Chapter 9

Innovative Multilingual CAPTCHA Based on Handwritten Characteristics

Maha Hamad Aldosari

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72599

Abstract

Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA) is a kind of test which is commonly used by different websites on the Internet to differentiate between humans and automated bots. Most websites require users to pass the CAPTCHA before signing up or filling out most forms. CAPTCHA today is even used on some mobile applications to provide a higher security level that can protect websites and mobile applications against malicious attacks by automated bots and spammers. The technique essentially relies on employing the human recognition ability, which is not available in automated bots or machines, through leveraging the handwriting characteristics in designing CAPTCHA. The novelty of the technique proposed in this work is that it adopts handwritten characters of four different languages (English, Arabic, Spanish, and French) to generate handwritten multilingual CAPTCHA text. The technique was duly tested and the initial experiments' results for the technique have shown a promising security level that each of the techniques would provide.

Keywords: CAPTCHA, handwritten CAPTCHA, web security, optical character recognition (OCR)

1. Introduction

Web applications have increased rapidly and become a daily necessity for most people [1]. Creating an email account, using social networking sites, and accessing websites are examples of day-to-day activities for Internet users. The fast evolution of the Internet means that the security aspect is being threatened [2]. The number of bots (robot) programs that attack websites has increased. These bots can bring down the site and cause a significant amount of damage. These attacks can take many forms such as DDoS attacks, viruses, worms, and many other malicious devices. They are also considered as the primary reason for email spam [3].

Therefore, it is obvious that stopping such bots by means of a reliable Completely Automated Public Turing Test to Tell Computers and Humans Apart (CAPTCHA) is inevitable. More so, in a multilingual world, multilingual CAPTCHAs are indispensable.

Completely Automated Public Turing Test to Tell Computers and Humans Apart () is considered one of the most common techniques that can be used to distinguish between humans and artificial agents (or bots). For time being, the exponential growth of free web services has led to the misuse of automated bots and spam [4], which has resulted in serious security issues in web services. Using CAPTCHA in its various types has proven to be effective in protecting websites, and the services they provide, from any harm caused by bots' attacks [1].

2. CAPTCHA technique based on handwriting

This technique adopts the handwritten text in the CAPTCHA images and applies a unique feature (separating handwritten characters). This feature can help in differentiating it from any previous handwritten CAPTCHA techniques, and prospectively enhances security level. Moreover, the CAPTCHA's text combines different text languages beside the default language (English) which makes it a multilingual CAPTCHA. The secondary language is selected from a set of languages (French, Spanish, and Arabic) based on the user's region. The main reason for providing multilingual CAPTCHA is that other OCR programs in other languages have not reached the professionalism level of the English OCR yet, and to expand the CAPTCHA usage scope to be used worldwide [1].

At the beginning, different handwritten characters were collected from 100 volunteers; each volunteer wrote the alphabet characters of the 4 adopted languages for the research, each using their own handwriting style. The handwritten characters were classified and stored in a database. These characters were used to synthesize random words that generate the CAPTCHA text, and users should recognize the words in order to pass the CAPTCHA. Furthermore, for the sake of adding a proper security level that will protect the website services from bots' attacks, some distortion methods are applied on each handwritten character separately at the generation process to increase the difficulty for bots to break the CAPTCHA, besides the handwritten characteristics that are fairly resistant for such bots to break.

In summary, this technique goes through two main phases as part of its generating process: the first phase is data gathering and preparation, and the second phase is CAPTCHA implementation with some steps included in each phase.

2.1. Data gathering and preparation

This phase goes through six steps. They are as follows:

- The first step is characters' samples creation. In this step, samples for each character in the four different adopted languages (English, Arabic, Spanish, and French) that will be used in the CAPTCHA text are made [1].

- The second step is samples distribution. In this step, the CAPTCHA characters' samples were distributed to 100 volunteers; each volunteer wrote the samples' characters of the 4 adopted languages by their own handwriting style [1]. As a result, we had a total of 100 different instances with different handwriting styles for each character in each language. So, we ended up with almost more than 10,700 samples' characters that need to be stored on a database in order to be used later in the CAPTCHA implementation phase.

