Displays 32 (2011) 81–86

Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

Usability study of text-based CAPTCHAs

Ying-Lien Lee *, Chih-Hsiang Hsu

Department of Industrial Engineering and Management, Chaoyang University of Technology, Taichung County 413, Taiwan

ARTICLE INFO

Article history:
Received 2 February 2010
Received in revised form 5 August 2010
Accepted 21 December 2010
Available online 13 January 2011

Keywords:
CAPTCHA
Age group
Distortion
Usability
Visual fatigue
Workload

ABSTRACT

Completely Automatic Public Turing test to tell Computers and Humans Apart, or CAPTCHA, is a security measure that guards a system from exploitation by the discrimination between a real human being and an automated computer program via the method of presenting to the unknown user the challenges that are hard for computer yet easy for human. Focusing on text-based CAPTCHA, this study conducted an experiment to study the effect of age groups and distortion types on the CAPTCHA task. Twenty-four participants were recruited to take part in the experiment, where twelve of them were in the elder group (aged 50–56) and twelve in the young group (aged 23–24). One type with no distortion and five types of common CAPTCHA distortion techniques were considered. The results of non-parametric analysis of the data revealed that the participants of two age groups differed significantly in terms of response time, error rate, and NASA-TLX score. Distortion type had significant effect on response time, error rate, Critical flicker fusion (CFF), and NASA-TLX score. Post-hoc analysis showed that Blot Mask and Line Mask were the hardest CAPTCHAs, while Thread Noise, Global Warp, and Geometry Noise were on a par with Normal Type (no distortion). Dependent variables were also correlated to each other. With the inevitability of the security measure and the increasing population of elder internet users, this study has important implications for the design of CAPTCHA systems.

1. Introduction

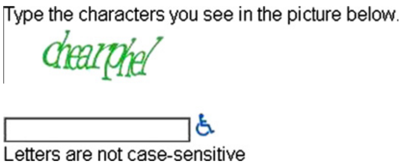
CAPTCHA, the acronym of “Completely Automatic Public Turing test to tell Computers and Humans Apart”, is a security measures that guards internet services against automated exploitations with abusive purposes [1]. CAPTCHA is a method widely employed to deter ill-motivated users from taking advantage of certain internet resources such as email accounts, online forums, blog comments, and online polls. The merit of a CAPTCHA system lies in the system’s capability to tell whether the user in question is a real human being or a robot program. A CAPTCHA process typically involves a session in which computer-generated questions are presented to users whose true identities are unknown to the system. Based upon the answers replied by the users, the CAPTCHA system determines whether the user in question is a human or not. To effectively tell computers and humans apart, the proposed questions have to be hard for computer to solve, yet easy for human to answer [2].

A variety of CAPTCHA question types have been proposed and implemented, such as picture labeling, text recognition, object identification, speech recognition, and puzzle solving [3]. Among them, text recognition, or text-based CAPTCHA, is the most common type because a computer can generate innumerable questions at very low cost so that the question–answer pairs are inexhaustible, an important mechanism underlying a successful CAPTCHA [4,5]. (For brevity reason, text-based CAPTCHA is referred to as CAPTCHA in the rest of the text unless being emphasized otherwise.) Fig. 1 is an example CAPTCHA used to guard the process of unlocking a locked Google account. An audio version of the question is also provided to make the service accessible to visually impaired users. One must type “chearphel” in the text field and submit the form to access the protected service. Another question will be presented if the answer is incorrect.

Among the techniques of breaking text-based CAPTCHA protection mechanism, Optical Character Recognition (OCR) program is one of the commonly deployed methods to defeat the protection [4]. Generally, OCR programs recognize characters contained in an image via three steps [6]: (1) pre-processing of the image to make the image suitable for further processing, (2) segmenting the image into regions in which each region contains only one character, and (3) identifying the character in each region. To lower the success rate of character recognition by the OCR programs, CAPTCHA systems usually distort the images in certain ways to complicate the steps OCR programs typically employ. For example, a CAPTCHA system may warp and/or segment the image globally or locally, or add extra noises (such as masks, lines of various slopes, or circles of various radii) to counter the attacks of OCR programs. With the distortion, however, CAPTCHA questions may become difficult to solve even for human being, which may stress

* Corresponding author. Address: No. 168, Jifong E. Rd., Wufong Township, Taichung County 413, Taiwan. Tel.: +886 4 23323000; fax: +886 4 23742327.
E-mail address: yinglienlee@gmail.com (Y.-L. Lee).

