Fixation and Confusion – Investigating Eye-tracking Participants' Exposure to Information in Personas

Joni Salminen
Qatar Computing Research Institute, Hamad Bin Khalifa University; and Turku School of Economics
jsalminen@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute, Hamad Bin Khalifa University
jjansen@acm.org

Jisun An
Qatar Computing Research Institute, Hamad Bin Khalifa University
jan@hbku.edu.qa

Soon-Gyo Jung
Qatar Computing Research Institute, Hamad Bin Khalifa University
sjung@hbku.edu.qa

Lene Nielsen
IT University of Copenhagen
lene@itu.dk

Haewoon Kwak
Qatar Computing Research Institute, Hamad Bin Khalifa University
hkwak@hbku.edu.qa

ABSTRACT

To more effectively convey relevant information to end users of persona profiles, we conducted a user study consisting of 29 participants engaging with three persona layout treatments. We were interested in confusion engendered by the treatments on the participants, and conducted a within-subjects study in the actual work environment, using eye-tracking and talk-aloud data collection. We coded the verbal data into classes of informativeness and confusion and correlated it with fixations and durations on the Areas of Interests recorded by the eye-tracking device. We used various analysis techniques, including Mann-Whitney, regression, and Levenshtein distance, to investigate how confused users differed from non-confused users, what information of the personas caused confusion, and what were the predictors of confusion of end users of personas. We consolidate our various findings into a confusion ratio measure, which highlights in a succinct manner the most confusing elements of the personas. Findings show that inconsistencies among the informational elements of the persona generate the most confusion, especially with the elements of images and social media quotes. The research has implications for the design of personas and related information products, such as user profiling and customer segmentation.

ACM Reference format:
J. Salminen, B.J. Jansen, J. An, S.-G. Jung, L. Nielsen, and H. Kwak. 2018. Fixation and Confusion: Investigating Eye-tracking Participants' Exposure to Information in Personas. In *CHIIR '18: Conference on Human Information Interaction and Retrieval, March 11-15, 2018, New Brunswick, NJ, USA.* ACM, NY, NY, USA, 10 pages.
DOI: https://doi.org/10.1145/3176349.3176392

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

ACM ISBN 978-1-4503-4925-3/18/03…$15.00
https://doi.org/10.1145/3176349.3176391

1 INTRODUCTION

Personas are fictitious information representations of core user groups [1]. They are used by professionals in marketing, product development, system design, and corporate decision making [2] [3] [4] [5]. Traditionally, personas have been created using manual methods, such as ethnography and interviews [6]. While these methods result in deep user insight, their feasibility is reduced by time and cost, making personas unavailable for many organizations with tight product deadlines or limited budgets. More recently, researchers have looked into automating the persona generation process, which is based on information on real user behavior in social media whose collection and processing has been automated [7] [8] [9] [10] [11].

While social media benefits persona generation in many ways, the task of compressing social media data into simple persona presentations is not a trivial one. First, among all information elements (e.g., demographics, psychographics, etc.), one has to choose the right elements for a particular user or use case. Second, one needs to determine how this information is presented to be helpful for the end users of personas, while minimizing negative cognitive effects, such as information overload and confusion. Achieving these goals requires an in-depth understanding of how users perceive an interface or system and what are their cognitive reactions to it. Such questions are best addressed by experimental studies, measuring constructs such as confusion, defined here as a state of cognitive disorientation. In particular, earlier studies have found that end users can react to personas with disbelief and perception of inconsistency [12] [13] [14], perceptions that are conceptually similar to confusion.

This research reports the interaction between automatic personas and users' cognitive state, although we believe the findings are applicable to personas generated by any method. Particularly, we are investigating the relationship between users and their perceived confusion. Our research questions are:

- How do confused users differ from non-confused users in terms of their eye fixation patterns?
- Which areas of automatic personas cause the most confusion?
- What are the most powerful predictors of confusion?

To answer the questions, we analyze the eye-tracking data from a user study consisting of 29 participants. For the first question, we first look at the quantity and duration of fixations between the groups. Then, we examine the structural differences between the transition paths from one area of interest (AOI) to another. AOIs are commonly used in eye-tracking studies to connect fixation observations to particular areas of the screen. After this, we look deeper into the participants' interactions to see what caused confusion. We do this by a mixed method approach, combining qualitative and statistical techniques, and present then our findings. Finally, we conclude by discussing the measurement and analysis of confusion from eye-tracking data in user studies and its implication for the design of personas, both automatically generated and traditionally developed, and related artifacts such as user profiles.

2 RELATED LITERATURE

Granka et al. [15] note that most of eye-tracking user studies are in fact focused on analyzing cognitive information processing, e.g. what the users are thinking of and what information they are paying attention to. The general problem is how to identify positive or desired cognitive states (i.e., interest) from negative ones (e.g., confusion), as the former indicate good designs and the latter bad ones. Earlier research has shown users may experience confusion relating to several reasons, such as poor information designs [16]. Adopting the eye-mind hypothesis [17], confused users should pay more attention to their points of confusion. The basic metrics relate to fixation quantity, duration, and screen position.

A fixation is defined as a relatively stable state of the eye, focusing on particular gazepoint, and lasting 100-600ms [18] [19]. Saccades, in turn, are shorter, rapid eye movements between fixations. Together the two form scanpaths [20]. Fixations capture the direction of a user's attention and therefore indicate where information acquisition and processing are possibly taking place [15]. In addition, fixations are related to the depth of information processing and its level of difficulty [21]. For example, Golberg and Kotval [16] found that an intentionally poorer user interface resulted in significantly more fixations than a better design. In a similar vein, long fixation durations may indicate participant confusion [16].

The studied persona profile layout is divided into AOIs, so that each fixation is targeting a specific area of interest that matches its coordinates (x, y). AOIs are commonly used to identify user's interest in the examined layout [15]. For example, listings on a search results page, or different focal elements in e-commerce site could be defined as AOIs by researchers [22] [23]. Scanpaths between different AOIs are seen as records of visual attention [24]. For example, a longer scanpath can be indicative of less efficient searching due to a poor layout, as users are resorting to more cognitive effort to find what they are looking for [25]. Albeit comparing scanpaths of different groups is generally seen more difficult than pairwise comparison [15], some approaches have been developed. For example, Eraslan et al. [26] introduced the scanpath trend analysis (STA) algorithm. Moreover, in experimental eye tracking studies, it is common to record the cognitive processes of participants through the talk-aloud method to enable deeper analyses [27]. Analyzing the talk-aloud records gives an understanding of users' the state of the mind, as they are explicitly telling about their perceived mental state [28].