- The third step is transforming samples into digital format. Here, all the collected samples were scanned and stored in digital formats (images).

- The fourth step is sorting data. In this step, we sorted out the collected data into four languages. So, at this point, we have 4 sets of images, each set belongs to one language, and there are 100 different images for each character in each language. Moreover, 4 tables were created on the database to store the images that will be used later to generate the CAPTCHA [1].

- The fifth step is classifying the worldwide countries into categories according to the spoken language there. The countries go with one of the four adopted languages; however, the rest of the countries where their main spoken language is not one from the four adopted languages, we classify them as English-speaking country. After that we stored the countries list with their matched languages on the database [1].

- The sixth step is identifying a list of inappropriate words in each language and storing it on the database as well. **Figure 1** summarizes the data gathering and preparation phase steps.

2.2. Algorithm technique

Figure 2 shows an abstract view of the technique process.

2.3. Handwriting characteristics

Choosing and utilizing the handwriting in designing new CAPTCHA technique was not decided randomly with any logical reasons. On the contrary, it was chosen after a quite long search and study of what characteristics the handwriting has, and how it could be utilized in security field.

Nevertheless, the handwriting in general has some characteristics that can only be utilized by humans. Due to the human brain's superior ability, the brain can analyze and recognize unclear handwritten characters and digits; it also can recognize various different

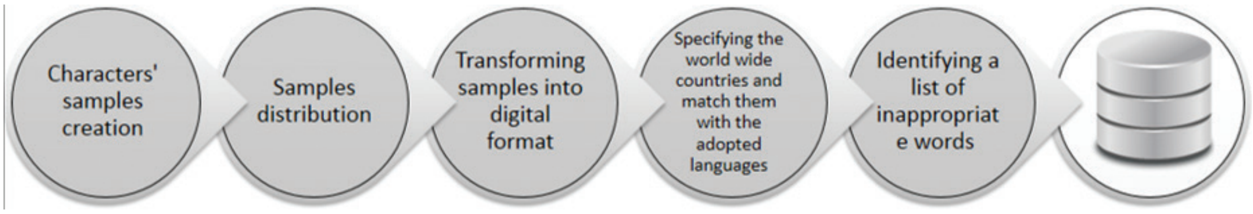
Characters' samples creation
Samples distribution
Transforming samples into digital format
Specifying the world wide countries and match them with the adopted languages
Identifying a list of inappropriat e words

Figure 1. Steps of the data gathering and preparation phase.

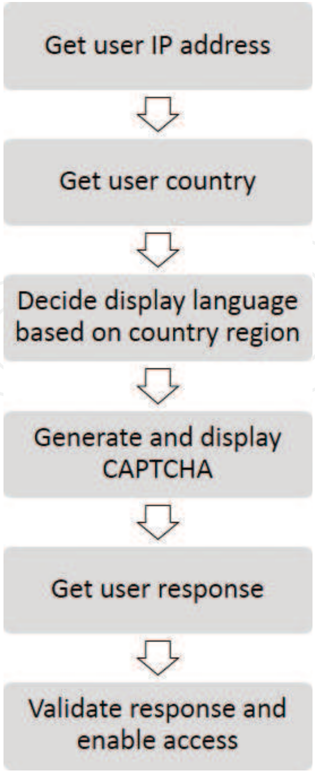
Get user IP address
Get user country
Decide display language based on country region
Generate and display CAPTCHA
Get user response
Validate response and enable access

Figure 2. Abstract view of the technique process.

handwriting styles written by different people. Moreover, the human brain has the privilege of using its experience to figure out uncompleted characters or uncompleted words that have missing letters. It even can read the Arabic words written without any dots on its letters, because the words' shapes can be enough for the brain to figure out the words, unlike OCR machines which mostly cannot recognize the words if they are not complete or without dots in Arabic words case.

