Evaluating User Satisfaction with Typography Designs via Mining Touch Interaction Data in Mobile Reading

Junxiang Wang[1], *__Jianwei Yin__[1], **Shuiguang Deng**[1], **Ying Li**[1], **Calton Pu**[2], **Yan Tang**[1], **Zhiling Luo**[1]
[1]Zhejiang University, Hangzhou, China [2]Georgia Institute of Technology, GA, USA
{junxiang, zjuyjw, dengsg, cnliying, allen_tung, luozhiling}@zju.edu.cn calton.pu@cc.gatech.edu

ABSTRACT
Previous work has demonstrated that typography design has a great influence on users' reading experience. However, current typography design guidelines are mainly for general purpose, while the individual needs are nearly ignored. To achieve personalized typography designs, an important and necessary step is accurately evaluating user satisfaction with the typography designs. Current evaluation approaches, e.g., asking for users' opinions directly, however, interrupt the reading and affect users' judgments. In this paper, we propose a novel method to address this challenge by mining users' implicit feedbacks, e.g., touch interaction data. We conduct two mobile reading studies in Chinese to collect the touch interaction data from 91 participants. We propose various features based on our three hypotheses to capture meaningful patterns in the touch behaviors. The experiment results show the effectiveness of our evaluation models with higher accuracy on comparing with the baseline under three text difficulty levels, respectively.

Author Keywords
Typography design; mobile reading; touch interaction data; user satisfaction; text difficulty; implicit feedback.

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces

INTRODUCTION
Today, people are still through reading texts to obtain various kinds of information. With the rise of mobile reading, a great challenge for designers is that how to provide a good reading experience on the small touch screen of mobile devices [13]. To address this challenge, one of the important suggestions for designers is to improve the design of typography [15, 23], which has been a classical problem in the field of HCI for a long history [2, 8, 27, 37]. However, traditional research on typography mainly focus on finding the optimal design, such as the best font size for text-heavy websites [33]. But the optimal design may not be for users who have individual needs. For example, the suggested font size for body content from iOS Human Interface Guidelines is 17 points [24], but for users who were diagnosed dyslexia may prefer bigger font size [34]. Moreover, different documents have different text difficulty which also has been demonstrated that has a great influence on users' reading performance [25, 28]. Thus, a best design may not always satisfy users' needs when they read texts vary in difficulty. Owing to the reasons above, user experience specialists suggest that designers should allow users to adjust the typography design by themselves, but actually most users seldom change the default design [31].

Our long-term goal is to provide better reading experience to the users with personalized typography designs under varying text difficulty levels. For achieving this goal, an indispensable first step is collecting user feedbacks for guiding the evaluation of user satisfaction with a particular typography design. Asking users directly for their opinions by a popping up dialog is a common method to collect user feedbacks on mobile devices [38]. But this method will abruptly interrupt users' reading, make them feel bad, and they may not express real feelings [32]. To address this problem, we propose a novel method to evaluate user satisfaction with the typography design by mining users' touch interaction data in reading. To our knowledge, we are not aware of any work specifically evaluating user satisfaction with the typography design by analyzing users' implicit feedbacks, such as touch behaviors.

So far, there has been a lot of research on the analysis of touch interaction data on mobile devices, such as search online [22], evaluate users' emotional states when playing games [20]. In this paper, we hypothesize that users' touch behaviors in reading can reflect their satisfaction with the typography design. Besides, we also assume text difficulty plays a critical role in affecting the touch behaviors in reading. Therefore, we collect and analyze the touch interaction data generated when users read to answer the following two questions:

- Do users behave differently on a touch-enabled smart phone when they reading with a satisfied typography design compared to an unsatisfied design?
- Does the text difficulty affect this behavior differences? If does, how?

To answer these two research questions, we analyze users' touch behaviors when they are reading on a smart phone in two controlled user studies, which involved 91 participants

*Corresponding author.

CHI'18, April 21–26, 2018, Montréal, QC, Canada

ISBN 978-1-4503-5620-6/18/04…$15.00
DOI: http://dx.doi.org/10.1145/3173574.3173687

in total and 494 unique reading sessions. The reason that we conducted two user studies is the high dimensionality of the typography design space: It can be argued that many typography variables, such as font size, line spacing and so on, which all have possible effects on user satisfaction and make the experiment complicated. Thus, we conduct a two-phase experiment: (1) we first perform a user study in which we majorly consider one typographic variable: font size, and take into account the influence of text difficulty. In a qualitative analysis, we identify behavioral patterns that correspond to reading with satisfied and unsatisfied typography designs respectively, and propose three hypotheses based on our observations. To validate the hypotheses, we design a set of touch behavioral features for capturing meaningful patterns and analyze the correlations of these features value with participants' satisfaction ratings. (2) We next validate these findings in a more complex and realistic situation, where we considered four typographic variables: font size, line spacing, paragraph spacing and page margin, and their combinations. Finally, we developed user satisfaction evaluation models based on our touch behavioral features and the experiment results showed a significant higher accuracy under three text difficulty levels than the baseline.

To summarize, our main contributions include:

- We conducted two mobile reading studies to collect touch interaction data as users' implicit feedbacks in reading. Utilizing these data, we developed models that can evaluate user satisfaction with the typography design more accurately than the baseline model.
- We proposed a variety of features to capture meaningful patterns in touch behaviors from the time dimension (the relationship between swipe distance and reading time), spatial dimension (the distribution of touch points on screen), and user expectation dimension (users' reading performance at the beginning of a reading session). Based on these features, we demonstrated that users' touch behaviors can reflect their satisfaction with the typography design.
- We provided a detailed study of the most discriminative features to evaluate user satisfaction with the typography design under three text difficulty levels, respectively. Our results indicated that text difficulty has an influence on users' touch behaviors in reading.

RELATED WORK

Typography Design

Previous work [4, 7, 21] have demonstrated that many typographic variables, such as font size and line spacing, have great influences on reading experience. In those works, researchers invited lots of participants to complete some well-designed reading tasks. And they quantitatively investigated the influence of one or several typographic variables on users' reading performances and satisfactions by some metrics. Those metrics can be divided into objective and subjective groups, the former includes reading time, comprehension score [18], fixation duration [5, 6] and so on, while the latter is mainly users' subjective satisfaction ratings.

Bernard et al. [3] performed a study with 60 participants for comparing 10, 12, and 14 points font sizes. They used reading time, preferences, and errors as metrics. The results revealed that font size of 12 points leads to significantly faster reading. Rello et al. [33] performed a study with 104 participants to investigate the effect of font size and line spacing on online reading. They compared the fixation durations, comprehension scores, and subjective perception ratings in the experiment. Based on their findings, they recommended designers to use font size of 18 points and the default line spacing.

As we can see, the goal of the most of those work is to find the optimal design, which are mainly for general purpose and may not satisfy users' personalized needs. To provide a user with a personalized typography design for better reading, one of the important problems that we attempt to address in this paper is to evaluate users' satisfaction with the typography design by analyzing users' implicit feedbacks.

Touch Interaction Data

Along with the increasing popularization of touch-enabled mobile devices, touch (includes gesture) has been a common interaction modality. A lot of touch interaction data are produced when users are using mobile devices. Collecting the touch interaction data can be implicitly performed without affecting the normal use [10]. Thus, through analyzing the touch interaction data to improve the accuracy of the prediction model in some application scenarios (e.g., mobile search [22]) has been a research hotspot in many fields, such as HCI [11, 26, 30, 36] and information security [14, 19, 35].