There are also studies that attempt to predict cognitive states from eye-tracking data using neural networks. In their pioneering work, Yamada et al. [29] define four emotion states as inputs for neural network learner, inferring these from individuals' voice signals. Harada et al. [30] propose a model for assessing the level of distraction, especially focused on drivers, and Grace et al. [31] explore the use of neural networks for detecting drowsiness and distraction in driving. Kuperberg and Heckers [32] investigate using neural networks for classification of schizophrenia.

However, inferring confusion from eye-tracking data is quite complicated because of three reasons: first, the fixation patterns tend to be complex, consisting of users' eye fixation jumping from one area of the screen to another. Beyond basic metrics, such as number and duration of fixations, one also should consider the sequence of AOIs that intuitively should matter for the prediction (assuming that gazing behaviors between confused and non-confused participants are different; see e.g. Eraslan et al. [35] for discussion). Second, earlier research has shown that basic features, such as duration spent fixated on an AOI can be interpreted differently depending on the use case and user in question. For example, the longer duration can indicate confusion in information retrieval tasks [24], i.e. individuals are having a tough time making sense of the information, but in website browsing, a higher fixation duration can indicate stronger interest [23]. Granka et al. [15] argue that in some tasks where the high focus is required, long fixation might not indicate confusion but the opposite.

Third, the problem lies in getting from the high-dimensional sequence and duration representation of confused users into a well-defined and evaluated predictive model. The users' eye fixation pattern should reveal they are confused, but this pattern is not easily analyzable by traditional methods [24]. Overall, the relative nature of this problem implies that general rules about the relationship between fixation durations and patterns with confusion cannot be easily formulated. Rather, we suggest that such relationships are better off being predicted from the data, given that we have labeled data on confusion, such as when cognitive measures are retrieved by the talk-aloud method [36]. Even though prior research has provided indications for the relationship between confusion and eye fixations, this relationship is not well known. We aim at targeting this research gap.

3 METHODOLOGY

3.1 Data collection

We conducted an eye-tracking user study to test different persona profile layouts for automatic persona generation (APG) which is both a system and a methodology for generating behaviorally accurate personas [7-11]. Personas are imaginary representations of core customers of a company or other organization [4]. They can be created automatically by retrieving social media data via application programming interfaces (APIs), and processing this data with non-negative matrix factorization and topic modeling [7] [8] [10]. This process has been described in detail in earlier work [7-11], and we refrain from repeating it here. An example of an automatically generated persona is shown in Figure 1.

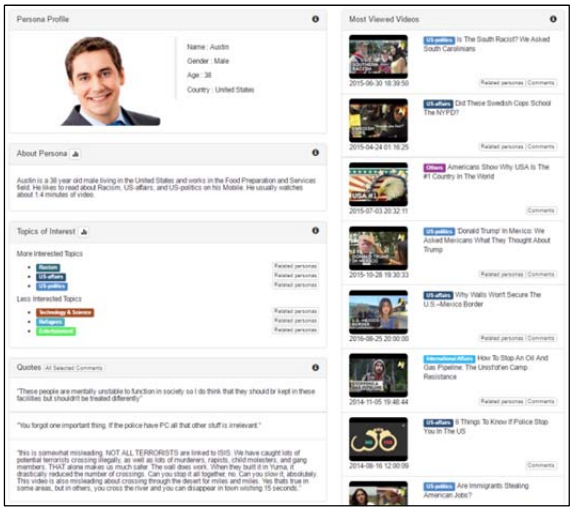

Figure 1: Example of automatically generated persona. It (he) has a picture, name, demographic information, topics of interest, descriptive quotes, and most viewed video content.

In particular, we wanted to know how many and what kind of images should be used in the automatically generated persona profiles. Previous persona literature does not provide an answer to this question, as there are only a few studies focused on persona images [37] [35] [36], none of which are measuring confusion. Therefore, we conducted the user study and chose eye-tracking as the form of data collection. The participants included 29 individuals working at the case company using our APG system, a large media company based in Doha, Qatar. The study consists of all 29 participants without grouping them, and all participants are exposed to the same layouts listed in Figure 2 (within-subjects design). Table 1 includes information about the participants.

Table 1: Information on the eye-tracking study participants.

	Male	Female	Total
Avg. age (years)	28.5	30.2	32.6
Avg. exp. in news industry (years)	7.1	7.5	7.3
Producers	11	7	18
Editors	3	5	8
Others	1	2	3
Total	15	14	29

We set up the eye-tracking device (EyeTribe, which comes with a cloud-based software for data processing) with desktop computers in the company's premises and, during five working days, conducted eye-tracking trials with the participants. We had each participant undergo three treatments (see Figure 2) at a random sequence and simultaneously collected talk-aloud data to connect the eye-tracking observations to cognitive processes of the individuals [17]. The participants were free to spend as much as time as they wanted with each treatment; after they were done, they clicked forward to the next one. The order of treatments was randomized. We recorded the voice-aloud comments during the experiment, and analyzed them later by dictionary-based cognitive discourse analysis (CDA) [40], in which we paid attention to verbal cues from the participants' speech to detect confusion.

During the experiment, the participants were encouraged to express their cognitive states as they were looking at the screen ("Where are you looking at? What do you see?"). We then adopted CDA [40] to code each AOI in terms of confusion expressed by the participant (e.g., "not understanding why three pictures are shown" indicated confusion targeting the images, and "seems confusing, not sure what quotes mean" targeting quotes). Following this technique, we paid attention to verbal cues of confusion when coding the perceived confusion expressed by the participants during the experiment. The cue words included e.g. "confusing", "did not understand", "difficult to say", etc. Table 2 includes examples of the confusion instances and cue words.

Table 2: Examples of confusion cue words used in coding.