0141-9382/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.displa.2010.12.004



Type the characters you see in the picture below.

Letters are not case-sensitive

Fig. 1. A screenshot of the CAPTCHA of Google account.

the user's cognitive system and vision system, and make a potential customer walk away [2,7].

The human reading process begins with the visual perception of the raw stimuli of characters. The stimuli are then analyzed to extract features (such as vertical lines, horizontal lines, diagonals, and curves) that make up a character. The perceived groups of characters are in turn recognized as words that make up sentences [8]. Through practice and constant mapping, this bottom-up process of stimulus-to-character or stimulus-to-word can be *unitized* to such an extent that it becomes automatic [9]. The recognition of characters can also be aided by the word the characters form. People are more accurate in recognizing a letter when the character is presented in the context of a meaningful word than when the character is presented alone, or when the character is presented within a non-word. Such phenomenon is called the *word superiority effect* (WSE), a form of top-down processing [10,11].

Although text-based CAPTCHA systems use distortion techniques to lower the success rate of the attack by OCR programs, some attacks may utilize dictionaries to disambiguate uncertain words formed by the unrecognizable characters. To counter such kind of enhanced attacks, most text-based CAPTCHA systems use non-words, in addition to image distortion techniques, as the questions. However, such countermeasure also prevents the human solvers from using their knowledge of words, hence eliminating top-down processing, to aid the CAPTCHA solving process. Human solvers have to go through the bottom-up process more than once if the presented characters are not easily recognizable. Such condition may stress the visual system and cognitive system of human solvers. In the bottom-up processing, the first step is to analyze and extract features from the raw stimuli of characters, and then followed by the second step to recognize the most probable character given the extracted features. When this process is complicated by the lack of context (i.e. no top-down processing) and by the distorted raw stimuli, human solvers must repeat the bottom-up processing until an answer emerges. This will put extra loading on the visual system and cognitive system of human solvers.

Although the usability issues of CAPTCHA systems have been studied before [3], empirical investigation of CAPTCHA's usability issues is still lacking. In addition, considering the increasing population of elder internet users and the difficulties such security measures might pose to this population [13], elder group must be considered to understand the security measures impact on the visual and cognitive systems of users. This study hypothesized that CAPTCHA type and age group would have effect on the task performance, visual system, and cognitive system. Thus, the main goal of this research was to compare the effects of different CAPTCHA distortion techniques on human's visual system and cognitive system of different age groups. This study used an experiment, in which participants of two age groups (young and elder) took part, to collect objective and subjective data, including task performance, visual fatigue, workload, and post-session interviews. Comparisons within and between age groups were made to study the usability issues of the CAPTCHA distortion techniques.

2. Method

2.1. Participants

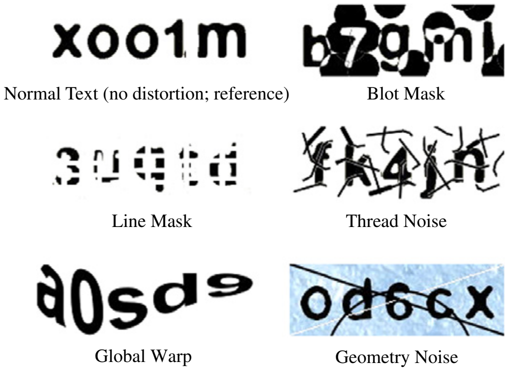
Two groups of participants were recruited to take part in the experiment. The 12 participants in the elder group were between 50 and 56 years old with an average of 53.2 years, while the other twelve in the young group were between 23 and 24 years old with an average of 23.6 years. There were four female participants; three of them were in the elder group and one in the young group. All participants had normal or corrected to normal visual acuity, and had normal color vision. Nineteen of them (seven in the elder group, 12 in the young group) had seen CAPTCHAs in action. All of them had experience with Internet via personal computers in the past.

