visual elements $\vec{A}$ and $\vec{B}$ based on an Euclidean distance where A_1 and A_2 are x and y coordinates of the centre of the visual element $\vec{A}$, and α is a type of normalisation parameter [Takeuchi and Habuchi 2007]. We used 0.001 as a value of the parameter as suggested by Takeuchi and Habuchi [2007].

$$f(\vec{A}, \vec{B}) = \alpha \sum_{i=1}^2 \sqrt{(\vec{A}_i - \vec{B}_i)^2} \quad (4)$$

The substitution matrix was then used with the String-edit algorithm to find the distance between two scanpaths by calculating the minimum total cost to transform one scanpath to another one. To find the similarity between two scanpaths as a percentage based on the String-edit distance [Underwood et al. 2008], the distance between two scanpaths was divided by the length of the longer scanpath to find a normalised score which allows preventing possible inconsistencies caused by different lengths. The normalised score was then subtracted from one and multiplied by 100.

When the similarity to the entire group was calculated for each smaller group, we used the curve estimation feature of SPSS¹ to select the best curve that fits into the average of these similarities for the six web pages for the browsing and searching tasks. In other words, we selected the curve with the minimum standard error of the estimate.

6 Results

The average, median, standard deviation, minimum and maximum similarities between the trending scanpaths of the entire group and the 100 random subsets of each group size on each web page for the browsing and searching tasks are provided in our external repository (See Open Data Section). Figure 4 shows the average estimated curves with their equations for the browsing and searching tasks where S is the similarity to entire group and i is the number of users. Both of these curves illustrate that it is possible to approximate almost the same results with a smaller group of users in scanpath analysis.

Based on our experimental results, 27 users can provide 75% similar results to the results of 65 users for searching tasks whereas 34 users are needed to achieve 75% similarity for browsing tasks. Table 1 also shows the numbers of users required to reach 75%, 80%, 85%, 90%, 95% and 100% similar results to the results of 65 users for browsing and searching tasks. As can also be seen from the table, after 52 users, the same trending scanpath is likely to be identified when users are asked to find specific information or items on web pages (See also Figure 4).

Table 1: The numbers of users required to reach 75%, 80%, 85%, 90%, 95% and 100% similar results to 65 users' results for browsing and searching tasks based on Figure 4

The similarity to the entire group	The number of users for browsing tasks	The number of users for searching tasks
75%	34	27
80%	39	31
85%	45	36
90%	51	41
95%	58	46
100%	65	52

In order to cross-validate the results, we divided the web pages into two equal-sized sets. Each set consists of three randomly chosen web pages with the low, medium and high complexity. These sets were named as the set A and the set B. The Babylon, AVG and Godaddy pages were included in the set A whereas the Apple, Yahoo and BBC pages were included in the set B. To estimate the best curve for these two sets for the browsing and searching tasks, we applied the same methodology described in Section 5.

To compare the results of the set A and the set B, we calculated the absolute difference in their curves for each group size. For example, as illustrated in Table 2, the difference between the set A and the set B for 15 users is equal to 7.30 (Set A: 55.99, Set B: 48.70) for the browsing task and 11.60 (Set A: 51.91, Set B: 63.51) for the searching task. Table 3 also shows the average, maximum, minimum and standard deviation of these differences between these two sets for the browsing and searching tasks. As can be seen from the table, the average difference between the curves of the set A and the set B is only 8.06% for the searching tasks and 11.98% for the browsing tasks.

¹http://www-01.ibm.com/support/knowledgecenter/SSLVMB_23.0.0/spss/base/curve_estimation_models.dita?lang=en

Figure 4: Curve estimation for the six web pages for the browsing and searching tasks on average

Table 2: The differences in the curves of the set A and the set B [i: The number of users, S_{setA} : The similarity to the entire group based on the set A, S_{setB} : The similarity to the entire group based on the set B] (See Open Data Section for the complete table)

Task	i	S_{setA}	S_{setB}	Difference
Browsing	...	..	..	..
	10	46.54	40.94	5.60
	...	...	...	...
	15	55.99	48.70	7.30
	...	...	...	...
Searching	...	...	...	...
	10	41.38	55.31	13.92
	...	...	...	...
	15	51.91	63.51	11.60
	...	...	...	...

Table 3: The average, maximum, minimum and standard deviation of these differences between the set A and the set B

Task	Average	Max	Min	Std. Deviation
Browsing	11.32	18.07	1.01	4.67
Searching	8.09	15.67	0.09	4.68

The cross-validation supports our findings illustrated in Figure 4 with some variations. In particular, the results of the cross-validation also suggest that 27 users can achieve $75 \pm 3\%$ similar results to the results of 65 users for searching tasks and 34 users can achieve $75 \pm 6\%$ similar results to the results of 65 users for browsing tasks.

