ISSN 2278-3091
Volume 10, No.3, May - June 2021
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse891032021.pdf
https://doi.org/10.30534/ijatcse/2021/901032021

CAPTCHA Design: A Novel Security Method using Sindhi Language

Abdullah Maitlo[1], Riaz Ahmed Shaikh[2], Haque Nawaz[3], Asad Hameed Soomro[4], Samar Abbas Mangi[5], Inayatullah Soomro[6]

[1]Department of Computer Science, Shah Abdul Latif University, Khairpur Mirs, Sindh, Pakistan, abdullah.maitlo@salu.edu.pk

[2]Department of Computer Science, Shah Abdul Latif University, Khairpur Mirs, Sindh, Pakistan, raiz.shaikh@salu.edu.pk

[3]Department of Computer Science, Sindh Madressatul Islam University, Karachi, Sindh, Pakistan, hnlashari@smiu.edu.pk

[4]Department of Computer Science, Shah Abdul Latif University, Khairpur, Sindh, Pakistan, asad.soomro31@yahoo.com

[5]Department of Computer Science, Shah Abdul Latif University, Khairpur Mirs, Sindh, Pakistan, mangisamar@gmail.com

[6] Department of Mathematics, Shah Abdul Latif University, Khairpur Mirs, Sindh, Pakistan, inayat.soomro@salu.edu.pk

ABSTRACT

CAPTCHA is one of the most widely used security on internet now a day's. This CAPTCHA test is used against the Bots that creep into website and perform automatic registration. Among different type of CAPTCHAs one of oldest and easiest method is text based. Text based CAPTCHAs are available in the market in different languages i.e. English, Arabic, Urdu. This paper focuses on design of text based CAPTCHA as security. The proposed design of the CAPTCHA uses Sindhi which is an Arabic scripting language named as CaSN i.e CAPTCHA as Sindhi Numbers. CaSN focuses on the Sindhi Language numbers to generate a CAPTCHA image for the users. An algorithm is used to generate the CaSN image. In this regard, the evaluation was performed on 290 participants from local areas of Sindh, Pakistan. The study produced prominent results after performing CaSN test among participants of the study. The study was limited among local areas of Sindh, Pakistan. The limitation of the study is to provide a new context in CAPTCHA especially for Sindhi speaking people.

Key words: CAPTCHA, Security Method, Sindhi Language

1. INTRODUCTION

Now a days secured online communication has become major issue [1] and Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) have been utilized for secure access to various websites and search engines since long time. The CAPTCHA test has been conducted to differentiate between computers and humans. CAPTCHA is become an essential part of the web over the years for preventing internet services from the bots. These bots create spams, falsified reports, Denail of Service (DoS) attacks, false online polling etc.[2]. Initially, the first CAPTCHA was projected over the search engine named as AltaVista to prevent the auto-form submission from bots. It was the first successful attempt for CAPTCHA implementation and it was report that about 95% of the spams were prevented [3]. DEC Systems Research Center invented such a secure mechanism and claimed its copy rights as a system that randomly approve the accessibility requests from users for a server [4] [5]. CAPTCHA is one of the primary part on the Internet which has been used by some high rating International companies Amazon, eBay, Facebook, Yahoo. There are different types of CAPTCHAs available such as text-based, image-based as shown in figure 1, audio CAPTCHAs have been created instead of visual CAPTCHAs for clients who are visually impaired or outwardly weakened as shown in figure 2, video-based as figure 3 depicts, 3D-based [6], Math [7], game-based [8], and Social-authentications [9]. Among all these types one of oldest and easiest type of CAPTCHA is text-based.

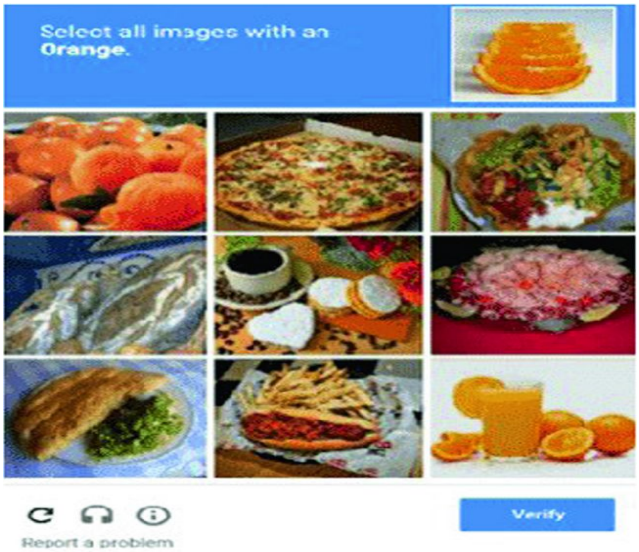

Figure 1: Example of Audio based CAPTCHA [10]