Overall, this confirms the human capability in utilizing the handwriting characteristics, which cannot be found in any OCR machine, and it encourages us to go through this CAPTCHA technique which is based on handwriting.

3. Technique implementation

The generation process of this technique starts by getting the user's IP address. Then, it gets the country's name where the user is located at the time of accessing which is obtained using the IP-API service. Consequently, a country language will be retrieved from the database using the country name, where a list of countries is sorted and classified into one of the adopted languages (Arabic, English, French, and Spanish). A list for each of the adopted language was created which contains the countries which speak the specified language. Hence, the countries classification is done based on the official spoken language in each country. However, if the country's name is not on any of the four countries' lists, then English will be the default language to use (English).

In addition, the user's website default language will be determined and compared to the retrieved country's language; if they are different, then the website default language will be used.

The following flowchart (**Figure 3**) illustrates the whole process of the first step in this CAPTCHA technique which decides the CAPTCHA language to be displayed to the user.

Furthermore, after the language has been decided on, the CAPTCHA generation process will move on to the next step which is choosing the CAPTCHA word length. The word length is chosen randomly from five to eight characters. Next, the word construction process will start by selecting the handwritten characters and distorting them separately. However, this step will be done little bit differently if the previously decided language is Arabic; the following flowchart (**Figure 4**) clarifies the CAPTCHA word construction process in detail.

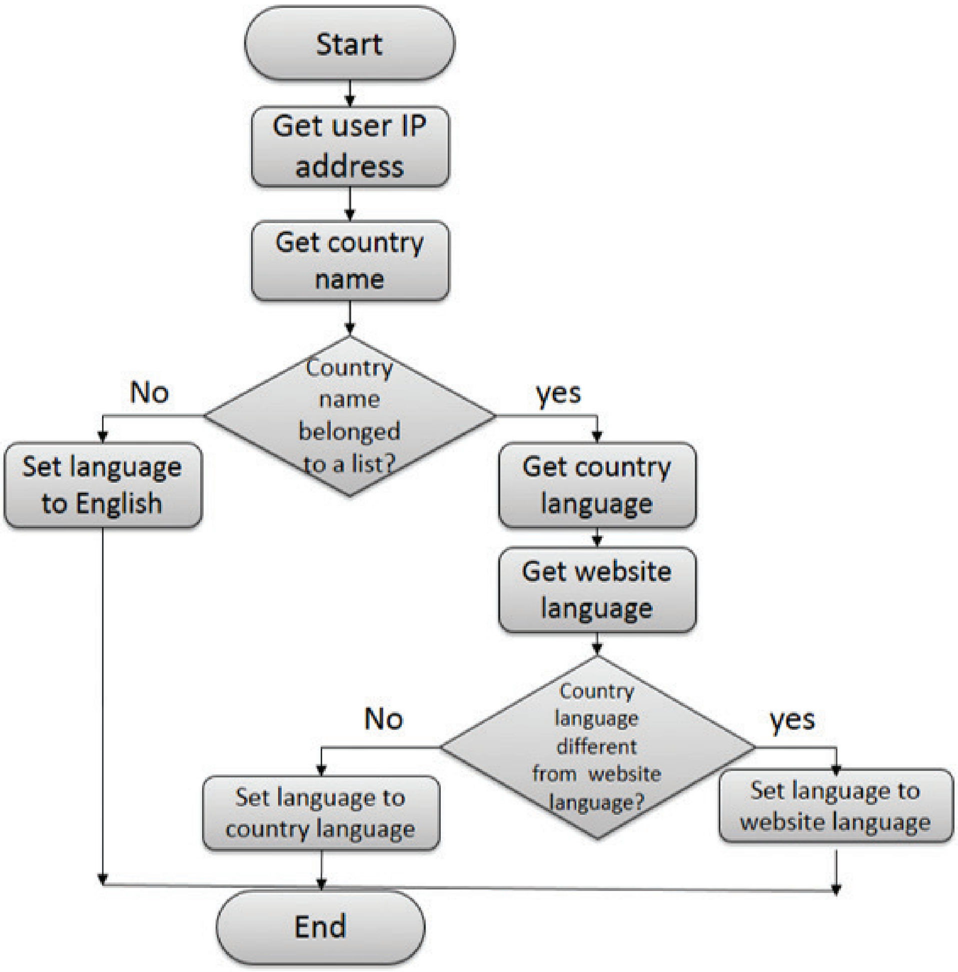
Start
Get user IP address
Get country name
Country name belonged to a list?
No
yes
Set language to English
Get country language
Get website language
Country language different from website language?
No
yes
Set language to country language
Set language to website language
End