Perceived confusion = TRUE	Cue words
"seems confusing, not sure what quotes mean"	Confusing, not sure
"there are different pictures, I don't understand"	Don't understand
"lost on here - conflicted profile"	Lost, conflicted
"weird information about videos - how are they related?"	Weird, how
"not sure what to think of the picture"	Not sure
"doesn't make sense"	Does not make sense

When a cue is found, the instance is coded as 1 (TRUE). If the notes lack cues, the instance is coded as 0 (FALSE). The coding is

done for each treatment of each participant based on the talk-aloud transcripts made during the eye tracking sessions.

We coded the confusion for each participant in each trial and verified the coding reliability by inter-rater test (Cohen's Kappa 0.86). In addition to the fixation observations, collected with the EyeTribe system, and the confusion coding, we asked background information from each participant, including age, gender, and experience in the industry. Treatments used in the study are shown in Figure 2.

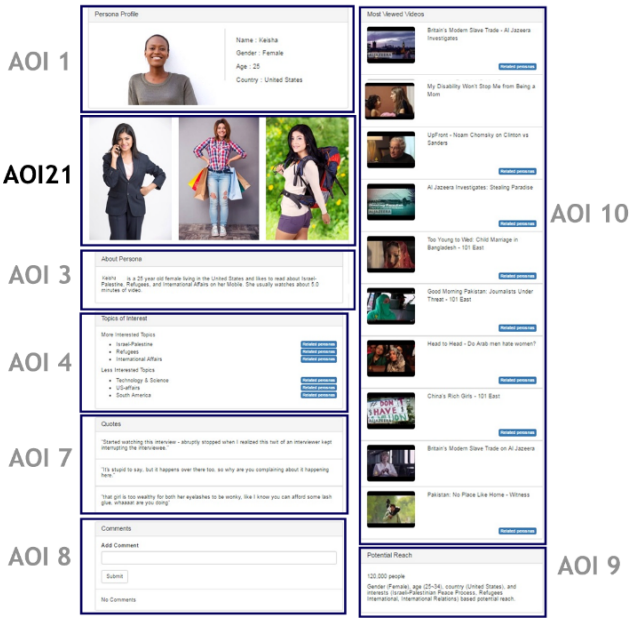

Figure 2: The tested layout. Between the three treatments, everything else was equal except T2 and T3 (in picture) had contextual pictures (AOI 21) added. As can be seen, different parts of the screen are defined as areas of interest (AOIs)

Automatic personas include six sections: *persona profile* with photo, name, age, gender, and country (AOI 1); *textual description* about persona (AOI 3); *topics of interest* the persona is most and least interested in (AOI 4); *descriptive quotes* aggregated from real social media users (AOI 7); *content* the persona has most interacted with (AOI 10); and *total audience size* retrieved from Facebook Marketing API [2] with the corresponding targeting criteria (AOI 9). T2 and T3 also included additional contextual pictures (AOI 21) that were manually added to explore their effect on users' stated confusion. In terms of other content, the treatments were identical.

3.2 Description of data

To select meaningful metrics for our study, we adopt Goldberg and Kotval's [16] suggested metrics. For temporal aspect, we are looking at fixation duration and dwell times. For spatial aspect, we look at heatmaps generated by the eye-tracking software, as well as the length of fixation and transition paths. Table 3 describes the studied variables.

Table 3: Data variables for quantitative analysis.

Variable	Description
Number of fixations	The number of fixation observations recorded by the measurement device. The sampling frequency of the device was 50Hz.
Avg. duration of fixations	The average duration of fixations. A fixation duration is typically 100-600ms [18] [19].
Total duration of fixations	Sum of fixation durations (e.g., by participant, treatment, AOI).
Number of transitions between AOIs (i.e., *transition paths*)	Indicates fixation movement from one AOI to another. The number of transitions can be computed by eliminating fixations targeting the same AOI repetitively from the total number of fixations (e.g. A1 → A1 → A2 would transform into A1 → A2).
Background information	Questionnaire answers: role (producer / editor / other), age (young / mature); gender (male / female); experience (novice / experienced).
Perceived confusion	Indication if the participant expressed confusion during a given treatment (TRUE/FALSE).

The measures applied in our study are commonly used in eye tracking studies to analyze the data. For example, Cowen et al. [41] analyze the number of fixations and total fixation duration. The number and duration of fixations (dwelling time) are commonly used in eye tracking studies to evaluate user engagement and sense of relevance [42]. In addition, Eraslan et al. [35] point out that individual variables, such as gender and user expertise tend to influence eye-tracking patterns, so we also include them.

Note that we are using fixations, but not saccades. Transition paths are different from the fixations paths; the former capture the movement from one AOI to another, while the latter includes also repetitive fixations to an AOI. It is important to examine both, as they might reveal different information about the viewing behavior of the user. In particular, a transition path is a higher-level description of viewing pattern than fixation path that includes also the repeated views on the same AOI. Fixation path is equal to the number of fixations subtracted by 1 (the start state).

Finally, the confusion data is available for each trial of each participant (T-P); and for each AOI at each trial of each participant. We use both levels of coding depending on the question we are answering to. If T-P level data suffices to answer a question, we prefer using it since the AOI-level data is sparser.

[2] https://developers.facebook.com/docs/marketing-apis/

However, for questions dealing with the AOI-level impact on confusion, we must use that level data.

4 FINDINGS

4.1 How do confused users differ from non-confused users?

There are 29 participants with usable data, of which 23 (79%) expressed confusion and 6 (21%) did not. The confusion varies greatly by treatment, and most confusion was expressed in T3 (60% of all confusion observations), the least in T1 (19%). Out of all participants, 8 (28%) expressed confusion in two or more treatments. Therefore, we find that the participants do differ by perceived confusion. Figure 3 shows the observed confusion among the participants.

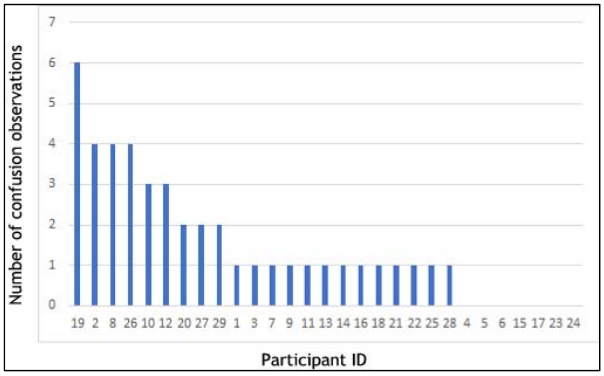

Figure 3: Observed confusion among participants. Confusion is calculated by each AOI of each participant in each trial. Not that a participant can express confusion toward several AOIs per trial.