7 Discussion

According to our experimental results, there are some differences in the results because of the type of tasks. In particular, more users were needed to reach 75% similar results for the browsing tasks. This is an expected situation because no specific requirement was given for the browsing tasks such as reading the first item of Sport

News on the BBC page or finding the telephone number for technical support on the Godaddy page. Without a specific requirement, the users freely browsed on the web pages and focussed on the visual elements that attracted their attention. Therefore, the variances between the individual scanpaths for the browsing tasks are likely to be higher than the variances for the searching tasks. Based on the results, we can suggest that when users are requested to find specific information or items on web pages, a smaller group of users is needed in comparison with when users freely browse on web pages to reach the desired results.

Pernice and Nielsen [2009] suggest to conduct eye tracking studies with 39 users for stable heat maps. Heat maps do not represent the sequences of fixations, but our study focuses on these sequences. As also mentioned in Section 2, eye movement sequences are more complicated for analysis in comparison with heat maps. However, based on our experimental results, we can suggest that our findings are still consistent with their suggestions, especially for browsing tasks (See Section 6).

The majority of the users included in our experiments were students. However, we tried to increase the diversity between the users by conducting the eye tracking study at both the University of Manchester and Middle East Technical University Northern Cyprus Campus. We also invited people to our study from different countries and different departments. More web pages could also be included from different categories such as a social networking website and a search engine. However, we believe that the results are not affected significantly because the focus here was the structures of web pages, not their contents. Although there are also some other existing approaches for identifying a common scanpath of multiple users (such as Eraslan and Yesilada [2015]), our STA algorithm provides the most representative common scanpath (See Section 3) [Eraslan et al. 2016b]. Because of this reason, we used the STA algorithm in our experiments.

In our experiments, we used the fifth granularity level of the extended and improved version of the VIPS algorithm to segment web pages [Akpınar and Yeşilada 2013]. We are now planning to investigate the effects of different granularity levels on our findings. In addition, the String-edit algorithm was only used in our experiments to calculate the similarities between the resulting scanpaths [Takeuchi and Habuchi 2007]. Therefore, we are also planning to use an alternative method to the String-edit algorithm (such as

the NeedlemanWunsch algorithm [Cristino et al. 2010]) to investigate whether our findings are still consistent. The experiments can also be conducted with different segmentation approaches and different scanpath analysis algorithms to cross-validate the results in the future. For example, a grid-layout segmentation [Underwood et al. 2008] can be used to segment web pages into their visual elements and a transition matrix can be constructed to find the transition probabilities between the visual elements [West et al. 2006]. Furthermore, Eraslan and Yesilada [2015] suggest that the results of scanpath analysis can be affected by individual factors, such as a familiarity with web pages. These factors can also be considered to investigate whether or not they can affect the curve estimations. We believe that our work is a promising step towards understanding how the number of users could be estimated for achieving similar results on scanpath analysis.

We also wanted to conduct our experiments with a third-party eye tracking dataset to cross-validate the results but we could not find a publicly available eye tracking dataset on real web pages – not on the images of web pages. The lack of publicly available datasets on web pages was also highlighted by Shen and Zhao [2014]. We also tried to contact some other researchers in related fields² if they have appropriate eye tracking datasets with a sufficient number of users to share with us. Unfortunately, we could not find any public dataset on real web pages for this particular study. Because of this reason, we decided to make our anonymised dataset available so that other researchers can use the dataset for their studies (See Open Data Section). We believe that this would also be a useful contribution to the eye tracking community. We are also planning to conduct another eye tracking study in the long run with different users on different web pages to cross-validate our findings. In that case, the current eye tracking dataset will be considered as a training set and the new dataset will be considered as a validation set.

Scanpath analysis is useful to understand how users interact with web pages, and therefore it contributes to usability research. In particular, Experiential Transcoding suggests that when web pages are transcoded by considering user behaviours, the pages become more accessible in constrained environments, such as on small screen devices [Yesilada et al. 2013; Akpinar and Yeşilada 2015]. Specifically, the visual elements can be re-arranged based on the trending scanpath. However, this should be further investigated with some additional studies. In addition to usability research, scanpath analysis on web pages can also be used in different fields. For instance, Information Retrieval (IR) researchers would also benefit from this analysis by getting some guidance on which parts of web pages should be provided to users and in which order. Scanpath analysis can also be beneficial for advertisers as they typically want to make users interested in their products as soon as possible the users look at web pages [Janoschka 2004].