Figure 2: Example of audio based CAPTCHA [11]

Figure 3: Example of audio based CAPTCHA [12]

Text-based CAPTCHA holds the character sequence in the pictorial form along with some kind of distortion and/or noise (i.e. lines, dots). The text and noise can be embedded together in such a manner that human user can easily understand and recognize it and the bot or computer could not identify the text. The reason behind its ubiquitous usage is because of its easy perception among human users. A novice level of user can also solve the text-based CAPTCHA easily and in accurate manner among other types of CAPTCHAs. Likewise, the brute-force attack will take more time because of large search space, and also higher the computing cost in the text based CAPTCHA [2] [13]. Text-based CAPTCHAs are available in the market in different languages i.e. English, Arabic, Urdu. This paper focuses on design of text-based CAPTCHA as security. The proposed design of the CAPTCHA uses Sindhi an Arabic scripting language.

Sindhi opted from Sindhu word, which is prehistoric name of the Indus River. Over 30 million Sindhi speaking people are living around the globe having third in number among spoken languages of India and Pakistan. In addition, here is huge number of Sindhi speaking persons the United States and the United Kingdom. Sindhi language has its own structure which makes it adaptable with such potential to grow and expand itself to satisfy the need of modern times. The utilization of the Sindhi language is found in all educational, social and communication platforms in Sindh province of Pakistan, such as, in educational institutes, history, official working, social-media and also in religious studies.

Proof for Sindhi as a composed language dates to a Sindhi interpretation of the Islamic Qur'an in 883 A.D., followed a century later by a Persian interpretation of the old Indian religious epic Mahabharata taken from a language thought to be Old Sindhi. The initial point of existing Sindhi Abjad utilizes to write Urdu is a form of Perso-Arabic writing, under British impact in 1852 the equivalent was received to compose Sindhi. In India Sindhi is additionally composed with the Devanagri script [14] [20].

Both Arabic and Sindhi utilize same composing framework for example Semitic Abjad that addresses consonants in addition to certain vowels, and use Naskh way of writing in which all content is composed on a fanciful even line called benchmark. The two dialects are cursive in nature in which letters are associated with one another in sub-words on the gauge. This is like Latin 'signed up' handwriting, which is likewise cursive. Arabic has also Sindhi characters that can be joined in both hand-written and printed text.

Sindhi has a sum of 52 letters, increasing the Persian with digraphs and eighteen new letters (ڄ ٺ ٽ ٿ ڀ ٻ ڙ ڍ ڊ ڏ ڌ ڇ ڃ ڦ ڻ ڳ ڱ ڪ) for sounds specific to Sindhi and other Indo-Aryan dialects. A few letters that are recognized in Arabic or Persian are homophones in Sindhi. In addition, Sindhi language numbers are also written same as other scripting languages. The Sindhi Numbers with Indo-Aryan style has been shown in table 1. The focus of the paper is on the design of the Sindhi Language number embedded CAPTCHA image for the Sindhi speaking users named as CaSN (CAPTCHA as Sindhi Numbers).

The structure of paper is as follows. Section II will provide literature review regarding previous studies, section III holds the proposed design of CaSN, section IV contains the evaluation of the paper, Section V has the evaluation results and findings regarding the proposed design and in the last, section VI conclusion and projected future research.

Table 1: Sindhi Language numbers with Indo-Aryan style

	Numeral	style
0	٠/٥	
1	٠١-Jan	پهريون
2	٠٢-Feb	ٻيون
3	٠٣-Mar	ٽيون
4	٠٤-Apr	چوٿون
5	٠٥-May	پنجون
6	٠٦-Jun	ڇهون
7	٠٧-Jul	ستون
8	٠٨-Aug	اٺون
9	٠٩-Sep	نائون

2. LITERATURE REVIEW

The possibility of CAPTCHA was first essentially carried out by AltaVista to keep computerized bots from naturally enrolling the sites [5]. The instrument behind the thought was to produce somewhat mutilated characters and to introduce it to the client. In [15], proposed the strategy for implanting numbers in text CAPTCHAs. They express that the labeled numbers in the content confounds the OCR and make it incapable to section the characters.