As can be seen, there are both confused and non-confused participants. Confusion is calculated by each AOI of each participant in each trial. The maximum number of confusion observations per participant is six, and minimum zero. The average is 1.48 confusion observations per participant. Table 4 shows basic eye-tracking metrics between the confused and non-confused users.

Table 4: Basic eye-tracking metrics comparing confused and non-confused users. The confusion is calculated by participant and trial.

	Confused	Non-confused
Avg. number of fixations	885	766
Avg. dur. of fixations (ms)	336	363

The difference of fixation durations between confused and non-confused users is small (confused have ~8% longer fixation duration). In turn, the confused have ~16% more fixations than non-confused, warranting further inspection. We do this by carrying out a Mann-Whitney-Wilcoxon test, in which we compare confused group (number of fixations from each confused trial of each participant) with the non-confused group, with the null hypothesis that there is no statistically significant difference.

It is clear from the test that there is a significant difference between the two groups (p-value = 0.012). This implies the number of fixations is statistically different across the two groups ($p<0.05$). The number of fixations for non-confused is smaller on average than for confused. Thus, we find evidence of differences and proceed to explore the data further.

4.2 How do confused and non-confused users differ by their fixation paths?

To answer this question, first, we measure the average length of transition paths. This is done by eliminating repetitive fixations targeting any given AOI, revealing the "bare" transition path between participants and AOIs (e.g., fixation path for a participant P1: A1 → A1 → A2 will become transition path P1: A1 → A2). Table 5 shows the average of transition paths of confused and non-confused users.

Table 5: Average length of transition paths in AOIs.

	Confused	Non-confused
Avg. length of transition paths	162.5	140.2

The transition paths are longer with the confused group (calculated as P-T), indicating they are "jumping" from one AOI to another more often. To find out if this difference is statistically significant, we conduct a Mann-Whitney-Wilcoxon test. It is clear from the results (p-value = 0.3091) that there is no significant difference between the two groups on the length of transition paths. Yet, based on standard deviations, the lengths of confused (σ=82.28) are more spread out than those of non-confused (σ=61.37), giving some indication of more sporadic behavior.

We are also interested in knowing if the paths vary by content, i.e., their AOI states. For this purpose, we use the Levenshtein distance to compare paths turned into strings against one another [43]. Such sequence alignment techniques are commonly used in measuring eye-tracking paths [44] [15].

First, we build four similarity matrices: M0, comparing all participants' fixation sequence strings to one another; M1, comparing confused participants to one another; M2, comparing non-confused participants to one another; and M3, comparing confused participants with non-confused participants. We then average the pairwise comparisons of each matrix to produce an overall score of fixation path similarity within a group. The results are shown in Table 6.

Table 6: Similarity between Confused and Non-confused fixation paths.

Similarity matrix	Fixation path similarity
Confused	660
Non-Confused	559
Compared Confused with Non-confused	628
All	604

As shown in Table 6, these are non-normalized edit-distances (i.e., Levenshtein distance). Here, the score refers to the number of operations needed to substitute one fixation path to another. For example, 660 means that 660 operations (i.e., delete, insert or substitute) are needed to change the fixation path to another fixation path. Thus, the higher the distance is, the less similar a pair of two participants are. From the fixation path similarity numbers, we observe that the confused are most different from one another, and non-confused are most similar to one another. Moreover, non-confused are more different relative to confused than to other non-confused. However, interestingly, confused are more different to one another than to non-confused. This seems to suggest that confusion is sporadic, i.e., consisting of random rather than systematic patterns. In addition, the relative differences are quite large (the confused are 17.9% more dissimilar than the non-confused). Figure 4 illustrates the similarity matrices.

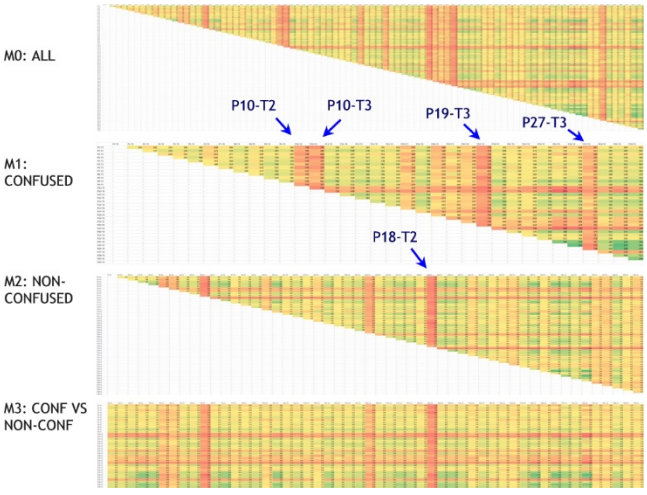

Figure 4: Levenshtein distance matrices; each column and row maps into each user's fixation path. Red indicates higher distance, green closer. Yellow is in between. In this matrix, each participant is compared with all the other participants once (e.g., comparing P1 and P2 in P1 row means we do not repeat the comparison with P2 column).

From Figure 4, we can detect some individual differences. Very distinct fixation paths from all others could indicate measurement errors. For example, P18 (T2) is distinctly different from other non-confused (red vertical line in M2). The transition path of this participant is displayed in Figure 5.