8 Concluding Remarks

Eye tracking studies have been widely used in usability research, especially for investigating users interactions with web pages to improve their usability [Ehmke and Wilson 2007]. However, researchers typically encounter difficulties in finding users for their eye tracking studies. These studies also usually need more time to complete. Hence, it is crucial for researchers to determine the ideal number of users based on their budget and time (See Section 1). In the literature, some numbers have been suggested but these numbers are typically associated with the studies which aim to find usability problems of a particular product (See Section 2).

Our objective was to investigate the effects of the number of users

in eye tracking studies conducted for analysing eye movement sequences of users on web pages. In order to achieve this, we conducted experiments with an eye tracking study, our scanpath analysis algorithm and the extended and improved version of the VIPS algorithm (See Sections 3-5). Our major findings show the possibility to reach almost the same results with a smaller group of users (See Section 6). From our research, we observed that we can approximate the results of 65 users with 75% similarity by including 27 users for searching tasks and 34 users for browsing tasks.

Since the number of users for usability studies has been a controversial issue over 30 years [Lewis 1982; Schmorrow 2012] and existing research does not focus on eye tracking studies which analyse user scanpaths on web pages, this paper contributes to both usability and eye tracking fields. These findings are helpful and useful for researchers who are interested in using eye tracking studies to analyse user scanpaths. Particularly, the researchers who have limited budget and time would benefit from this work.

OPEN DATA

The experiments' data including the materials (the saved versions of the Apple, Babylon, AVG, Yahoo, Godaddy and BBC web pages, information sheet, consent form and questionnaire), the individual scanpaths in terms of the visual elements of the web pages can be accessed at http://iam-data.cs.manchester.ac.uk/data_files/19. This link also allows to access the average, median, standard deviation, minimum and maximum similarities between the trending scanpaths of the entire group and the 100 random subsets of each group size on each web page for the browsing and searching tasks. Moreover, it includes the differences in the curves of the set A and the set B.

References

- AKPINAR, M. E., AND YEŞİLADA, Y. 2013. Vision Based Page Segmentation Algorithm: Extended and Perceived Success. In *Current Trends in Web Engineering - ICWE 2013 Workshops and PhD Symposium.*, Springer, Q. Z. Sheng and J. Kjeldskov, Eds., vol. 8295 of *LNCS*, 238–252.
- AKPINAR, E., AND YEŞİLADA, Y. 2015. “Old Habits Die Hard!”: Eyetracking Based Experiential Transcoding: A Study with Mobile Users. In *Proceedings of the 12th Web for All Conference*, ACM, New York, NY, USA, W4A '15, 12:1–12:5.
- ALROOBAAE, R., AND MAYHEW, P. 2014. How many participants are really enough for usability studies? In *Science and Information Conference (SAI), 2014*, 48–56.
- BEVAN, N., BARNUM, C., COCKTON, G., NIELSEN, J., SPOOL, J., AND WIXON, D. 2003. The “Magic Number 5”: Is It Enough for Web Testing? In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI EA '03, 698–699.
- CAULTON, D. A. 2001. Relaxing the homogeneity assumption in usability testing. *Behaviour and Information Technology* 20, 1, 1–7.
- CRISTINO, F., MATHÔT, S., THEEUWES, J., AND GILCHRIST, I. D. 2010. ScanMatch: a Novel Method for Comparing Fixation Sequences. *Behavior Research Methods* 42, 3, 692–700.
- EHMKE, C., AND WILSON, S. 2007. Identifying Web Usability Problems from Eye-tracking Data. In *Proceedings of the 21st British HCI Group Annual Conference on People and Comput-*