2.2. Experiment design

The experiment was a two-by-six mixed factorial design in which the two factors were age group (encoded as AGE) and CAPTCHA type (encoded as TYPE). The factor AGE was a between-subject factor with two levels, young and elder groups. The factor TYPE was a within-subject factor with six levels, each representing a kind of common CAPTCHA distortion, which are summarized in Fig. 2. "Normal Text" was the reference type in which no distortion was applied. "Blot Mask", which is akin to BaffleText [14], used solid blots as masks to distort the image. "Line Mask" used horizontal and vertical white stripes to mask the image, which is similar to ScatterType [15]. "Thread Noise" used irregular threads as noises to distort the image, which is also a common CAPTCHA type. "Global Warp" globally applied warp effect to the image, which is used by Yahoo and Facebook. The last type, "Geometry Noise", which is employed in Wretch (http://www.wretch.cc/), a popular social web site owned by Yahoo Taiwan, used a combination of diagonal lines, arcs, and patterned color background. Each type had a set of ten non-words (thus there were 60 sets of CAPTCHA in the experiment), and each non-word was formed by five lowercased alphanumeric characters whose frequencies of appearing in all the non-words were equal.

2.3. Environment setup and apparatus

The experiment took place in a laboratory in which the illumination level was 500 lux (measured at 1 m in front of the monitor center). A personal computer with mouse, keyboard, and monitor was setup for the experiment. The monitor was a 17 in. liquid crystal display (LCD) set to 1024 ∗ 768 resolution. Participants performed the experiment task in a seated position; the tilt angle of



Normal Text (no distortion; reference)

Blot Mask

Line Mask

Thread Noise

Global Warp

Geometry Noise

Fig. 2. CAPTCHA types used in the experiment.

the monitor and the height of the chair were adjusted to their preferences. The experiment system in charge of presenting CAPTCHA stimuli and of gathering various data was developed with PHP (a programming language) using MySQL (a database management system) as the data store.

2.4. Procedure, tasks, and dependent variables

The experiment was conducted by one experimenter in the settings mentioned above. A consent form with brief description of the study was given to each participant before the experiment could begin. Once the form was signed, the experimenter introduced the procedure to the participant and helped the participant familiarize oneself with the tasks involved. As described before, a participant had to finish six sessions to conclude the experiment. The first session was always the reference type (Normal Type), while the types of the latter five sessions were randomly determined for each participant. The flowchart of a session is depicted in Fig. 3. Before a formal session began, the participant could practice the kind of CAPTCHA stimuli in a practice session. Stimuli used in the practice and formal sessions underwent the same kind of distortion, but the CAPTCHA texts of the two were different.

When a session began, objective visual fatigue data (assessed via CFF, which will be described later) of a participant were taken, after which the participant had to home the mouse cursor to a pair of black crosshairs on the monitor and click on it to make a stimulus appear. Since mouse movement was guided by eye movement, this step was to ensure that all participants had the same initial point for eye gaze and mouse movement. Once the stimulus was visible, the participant had to decide and memorize what the distorted characters were, and then click an on-screen button captioned “Next” to make the stimulus disappear and to make a text field appear in which the answer should be typed. The participant then clicked a submit button to submit the answer. After which, a pair of black crosshairs would show again to usher in the next set of stimulus. A participant had to do ten sets of stimuli to finish a session. Once a session was done, objective visual fatigue (i.e., CFF) and subjective workload (assessed via NASA-TLX) of the participant were taken. The experimenter then showed a summary of the session and asked the participant about the difficulties he or she encountered during the session. After the interview, the participant took a break of two to three minutes. The total length of a session varied between 60 and 150 min.

Unlike the CAPTCHAs seen on most web sites, the stimulus and the text field in the experiment were not visible at the same time. The reason for such arrangement was to exclude the difference of motor response capability (or to be specific, typing) among the participants, and to include the possible effect of distortion on the rehearsing of the recognized characters.