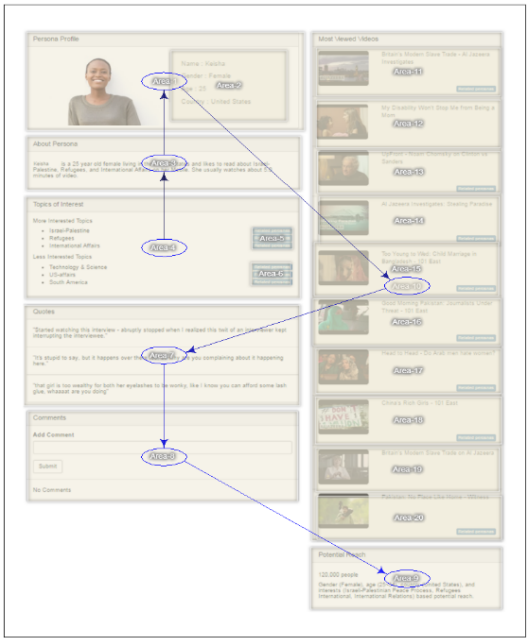

Figure 5: Transition path of P18 (male, 39, other). The line is drawn from the center of one AOI to another and thus does not depict the actual scanpath.

We can observe that, uncommon to most participants who start from the top-left of the screen and the move down and to topics of interest and then to right-side column videos, this participant firstly focused on topics of interest, and then moves upward. From confused ones, P10 (T2 and T3), P19 (T3), and P27 (T3) clearly differ from the others (red vertical lines in M1 in Figure 4). Table 7 shows the dissimilarity of these fixation paths.

Table 7: The most dissimilar fixation paths. Difference is from the mean Levenshtein distance of other confused participants.

Participant/Treatment	Avg. distance	Diff. from mean
P10-T2	702	+6%
P10-T3	720	+9%
P19-T3	736	+12%
P27-T3	692	+5%

In this case, we observe some individual patterns in Levenshtein distances of grouped users. For example, the Mann-Whitney test shows that the distribution of dissimilarity scores from P19-T3 is significantly different the distribution of others ($W = 18295$, p-value $< 2.2e-16$). Thus, one can use the dissimilarity matrix as a basis of visualization, and then explore the visible differences between statistical and qualitative means. Both group-level and individual insights can be found by analyzing the distance matrix.

4.3 Which areas of personas profiles cause the most confusion?

Next, we investigate the relationship between confusion and areas of interest. Figure 7 shows confusion observations by AOIs.

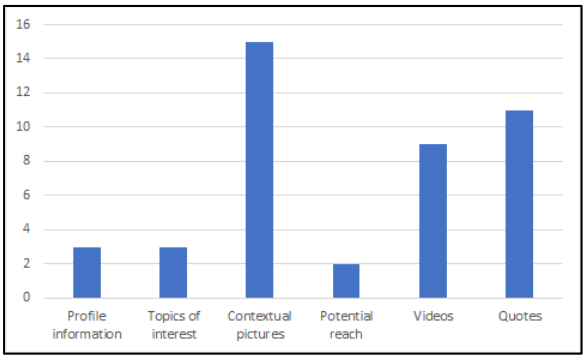

Figure 7: AOIs with the most confusion. AOIs without confusion observations are not included.

Contextual pictures were the target of most confusion (35%), (see Figure 8 – A), followed by persona quotes (26%) (Figure 8 – B) videos (21%) (Figure 8 – C), and topics of interest (7%) (Figure 8 – D). Yet, the number and duration of fixations indicate that videos are most looked at. Two areas in the layout did not gather any confusion: Description which includes the textual description for the persona, and the Comments section are largely ignored.

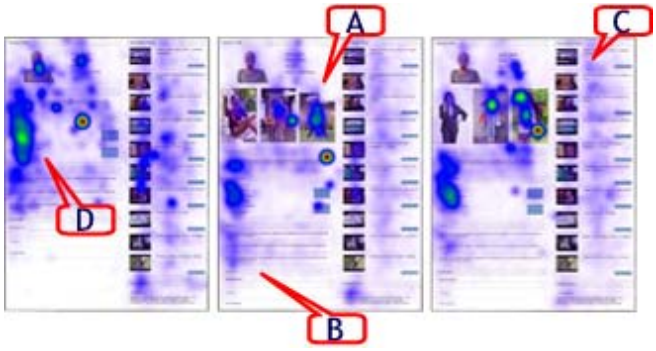

Figure 8: Heatmaps of visual attention of participants (includes all participants). Brighter color indicates more fixations.

From the heatmaps in Figure 8, we can see that in T2 (middle) and T3 (right), which overall had the strongest confusion, the attention seems to focus on a) the extra pictures and b) topics of interest. Confusion was also highest for the extra pictures among the AOIs, indicating a relationship between confusion and number of fixations. In contrast, quotes and videos, which also had a substantial share of confusion, were paid less attention to. It, therefore, seems that confusion targeting each AOI should be evaluated relative to the attention it is receiving.

For this purpose, we compute the "confusion ratio", a metric which we invented here, to evaluate the relative intensity of confusion per AOI: this is calculated by dividing the amount of time targeted by confused users to an AOI with the amount of time from non-confused users targeting the same AOI. That is, if the ratio is high, the relative confusion of that AOI is higher than otherwise. Table 8 shows the results of the calculations.

Table 8: Confusion ratio. The order of the rows is based on the number of confusion observations.

AOI	Confusion ratio	Name of AOI
A21	1.160	Contextual pictures
A7	1.711	Quotes
A10	1.072	Videos
A1	1.144	Profile information
A4	2.422	Topics of interest
A9	2.189	Potential reach

We can see that the "ranking" of AOIs in terms of confusion changes from the ranking with pure observations (Figure 7) when we account for fixation duration targeting that same AOI. This captures the fact that time spent in AOIs is not equally distributed. Thus, even though contextual pictures have the largest share of confusion observations, their confusion ratio is actually lower than for potential reach, which is rarely looked at but makes users more confused when it is being looked at. Videos, in turn, have a low confusion ratio because they are looked at often, but the relatively lower number of participants found them confusing. Figure 9 shows the average dwell time (sum of fixation durations) of confused and non-confused participants.

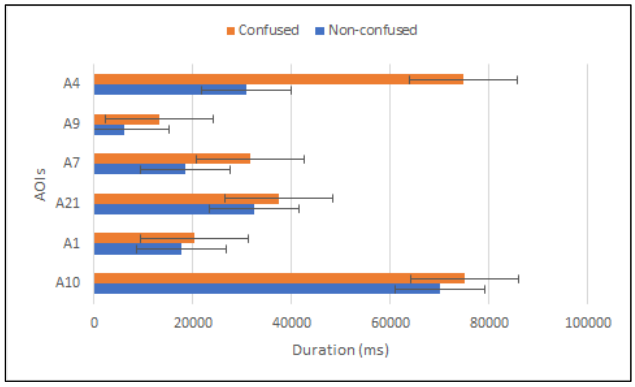

Figure 9: Average dwell times per AOI by confused and non-confused users. The line on each bar indicates standard error. Topics of interest (A4) have the clearest difference, and it is also ranked highest by confusion ratio metric.