Guo et al. [22] performed a study with 26 participants to model touch interactions on a smart phone for improving Web search ranking. They investigated a variety of touch interactions, such as swiping and zooming gestures, as implicit document relevance feedback, and identified novel patterns of touch interactions for predicting document relevance. The results demonstrated significant improvements to search ranking quality by mining touch interaction data. Luca et al. [16] proposed an implicit authentication approach to enhance password patterns which are easy to steal and reproduce. They used touch screen data of smartphones (pressure, coordinates, size, speed, time etc.) to distinguish between the rightful user and an attacker. The results proved that this implicit authentication approach actually works.

An important step that we can find in those work is to extract a set of task-related behavioral features from the touch interaction data. These features are exactly the key of improving the accuracy of prediction models. Moreover, previous work have also demonstrated that touch behavioral features are able to reflect users' subjective feelings. Gao et al. [20] performed a study with 15 participants to investigate whether touch behaviors reflect players' emotional states. They extracted finger-stroke features during gameplay on an iPod and analyzed the discriminative power of these features. The results show a good accuracy for discriminating between four affective states using stroke behavior. Similarly, the assumption that we make in this study is that there are some touch behavioral features can reflect users' satisfaction with the typography design. Thus, we will compare the touch behaviors collected from the reading sessions which get different satisfaction ratings to find the discriminative features.

USER STUDY 1

The goal of this study is to capture meaningful patterns from users' touch behaviors when they read under varying text difficulty levels to evaluate user satisfaction with the typography designs. We majorly consider one typography variable: font size, which is one of the crucial factors for reading, to reduce the complexity of the experiment and remain other design factors unchanged.

Design

We use a mixed-measure design. Text difficulty with 2 levels (*easy* and *hard*) is a between subject variable and font size with 5 levels (11, 14, 17, 21, 26 points) is a within subject variable. In this experiment, each participant read five documents with different font sizes but under a same difficulty level. So that we can avoid the influence of text difficulty on each participant, and we can find the discriminative features more easily. Hence, for font size, we collected repeated measures, while for text difficulty, we obtained between-group data. Text difficulty was assigned to each participant randomly and the order in which the font sizes were presented was counter-balanced.

Among the five font sizes, 17 point is recommended by the iOS human interface guidelines for the body text using [24]. We choose 11 and 14 points because these relative small font sizes can contain more information in one screen, so that people may need less swipes to read the whole content. And we choose the larger font sizes, 21 and 26 points, to cover a wide range of sizes, as previous work has indicated that larger font sizes can improve the readability on online materials [33].

Ten documents written in Chinese which is participants' native language were selected and divided evenly into two text difficulty levels (easy texts and hard texts). We measure the text difficulty level of Chinese texts according to the formula in [12] and our research group members also verified the results. The topic of easy texts and hard texts are family story and fruit wiki, respectively. These two topics are of general interest, not technical or academic, so that participants would not feel boring [33]. Documents are of nearly the same length, around 1000 words (mean length = 1012 words, SD = 24 words), which is common on mobile devices. This length is able to make sure that participants should take several minutes to read, and not make them feel fatigue or boring.

Satisfaction Rating

Participants' subjective satisfaction ratings are collected, as the ground-truth, to evaluate the typography design. Hence, after reading all documents, participants are asked to rate the typography design of each document they have read on a five-point Likert scale (1 - very bad, 5 - very good).

In this study, satisfaction ratings greater or equal to 4 ("good") were considered "user-satisfied", while ratings less than 4 were considered "user-unsatisfied".

Comprehension Score

To motivate the participants to read the documents carefully, we prepare some easy comprehension questions which are content-relevant and easy to find the answer. Each document has five true or false questions with three alternative choices:

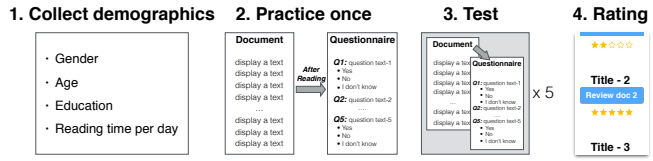


Figure 1. The procedure of the experiment.

"yes", "no", and "I don't know". After reading a document, the participants should immediately answer the corresponding questions without reading the document again. If the answer for a question is correct, the participant get one score. Thus, for each document, a participant can get five scores at most. We are interested in finding a possible relationship between satisfaction rating and comprehension score.

Touch Behavior Log

In this experiment, the participants use their fingers to swipe on the screen to change the displayed content, which is a typical way to read on mobile devices [37]. Thus, we mainly record swipe gesture behavior data. For each swipe gesture event, the mobile application, which is developed for showing the documents, automatically collects the following information:

- the start and end time of each swipe.
- x and y coordinates of touch points of each swipe.
- d: the offset along the y axis.

Participants and Apparatus

We invite 52 participants (mean age = 23.7 years, SD = 1.4 years; 17 female), including 9 undergraduate students, 35 master students, and 8 doctoral students. All participants have normal vision or are corrected to normal vision. All participants are familiar with smartphones and 34 are iPhone users. In terms of the time spent on mobile reading each day, 22 participants spend <= 1 hour, 22 participants spend > 1 hour but <= 3 hours, and 7 participants spend > 3 hours; only 1 participant spends <= 15 minutes on mobile reading.

We use an iPhone 5S with a 4-inch screen size (1136×640) as the test smartphone. Each participant has a chance to be familiar with the test smartphone before the experiment. We develop a mobile application for showing the documents and deploy it on the test smartphone.

Procedure

We conduct the experiment at our laboratory, and it lasts around 20 minutes for each participant. Each experiment takes place in a quiet, clean, and well-lit room, where only the interviewer (the first author) is present, to ensure that the participants could concentrate. Participants are asked to sit on a chair with their preferred posture holding the smartphone. The brightness of the smartphone is set to the moderate level and all notifications are disabled.

Participants are asked to perform the following four steps (see Figure 1). First, they show us their basic demographic information. Then, they read a sample article and answer the questions for a practice. Next, they read five articles in silence, and answer the corresponding questions after finishing each article. At the end, all articles are presented together to the participants for rating.

Font Size (points)	User-Unsatisfied		User-Satisfied	
	hard	*easy*	*hard*	*easy*
11	26	26	0	0
14	21	15	5	11
17	1	3	25	23
21	3	8	23	18
26	18	21	8	5

Table 1. The number of "user-unsatisfied" and "user-satisfied" with different font sizes under two text difficulty levels.

Measures	Hard	Easy
Satisfaction rating	3.2 (1.5)	3.0 (1.5)
Comprehension score	3.7 (1.1)	3.8 (0.9)
Number of swipes	35.8 (26.5)	30.9 (28.3)
Reading time	177.3 (55.0)	108.7 (31.1)

Table 2. Means and standard deviations of four measures under two text difficulty levels.

Results

Across all participants, we gathered 5 docs$\times$52 participants = 260 sessions (easy texts: 130, hard texts: 130). All of these reading sessions were completed successfully.

Table 1 shows the number of "user-unsatisfied" and "user-satisfied" with different font sizes under two text difficulty levels. As we can see, except for font size of 11 points, which is unsatisfied by all participants, other font sizes are all satisfied or unsatisfied by some participants. It indicates that the participants' preferences are different, which means they have personalized needs. Thus, we can recognize that using the recommended design, such as font size 17 points which is also the most popular font size in this experiment, may be the most convenient way for designers to present a good typography design, but not the way to satisfy users' individual needs for providing the best reading experience.