Critical flicker fusion (CFF), also known as flicker fusion threshold or flicker fusion rate, is defined as the threshold frequency at which flickering or steady light stimulus just appears steady or flickering to the eyes of a human beholder. Researchers have been using the change of CFF values to assess visual fatigue in visual display terminal (VDT) tasks [16–18]. This study used the difference between the CFF values measured by a handy flicker (Handy Flicker HF by Neitz Instruments, Japan) before and after a session to assess the visual fatigue objectively. Each CFF value was the average of the CFF readings of turning up (i.e., flickering to steady) and turning down (i.e., steady to flickering) of the flicker frequency. The handy flicker has three LEDs of color red, green, and yellow, and controls to turn up or down the flicker frequency of the LEDs within the range of 1–79 Hz with the accuracy of 0.01%. In this study, the measurement (using green LED) was taken at the distance about 50 cm [19] in front of the handy flicker panel by having the participant turn the frequency dial until the flicker just appeared steady or flickering.

Response time and error rate were also recorded. The former was defined as the time (in millisecond) elapsed between the time when the pair of black crosshair was clicked and the time when the “Next” button was clicked. The latter was defined as the ratio (in percentage) of the number of the incorrect characters to the number of total characters (i.e., 50) in a session.

3. Results

The means and standard deviations (enclosed in parentheses) of the dependent variables are summarized in Table 1 for distortion types, and in Table 2 for distortion types and age groups (elder and young). Generally, Blot Mask and Line Mask are the harder types. The elder group has longer response time, higher error rate, greater CFF difference, and higher NASA-TLX score than the young group does. Because the data do not conform to the assumptions of mixed design analysis of variance (ANOVA), the effects of the factors Type and Age group are analyzed using non-parametric tests. The factor Type, a within-subject factor, is analyzed using Friedman’s test, while the factor Age group, a between-subject factor, is analyzed using Mann–Whitney test. For significant factors, post-hoc analysis is applied to find out where the differences are from.

The test results of the main factors are summarized in Table 3. The dependent variables are all significantly different among the

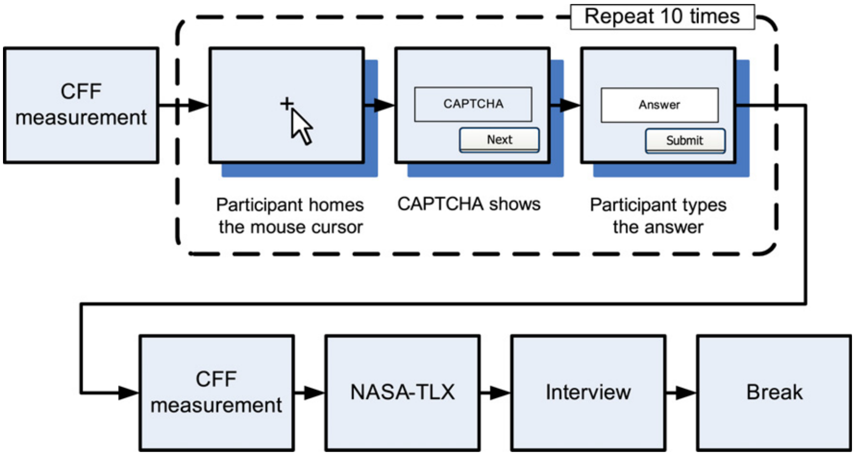


Fig. 3. Experiment procedure of a session. Rectangles with shadows represent computer screens.

Table 1
Descriptive statistics of the dependent variables by types.

Mean (SD)	Response time in ms	Error rate	CFF in Hz	NASA-TLX score
Normal Text	4724.79 (3080.46)	0.03 (0.04)	0.56 (0.56)	15.5 (2.93)
Blot Mask	7312.56 (3816.59)	0.1 (0.06)	2.19 (0.59)	18.46 (2.8)
Line Mask	9354.17 (4359.26)	0.22 (0.07)	2.65 (0.62)	19.88 (2.49)
Thread Noise	4615.89 (2581.83)	0.05 (0.02)	1.15 (0.71)	16.08 (2.7)
Global Warp	4484.76 (2545.74)	0.04 (0.03)	1.44 (0.45)	16.08 (2.78)
Geometry Noise	4398.54 (2400.69)	0.03 (0.03)	1.08 (0.6)	16 (2.9)

Table 2
Descriptive statistics of the dependent variables by types and age groups.