4.4 Reasons for confusion

For any type of user study focusing on confusion and informativeness, it is useful to know why users were confused so that proper conclusions can be drawn for informing information design. For this purpose, we also highlight reasons to why the participants expressed confusion (Table 9)

Table 9 Examples of confusion reasons.

AOI	Explicated reason for confusion
Contextual pictures (A in Fig. 8)	"confused – don't get the photos" (P13); "don't understand why these photos are there" (P8)
Quotes (B in Fig. 8)	"confused; talks about video maybe [means] something on the refugee situation international community" (P2)
Videos (C in Fig. 8)	"looking at the videos – can't find stories tie into the topics" (P18)
Topics of interest (D in Fig. 8)	"[the persona is] not interested in South America although closer to her location" (P29)

We can thus see that there are various underlying reasons for confusion. For example, conflicting information. The photos that represented other, similar individuals were perceived confusing. The participant could not understand the linkage of them being similar to the person depicted in the mugshot ("[I am] a little confused, all different women" (P14). One participant assumed they are "pictures of her friends" (P19). In other cases, information definitions are not clear to the user ("don't know what the quote section is; don't know if it's about her or by her" (P8)). Overall, the findings suggest that AOIs are processed relative to one another, so that inconsistent information becomes a major source of confusion.

In addition, confusion coding revealed insights useful for system development, e.g. that "potential reach" was not understood by the news producers the same way as marketers would understand it (i.e., as potential audience size), but instead as the reach of the persona. Consequently, we clarified the definitions of the titles accordingly in the system (see Figure 10).

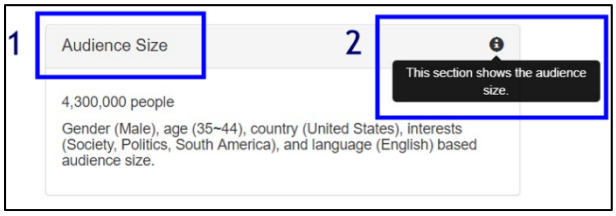

Figure 10: Example of changes made based on the user study. Users did not understand potential reach, so we decided to change to "audience size" (1), which is a more unambiguous term. Additionlly, we included tooltip definition (2).

Finally, some users questioned the topics chosen for topic classification ("I feel international affairs is too broad unless I knew more about what exactly she's interested in -- too vague") (P19). For example, the concept of *Human story* raised some questions. This goes to show that when defining topics for data analysis of labels for user interfaces, researchers should ensure they are "speaking the same language" as the end users.

4.5 What are the best predictors of confusion?

To answer this question, we use a binary classification model to test the predictive power of the variables. Binary classification can be used for mapping instances between certain classes/groups to determine the best predictors for a given result, in this case confusion. In our case, the variables are the previously mentioned variables we are working with. The accuracy of the model is expressed by AUC (Area Under the Curve) metric. Table 10 shows the AUC scores of each variable.

Table 10 Accuracy of variables; higher is greater accuracy.

Predictor	AUC
Length of transition path (T3)	0.66
Experience Group	0.65
Number of fixations	0.64
Age Group	0.61
Total duration of fixations	0.60
Gender	0.56
Length of transition path (T1)	0.54
Avg. duration of fixations	0.53
Role	0.51
Length of transition path (T2)	0.49

The most predictive factors are a) Age Group, b) Experience Group, c) Length of transition path (for Treatment 3 with extra pictures), and d) Number of Fixations. The proposed model, based on these four variables, gives a good accuracy, giving the right prediction about 8 times out of 10 (AUC=0.812). We conclude that the four most significant factors are good predictors of user confusion. To examine the influence of user-level variables more closely, we plot them in one visualization. We choose T3 as a filter because it has the most confusion observations (Figure 11).

Finally, because the binomial classification model predicts but does not provide significance analysis beyond the AUC metric, we conduct a regression analysis. As the results of the binomial model indicate that the Treatment 3 (T3) is an important factor for confusion, we test each treatment separately. Since there are only 29 subjects, we used the stepwise regression method to get the final model by reducing non-significant variables.

To choose the variables for the reduced model, we are using a procedure called backward selection that finds any significant variable by Akaike information criteria (AIC) by dropping one variable at a time and seeing which one minimizes the AIC most, and moving forward until the AIC change is insignificant.

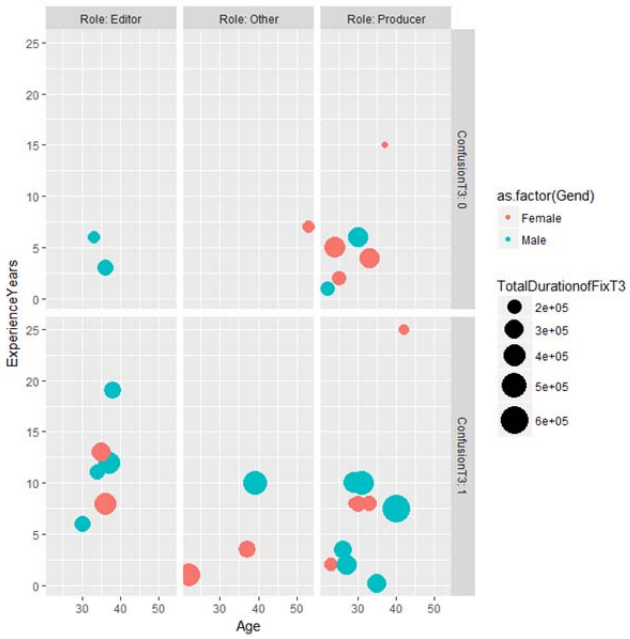

Figure 11: Confusion of participants by background information. It seems that mature males are most prone to confusion in this case.