We also compare the mean and the standard deviation of four measures under two text difficulty levels, which are summarized in Table 2. No significant difference was found on these measures under the two levels, except for reading time. A t-test shows a significant effect of text difficulty on reading time ($t(258) = -12.39, p < .001$). Participants spent more time reading the hard texts, which makes sense and demonstrates that our text difficulty design is reasonable.

For the easy texts, there is no significant correlation between satisfaction rating and comprehension score ($r(128) = -.01, p = .87$). While, for the hard texts, we find a significant and positive correlation between rating and score ($r(128) = .18, p < .05$). It indicates that participants are more easily affected by the typography design when they read the hard texts. This makes sense because participants obviously need more time and efforts to read the hard texts, and a satisfied typography design can help them read the text more efficiently and improve the comprehension of the content.

Besides, we have 8567 touch interaction records collected from the 260 reading sessions. In the next section, we will analyze the touch interaction data to explore users' reading processes in a qualitative way.

ANALYZE READING BEHAVIORS QUALITATIVELY

In this section, we qualitatively analyze the touch interaction data from two aspects. The detailed descriptions are follows.

Relationship between Swipe Distance and Reading Time

When users use fingers to swipe on the screen of mobile devices to change the displayed content, their reading behaviors are usually a mix of swiping behaviors and inactivity behaviors (the interval between two consecutive swipes). We assume that user satisfaction with the typography design will affect users' emotions and reflect in their reading behaviors. We compare the reading behaviors from two reading sessions, in which the typography designs get different satisfaction ratings from a same participant, and provide some illustrative examples under two text difficulty levels respectively in Figure 2.

Figure 2 (a) and (b) present two reading sessions of the hard texts by a same participant (*uid=31*). The page offset (y-axis), which is constantly altered by the swiping event, is charaterized as the function of the reading time (x-axis). Figure 2 (c) and (d) present another two reading sessions of the easy texts in the same way but from another participant (*uid=10*).

As shown in Figure 2 (b) and (d), the relationship between swipe distance and reading time seems like a good linear relationship. The typography designs in these two sessions both get a high satisfaction rating (which indicates "user-satisfied"). While this observation does not apply for Figure 2 (a) and (c) and the typography designs in these two sessions both get a low rating (which indicates "user-unsatisfied"). In fact, the participants' reading behaviors change frequently in Figure 2 (a) and (c). For example, the durations of inactivity are range from a few seconds to more than 20 seconds, which implies that the participants try to read carefully but they can not immerse themselves in reading, because the terrible typography design may make them feel impatient. Besides, if we take a closer look at these reading sessions, we can divide the swiping behaviors and inactivity behaviors into groups according to different swipe lengths and inactivity durations respectively. So that we may find some frequent behavior state transitions which probably occur in the reading sessions that users are satisfied or unsatisfied with the typography designs. Based on the above observations, we proposed the first hypothesis:

H1. If users are satisfied with the typography design in a reading session, they probably read the text at a moderate and steady speed without changing reading behaviors frequently.

Moreover, we also find the participants tend to take more time at the beginning of a reading session when they are unsatisfied with the typography design. For example, in Figure 2 (c), the participant spent more than 1/3 of the total reading time at the beginning of the reading session, but only swiped less than 1/5 of the total distance. It implies that the typography design is different from the participant's expectation and had a negative influence on the participant's reading efficiency, so that the participant needed more time to adapt to the design. According to this observation, we propose the second hypothesis:

H2. If users are not satisfied with the typography design in a reading session, they probably spend more time on adapting to the design at the beginning of the session.

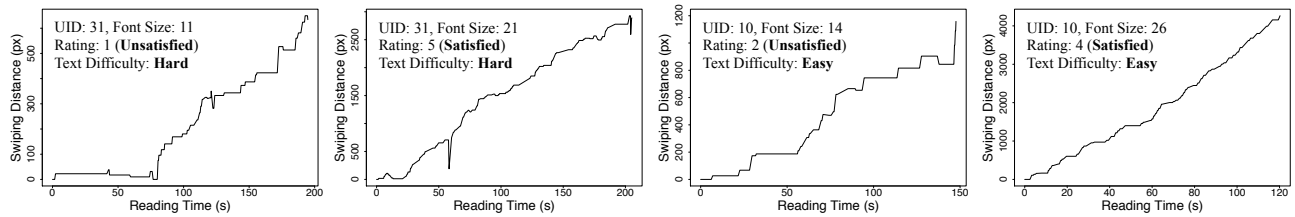


Figure 2. Examples of the relationship between swipe distance and reading time, represented by the vertical swipe coordinates (black lines) over time. The x-axis represents the time from the document displayed in seconds, and the y-axis represents the vertical page offset resulted by the swiping event. (a) a user (uid = 31) reading a document which is from the hard texts with an unsatisfied (rating = 1) typography design; (b) the same user (uid = 31) reading a document which is also from the hard texts with a satisfied (rating = 5) typography design; (c) another user (uid = 10) reading a document which is from the easy texts with an unsatisfied (rating = 2) typography design; (d) the same user (uid = 10) reading a document which is also from the easy texts with a satisfied (rating = 4) typography design.

Distribution of Touch Points

As mentioned above, participants use fingers to swipe on the screen to change the displayed content. A swipe gesture is a sequence of touches with x and y coordinates, starting with a "down" touch which leaves a begin touch point. We assume the distribution of begin touch points of all swipes in a reading session can reflect the participant's satisfaction with the typography design. In order to find significant distribution differences, we compare the distributions of begin touch points from two reading sessions, in which the typography designs get different ratings, and provide some illustrative examples under two text difficulty levels respectively in Figure 3.

Figure 3 (a) and (c) present the distributions of touch points from two reading sessions of the hard texts by a same participant (*uid = 47*), each of which is represented by the coordinate system of the test smartphone with the touch points on it. The touch points are divided evenly into three groups according to the chronological order of touch events. Red, green and blue points respectively represent the first, second and third groups. Figure 3 (b) and (d) also present the distributions of touch points from another two reading sessions of easy texts in the same way but from another participant (*uid = 16*).

As shown in Figure 3 (a), the touch points are scattered widely on the bottom of the screen and the participant is not satisfied with the typography design. While in Figure 3 (b), the touch points are clustered on the bottom of the screen and the participant is satisfied with the typography design. A same observation appears in the comparison between Figure 3 (c) and (d), which implies the participant's satisfaction with the typography design may affects the spatial distribution of the touch points. Besides, in Figure 3 (c), some touch points in green color are far away from the touch points in other two colors, which indicates the participant's satisfaction with the typography design may also affects the spatio-temporal distribution of the touch points. Thus, based on the above observations, we propose the third hypothesis:

H3. If users are satisfied with the typography design in a reading session, the distribution of touch points are probably more clustered.

The reading sessions shown in Figure 2 and 3 are not special cases. Actually, for Figure 2, there are 26 (100%) participants and 59 (45.4%) reading sessions show the similar patterns under easy level. And there are 26 (100%) participants and 66 (50.8%) sessions show the similar patterns under hard level. For Figure 3, there are 24 (92.3%) participants and 62 (47.7%) sessions show the similar patterns under easy level. And there are 25 (96.2%) participants and 74 (56.9%) sessions show the similar patterns under hard level. These demonstrate that our findings can be observed among all sets of experiments.