Another study by [16], proposed a technique in which the client needs to choose the picture of a feline from the assortment of pictures of felines and canines. For the picture assortment they utilized the data set of in excess of 3 million photographs from a site used to discover homes for destitute pets, i.e., petfinder.com. Ahn gave mechanized Turing test that did not depend on the trouble of the OCR, rather on hard man-made brainpower issues [17]. Their proposed strategy utilizes hard AI issues to build CAPTCHAs to separate people from PCs.

Despite the fact that, getting web from mechanized bots has been given a solid consideration in the examination local area, notwithstanding, the accentuation is for the most part on the CAPTCHAs utilizing Sindhi content which has numerous inborn shortcomings in it. In this manner, we consider our work an importance progress in the field of security utilizing CAPTCHAs dependent on Sindhi number. This paper presents a CAPTCHA conspire dependent on the Sindhi Numbers. The Sindhi numbers are introduced as picture and misshaped with various methods so it is unbreakable by OCRs, yet effectively decipherable by peoples

3. PROPOSED DESIGN

There are various methods existed to provide security using CAPTCHA. These existing methods utilize several dissimilar tactics to build CAPTCHA as security (i.e image, video, text). We present text based CAPTCHA based on Sindhi Language text. This type of CAPTCHA holds the character of sequence in the pictorial form along with some kind of distortion and/or noise (i.e lines, dots) for making hard for OCR programs to recognizing the text. Likewise, the brute-force attack will take more time because of large search space, and also higher the computing cost in the text based CAPTCHA [18].

The concept behind the paper is CaSN (CAPTCHA as Sindhi Numbers). CaSN is totally based on image including Sindhi numbers and lines as distortions. The user will be presented with an image in which Sindhi numbers will be embedded along with cursive lines using rand function to make hard for OCR programs to recognize the numbers. These numbers will be selected randomly from 0-9 for CAPTCHA image generation. Figure 4, shows the generated image of CaSN. The user after recognizing the numbers will write the numbers in the number-box and click verify. The numbers written in the number-box by the user will be verified by the system. After verification, the user will be allowed or denied for further proceedings (i.e login, form submission). If the numbers hold by number-box is detected wrong by the system a new CaSN image will be provided for the user to attempt again.

Figure 4: Generated CaSN image

In this manner, a new series of numbers will be selected randomly from 0-9 Sindhi Numbers after first attempt of failure. These letters will be shuffled and redesigned for the user and Re-CAPTCHA will be performed. In addition, lines will also be shuffled and presented using rand function. Total three attempts will be available for the user and after these attempts the system will be blocked for 2 minutes for security reasons to stop any brute-force attack and to reduce computational cost. CaSN is a novel CAPTCHA that will be implemented using Sindhi Numbers. This is one of the first step towards Sindhi number based CAPTCHA to provide security to the user.

The algorithm used to generate new image is as follows:

3.1 Algorithm:-

Step 1: Start

Step 2: Image will be generated of 200x200 pixels

Step 3: Distortion of lines will be created using rand function

Step 4: Sindhi Numbers will be incorporated in the created image

Step 5: User recognizes the numbers

Step 6: User write the CAPTCHA numbers in the number-box

Step 7: Click Verify

Step 8: Is written numbers are correct or not?

If

Correct

"Successful"

Wrong

Count ++

If

Count == 3

Block User using Cookies

Start timer for 2 minutes

If

Timer == 0

Unblock User

Go to Step No. 2

Go to **Step No. 2**

Step 9: End

4. IMPLEMENTATION AND RESULT ANALYSIS

The projected algorithm has been evaluated using Php (web programming) language as discussed in section 3.1 in detail. The proficiency of the algorithm can be measured in two terms i.e. readability and strength. Strength can be measured when OCR programs are unable to recognize the text of the CAPTCHA. In the Contrast if the text is not enough readable by the user then it is useless. In this manner, it is important for CAPTCHA to be human readable. An observation regarding the readability has been perceived that the high complexity decreases the readability of the CAPTCHA [19]. In this section, the readability of the CAPTCHA has been measured and discussed.