We thus start from a large number of variables, including background information, fixation information, as well as calculated metrics (e.g., transition path length), and find that the number of fixations has a significant positive relationship with confusion, and the relationship holds across treatments (T1: p-value = 0.041; T2: p-value = 0.059; T3: p-value = 0.037). This corroborates our previous analysis comparing confused and non-confused users in terms of number of fixations. The results of the regression analysis for T3 are shown in Table 11.

Table 11: Reduced regression model for T3.

	Estimate	Std. error	Z	Pr(>|z|)
(Intercept)	-7.93e+00	4.23e+00	-1.88	0.061
No. of fixations	3.440e-02	1.663e-02	2.068	0.039*
No. of transitions	1.000e-01	6.496e-02	1.540	0.124
Total duration	-3.78e-05	1.996e-05	-1.90	0.058
Transition ratio	-4.98e+01	2.52e+01	-1.97	0.048*
Experience	2.627e-01	1.611e-01	1.631	0.103

'' 0.05 significance*

The number of fixations is a significant variable in predicting confusion. Another significant variable in T3 is the transition ratio (total number of transitions/total number of fixations), which provides information on how frequently the user switched from one AOI to another relative to overall fixation activeness. Total duration (i.e., dwell time) is close to being significant but is not at 5% significance level. Background variables (gender, age, role) were eliminated from the reduced model as they did not improve the explanatory power of the model.

5 CONCLUSION AND DISCUSSION

Overall, our study responds to the call of Blascheck et al. [36] for correlating eye tracking, think-aloud, and other data for analysis of users' cognitive states. We find that personas seem to raise a considerable amount of confusion. Confusion mostly relates to pictures and quotes. While we did not find significant differences in duration of fixation paths between confused and non-confused users, a positive relationship between the number of fixations and confusion is implied both by regression analysis and binomial classification. The fixation paths of confused users are longer and more varied than those of non-confused users. In addition, the binomial classification showed that the most notable predictors for confusion were age, experience, and the number of transitions and fixations. These results confirm earlier findings on the impact of user-level characteristics on users' mental state [35]. In particular, it seemed that older men had more trouble with the more complex layout, supporting findings that gender and age play a role in information processing [45] [46].

The study provides several practical insights for persona development, especially when automated. Most importantly, consistency is a problem when automatically generating quotes and pictures. This concern has also been raised in earlier persona literature [12], and we dub it here as the *consistency problem*. From the explicit feedback, we can see that perceived inconsistency between different informational elements is associated with confusion. Confusion arising from inconsistency could be reduced e.g. by contextualization of the data (i.e., presenting numbers or diversity in the underlying group the persona is based on), and manual verification of different informational elements to ensure they make sense. Consistency is more acute when integrating data from different sources, such as quotes from different users or social network. Further research could find ways to measure and improve consistency automatically, which would help improve the information design of personas profiles.

Finally, we found confusion to vary highly across AOIs, and introduce a metric, confusion ratio, that takes into consideration the relative difference of attention paid to each AOI (dwell time) when determining the criticality of the confusion for the users. By considering this relativity, AOIs that appear confusing can be actually less confusing than what the absolute numbers claim.

We find the talk-aloud technique useful because it helps finding both evidence for confusion for a given user-trial and the reasons behind confusion, supporting quantitative and qualitative analysis. Yet, it is not a perfect technique, as the individuals may differ in their accounts, so that not all users are equally vocal about their experience confusion. In addition, talk-aloud may influence the actual viewing behavior. Therefore, we

suggest corroborating talk-aloud records with more robust data analysis in future works.

REFERENCES

[1] A. Cooper, *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*, 1 edition. Indianapolis, IN: Sams - Pearson Education, 2004.

[2] J. Pruitt and J. Grudin, "Personas: Practice and Theory," in *Proceedings of the 2003 Conference on Designing for User Experiences*, New York, NY, USA, 2003, pp. 1–15.

[3] J. Pruitt and T. Adlin, *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*, 1 edition. Amsterdam ; Boston: Morgan Kaufmann, 2006.

[4] L. Nielsen and K. Storgaard Hansen, "Personas is applicable: a study on the use of personas in Denmark," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 1665–1674.

[5] L. Nielsen, *Personas-user focused design*, vol. 15. Springer Science & Business Media, 2012.

[6] L. Nielsen, K. S. Nielsen, J. Stage, and J. Billestrup, "Going Global with Personas," in *Human-Computer Interaction – INTERACT 2013*, 2013, pp. 350–357.

[7] S.-G. Jung, J. An, H. Kwak, M. Ahmad, L. Nielsen, and B. J. Jansen, "Persona Generation from Aggregated Social Media Data," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2017, pp. 1748–1755.

[8] J. Salminen *et al.*, "Generating Cultural Personas From Social Data: A Perspective of Middle Eastern Users," presented at the The Fourth International Symposium on Social Networks Analysis, Management and Security (SNAMS-2017), Prague, Czech Republic, 2017.

[9] J. An, H. Kwak, and B. J. Jansen, "Validating Social Media Data for Automatic Persona Generation," presented at the The Second International Workshop on Online Social Networks Technologies (OSNT-2016), 13th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2016. 29 November - 2 December, 2016.

[10] J. An, K. Haewoon, and B. J. Jansen, "Personas for Content Creators via Decomposed Aggregate Audience Statistics," presented at the Advances in Social Network Analysis and Mining (ASONAM 2017). July 31., 2017.

[11] J. An, H. Kwak, and B. J. Jansen, "Towards Automatic Persona Generation Using Social Media," presented at the The Third International Symposium on Social Networks Analysis, Management and Security (SNAMS 2016), The 4th International Conference on Future Internet of Things and Cloud. 22-24 August, 2016.

[12] C. N. Chapman and R. P. Milham, "The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 5, pp. 634–636, Oct. 2006.

[13] T. Matthews, T. Judge, and S. Whittaker, "How Do Designers and User Experience Professionals Actually Perceive and Use Personas?," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2012, pp. 1219–1228.

[14] K. Rönkkö, M. Hellman, B. Kilander, and Y. Dittrich, "Personas is Not Applicable: Local Remedies Interpreted in a Wider Context," in *Proceedings of the Eighth Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices - Volume 1*, New York, NY, USA, 2004, pp. 112–120.

