The Influence of the CAPTCHA Types to Its Solving Times

D. Brodić*, S. Petrovska *, M. Jevtić* and Z. N. Milivojević**
* University of Belgrade, Technical Faculty in Bor, V.J. 12, 19210 Bor, Serbia
* College of Applied Technical Sciences, Aleksandra Medvedeva 20, 18000 Niš, Serbia
dbrodic@tf.bor.ac.rs

Abstract - In this paper, we present a survey of different CAPTCHA types. Then we address the problem of the CAPTCHA userability. In the experiment, the Internet user response to solve different types of CAPTCHA is tested. The obtained results are given and statistically processed. They are discussed leading to the conclusion different types of CAPTCHA usability and suitability for Internet users.

I. INTRODUCTION

CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) is a program created in order to differentiate between humans and bots during the logging to the website [1]. In this sense, the bot is a software robot, which tries to emulate human users. Hence, it includes elements of artificial intelligence as well as ability of automated reasoning. In fact, CAPTCHA represents a test program which gives a task to be solved. If the correct answer is obtained, then the program classifies the user as a human. The aim of the program is to stop the attacks of bots. Today's research on CAPTCHA is focusing on the development of the test program, which will be easily solved by people and represents a heavy problem to bots. The reasons for using CAPTCHA are as follows [2]:

- Prevention of spams on forums.
- Prevention to open a large number of orders by users on sites that offer free services like Gmail, etc.
- User accounts protection from attacks through which bots are discovering user passwords.
- Validity of online surveys by determining whether the humans or bots answering the questionnaire.

To develop a CAPTCHA that incorporates a high level of security, it has to meet the following requirements [3]:

- The solution must not be conditional, which depends on the user's language and age. It means that it should be intuitive.
- It has to be hard to solve the CAPTCHA test except for the humans in order to differentiate humans from bots.
- It has to be created to not disturb the user privacy. Hence, it has to be not user related.

First CAPTCHA was designed by Broder's team in 1997 for Altavista, to prevent automatic adding URL to a database of a web browser [4]. CAPTCHA may be based on the [5]: (i) Image elements, (ii) Text elements, and (iii) Audio and video elements.

Text-based CAPTCHA is the most common form. This type of CAPTCHA asks the user to decrypt the text which is usually distorted in some way [6]. Unfortunately, this type of CAPTCHA can be successfully attacked by bot due to the existence of good decoders. Figure 1 shows the example of the text-based CAPTCHA.

Image-based CAPTCHA is usually considered as the most advanced and safest one. This type of CAPTCHA requires users to find a desired image between the list of images. Because it is based on image details, it represents an almost impossible task to be solved by bots. Figure 2 shows an example of the image-based CAPTCHA.