We have proposed three hypotheses above to reveal significant behavior differences for evaluating satisfaction. Essentially, the first hypothesis (**H1**) is proposed from the time dimension, while the third hypothesis (**H3**) is proposed from the spatial dimension. And the second hypothesis (**H2**) is proposed from the user expectation dimension. In the next section, we will present a set of features to validate these hypotheses.

ANALYZE READING BEHAVIORS QUANTITATIVELY

In this section, we present a detailed description of a variety of features for supporting the three hypotheses and show the correlations between these features and satisfaction rating.

Feature Description

Except for the features used to validate those hypotheses, we also consider reading time and user efforts, which are typically common measures used to evaluate user satisfaction. The detailed descriptions of these features are as follows.

Reading Time

Task completion time has been proved to be a strong predictor of user satisfaction in many tasks [22]. Participants will be more satisfied when they take less time to finish a specified task. In our task, *reading time* is defined as the interval, in seconds, between the time the document is presented and the time the participant finishes the reading.

User Efforts

Due to the small screens of mobile devices, users need to swipe many times to read a document on such devices. Thus, to evaluate the efforts of a user should take for reading a document, we present two features: *swipe count per distance* and *swipe frequency*, which were proven to be negatively correlated with user satisfaction in previous study [18]. In our task, *swipe count per distance* is defined as the ratio of the number of swipes to the total swipe distance, and *swipe frequency* is defined as the ratio of the number of swipes to the total duration of swipes.

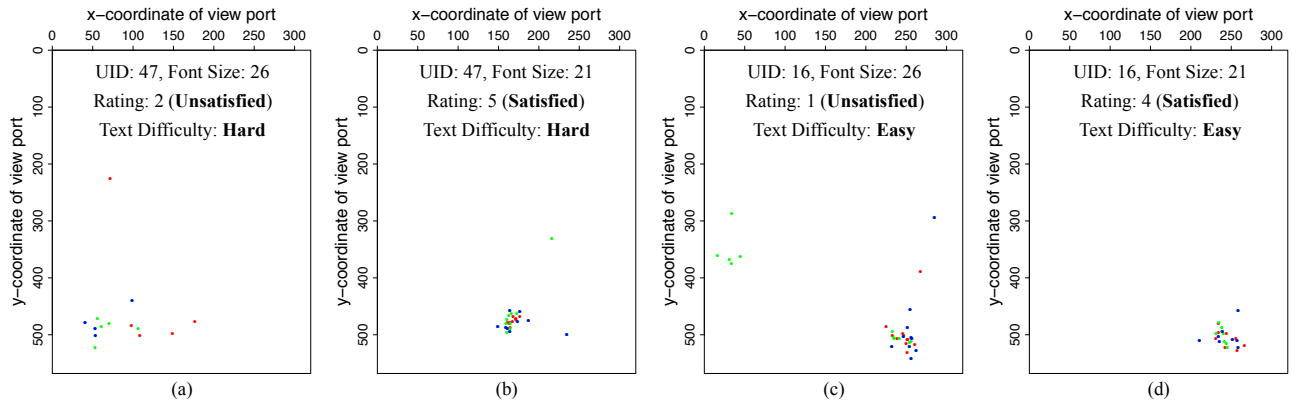


Figure 3. Examples of the distribution of touch points, represented by colored dots on the coordinate system of a smartphone. These touch points are evenly divided into three groups according to the chronological order of touch events. Red, green and blue points respectively represent the first, second and third group. (a) a user (uid = 47) reading a document which is from the hard texts with an unsatisfied (rating = 2) typography design; (b) the same user (uid = 47) reading a document which is also from the hard texts with a satisfied (rating = 5) typography design; (c) another user (uid = 16) reading a document which is from the easy texts with an unsatisfied (rating = 1) typography design; (d) the same user (uid = 16) reading a document which is also from the easy texts with a satisfied (rating = 4) typography design.

State	Description
SDVS	very short length of swipe down (< 1 line)
SDS	short length of swipe down (1∼2 lines)
SDM	medium length of swipe down (2∼5 lines)
SDL	long length of swipe down (> 5 lines)
SU	swipe up
IVS	very short period of inactivity (< 1s)
IS	short period of inactivity (1∼3s)
IM	medium period of inactivity (3∼10s)
IL	long period of inactivity (> 10s)

Table 3. Behavior states used in sequence modeling of a reading session.

Features for Hypothesis 1

To validate the first hypothesis, we provide a set of features based on a three-level (macro, meso and micro level) analysis.

In the macro level analysis, we consider the *linear correlation* between reading time and swipe distance. According to our observations in the above section, we assume that the stronger the linear correlation, the more satisfied the user will be.

In the meso level analysis, we focus on swiping behavior and inactivity behavior. For the swiping behavior, we consider the standard deviation of the length, duration and speed of the swipes. The speed of a swipe comes from the length of the swipe divided by the duration of the swipe. For the inactivity behavior, we consider the standard deviation of inactivity duration. We assume that the lower the standard deviation value, the more satisfied the user will be, which is in line with previous work [29].

In the micro level analysis, we define the fine-grained user behaviors during a mobile reading session as a sequence of observed behavior states: $S = START \rightarrow s_1 \rightarrow ... \rightarrow s_n \rightarrow END$, where $s_i \in \{SDVS, SDS, SDM, SDL, SU, IVS, IS, IM, IL\}$. We have summarized the behavior states in Table 3. We assume some behavior state transitions are likely to occur in the reading session that the participants are satisfied or unsatisfied with the typography design [22]. For example, a long period of inactivity (> 10s) followed by the start of a reading session is an indicator of encountering difficulties, while short first inactivity (< 3s) may suggests the document were easy to read. Therefore, the features from this analysis aim to capture these sequential patterns. When we extract the sequence of behavior states from a reading session, we first count the occurrences of transitions $s_i \rightarrow s_j$, and then filter the transitions which occurs less than three times. After finishing the extractions and filtrations, we calculate the likelihoods of the transitions occur in the satisfied and unsatisfied situations, and get the corresponding likelihoods ratio.

Features for Hypothesis 2

To validate the second hypothesis, we propose two features, *initial time to distance* (*iT2D*) and *initial distance to time* (*iD2T*), for evaluating a participant's reading performance at the beginning of a reading session. The definitions of these two features are given in Equation 1.

$$iT2D = \frac{D_{T_\alpha}}{D}, \quad iD2T = \frac{T_{D_\alpha}}{T} \tag{1}$$

- D: The total swipe distance in a reading session.
- T: The total reading time in a reading session.
- D_{T_α}: the sum of distance that a user swipes at the beginning of a reading session when it takes the total time equal to or nearly equal to $\alpha \times T$ $(0 < \alpha < 1)$.
- T_{D_α}: the time that a user takes at the beginning of a reading session to finish the sum of swipe distance which is equal to or nearly equal to $\alpha \times D$ $(0 < \alpha < 1)$.

We set $\alpha = 0.3 \pm 0.05$. We assume that the higher the value of *iT2D*, which means the user reads a lot of content at the beginning of the reading session and indicates that it is easy for the user to read the document, so the more satisfied the user will be. Similarly, we also assume that the higher the value of *iD2T*, which means the user spends much time at the beginning of the reading session and implies the user probably encounters some difficulties in reading, so the more unsatisfied the user will be.