Figure 3. Deciding the CAPTCHA language process.

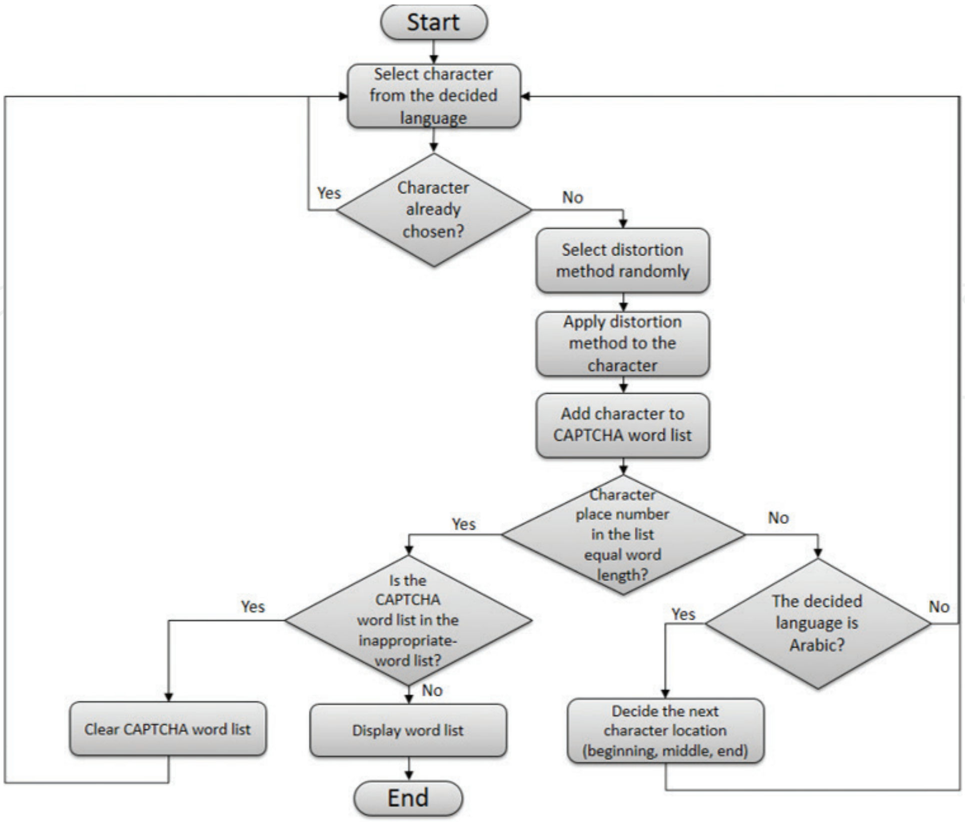
Start
Select character from the decided language
Character already chosen?
Yes
No
Select distortion method randomly
Apply distortion method to the character
Add character to CAPTCHA word list
Character place number in the list equal word length?
Yes
No
Is the CAPTCHA word list in the inappropriate-word list?
Yes
No
The decided language is Arabic?
Yes
No
Clear CAPTCHA word list
Display word list
Decide the next character location (beginning, middle, end)
End

Figure 4. CAPTCHA word construction process.

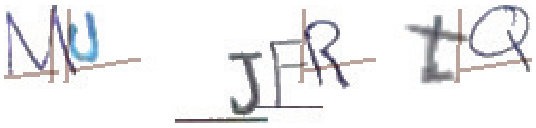

Figure 5. English CAPTCHA "M u J F R t Q".