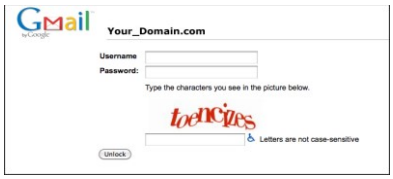


Figure 1. An example of text-based CAPTCHA

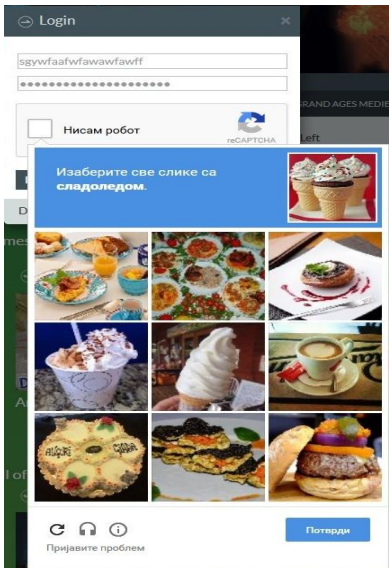


Figure 2. An example of image-based CAPTCHA

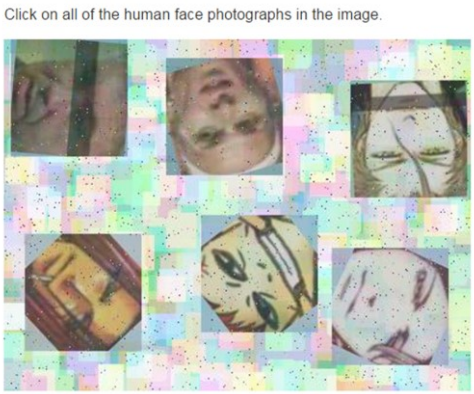


Figure 3. An example of FaceDCAPTCHA

Although, CAPTCHA protects user accounts and passwords, it often represents a firm obstacle not just the bots, but to the humans too. As an extension to image-based CAPTCHA the FaceDCAPTCHA is used [7]. It is a CAPTCHA that incorporates the elements of a face detection. It is one of the newer CAPTCHA types that includes a high level of security. It exploits a research about the human brain, which is very effective in the process of natural face segmentation in spite of used complex backgrounds. Figure 3 shows an example of the FaceDCAPTCHA.

Video and audio-based CAPTCHA refers to the auditory reproducible characters that the user have to input. Although, this type of CAPTCHA is typically attacked in approximately 70% of cases, its development and innovation is essential for the blind users [8].

All articles about CAPTCHA have researched the safety and security standpoint ignoring the difficulties of users to solve its task. In this paper, we explore the complexity of CAPTCHA tasks from the user's viewpoint. Hence, we conducted the experiment based on four different types of CAPTCHA, which are tested in the community of 100 Internet users differentiated by the level of Internet use and gender. The aim of our research was to identify the following:

- The user's response rate of solving different types of CAPTCHA.
- The suitability of different types of CAPTCHA to the certain group of Internet users.

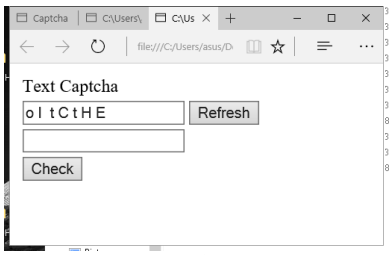
As the final results, we suggest the use of certain types of CAPTCHA and their implementation in different software environment.

The paper is organized as follows. Section 2 describes the elements of the experiment. Section 3 presents the results and discusses them. Section 4 gives the conclusions.

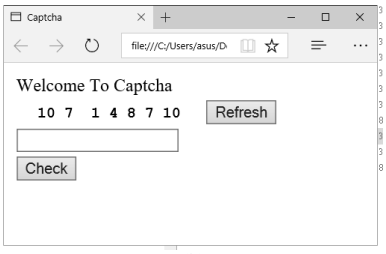
II. EXPERIMENT

The experiment is performed on four different CAPTCHAs. Figure 4 illustrates these four CAPTCHAs.

The aim was to solve these types of CAPTCHAs and to measure the user's response rate (in seconds) to successfully solve the task of each CAPTCHA. In this



(a)



(b)

(c)

(d)

Figure 4. Four different CAPTCHAs: (a) Text based CAPTCHA, (b) Text-number based CAPTCHA, (c) Image-based CAPTCHA – Animals in Wild, (d) Image-based CAPTCHA – Picture of CAPTCHA

sense, the response rate of each user for each of four CAPTCHAs represents the dependent variable. It is given depending on the Internet years of use as well as the gender. The experiment is carried out on the sample of 100 Internet users. The Internet users are classified according to: (i) years of Internet use (from 1 to 9 years) and (ii) gender (male of female). The population has 50 female and 50 male users.

Furthermore, the tested population is chosen to have a Gaussian distribution. Figure 5 shows the distribution of testing population, according to the years of Internet use.

Three hypotheses are in the focus of our experiment:

- Hypotheses 1: Subjects who use the Internet more time will have a shorter response rate.
- Hypotheses 2: Faster response rate is presumable for image-based compared to text-based CAPTCHA.
- Hypothesis 3: Does the gender have any influence on solving the CAPTCHA?