Feature	Hard	Easy
reading time	.034	.022
swipe count per distance	-.106	-.124
swipe frequency	.022	-.060
linear correlation	.145	.143
std of swipe length	-.001	-.094
std of swipe duration	-.137	-.144
std of swipe speed	.025	-.122
std of inactivity duration	**-.337***	**-.180***
SDS → IS	**.255***	.089
IS → SDS	**.220***	-.012
initial time to distance	.114	-.007
initial distance to time	**-.265***	-.125
average touch point distance	-.098	-.001
std of touch point distance	-.107	.056
temporal touch point distance	-.158	-.090

Table 4. Pearson's correlation between satisfaction rating and the value of features (* indicates < .05 statistical significance).

Features for Hypothesis 3

To validate the third hypothesis, we propose three features to measure the dispersion degree of the distribution of touch points. To describe these three features formally, we first present the definitions of two basic variables, *center point* and *touch point distance* (TPD), in Equation 2 and 3.

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n}, \quad y_c = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2}$$

- x_c, y_c: The x and y coordinates of the center point.
- x_i, y_i: The x and y coordinates of the i-th touch point.
- n: the number of touch points.

$$TPD_i = \sqrt{(x_c - x_i)^2 + (y_c - y_i)^2} \tag{3}$$

- TPD_i: the distance between the i-th touch point and the center point.

From the spatial dimension, we present two features, *average touch point distance* and *standard devidation of touch point distance*, which are the average and the standard deviation of the *touch point distance* of all touch points. Based on our observations in the above section, we assume that the higher the two features value, the more unsatisfied the user will be.

Moreover, as shown in Figure 3, touch points are evenly divided into three groups according to the chronological order of touch events. Thus, from the spatio-temporal dimension, we present a feature, *temporal touch point distance* ($tTPD$), and show its definition in Equation 4. We assume that the higher the value of $tTPD$, the more unsatisfied the user will be.

$$\begin{aligned} tTPD = &\sqrt{(x_{c_1} - x_{c_2})^2 + (y_{c_1} - y_{c_2})^2} \\ &+ \sqrt{(x_{c_2} - x_{c_3})^2 + (y_{c_2} - y_{c_3})^2} \end{aligned} \tag{4}$$

- x_{c_i}, y_{c_i}: the center point of touch points in the i-th group.

Difficulty	*Transition*	*Satisfied*	*Unsatisfied*	*Ratio*
Hard	SDVS → IS	0.115	0.021	5.476
	SDS → IS	0.344	0.064	5.375
	IS → SDS	0.311	0.106	2.934
	IM → SDM	0.639	0.340	1.879
Easy	IL → SDL	0.175	0.093	1.882
	SDS → IS	0.211	0.130	1.623
	IVS → SU	0.070	0.148	0.473

Table 5. The satisfied and unsatisfied likelihoods, and the likelihood ratio of behavior state transitions under two text difficulty levels.

Feature Results

We summarize the Pearson's correlation between the value of these representative features and satisfaction rating under two text difficulty levels respectively in Table 4.

As we can see, *reading time* is found to have a positive weak correlation with rating under the two levels, which is contrary to our expectation of a negative correlation. *Swipe count per distance* is found to be a indicator of reader unsatisfaction, exhibiting a negative but insignificant correlation around -0.11 with satisfaction rating under the two levels. This makes sense because the more efforts the users need to make to read the text, the more likely they are unsatisfied with the typography design. In contrast, *swipe frequency* is found to have a weak correlation with satisfaction rating under the two levels. Somewhat surprsing is the positive correlation (instead of a negative correlation as assumed) for the hard texts. One possible reason is that the high swipe frequency may indicate the participant's good comprehension of the text content.

The linear correlation between swipe distance and reading time is found to have a positive correlation around 0.14 with satisfaction rating under the two levels. The swiping and inactivity behavior exhibit negative correlations with reader satisfaction. This makes sense because the more frequent the user changes swiping and inactivity behavior, the more likely the user is impatient and unsatisfied with the typography design. One subtle difference is that the standard deviation of swipe speed has a positive correlation (instead of a negative correlation as assumed) with rating for the hard texts. A possible reason is that the participants' reading speed tend to be faster when they get familiar with the main idea of the document.

Table 5 shows examples of behavior state transitions that are indicative of user satisfaction, with each row reporting the likelihoods of a state transition that appears in the "user-satisfied" and "user-unsatisfied" situations and the corresponding likelihood ratio. *SDS → IS* is an indicator of user satisfaction for the hard texts, since it is 5.4 times more likely to happen in reading with a satisfied design as compared to an unsatisfied design. Its opposite transition *IS → SDS*, similarly, exhibiting a 2.93 likelihood ratio. The combination of these two transitions shows that a user swipes down 1∼2 lines, then rests around 3 seconds, and this process repeats a few times, which implies that the user reads fluently with a satisfied design.

The initial reading performances of users are found to be indicators of reader satisfaction, especially for the feature *initial distance to time*, exhibiting a negative correlation with rating under the two levels. This makes sense because if users

spent too much time at the beginning of a reading session, they probably encountered some difficulties and the terrible typography design may contributed a lot. Somewhat surprising is that the feature *initial time to distance* does not work for the easy texts. One possible reason is that users are not familiar with the background of the story at the first, and when they understand the main idea, their reading speeds lead to faster.

Unsurprisingly, features of the distribution of touch points are found to exhibit negative correlations with satisfaction rating. But the correlations are weak and insignificant, which may be explained by the short length of the documents, so that the number of touch points is not adequate. One subtle difference is that, for the easy texts, the feature *average touch point distance* does not work, and the feature *std of touch point distance* exhibits a positive corelation (not a negative corelation as assumed) with rating. One possible reason is that users usually read the story faster after they get familiar with the content, thus the distribution of touch points will change over time. Due to the same reason, we can see that the feature *temporal touch point distance* works under the both levels.

Discussion

Our findings support the assumption that user satisfaction with the typography design will affects users' touch behaviors in reading. The results also imply that text difficulty has an influence on users' reading behaviors. The validation results of the three hypotheses are summarized in Table 6. As we can see, for the hard texts, all the hypotheses are supported, while for the easy texts, the first hypothesis is supported and the other two hypotheses are partly supported. This makes sense because compared to the easy texts, users probably need more time and efforts to read the hard texts. Thus, users' reading behaviors are more easily affected by the typography design when they read the hard texts.

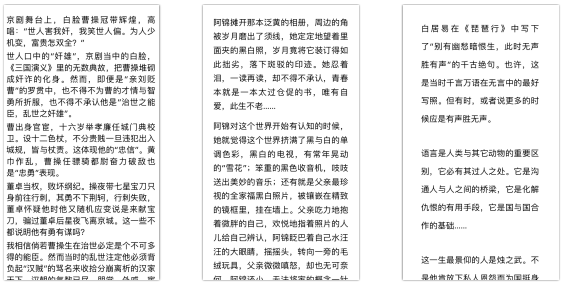
Hypothesis	***Hard***	***Easy***
H1	Supported	Supported
H2	Supported	Partly Supported
H3	Supported	Partly Supported

Table 6. The validation results of the three hypotheses under the two text difficulty levels.

In the user study 1, we majorly consider changing font size as different typography designs, which simplify the experiment so that we can find the meaningful patterns more easily. But it is not in line with the actual situation. In the real world, to display texts well, designers will consider many other design factors, such as line spacing, page margin, and combinations of these design factors. Moreover, each participant only read documents of a similar text difficulty level, which is also not in line with the actual situation. Hence, to validate our findings in a more realistic situation, we consider more typography design factors and their combinations and more text difficulty levels in the study 2. And we evaluate user satisfaction with the typography design by modeling the touch interaction data.