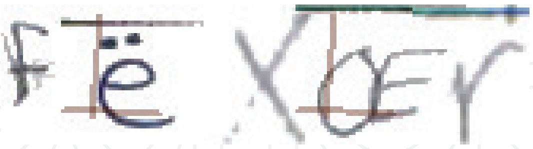

Figure 6. French CAPTCHA "F ë x Œ r".

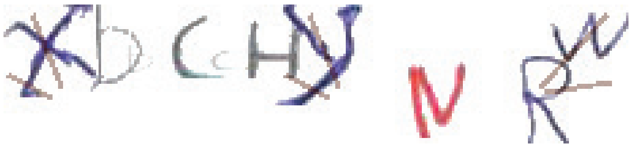

Figure 7. Spanish CAPTCHA "X b CH y N R w".

As shown above, when the CAPTCHA word is generated, it is displayed to the user in one of adopted languages (Arabic, English, French, and Spanish). **Figures 5–8** show examples of the handwritten CAPTCHA technique with each adopted language.

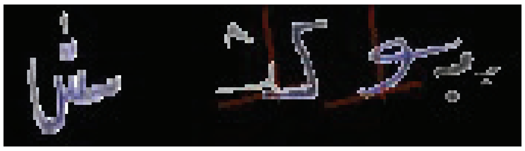

Figure 8. Arabic CAPTCHA "بوكنش".

4. Experiment techniques

In the conducted experiments, six different OCRs were used to test the technical performance of the proposed CAPTCHA techniques. The used OCRs have a good review from some technical experts and they provide good results when used to extract the regular text.

Moreover, other methods were used to test the usability, such as surveys and local web pages to get the users' responses and analyze them from different perspectives.

4.1. First OCR

The first OCR used in the experiments is an application called Free OCR. This application utilizes the most recent version of the Tesseract OCR engine (v3.01), which can ensure a reliable level of text-extracting accuracy. Tesseract is an open-source OCR engine maintained by Google. It offers support for different languages, with a level of accuracy potentially reaching 98% [1, 5].

4.2. Second OCR

Capture2Text is the second technique used in our experiments. It is an open-source OCR tool, like the first OCR; it uses the Tesseract engine introduced by Google to capture the written text in images and then copies it to the clipboard.

4.3. Third OCR

The third OCR used is a free online OCR called i2OCR. It is available in the following link: http://www.i2ocr.com/. This online OCR supports various recognition languages; it also has the ability to extract text from various columns in the images.

4.4. Fourth OCR

FreeOCR is the fourth OCR tool used in the experiments. As the name suggests, it is available online as a free service, which is available in the following link: http://www.free-ocr.com/. Moreover, the extraction process speed for this OCR site is considered fast in comparison with other online OCRs, and it produces the extracted text fairly quickly [6].

4.5. Fifth OCR

The fifth OCR we adopted in the experiments is an online OCR software called OnlineOCR. This OCR software is available in the following link: http://www.onlineocr.net/. Additionally, this

OCR software supports 46 recognition languages and it is able to extract texts in any of these languages. It also can detect text written in more than one language in the same image or document.

4.6. Sixth OCR

NewOCR is a free online OCR service that we used as the sixth technique in our experiments. The NewOCR service is available in the following link: https://www.newocr.com/. This online service supports more than 100 recognition languages and different fonts supports. In addition, the NewOCR service works using Tesseract OCR engine which is considered the best accurate OCR engine available at this time. It also supports the low-resolution images and can extract the text written in these images.

5. Technical performance testing

First of all, we started with the technical performance testing to test the technical aspects of the proposed handwritten CAPTCHA. A total of 500 different CAPTCHA images were generated for each language. Each of the 500 CAPTCHA images of the first 3 languages (English, French, and Spanish) were tested on 6 different OCRs, while the 500 Arabic CAPTCHA images were tested on the second and sixth OCRs. **Table 1** illustrates the testing results for the six different OCRs on each adopted language.