III. RESULTS AND DISCUSSION

First of all, it is important to know that all users successfully solve all four CAPTCHAs. However, the

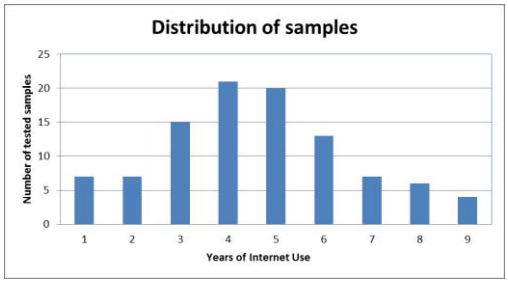


Figure 5. The distribution of testing population, according to the years of Internet use

time of solving CAPTCHA is quite different between them. Figures 6-7 show the results representing the time to solve the CAPTCHA in response to the years of Internet use or gender, respectively.

The measures that characterize each CAPTCHA are: (i) Minimum time to solve a CAPTCHA, (ii) Maximum time to solve a CAPTCHA, (iii) Mean time to solve a CAPTHA, (iv) Standard deviation, and (v) Correlation coefficient *R*. The statistical measures like standard deviation, variance and correlation coefficient are defined respectively as follows [9]:

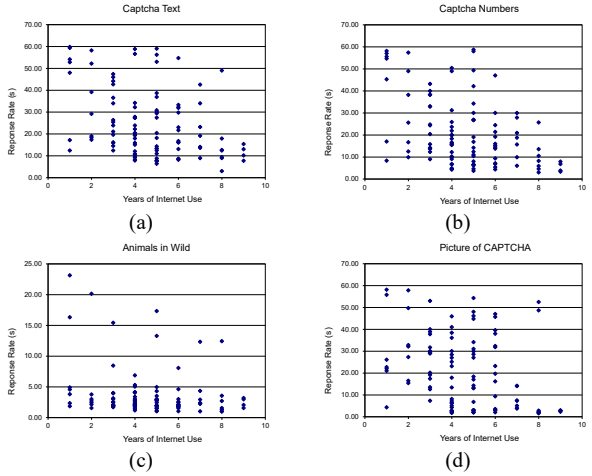


Figure 6. The distribution of the CAPTCHA solving times in accordance to the years of Internet use

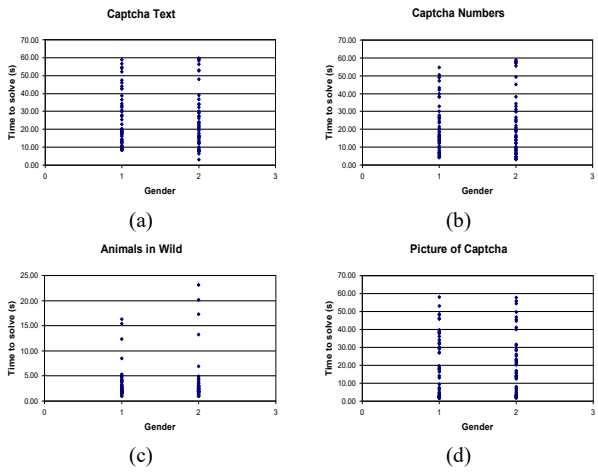


Figure 7. The distribution of the CAPTCHA solving times in accordance to the gender

$$\sigma = SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}\ , \tag{1}$$

$$R = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}\ . \tag{2}$$

The correlation coefficient *R* expresses the strength and direction of a linear relationship between two variables *x* and *y*, where x_i is the value of the variable *x* for the instance *i*, $\bar{x}$ is the mean of *x* value, y_i the value of the variable *y* for the instance *i*, and $\bar{y}$ is the mean of *y* value [9]. Consequently, *R* measures the strength and direction of a linear relationship between two variables on a scatterplot [10]. It can receive the value from -1 to +1. The larger the absolute value of the coefficient, the stronger the relationship between the variables. An absolute value of 1 indicates a perfect linear relationship. A correlation close to 0 indicates no linear relationship between the variables. If *R* is positive, then the two variables tend to increase or decrease together. In contrast, if *R* is negative, then one variable increases as the other decreases.