USER STUDY 2

In this study, we consider four typography design factors and their combinations for validating our findings. And we also take into account documents of three text difficulty levels.



(a) (b) (c)

Figure 4. Examples of typography designs in font size of 17 points of three levels of text density: (a) compact; (b) moderate; (c) loose.

Method

In this experiment, we invite 39 participants (mean age = 23.8 years, SD = 2.3 years; 6 female), including 15 undergraduate students, 15 master students, and 9 doctoral students. All participants have normal vision or are corrected to normal vision. We used the same experimental set-up as in the first study with two exceptions. One is that participants are asked to read six documents under different text difficulty levels in a random sequence with a random assignment of six typography designs. The other one is that the number of questions is reduced to two since the length of documents become shorter.

We prepare six typography designs, which are combinations of four typographic variables, including font size, line spacing, paragraph spacing and page margin. These four design factors are all proven to have a great influence on user's reading efficiency in prior work [18, 33]. Among the six designs, we consider two levels of font size (17 points and 21 points), and three levels of text density (loose, moderate and compact), which are combinations of line spacing, paragraph spacing and page margin. We present the design specification of the three levels of text density in Table 7. We also show some examples of typography designs of the three levels of text density in font size of 17 points in Figure 4.

Text density	Design		
Loose	*line spacing*	=	$1.0 \times$ *font size*
	para. spacing	=	$2.0 \times$ *line spacing*
	page margin	=	$0.1 \times$ *screen width*
Moderate	*line spacing*	=	$0.5 \times$ *font size*
	para. spacing	=	$1.5 \times$ *line spacing*
	page margin	=	$0.05 \times$ *screen width*
Compact	*line spacing*	=	$0.2 \times$ *font size*
	para. spacing	=	$1.2 \times$ *line spacing*
	page margin	=	0

Table 7. The design specification of three levels of text density.

We prepare six documents on topics of fairy tale, biography and philosophical essay, which are of general interest. These documents are divided into three text difficulty levels (easy texts, medium texts and hard texts), each has two documents. We measure the text difficulty of these documents using the same method as in the first study, and the results are verified by our research group members. The length of documents are about 800 words (mean length = 795 words, SD = 43 words).

Data Preparation and Analyses

There are 39 participants each reading 6 documents in this user study, thus we get 234 reading sessions in total. All of these reading sessions were completed successfully. Table 8 shows the mean and the standard deviation of reading time under three text difficulty levels. We find a significant effect of text difficulty on reading time ($F(2,231) = 11.41, p < .001$), which indicates that our text difficulty design is reasonable.

Text difficulty	Mean	SD
Easy	88.6	26.5
Medium	101.4	33.9
Hard	113.8	37.6

Table 8. Means and standard deviations of reading time per difficulty.

According to the text difficulty levels, we divide the touch interaction data extracted from the 234 reading sessions into three small datasets, each involving the data from 78 reading sessions. The following results on evaluating user satisfaction with the typography design under three text difficulty levels respectively are based on the modeling of these data.

Evaluation Models

We treat this user satisfaction evaluation problem as a binary classification problem, "user-satisfied" or "user-unsatisfied". We select KNN, Random Forest, SVM, GBDT and AdaBoost as our evaluation models. Random Forest and SVM are proven to be the best classification models on solving classification problems [17]. We choose the GBDT and AdaBoost, since they are two good ensemble learning models, which have a potential high classification accuracy. The advantage of KNN model is simple and efficient.

Because our datasets are small, we use a leave-one-out cross-validation method to train our models. A reading session is selected in each iteration as a test sample, and all remaining reading sessions are regarded as a training set. Each reading session is selected only once and all other reading sessions of the participant of the test sample are removed from the training data for avoiding overfitting.

We use a classical approach as the baseline, which selects the class that has the most observations in the training data and uses that class as the prediction of the test sample [9].

Evaluation Results

Figure 5 shows the comparison of the results of our models with the baseline. We can see that, for the hard texts, our evaluation models acquire statistically significant higher accuracy than the baseline ($p < .05$). For the medium texts, the results show that our features capture meaningful patterns in reading behaviors, but only GBDT and AdaBoost models have statistically significant higher accuracy than the baseline ($p < .05$). For the easy texts, the results also indicate that our features find meaningful patterns in reading behaviors, but no models get a statistically significant higher accuracy than the baseline.

Table 9 shows the comparison of the best performance of our models with the baseline under three text difficulty levels. As we can see, the best classification accuracy that our models can get under the three levels are all more than 72%, which is a huge increase as compared to the baseline.

	Easy	*Medium*	*Hard*
Best Accu.	0.72(+24%)	0.74(+48%)	0.74(+71%)
Baseline	0.58	0.50	0.44

Table 9. The comparison of the best classification accuracy (Best Accu.) of our models with the baseline under three text difficulty levels.

The Importance of Features

To determine the most discriminative behavioral features for the evaluation of user satisfaction under the three text difficulty levels, we used a multiple linear regression with backward elimination. Table 10 displays the top-5 most discriminative features for evaluating user satisfaction with the typography design under the three levels.

Text difficulty	No.	Feature	$+/-$
Easy	1	*std of TPD*	-
	2	*average TPD*	+
	3	*temporal TPD*	-
	4	*swipe count per distance*	-
	5	*std of swipe duration*	+
Medium	1	*reading time*	-
	2	*std of inactivity duration*	+
	3	*temporal TPD*	-
	4	*average TPD*	+
	5	$IL \rightarrow SDM$	+
Hard	1	*linear correlation*	-
	2	$IM \rightarrow SDS$	-
	3	$SDS \rightarrow IVS$	-
	4	*std of swipe speed*	-
	5	$IM \rightarrow SDL$	+

Table 10. The top-5 most discriminative features for evaluating user satisfaction with the typography design under three text difficulty levels ($+$: positive correlation, $-$: negative correlation).

For the easy texts, as we expected, *std of touch point distance*, *temporal touch point distance* and *swipe count per distance* all have a negative correlation with user satisfaction. It is in line with our findings in the first study, supporting the third hypothesis (**H3**). However, *average touch point distance* and *std of swipe duration* both exhibit a positive correlation (not a negative corelation as assumed), which may be explained by the existence of paragraph spacing and large line spacing – for example, users usually swipe a large distance with a fast speed to skip the paragraph spacing or large line spacing, so that the distribution of the touch points tends to be scattered and the swipe duration tends to have a greater change.

For the hard texts, the most discriminative features include $IM \rightarrow SDS$, $SDS \rightarrow IVS$, *std of swipe speed* and $IM \rightarrow SDL$, which are relate to the findings in the first study, supporting the first hypothesis (**H1**). $IM \rightarrow SDS$ is found to have a negative correlation with user satisfaction. This makes sense because when the user has rested around 10 seconds, but only swipe down one or two lines, which indicates the user may encounter difficulties in the reading. While its counterpart $IM \rightarrow SDL$ is found to have a positive correlation with user satisfaction. It indicates that the user swipe down about 5 lines after a rest of around 10 seconds, which implies that it is easy for the user to read the document.