[15] L. Granka, M. Feusner, and L. Lorigo, "Eye Monitoring in Online Search," in *Passive Eye Monitoring*, R. I. Hammoud, Ed. Springer Berlin Heidelberg, 2008, pp. 347–372.

[16] J. H. Goldberg and X. P. Kotval, "Computer interface evaluation using eye movements: methods and constructs," *International Journal of Industrial Ergonomics*, vol. 24, no. 6, pp. 631–645, Oct. 1999.

[17] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, Oct. 1976.

[18] W. Barfield and T. A. Furness, *Virtual Environments and Advanced Interface Design*. Oxford University Press, 1995.

[19] L. Granka and K. Rodden, "Incorporating eyetracking into user studies at Google," in *Workshop Position paper presented at CHI*, 2006.

[20] D. D. Salvucci and J. H. Goldberg, "Identifying Fixations and Saccades in Eye-tracking Protocols," in *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, New York, NY, USA, 2000, pp. 71–78.

[21] B. Follet, O. Le Meur, and T. Baccino, "New Insights into Ambient and Focal Visual Fixations using an Automatic Classification Algorithm," *i-Perception*, vol. 2, no. 6, pp. 592–610, Aug. 2011.

[22] D. Beymer, P. Z. Orton, and D. M. Russell, "An Eye Tracking Study of How Pictures Influence Online Reading," in *Human-Computer Interaction – INTERACT 2007*, 2007, pp. 456–460.

[23] E. Cutrell and Z. Guan, "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*, New York, NY, USA, 2007, pp. 407–416.

[24] J. H. Goldberg and J. I. Helfman, "Scanpath Clustering and Aggregation," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, New York, NY, USA, 2010, pp. 227–234.

[25] C. Ehmke and S. Wilson, "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*, Swinton, UK, UK, 2007, pp. 119–128.

[26] S. Eraslan, Y. Yesilada, and S. Harper, "Scanpath Trend Analysis on Web Pages: Clustering Eye Tracking Scanpaths," *ACM Transactions on the Web*, vol. 10, no. 4, pp. 1–35, Nov. 2016.

[27] T. Blascheck, K. Kurzhals, M. Raschke, S. Strohmaier, D. Weiskopf, and T. Ertl, "AOI hierarchies for visual exploration of fixation sequences," 2016, pp. 111–118.

[28] P. Balatsoukas and I. Ruthven, "An eye-tracking approach to the analysis of relevance judgments on the Web: The case of Google search engine," *J Am Soc Inf Sci Tec*, vol. 63, no. 9, pp. 1728–1746, Sep. 2012.

[29] T. Yamada, H. Hashimoto, and N. Tosa, "Pattern recognition of emotion with neural network," in *, Proceedings of the 1995 IEEE IECON 21st International Conference on Industrial Electronics, Control, and Instrumentation, 1995*, 1995, vol. 1, pp. 183–187 vol.1.

[30] T. Harada, H. Iwasaki, K. Mori, A. Yoshizawa, and F. Mizoguchi, "Evaluation model of cognitive distraction state based on eye-tracking data using neural networks," in *2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing*, 2013, pp. 428–434.

[31] R. Grace *et al.*, "A drowsy driver detection system for heavy vehicles," in *17th DASC. AIAA/IEEE/SAE. Digital Avionics Systems Conference. Proceedings (Cat. No.98CH36267)*, 1998, vol. 2, p. I36/1-I36/8 vol.2.

[32] G. Kuperberg and S. Heckers, "Schizophrenia and cognitive function," *Current Opinion in Neurobiology*, vol. 10, no. 2, pp. 205–210, Apr. 2000.

[33] R. Chai *et al.*, "Classification of EEG based-mental fatigue using principal component analysis and Bayesian neural network," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 4654–4657.

[34] S.-K. Jang, S. Kim, C.-Y. Kim, H.-S. Lee, and K.-H. Choi, "Attentional processing of emotional faces in schizophrenia: Evidence from eye tracking," *J Abnorm Psychol*, vol. 125, no. 7, pp. 894–906, 2016.

[35] S. Eraslan, Y. Yesilada, and S. Harper, "Eye tracking scanpath analysis techniques on web pages: A survey, evaluation and comparison," *Journal of Eye Movement Research*, vol. 9, no. 1, Dec. 2015.

[36] T. Blascheck, M. John, S. Koch, L. Bruder, and T. Ertl, "Triangulating User Behavior Using Eye Movement, Interaction, and Think Aloud Data," in *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, New York, NY, USA, 2016, pp. 175–182.

[37] J. E. Nieters, S. Ivaturi, and I. Ahmed, "Making Personas Memorable," in *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2007, pp. 1817–1824.

[38] F. Long, "Real or imaginary: The effectiveness of using personas in product design," in *Proceedings of the Irish Ergonomics Society Annual Conference*, 2009, vol. 14.

[39] C. G. Hill *et al.*, "Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways?," 2017, pp. 6658–6671.

[40] T. Tenbrink, "Cognitive Discourse Analysis: accessing cognitive representations and processes through language data," *Language and Cognition*, vol. 7, no. 1, pp. 98–137, Jul. 2014.

[41] L. Cowen, L. J. s Ball, and J. Delin, "An Eye Movement Analysis of Web Page Usability," in *People and Computers XVI - Memorable Yet Invisible*, Springer, London, 2002, pp. 317–335.

[42] M. Lalmas, H. O'Brien, and E. Yom-Tov, *Measuring User Engagement*. Morgan & Claypool Publishers, 2014.

[43] A. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Phys. Dokl.*, vol. 10, pp. 707–710, 1966.

[44] C. M. Privitera, "The scanpath theory: its definition and later developments," presented at the Human Vision and Electronic Imaging XI, 2006, vol. 6057, p. 60570A.

[45] J. Meyers-Levy and D. Maheswaran, "Exploring Differences in Males' and Females' Processing Strategies," *Journal of Consumer Research*, vol. 18, no. 1, pp. 63–70, 1991.

[46] L. Lorigo, B. Pan, H. Hembrooke, T. Joachims, L. Granka, and G. Gay, "The influence of task and gender on search and evaluation behavior using Google," *Information Processing & Management*, vol. 42, no. 4, pp. 1123–1131, Jul. 2006.