Mean (SD)		Response time in ms	Error rate	CFF in Hz	NASA-TLX score
Normal	Senior	6694.56 (3177.19)	0.04 (0.04)	0.67 (0.69)	16.25 (2.14)
Text	Young	2755.03 (1131.92)	0.02 (0.03)	0.46 (0.4)	14.75 (3.49)
Blot	Senior	9960.46 (3645.76)	0.13 (0.06)	2.38 (0.61)	19.67 (2.31)
Mask	Young	4664.66 (1366.68)	0.07 (0.04)	2 (0.52)	17.25 (2.8)
Line	Senior	11974.85 (4385.73)	0.22 (0.06)	2.67 (0.39)	20.67 (2.02)
Mask	Young	6733.49 (2348.31)	0.23 (0.08)	2.63 (0.8)	19.08 (2.75)
Thread	Senior	6158.93 (2626.14)	0.05 (0.02)	1.29 (0.69)	16.92 (2.11)
Noise	Young	3072.86 (1358.78)	0.04 (0.03)	1 (0.74)	15.25 (3.05)
Global	Senior	6261.75 (2394.85)	0.06 (0.04)	1.54 (0.4)	17.25 (2.22)
Warp	Young	2707.77 (962.26)	0.02 (0.01)	1.33 (0.49)	14.92 (2.87)
Geometry	Senior	6104.62 (2082.41)	0.05 (0.04)	1.21 (0.58)	17.17 (2.69)
Noise	Young	2692.46 (1167.67)	0.01 (0.01)	0.96 (0.62)	14.83 (2.72)

Table 3
Summary of the test of the main factors Type and Age group.

Significance	Response time	Error rate	CFF	NASA-TLX score
Type[a]	**	**	**	**
Age group[b]	**	**	NS	**

NS: Not significant
** $p < 0.01$.
[a] Hypothesis test using Friedman's test.
[b] Hypothesis test using Mann–Whitney test.

six types (marked as $**$ in the table). For age groups, all dependent variables except CFF are significantly different between the two groups. The elder group has significantly longer response time, higher error rate, and higher NASA-TLX score than the young group. The CFF value of the elder, though higher in Table 2, is not significantly higher than the young group.

Post-hoc comparisons of the type pairs for each group are conducted using the critical difference (i.e., the z value) corrected for the number of comparisons being done [20]. The corrected critical difference is 2.24, and the test results for elder and young groups are summarized in Tables 4 and 5, respectively.

Since the current study collected four dependent variables, the correlations between pairs of them are also investigated. The correlation analysis of the dependent variables is summarized in Table 6. The Pearson's correlation coefficients of pairs of dependent variables are of medium to large effect.

4. Discussion

The results of the present study verified that participants of different age groups differ significantly in terms of response time, error rate, visual fatigue, and workload. In addition, CAPTCHA type has significant impacts on the dependent variables. Gender and prior exposure to CAPTCHAs had been tested to ensure that they did not affect the interpretation of the statistical results.

Participants were using three basic steps to perform the experiment task: recognition, rehearsal, and typing. To eliminate the effect of individual differences in typing speed, the response time was defined as the time (in millisecond) elapsed between the time when the pair of black crosshair was clicked and the time when the "Next" button was clicked. When the button was clicked, the stimulus would disappear to give way to a text field in which the answer should be typed. As a result, the typing time was excluded when measuring the response time, leaving only recognition and rehearsal time. As shown in Tables 2 and 3, participants in the elder group spent longer time in recognizing and rehearsing the answers than those in the young group did.

Table 4
Summary of the post-hoc comparison of type pairs for senior group. Types are numbered for brevity's sake, in which Normal Text is number as I, Blot Mask as II, Line Mask as III, Thread Noise as IV, Global Warp as V, and Geometry Noise as VI.