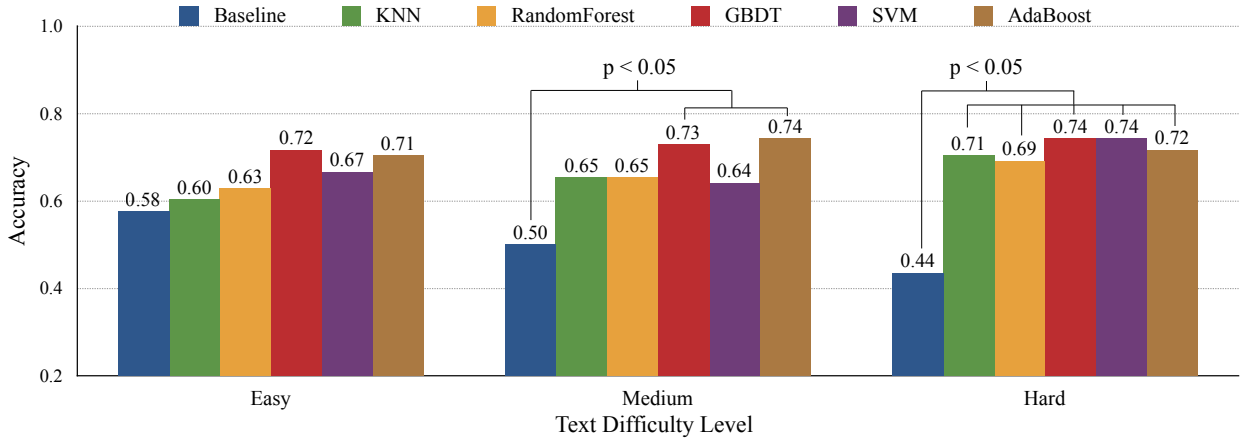


Figure 5. The classification accuracy of user satisfaction with the typography design under three text difficulty levels using baseline, KNN, Random Forest, GBDT, SVM and AdaBoost. Compared to baseline, statistically significant accuracy increments are labeled with p-value < 0.05. Our models considerably promote the accuracy when evaluating user satisfaction for the medium and the hard texts.

For the medium texts, to our surprise, *reading time* is proved to be a very discriminant feature, which is contrary to our findings (only a weak indicator of user satisfaction) in the first study. However, as we expected, *reading time* exhibits a negative correlation with user satisfaction. Other discriminative features for the medium texts also appear in the lists of the most discriminative features for the easy or hard texts. It indicates similar text difficulty levels may share some of the same discriminative features.

DISCUSSION AND LIMITATION

The evaluation models which based on our features acquire significant higher accuracy than the baseline in evaluating user satisfaction with the typography design. It shows that our features can capture meaningful patterns in the touch behaviors and demonstrates that the touch behaviors can reflect user satisfaction with the typography design. The detailed study of the most discriminative features under three text difficulty levels shows that text difficulty has a great influence on touch behaviors. And the model results show that compared to the baseline, our models get a better evaluation accuracy for the hard and the medium texts than for the easy texts. It implies that users' touch behaviors are more easily affected in reading the texts of the harder level.

A limitation of our study is that we only use the texts in Chinese as reading materials and ignore the texts of other languages. But, we think our work would be applicable to other languages for the following reasons. First, Chinese texts used in this study are read in the same direction as other languages, such as English and Spanish. They are all read from left to right in each line and from top to bottom in each page. Second, users usually read the texts by using fingers to swipe on the screen to change the displayed text. This reading manner is not affected by the language. Third, the behavioral features (e.g., swipe frequency) proposed in this study are independent with the language and mostly up to readers' personal habits. Fourth, varying text difficulty levels also exist in other languages and it also can affect users' reading performance (e.g., reading speed). So we believe that our experimental results are unbiased and will not only hold for Chinese. Due to the above reasons, we believe that our work is applicable to other languages. In future work, we will use different language texts and invite more participants of different backgrounds to validate the applicability of our method.

Besides, we did not consider more touch events, move event [1] for example, in this study. In fact, we investigated this aspect before in our pilot study, but the results were not good. The reason is that there are too many related variables (e.g., direction, length, angle, etc.) and our dataset is not large enough to cover these variables. Thus, we mainly recorded the vertical distance of touch points and the results were good which demonstrates our features do capture meaningful patterns. But when we conduct a large-scale in-situ study in future work, we should consider more touch events for capturing more patterns.

An obvious direction of our future work is improving the existing features. For example, the definitions of behavior states should be individualized. Moreover, we shall consider more typography factors (e.g., font family) and documents with multimedia factors (e.g., images and videos) to validate the generalizability of our method to the real world.

CONCLUSION

As the first step to provide personalized typography designs, evaluating user satisfaction without interrupting reading is still a challenge for designers. In this study, we conduct two reading studies and demonstrate that users' touch behaviors in reading can reflect their satisfaction with the typography design. We also find out text difficulty has an influence on the touch behaviors. The results show that evaluation models based on our features lead to higher accuracy than the baseline.

ACKNOWLEDGMENTS

We would like to sincerely thank all of the participants of the study for their generosity of time. Our work was partially supported by National Natural Science Foundation of China (No.61772459, No.61772461), National Key Research and Development Program of China (2017YFB1400600), Key Research and Development Project of Zhejiang Province (No.2015C01027, No.2017C01015), Natural Science Foundation of Zhejiang Province (No.LR18F020003, No.LY17F020014).

REFERENCES

1. Joanna Bergstrom-Lehtovirta and Antti Oulasvirta. 2014. Modeling the Functional Area of the Thumb on Mobile Touchscreen Surfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1991–2000. DOI: http://dx.doi.org/10.1145/2556288.2557354
2. Michael Bernard, Chia Hui Liao, and Melissa Mills. 2001. The Effects of Font Type and Size on the Legibility and Reading Time of Online Text by Older Adults. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems (CHI EA '01)*. ACM, New York, NY, USA, 175–176. DOI:http://dx.doi.org/10.1145/634067.634173
3. Michael Bernard, Bonnie Lida, Shannon Riley, Telia Hackler, and Karen Janzen. 2002. A comparison of popular online fonts: Which size and type is best. *Usability news* 4, 1 (2002), 2002.
4. Michael L Bernard, Barbara S Chaparro, Melissa M Mills, and Charles G Halcomb. 2003. Comparing the effects of text size and format on the readibility of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies* 59, 6 (2003), 823–835. DOI: http://dx.doi.org/10.1016/S1071-5819(03)00121-6
5. David Beymer, Daniel Russell, and Peter Orton. 2008. An Eye Tracking Study of How Font Size and Type Influence Online Reading. In *Proceedings of the 22Nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 2 (BCS-HCI '08)*. BCS Learning & Development Ltd., Swindon, UK, 15–18. http://dl.acm.org/citation.cfm?id=1531826.1531831
6. David Beymer and Daniel M. Russell. 2005. WebGazeAnalyzer: A System for Capturing and Analyzing Web Reading Behavior Using Eye Gaze. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1913–1916. DOI: http://dx.doi.org/10.1145/1056808.1057055
7. Sanjiv K Bhatia, Ashok Samal, Nithin Rajan, and Marc T Kiviniemi. 2011. Effect of font size, italics, and colour count on web usability. *International journal of computational vision and robotics* 2, 2 (2011), 156–179. DOI:http://dx.doi.org/10.1504/IJCVR.2011.042271
8. Dan Boyarski, Christine Neuwirth, Jodi Forlizzi, and Susan Harkness Regli. 1998. A Study of Fonts Designed for Screen Display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '98)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 87–94. DOI: http://dx.doi.org/10.1145/274644.274658
9. Jason Brownlee. 2014. How To Get Baseline Results And Why They Matter. https://machinelearningmastery.com/how-to-get-baseline-results-and-why-they-matter/. (2014). Accessed: Sep 17, 2017.
10. Daniel Buschek, Alexander De Luca, and Florian Alt. 2016. Evaluating the Influence of Targets and Hand Postures on Touch-based Behavioural Biometrics. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1349–1361. DOI: http://dx.doi.org/10.1145/2858036.2858165
11. Youli Chang, Sehi L'Yi, Kyle Koh, and Jinwook Seo. 2015. Understanding Users' Touch Behavior on Large Mobile Touch-Screens and Assisted Targeting by Tilting Gesture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1499–1508. DOI: http://dx.doi.org/10.1145/2702123.2702425
12. Alin Chen and Su Zhang. 1999. On the Difficulty Model of Chinese Text Reading and Readability Formula. *Computer Science* 11 (1999), 42–44. In Chinese.
13. Mark West & Han Ei Chew. 2014. Reading in the mobile era: A study of mobile reading in developing countries. http://www.unesco.org/new/en/unesco/themes/icts/m4ed/mobile-reading/reading-in-the-mobile-era/. (2014). Accessed: Sep 7, 2017.
14. Hsin-Yi Chiang and Sonia Chiasson. 2013. Improving User Authentication on Mobile Devices: A Touchscreen Graphical Password. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '13)*. ACM, New York, NY, USA, 251–260. DOI:http://dx.doi.org/10.1145/2493190.2493213
15. Iain Darroch, Joy Goodman, Stephen Brewster, and Phil Gray. 2005. The effect of age and font size on reading text on handheld computers. *Human-Computer Interaction-INTERACT 2005* (2005), 253–266. DOI: http://dx.doi.org/10.1007/11555261_23
16. Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. 2012. Touch Me Once and I Know It's You!: Implicit Authentication Based on Touch Screen Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 987–996. DOI:http://dx.doi.org/10.1145/2207676.2208544
17. Manuel Fernández Delgado, Eva Cernadas, Senén Barro, and Dinani Gomes Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15 (2014), 3133–3181. http://jmlr.org/papers/v15/delgado14a.html
18. Mary C. Dyson and Marf Haselgrove. 2001. The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies* 54, 4 (2001), 585 – 612. DOI:http://dx.doi.org/10.1006/ijhc.2001.0458
19. Mario Frank, Ralf Biedert, Eugene Ma, Ivan Martinovic, and Dawn Song. 2013. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE transactions on information forensics and security* 8, 1 (2013), 136–148. DOI:http://dx.doi.org/10.1109/TIFS.2012.2225048

20. Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What Does Touch Tell Us About Emotions in Touchscreen-Based Gameplay? *ACM Trans. Comput.-Hum. Interact.* 19, 4, Article 31 (Dec. 2012), 30 pages. DOI:http://dx.doi.org/10.1145/2395131.2395138

21. Jerzy Grobelny and Rafał Michalski. 2015. The role of background color, interletter spacing, and font size on preferences in the digital presentation of a product. *Computers in Human Behavior* 43 (2015), 85–100. DOI: http://dx.doi.org/10.1016/j.chb.2014.10.036

22. Qi Guo, Haojian Jin, Dmitry Lagun, Shuai Yuan, and Eugene Agichtein. 2013. Mining Touch Interaction Data on Mobile Devices to Predict Web Search Result Relevance. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '13)*. ACM, New York, NY, USA, 153–162. DOI: http://dx.doi.org/10.1145/2484028.2484100

23. Ding Long Huang, Pei Luen Patrick Rau, and Ying Liu. 2009. Effects of font size, display resolution and task type on reading Chinese fonts from mobile devices. *International Journal of Industrial Ergonomics* 39, 1 (2009), 81–89. DOI: http://dx.doi.org/10.1016/j.ergon.2008.09.004

24. Apple Inc. 2016. iOS Human Interface Guidelines: Typography. https://developer.apple.com/ios/human-interface-guidelines/visual-design/typography/. (2016). Accessed: Sep 7, 2017.

25. Yu-Cin Jian and Hwa-Wei Ko. 2017. Influences of text difficulty and reading ability on learning illustrated science texts for children: An eye movement study. *Computers & Education* 113 (2017), 263–279. DOI: http://dx.doi.org/10.1016/j.compedu.2017.06.002

26. Huy Viet Le. 2016. Modeling Human Behavior During Touchscreen Interaction in Mobile Situations. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '16)*. ACM, New York, NY, USA, 901–902. DOI: http://dx.doi.org/10.1145/2957265.2963113

27. Gordon E. Legge and Charles A. Bigelow. 2011. Does print size matter for reading? A review of findings from vision science and typography. *Journal of vision* 11, 5 (2011), 8. DOI:http://dx.doi.org/10.1167/11.5.8

28. Lin-Miao Lin, Karen M Zabrucky, and Dewayne Moore. 2002. Effects of text difficulty and adults' age on relative calibration of comprehension. *The American journal of psychology* 115 2 (2002), 187–98.

29. Pascual Martínez-Gómez and Akiko Aizawa. 2014. Recognition of Understanding Level and Language Skill Using Measurements of Reading Behavior. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*. ACM, New York, NY, USA, 95–104. DOI: http://dx.doi.org/10.1145/2557500.2557546

30. Mohammad Faizuddin Mohd Noor, Simon Rogers, and John Williamson. 2016. Detecting Swipe Errors on Touchscreens Using Grip Modulation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 1909–1920. DOI: http://dx.doi.org/10.1145/2858036.2858474

31. Jakob Nielsen. 2005. Lower-Literacy Users: Writing for a Broad Consumer Audience. https://www.nngroup.com/articles/writing-for-lower-literacy-users/. (2005). Accessed: Sep 7, 2017.

32. Thomas Peham. 2016. 6 mistakes to avoid when collecting design feedback. https://usersnap.com/blog/mistakes-collecting-design-feedback/. (2016). Accessed: Sep 12, 2017.

33. Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make It Big!: The Effect of Font Size and Line Spacing on Online Readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3637–3648. DOI: http://dx.doi.org/10.1145/2858036.2858204

34. Luz Rello, Martin Pielot, Mari-Carmen Marcos, and Roberto Carlini. 2013. Size Matters (Spacing Not): 18 Points for a Dyslexic-friendly Wikipedia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility (W4A '13)*. ACM, New York, NY, USA, Article 17, 4 pages. DOI: http://dx.doi.org/10.1145/2461121.2461125

35. Chao Shen, Yong Zhang, Zhongmin Cai, Tianwen Yu, and Xiaohong Guan. 2015. Touch-interaction behavior for continuous user authentication on smartphones. In *2015 International Conference on Biometrics (ICB)*. IEEE, 157–162. DOI: http://dx.doi.org/10.1109/ICB.2015.7139046

36. Philipp Tiefenbacher, Amir Chouchane, Daniel Merget, Simon Schenk, and Gerhard Rigoll. 2016. Biomechanics of Thumb Touch Gestures on Handheld Devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 3227–3233. DOI: http://dx.doi.org/10.1145/2851581.2892294

37. Wang-Chin Tsai, Yi-Lin Ro, Ya-Tzu Chang, and Chang-Franw Lee. 2011. The effects of font size and page presentation method of e-book reading on small screens for older adults. *Universal Access in Human-Computer Interaction. Context Diversity* (2011), 94–101. DOI: http://dx.doi.org/10.1007/978-3-642-21666-4_11

38. Jianing Zheng. 2017. How to Effectively Collect User Feedback in Mobile Application. https://www.infoq.com/articles/effectively-collect-user-feedback-mobile-apps. (2017). Accessed: Sep 12, 2017.