The evaluation was performed for which 290 participants were selected for analyzing the readability of CaSN. These participants were from local areas of Sindh and frequent in speaking and writing Sindhi Language. Further, the selection was made according to the user classification i.e. Novice, Knowledgeable, and Expert. Among 290 participants 160 were of Novice level. These participants were new users to solve CAPTCHAs. The remaining 130 were of expert level and vast knowledge about CAPTCHA and its functionalities.

5. RESULTS DISCUSSION

The evaluation was analyzed in two different perspectives to analyze the readability of CaSN. First in terms of time and second in terms of accuracy. As Figure 5, depicts that time taken to solve the CaSN by novice and expert type participants has little variation. Average time taken by the expert type users is 4.4 seconds. These participants took less amount of time because of their experience to solve different type of CAPTCHAs. On the other hand, the Novice users took 6.2 seconds on an average to solve the CaSN. These novice level users were of no experience to solve CAPTCHA but these novice level participants were from local areas of Sindh, Pakistan and very fluent in reading and writing Sindhi Language, which makes them easy to solve CAPTCHA test.

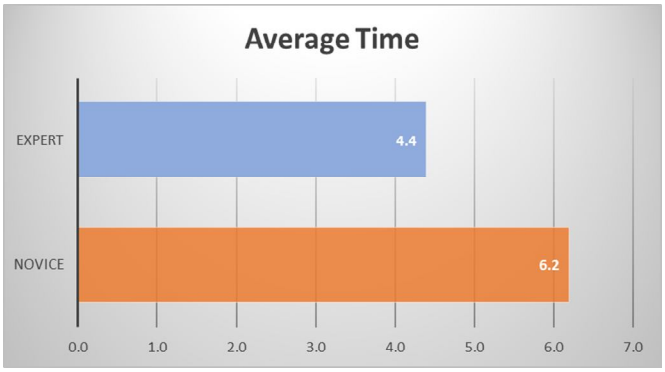

Figure 5: Average time taken by the participants

The accuracy of the CaSN were also analyzed during evaluation. Figure 6, depicts the results of the accuracy to solve the CaSN by participants. The figure clearly shows that average accuracy by both level of participants has very slight variation. It was observed that the negligible variation of the accuracy was because of the Language used in the CAPTCHA. The CaSN was easily recognized by both level of participants

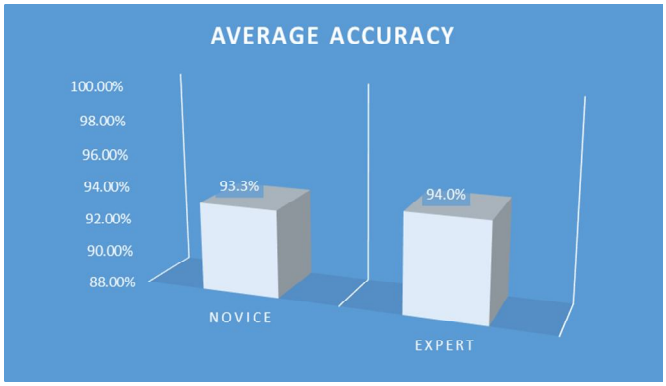

Figure 6: Average accuracy rate to solve CaSN

6. CONCLUSION

In this paper, a novel CAPTCHA method CaSN is proposed. The method uses Sindhi Language to generate a CAPTCHA image. The image is distorted by adding noises to the background in the form of lines. The evaluation was performed in which 290 participants took part of novice and expert level from local areas of Sindh, Pakistan. The study was conducted to analyze the readability of CaSN for which the time and accuracy of the proposed method was observed. The result shows that CaSN was solved by both level of users because of the understanding level of Sindhi language. The proposed methods show prominent results in the local context of Sindh. Future work of the study will be to analyze the effectiveness of the CaSN against OCR programs (robustness). In addition, the dataset will be extended to other areas of country to analyze the performance of the CaSN. This study is based on Sindh region and it will be further extended to bring in the participant from other countries. The proposed novel CAPTCHA method CaSN has been tested during study and this study has produced tremendous result however this method needs to be implemented in real-world Sindhi based computer applications and websites.