Significance	Response time	Error rate	CFF	NASA-TLX score
I vs. II	NS	*	*	*
I vs. III	*	*	*	*
I vs. IV	NS	NS	NS	NS
I vs. V	NS	NS	NS	NS
I vs. VI	NS	NS	NS	NS
II vs. III	NS	NS	NS	NS
II vs. IV	*	NS	NS	NS
II vs. V	*	NS	NS	NS
II vs. VI	*	*	*	NS
III vs. IV	*	*	*	*
III vs. V	*	*	*	*
III vs. VI	*	*	*	*
IV vs. V	NS	NS	NS	NS
IV vs. VI	NS	NS	NS	NS
V vs. VI	NS	NS	NS	NS

NS: Not significant
* Significant using the corrected critical difference of 2.24.

Most CAPTCHA systems give a user (be it a real human or an automated program) another set of question to answer when an incorrect response is submitted. The length of the intervening time is not considered (except for some technical issues such as session timeout). In real-world applications of CAPTCHA, response time may not be as critical as error rate. In this study, the elder group generally had higher error rate than the young group did.

Blot Mask and Line Mask deserve special attention. For Blot Mask, the elder group has the error rate of 0.13 while the young

Table 5
Summary of the post-hoc comparison of type pairs for young group. Types are numbered for brevity's sake, in which Normal Text is number as I, Blot Mask as II, Line Mask as III, Thread Noise as IV, Global Warp as V, and Geometry Noise as VI.

Significance	Response time	Error rate	CFF	NASA-TLX score
I vs. II	*	NS	*	*
I vs. III	*	*	*	*
I vs. IV	NS	NS	NS	NS
I vs. V	NS	NS	NS	NS
I vs. VI	NS	NS	NS	NS
II vs. III	NS	NS	NS	NS
II vs. IV	NS	NS	*	NS
II vs. V	*	NS	NS	*
II vs. VI	*	*	*	*
III vs. IV	*	*	*	*
III vs. V	*	*	*	*
III vs. VI	*	*	*	*
IV vs. V	NS	NS	NS	NS
IV vs. VI	NS	NS	NS	NS
V vs. VI	NS	NS	NS	NS

NS: Not significant
* Significant using the corrected critical difference of 2.24.

Table 6
Summary of correlation analysis of dependent variables.

Correlation coefficient	Response time	Error rate	CFF	NASA-TLX score
Response time	–	0.656**	0.457**	0.506**
Error rate	–	–	0.584**	0.563**
CFF	–	–	–	0.421**
NASA-TLX score	–	–	–	–

** $p < 0.01$.

group has 0.07; for Line Mask, both groups exhibited high error rates, 0.22 and 0.23 for the elder and young groups, respectively. Blot Mask and Line Mask also had different response time patterns from other types. Post-hoc comparisons revealed that, Line Mask is significantly different from other types for both groups in terms of response time and error rate, as shown in Tables 4 and 5. Blot Mask is also significantly different from other types in most comparisons. The CAPTCHA types considered in the present study can be divided into two groups: the *hard* group being Blot Mask and Line Mask, and the *easy* group being Normal Type, Thread Noise, Global Warp and Geometry Noise. Distinction of these two groups can be linked to a study of the recognition of partially occluded objects by Nakayama et al. [12]. It is easier to recognize the occluded objects when the *occluders* are visible than when they are not. The occluders are interpreted as foreground objects, making the fragments in the background more recognizable. Noises in Thread Noise and Geometry Noise were visible and were interpreted as foreground objects, which made them easier to recognize.

Of special interest is the resemblance of Normal Type to Thread Noise, Global Warp, and Geometry Noise in terms of response time and error rate. Normal Type was the reference type in which no distortion was applied. As effective CAPTHA designs, Thread Noise, Global Warp, and Geometry Noise can also provide good usability comparable to the one without any distortion. Such property makes them very ideal for effective and usable CAPTCHA systems.

CFF differed significantly among the CAPTCHA types, as shown in Table 3. Post-hoc comparisons also reveal a similar grouping mentioned before. Blot Mask and Line Mask placed heavier load to the visual systems of the participants than did Normal Type, Thread Noise, Global Warp, and Geometry Noise. However, CFF did not differ significantly between the two age groups. The two age groups underwent similar load on their visual systems when performing the tasks of different CAPTCHA types.

The significant difference among CAPTCHA types in terms of CFF should be taken into account when designing CAPTCHA systems. The time it takes to answer a CAPTCHA question is within couple minutes (in the present study, participants took about 1 min), and such a short time span should not have incur significant CFF changes [21]. However, when a CAPTCHA design uses improper distortion technique, the visual system of users will be overloaded.