The experimental result of these measures is given in Tables I-II. Figure 8 shows the level of the correlation coefficient *R* in accordance to the years of Internet use or gender of each examinee.

TABLE I. SOLVING CAPTCHA VS. INTERNET USE

Measures	Internet use	Text	Number	Animals in Wild	Picture of CAPTCHA
Population	100	100	100	100	100
Minimum	1	3.00 s	3.00 s	1.00 s	1.56 s
Maximum	9	59.78 s	58.71 s	23.14 s	58.14 s
Mean	-	24.71 s	21.40 s	3.61 s	20.78 s
SD	-	15.56	15.68	3.94	16.63
R(Internet use)	-	-0.42	-0.45	-0.27	-0.34

TABLE II. SOLVING CAPTCHA VS. GENDER (1-MALE, 2-FEMALE)

Measures	Gender	Text	Number	Animals in Wild	Picture of CAPTCHA
Population	100	100	100	100	100
Minimum	1-male	8.28 s	4.30 s	1.00 s	1.56 s
Maximum	1-male	58.80 s	54.67 s	16.32 s	58.14 s
Minimum	2-female	3.00 s	3.00 s	1.00 s	1.99 s
Maximum	2-female	59.78 s	58.71 s	23.14 s	57.81 s
Mean	1-male	25.09 s	21.28 s	3.48 s	21.85 s
Mean	2-female	25.01 s	22.20 s	3.83 s	19.74 s
R(Gender)	-	0.00	0.03	0.04	-0.01

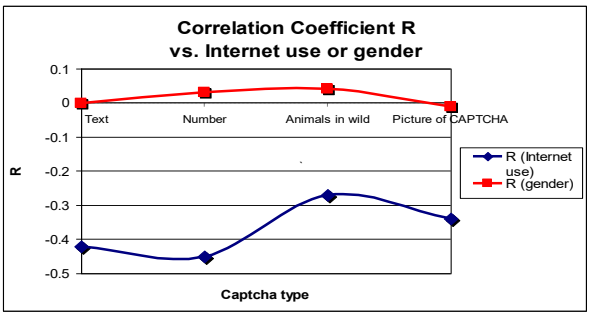


Figure 8. Correlation coefficient *R* for all four CAPTCHAs in accordance to: (a) years of Internet use, (b) gender

In our case, the *R* is negative for all three CAPTCHA. It means that if the users have more years of Internet experience, then CAPTCHA will be solved in shorter time. However, the *R* is -0.42 for text based CAPTCHA and -0.45 for number based CAPTCHA. It can be qualified as a moderate downhill or negative relationship. On the contrary, the image based CAPTCHA has *R* equal to -0.27, which represents a weak downhill or negative linear relationship. From the given *R* values, it is obvious that hypotheses 1 is confirmed.

If we take into account aforementioned hypothesis about CAPTCHA that its solution must not be conditional and it should be intuitive, then image based CAPTCHA is better qualified for the right, i.e. ideal CAPTCHA choice.

The obtained time to successfully solve CAPTCHA is as follows: (i) Text based CAPTCHA from 3.00 s to 59.78 s, (ii) Number based CAPTCHA from 3.00 s to 58.71 s, (iii) Image based CAPTCHA (Animals in the wild) from 1.00 s to 23.14 s, and (iv) Image based CAPTCHA (Picture of CAPTCHA) from 1.56 s to 58.14 s. The mean time of solving CAPTCHA brings great advantage of image based CAPTCHA, but only if the CAPTCHA image is carefully chosen. It means that image in CAPTCHA should be clearly defined and chosen. Hence, image based CAPTCHA (Animals in the wild) receives 3.61 s compared to 24.71 s and 21.40 s for text and number based CAPTCHA, respectively. However, badly chosen image CAPTCHA receives 20.78 s. It is obvious that image based CAPTCHA has clear advantages compared to text or number based CAPTCHA, if it is chosen adequately. In this way, hypotheses 2 were also confirmed. Furthermore, it is clear that solving time of CAPTCHA in accordance to gender choice are quite similar with the margin inside 10%. Hence, hypotheses 3 were not confirmed, because there is no evident difference between solving time of a male or female. It is also confirmed with correlation coefficient *R*, which receive the values between -0.01 and 0.04.