²chi-web@listserv.acm.org

- ers: *HCI... but not as we know it*, British Computer Society, Swinton, UK, vol. 1 of *BCS-HCI '07*, 119–128.
- ERASLAN, S., AND YESILADA, Y. 2015. Patterns in eyetracking scanpaths and the affecting factors. *Journal of Web Engineering* 14, 5&6, 363–385.
- ERASLAN, S., YESILADA, Y., AND HARPER, S. 2014. Identifying patterns in eyetracking scanpaths in terms of visual elements of web pages. In *Web Engineering*, S. Casteleyn, G. Rossi, and M. Winckler, Eds., vol. 8541 of *LNCS*. Springer International Publishing, 163–180.
- ERASLAN, S., YESILADA, Y., AND HARPER, S. 2016. Eye Tracking Scanpath Analysis Techniques on Web Pages: A Survey, Evaluation and Comparison. *Journal of Eye Movement Research* 9, 1:2, 1–19.
- ERASLAN, S., YESILADA, Y., AND HARPER, S. 2016. Scanpath Trend Analysis on Web Pages: Clustering Eye Tracking Scanpaths. *ACM Transactions on the Web (Accept subject to minor revisions)* ..., ...
- FAULKNER, L. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers* 35, 3, 379–383.
- FOLLET, B., MEUR, O. L., AND BACCINO, T. 2011. New insights into ambient and focal visual fixations using an automatic classification algorithm. *iPerception: Open-access Journal of Human, Animal, and Machine Perception* 2, 6, 592–610.
- GOLDBERG, J. H., AND HELFMAN, J. I. 2010. Scanpath Clustering and Aggregation. In *Proceedings of the 2010 Symposium on Eye Tracking Research and Applications*, ACM, New York, NY, USA, ETRA '10, 227–234.
- HWANG, W., AND SALVENDY, G. 2010. Number of People Required for Usability Evaluation: The 10 ± 2 Rule. *Communications of the ACM* 53, 5 (May), 130–133.
- JANOSCHKA, A. 2004. *Web Advertising: New Forms of Communication on the Internet*. John Benjamins Publishing Company.
- JERGER, S., DAMIAN, M. F., TYE-MURRAY, N., AND ABDI, H. 2014. Children use visual speech to compensate for non-intact auditory speech. *Journal of Experimental Child Psychology* 126, 2014, 295312.
- JOSEPHSON, S., AND HOLMES, M. E. 2002. Visual Attention to Repeated Internet Images: Testing the Scanpath Theory on the World Wide Web. In *Proceedings of the 2002 Symposium on Eye Tracking Research and Applications*, ACM, New York, NY, USA, ETRA '02, 43–49.
- LAW, E. L.-C., AND HVANNBERG, E. T. 2004. Analysis of Combinatorial User Effect in International Usability Tests. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '04, 9–16.
- LEWIS, J. R. 1982. Testing small system customer set-up. In *The Annual Meeting of the Human Factors and Ergonomics Society. Human Factors Society*, 718–720.
- LEWIS, J. R. 2001. Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction* 13, 445–479.
- MARCHIONINI, G. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (Apr.), 41–46.
- MICHAILIDOU, E. 2010. *ViCRAM: Visual Complexity Rankings and Accessibility Metrics*. PhD thesis, University of Manchester.
- NIELSEN, J., AND LANDAUER, T. K. 1993. A Mathematical Model of the Finding of Usability Problems. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '93, 206–213.
- NIELSEN, J., 2012. How Many Test Users in a Usability Study? <http://www.nngroup.com/articles/how-many-test-users/>, June.
- PERNICE, K., AND NIELSEN, J. 2009. How to Conduct Eyetracking Studies. Tech. rep., Nielsen Norman Group.
- RENKEMA-PADMOS, A., VOLKAMER, M., AND RENAUD, K. 2014. Building Castles in Quicksand: Blueprint for a Crowdsourced Study. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, NY, USA, 643–652.
- RÖSLER, A. 2012. Using the Tobii Mobile Device Stand in Usability Testing on Mobile Devices. Tech. rep., Tobii Technology.
- SCHMETTOW, M. 2012. Sample Size in Usability Studies. *Communications of the ACM* 55, 4 (Apr.), 64–70.
- SHEN, C., AND ZHAO, Q. 2014. Webpage Saliency. In *Computer Vision ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8695 of *LNCS*. Springer International Publishing, 33–46.
- SPOOL, J., AND SCHROEDER, W. 2001. Testing Web Sites: Five Users is Nowhere Near Enough. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, ACM, New York, NY, USA, 285–286.
- TAKEUCHI, H., AND HABUCHI, Y. 2007. A Quantitative Method for Analyzing Scan Path Data Obtained by Eye Tracker. In *IEEE Symposium on Computational Intelligence and Data Mining, 2007.*, 283–286.
- TOBII TECHNOLOGY. 2010. *Tobii Studio 2.X User Manual (Version September 2010)*. Tobii Technology AB.
- UNDERWOOD, G., HUMPHREY, K., AND FOULSHAM, T. 2008. Knowledge-Based Patterns of Remembering: Eye Movement Scanpaths Reflect Domain Experience. In *HCI and Usability for Education and Work*, A. Holzinger, Ed., vol. 5298 of *LNCS*. Springer Berlin Heidelberg, 125–144.
- VELICHKOVSKY, B. M., ROTHERT, A., KOPF, M., DORNHÖFER, S. M., AND JOOS, M. 2002. Towards an express-diagnostics for level of processing and hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour* 5, 2, 145–156.
- WEST, J. M., HAAKE, A. R., ROZANSKI, E. P., AND KARN, K. S. 2006. eyePatterns: Software for Identifying Patterns and Similarities Across Fixation Sequences. In *Proceedings of the 2006 Symposium on Eye Tracking Research and Applications*, ACM, New York, NY, USA, ETRA '06, 149–154.
- YESILADA, Y., HARPER, S., AND ERASLAN, S. 2013. Experimental transcoding: An Eye-Tracking approach. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, ACM, 30.