Both Type and Age group are significant factors in terms of NASA-TLX score, as shown in Table 3. Post-hoc comparisons show a similar grouping. There should be correlation among the dependent variables. As Table 6 summarizes, the correlation between response time and CFF, and between response time and NASA-TLX score, are of medium to large effect. It is postulated that the repetitive use of visual and cognitive systems not only lengthens the response time, but also increases the workload and the loading on the visual system.

When designing a CAPTCHA system, one should take into account the basic steps users take during the interaction with such system. A CAPTCHA system may be so designed that the recognition step is as easy as an undistorted one for human users; Thread Noise, Global Warp, and Geometry Noise are examples of such design. Yet, not only distortion technique, but also the character set, that can impact the recognition and the rehearsal steps. Obviously, *shape ambiguity* may play a role in the recognition step; for example, 1 (number one) vs. l (lowercased alphabet L), and b vs. p. On the other hand, *phonetic ambiguity* plays an even more important role in the rehearsal step and motor response step. In the present study, participants rehearsed the recognized alphanumeric characters either mentally or aloud. When characters being rehearsed had phonetically similar counterparts, participants were prone to err. One should also consider the native language of the target users when dealing with the ambiguity problem. For example, the participants in the present study spoke Chinese and learned English as a second language. When reading a string of alphanumeric characters, the participants spoke the alphabets in English, but the numbers in Chinese. Yet, the number *one* in Chinese sounds like *e* in English (the long e vowel). Some participants recalled during the post-session interviews that they recognized the characters correctly, but got confused when they were rehearsing the answers. The ambiguity issues have to be considered when using alphanumeric characters in CAPTCHA systems since most major web sites have global reach.

5. Conclusion

This study conducted an experiment to study the effect of age group and distortion type on CAPTCHA task performance, objective visual fatigue, and subjective workload. Two age groups and six distortion types were considered. Distortion type was found to be a significant factor in terms of response time, error rate, CFF, and NASA-TLX score. Age group also had significant effect on response time, error rate, NASA-TLX score. CFF values of the two groups were not significantly different. Post-hoc comparison of type pairs for two age groups revealed that, Blot Mask and Line Mask were in a homogeneous subset in terms of the dependent variables, while others were in another subset. Since Normal Type was the reference type (i.e., no distortion), Thread Noise, Global Warp, and Geometry Noise were ideal choices for usable and secure CAPTCHA distortion techniques. Correlation analysis also revealed that the correlation between pairs of the dependent variables were of medium to large effect. Increased response time is positively correlated with error rate, CFF, and NASA-TLX.

The results of this study have implications for the design of CAPTCHA systems. To counter the automated attacks of computer programs, the security of CAPTCHA systems relies on image distortion techniques and the use of non-words to test the fact of a remote user being human. Yet, some distortion techniques are hard for automated computer programs as well as for real human beings. It is then important to verify the usability of a CAPTCHA design before bringing it to the public. Elder users may take longer time and have higher error rate in responding to a CAPTCHA question. Such situation may contribute to increased workload and overloading of the visual system. Designers of CAPTCHA systems should choose a distortion technique that is both secure and user-friendly to users of different ages. The choice of character set used in a CAPTCHA system is also important when the user base is multinational. The process of a user solving a CAPTCHA question is composed of three general steps: recognition, rehearsal, and motor response. In the recognition step, users rely on visual stimuli. In the rehearsal step, phonetic coding is used either mentally or verbally. The choice of character set should avoid ambiguity for these steps. For example, as pointed out in this research, letter *e* and number *1* have similar pronunciation in Chinese, which are easily confused during mental or verbal rehearsal.

The distinction of the two groups of CAPTCHAs in this study provides useful implications for future CAPTCHA designs. As pointed out in Discussion, researchers found that it is easier to recognize the occluded objects when the occluders are visible than when they are not, which is attributable to the fact that the occluders are interpreted as foreground objects, making the fragments in the background more recognizable. In this study, Blot Mask had occluders that were not coherent objects, while Line Mask had occluders that were not visible. However, both Thread Noise and Geometry Noise had occluders that were visible. And the findings of the current study showed that the former two types were harder than the latter two. It is suggested that CAPCHAs can use coherent and visible noise to challenge machines while remain friendly to human cognition system.