As shown in **Table 1**, the testing results were divided into four patterns [1]:

1. Correctly recognized pattern: it is when all the text in the CAPTCHA image has been correctly recognized.
2. Partially correctly recognized pattern: it is when the OCR has recognized three or more characters in the CAPTCHA text.
3. Incorrectly recognized pattern: it is when no characters have been correctly recognized in the CAPTCHA text.
4. No text pattern found: it is when the OCR was not able to recognize the text or any character in the CAPTCHA image.

In the English CAPTCHA images, the six OCRs have failed to recognize the full text in 99% of the images, while only 1% was correctly recognized. Nevertheless, the 99% includes 6% partially correctly recognized patterns, 46% no text found in the CAPTCHA image, and 47% totally incorrect text recognition [1].

Moreover, the other languages testing outcomes resulted in a lower recognition percentage compared to the English one. In Spanish language case, all the six OCRs have failed to correctly recognize 99.97% of the Spanish CAPTCHA images; this 99.97% includes 4.23% partially

Pattern		Correctly recognized	Partially correctly recognized	Incorrectly recognized	No text found
First OCR	English	5	54	431	10
	Spanish	1	16	475	8
	French	0	2	486	12
Second OCR	English	0	30	390	80
	Spanish	0	24	403	73
	French	0	18	398	84
	Arabic	0	3	391	106
Third OCR	English	0	65	250	185
	Spanish	0	47	277	176
	French	0	23	201	276
Fourth OCR	English	0	40	150	310
	Spanish	0	37	84	379
	French	0	29	63	408
Fifth OCR	English	0	0	115	385
	Spanish	0	0	81	419
	French	0	0	97	403
Sixth OCR	English	0	5	85	410
	Spanish	0	3	118	379
	French	0	1	68	431
	Arabic	0	0	73	427

Table 1. The testing results for the six different OCRs on each adopted language.

correctly recognized, 47.8% no text found in the CAPTCHA image, and 47.9% totally incorrect text recognition. Likewise, in the French language case, the six OCRs used in the experiments did not succeed in correctly recognizing any of the French CAPTCHA images, while only [7] 0.4% of the whole French CAPTCHA images were partially correctly recognized, and 53.8% of the images resulted in no text found, and the remained 43.8% of the images were incorrectly recognized.

Similarly, the experiments result for the Arabic language shows that the two used OCRs have failed to correctly recognize any of the Arabic CAPTCHA images. However, the two used OCRs were able to partially correctly recognize only 0.3% from the whole Arabic CAPTCHA images, whereas 53.3% of the images resulted in no text found and 46.4% of the images were incorrectly recognized.

6. Usability testing

In order to infer the usability of the proposed CAPTCHA, two users' acceptance tests have been conducted on this technique. The first test was aimed at taking a large number of responses from different users, while the second test was aimed at testing a large number of the produced CAPTCHA images.

6.1. First usability test

The first test targeted 100 users through an online survey that aimed at understanding how users will interpret five different CAPTCHA images. These five images were chosen on the basis of different aspects to study users' responses regarding characters' distortions and unclear handwriting styles (see Appendix B for CAPTCHA images used in the survey).

Table 2 illustrates the results of the conducted survey and shows the answers' patterns for each CAPTCHA image used in the survey.

According to the results shown in **Table 2**, 82% of the users were able to correctly recognize the CAPTCHA characters of the first image, the remaining 18% failed and were confused between characters, and noise, and distortion.

As for the second and third images, they were recognized by 85 and 75% of the users, respectively, while the rest of the users were confounded by one character, due to the warping distortion method applied on that character.

Moreover, users succeed in correctly recognizing the fourth CAPTCHA image with a percentage of 74%, while 61% of them correctly recognized the fourth and fifth CAPTCHA images. However, the remain percentages of the users who failed to correctly recognize the last two CAPTCHA images have failed because of the unclear handwriting style which was selected on purpose to reflect the worst cases that could be produced from the collected database.