In some application, the time to solve a CAPTCHA can represent a critical value. Hence, the quick solving an easily differentiate humans from bots. Also, image-based CAPTCHA is more universal, because it is more appropriate to different types of electronic devices like computer, tablet or smartphone. From all aforementioned, the results of the experiment and comparison of four types of CAPTCHA give the clear winner in the competition of three different CAPTCHA. Image based CAPTCHA has a clear advantage in the time of its successfully solving. Furthermore, it is more intuitive and less conditional to be solved compared to other two tested CAPTCHA.

IV. CONCLUSION

The paper described the research conducted to evaluate different CAPTCHA according to the experience of Internet users (years of Internet use) and gender (male or female). The result of the experiment showed that users more easily solve (in less time) image-based CAPTCHA compared to text or number based CAPTCHA. Furthermore, the value of correlation coefficient *R* (negative value) proved that the years of Internet use can help users to solve CAPTCHA in a less time. However, one of the CAPTCHA postulate is that CAPTCHA should be intuitive. Using this premise, the ideal model of CAPTCHA can be equally easily solved by experience and inexperienced user. Because the correlation coefficient *R* of the text or number based CAPTCHA is approximately 50% higher than in the image based CAPTCHA, it proved that the image-based CAPTCHA is much less dependent of the user experience with the Internet. Hence, image based CAPTCHA better satisfied the aforementioned premise. The study also proved that there is small difference in CAPTCHA's solving time between male or female users.

Further research will be toward inclusion of higher number of different CAPTCHA's samples in the experiment as well as the exploration of education level and age of users as a parameter, too.

REFERENCES

[1] L. von Ahn, M. Blum, J. Langford, "Telling Humans and Computers Apart Automatically", Communications of the ACM, vol. 47, no. 2, pp. 56–60, 2004.

[2] www.google.com,http://en.wikipedia.org/wiki/CA PTCHA

[3] L. von Ahn, M. Blum, N. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security", Proceedings of Eurocrypt, Warsaw, Poland, May 4-8, 2003, pp. 294-311

[4] M. Lillibridge, M. Abadi, K. Bharat, A. Broder, "Method for Selectively Restricting Access to Computer Systems", United States Patent 6195698. Applied 1998 and Approved 2001.

[5] K. Kukade1, M.S. Deshmukh, "CAPTCHA Problems on AI Based For Protection from Automated Hacking Tool", International Journal of Computer Science and Mobile Computing, vol.4, no. 5, 2015, pp. 714-719.

[6] X. Ling-Zi and Z. Yi-Chun, "A Case Study of Text-Based CAPTCHA Attacks", Proceedings of International Conference on Cyber Enabled Distributed Computing and Knowledge Discover, Sanya, China, 2012, pp. 121-124.

[7] G. Goswami, B. M. Powell, M. Vatsa, R.a Singh, A. Noore, "FaceDCAPTCHA: Face detection based color image CAPTCHA", Future Generation Computer Systems, vol. 31, no. 2, 2014, pp. 59-68.

[8] H. Gao, H. Liu, Dan Yao, X. Liu, "An audio CAPTCHA to distinguish humans from computers," Proceedings of IEEE Symposium on Security Privacy, Oakland, U.S.A., May 16-19, 2010, pp. 399–413.

[9] J. Higgins, The Radical Statistician: A Beginners Guide to Unleashing the Power of Applied Statistics in The Real World (5th Ed.), Jim Higgins Publishing, 2006.

[10] Chi Yau, R Tutorial with Bayesian Statistics Using OpenBUGS, Kindle Book, 2015.