While this study focused on text-based CAPTCHAs, exploration of other types of CAPTCHA is also worthwhile, such as auditory and pictorial ones. Gender difference, learning curve, and the interaction between users native language and the character set used in a CAPTCHA are also important issues. Further studies can be conducted in these directions.

References

[1] L. von Ahn, et al., CAPTCHA: Using Hard AI Problems for Security, in: E. Biham (Ed.), Advances in Cryptology – EUROCRYPT, 2003, pp. 294–311.

[2] K. Chellapilla, et al., Designing human friendly human interaction proofs, presented at the ACM CHI'05, 2005.

[3] J. Yan, A.S.E. Ahmad, Usability of CAPTCHAs or usability issues in CAPTCHA design, Presented at the 4th Symposium on Usable Privacy and Security, Pittsburgh, Pennsylvania, 2008.

[4] A. Kolupaev, J. Ogijenko, CAPTCHAs: Humans vs. bots, Security and Privacy, IEEE 6 (2008) 68–70.

[5] L.v. Ahn et al., reCAPTCHA: human-based character recognition via web security measures, Science 321 (2008) 1465–1468.

[6] K. Chellapilla, et al., Building segmentation based human-friendly human interaction proofs, Presented at the 2nd International Workshop on Human Interaction Proofs, 2005.

[7] J. Elson, et al., Asirra: A CAPTCHA that exploits interest-aligned manual image categorization, Presented at the 14th ACM Conference on Computer and Communication Security, Alexandria, VA, USA, 2007.

[8] U. Neisser, Cognitive psychology, Prentice Hall, Englewood Cliffs, NJ, 1967.

[9] D. Broadbent, M.H. Broadbent, Priming and the passive/active model of word recognition, in: R. Nikerson (Ed.), Attention and performance, vol. VIII, Academic Press, New York, 1980.

[10] J.L. McClelland, J.C. Johnston, The role of familiar units in perception of words and nonwords, Perception and Psychophysics 22 (1977) 249–261.

[11] J.C. Johnston, J.L. McClelland, Visual factors in word perception, Perception and Psychophysics 14 (1973) 365–370.

[12] K. Nakayama et al., Stereoscopic depth: its relation to image segmentation, grouping and the recognition of occluded objects, Perception 18 (1989) 55–68.

[13] M. Karavidas et al., The effects of computers on older adult users, Computers in Human Behavior 21 (2005) 697–711.

[14] M. Chew, H.S. Baird, BaffleText: a Human interactive proof, in: Proceedings of 10th SPIE/IS&T Document Recognition and Retrieval Conference, vol. 5010, 2003, pp. 305–316.

[15] H.S. Baird, T.P. Riopka, Scattertype: a reading CAPTCHA resistant to segmentation attack, Proceedings of SPIE 5767 (2005) 197–201.

[16] T. Iwasaki et al., The changes in colour critical flicker fusion (CFF) values and accommodation times during experimental repetitive tasks with CRT display screens, Ergonomics 32 (1989) 293–305.

[17] K.H.E. Kroemer, E. Grandjean, Fitting the Task to the Human, fifth ed., Taylor & Francis, London, 1997.

[18] T. Iwasaki, S. Akiya, The significance of changes in CFF values during performance on a VDT-based visual task, in: M. Kumashiro, E.D. Megaw (Eds.), Towards Human Work: Solutions to Problems in Occupational Health and Safety, Taylor & Francis, London, 1991.

[19] T. Hosokawa et al., Basic study of the portable fatigue meter: effects of illustration, distance from eyes and age, Ergonomics 40 (1997) 887–894.

[20] S. Siegel, N.J. Castellan, Nonparametric statistics for the behavioral sciences, second ed., McGraw-Hill, New York, 1988.

[21] H.-C. Wu et al., Ergonomic evaluation of three popular Chinese e-book displays for prolonged reading, International Journal of Industrial Ergonomics 37 (2007) 761–770.