REFERENCES

[1] A. Maitlo, N. Ameen, H. R. Peikari, and M. Shah, “Preventing identity theft: Identifying major barriers to knowledge-sharing in online retail organisations,” Inf. Technol. People, 2019.

[2] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski, “Designing human friendly human interaction proofs (HIPs),” in Proceedings of the SIGCHI conference on Human factors in computing systems, 2005, pp. 711–720.

[3] H. S. Baird and K. Popat, “Human interactive proofs and document image analysis,” in International Workshop on Document Analysis Systems, 2002, pp. 507–518.

[4] K. Chellapilla, K. Larson, P. Y. Simard, and M. Czerwinski, "Building segmentation based human-friendly human interaction proofs (HIPs)," in International Workshop on Human Interactive Proofs, 2005, pp. 1–26.
[5] M. D. Lillibridge, M. Abadi, K. Bharat, and A. Z. Broder, "Method for selectively restricting access to computer systems." Google Patents, 27-Feb-2001.
[6] W. Susilo, Y.-W. Chow, and H.-Y. Zhou, "Ste3d-cap: Stereoscopic 3d captcha," in International Conference on Cryptology and Network Security, 2010, pp. 221–240.
[7] C. J. Hernandez-Castro and A. Ribagorda, "Pitfalls in CAPTCHA design and implementation: The Math CAPTCHA, a case study," Comput. Secur., vol. 29, no. 1, pp. 141–157, 2010.
[8] M. Mohamed et al., "A three-way investigation of a game-captcha: automated attacks, relay attacks and usability," in Proceedings of the 9th ACM symposium on Information, computer and communications security, 2014, pp. 195–206.
[9] S. Yardi, N. Feamster, and A. Bruckman, "Photo-based authentication using social networks," in Proceedings of the first workshop on Online social networks, 2008, pp. 55–60.
[10] M. P. Bhat and R. N. Raj, "Two-Way Image Based CAPTCHA," in Advances in Communication, Signal Processing, VLSI, and Embedded Systems, Springer, 2020, pp. 471–483.
[11] S. Yilmaz, S. Zavrak, and H. Bodur, "Distinguishing Humans from Automated Programs by a novel Audio-based CAPTCHA," Int. J. Comput. Appl., vol. 975, p. 8887, 2015.
[12] T. Ahmed, K. A. Tushar, S. I. Nova, and M. M. Rahman, "Simple, Robust & User Friendly CAPTCHA 'InstaCap'for Web Security," Int. J. Hybrid Inf. Technol., vol. 9, no. 1, pp. 163–182, 2016.
[13] J. Yan and A. S. El Ahmad, "Usability of CAPTCHAs or usability issues in CAPTCHA design," in Proceedings of the 4th symposium on Usable privacy and security, 2008, pp. 44–52.
[14] U. Pal and B. B. Chaudhuri, "Automatic separation of machine-printed and hand-written text lines," in Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318), 1999, pp. 645–648.
[15] A. Gupta, A. Jain, A. Raj, and A. Jain, "sequenced tagged Captcha: generation and its analysis," in 2009 IEEE International Advance Computing Conference, 2009, pp. 1286–1291.
[16] J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: a CAPTCHA that exploits interest-aligned manual image categorization.," in ACM Conference on Computer and Communications Security, 2007, vol. 7, pp. 366–374.
[17] L. Von Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in International conference on the theory and applications of cryptographic techniques, 2003, pp. 294–311.
[18] Y.-W. Chow, W. Susilo, and P. Thorncharoensri, "CAPTCHA design and security issues," in Advances in Cyber Security: Principles, Techniques, and Applications, Springer, 2019, pp. 69–92.
[19] B. Khan, K. Alghathbar, M. K. Khan, A. M. AlKelabi, and A. Alajaji, "Cyber security using arabic captcha scheme.," Int. Arab J. Inf. Technol., vol. 10, no. 1, pp. 76–84, 2013.
[20] R. B. Palh, H. Nawaz, and Z. A. S. and A. A. Wagan, "Design and Develop CMS for Sindhi E-News Papers," INDJST, vol. 12, no. 46, pp. 1–6, Dec. 2019, doi: 10.17485/ijst/2019/v12i46/148128.