In general, it must be mentioned that the partially correctly recognized pattern indicates that the user misinterpreted three characters or fewer from the CAPTCHA word, otherwise it will be considered that the user has incorrectly recognized the whole CAPTCHA word.

Additionally, the average time taken by each participant to solve the survey was 2 minutes and 38 seconds.

Pattern	First image	Second image	Third image	Fourth image	Fifth image
Correctly recognized	82	85	75	74	61
Partially correctly recognized	18	15	25	26	39
Incorrectly recognized	0	0	0	0	0

Table 2. Results of the survey conducted for the first technique.

6.2. Second usability test

The second usability test was carried out using the implemented web pages to give users a real experience similar to solving a real CAPTCHA on any website. Therefore, as mentioned earlier, the implemented web pages produce CAPTCHA images at the real time and responds to users as soon as they click on the “validate” button to inform them either they solve the CAPTCHA correctly or not.

Consequently, the test was conducted on 20 users with five CAPTCHA images average for each user; each user viewed the CAPTCHA on the web page and solved five different CAPTCHAs, and in the meanwhile the users’ answers and the web page responses to their answers were recorded.

Correspondingly, the testing results showed that 92% of the tested images were correctly recognized by users, while in the remaining 8% of the images fewer than three characters of each image were misinterpreted by the users.

However, the percentages of the results have proven a fair usability percentage, which also could be further improved with little adjustments on the distortion methods and on the collected handwritten characters database [1].

7. Conclusion

The rapid evolution of web and mobile applications turn these applications into an important part in people’s daily life, where people rely on them to accomplish most of their activities. On the other hand, all the rapid improvement in these applications comes with a rapid increase in the number of malicious bots and applications that threatens the security of web and mobile applications.

Therefore, we introduced a new CAPTCHA technique that utilizes handwriting styles and we have put it through several experiments to adjust, improve, and test the technique while trying to reflect every needed adjustment to the technique immediately.

The introduced technique is a novel handwritten CAPTCHA, which basically relies on employing the handwriting characteristics that can only be interpreted by humans while being comparatively hard for OCRs to recognize. The proposed approach adopts four different languages (English, Arabic, Spanish, and French); each language has its own handwritten characters used in synthesizing the CAPTCHA text.

Moreover, few testing experiments have been conducted on the proposed CAPTCHA to test its robustness as well as the level of security it provides. The experiments were done using six different OCRs on 500 different CAPTCHA samples. Nevertheless, the results of the experiments manifest the significant benefits of utilizing handwriting characteristics with CAPTCHA samples [1].

Author details

Maha Hamad Aldosari

Address all correspondence to: aldosari.maha@gmail.com

King Saud University, Saudi Arabia

References

[1] Aldosari MH, Al-Daraiseh AA. Strong multilingual CAPTCHA based on handwritten characters. In: 2016 7th International Conference on Information and Communication Systems (ICICS). 2016. pp. 239-245

[2] ur Rizwan R. Survey on captcha systems. Journal of Global Research in Computer Science. 2012;**3**:54-58

[3] Gummadi R, Balakrishnan H, Maniatis P, Ratnasamy S. Not-a-Bot: Improving service availability in the face of botnet attacks. In: NSDI. 2009. pp. 307-320

[4] Mehrnejad M, Ghaemi Bafghi A, Harati A. SEIMCHA: A new semantic image CAPTCHA using geometric transformations. ISeCure. 2012;**4**:63-76

[5] Patel C, Patel A, Patel D. Optical character recognition by open source OCR tool tesseract: A case study. International Journal of Computer Applications. 2012;**55**:50-56

[6] Kishore A. (2015, December 2nd). 5 Free Online OCR Services Tested and Reviewed. Retrieved from https://www.online-tech-tips.com/cool-websites/convert-image-to-text-using-free-online-ocr-software/

[7] Yan J, El Ahmad AS. Usability of CAPTCHAs or usability issues in CAPTCHA design. In: Proceedings of the 4th Symposium on Usable Privacy and Security. 2008. pp. 44-52