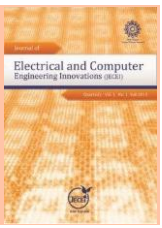
SRTTU

Journal of Electrical and Computer Engineering Innovations

JECEI, Vol. 4, No. 2, 2016

Regular Paper



Introducing New Trends for Persian CAPTCHA

Farzin Yaghmaee[1,*] and Mohadese Kamyar[1]

[1] Electrical and Computer Engineering Department, Semnan University, Semnan, Iran.

[*]Corresponding Author's Information: F_yaghmaee@semnan.ac.ir

ARTICLE INFO

ARTICLE HISTORY:
Received 17 September 2016
Revised 28 October 2016
Accepted 30 October 2016

KEYWORDS:
CAPTCHA
Persian language
Persian CAPTCHA
Image processing

ABSTRACT

To distinguish between human user and computer program to enhance security, a popular test called CAPTCHA is used on Web. CAPTCHA has an important role in preventing Denial Of Service (DOS) attacks in computer networks. There are many different types of CAPTCHA in different languages. Due to the expansion of Persian-language and documents on the internet, creating a suitable Persian CAPTCHA seems to be necessary. In this paper, we introduce three different types for Persian CAPTCHA in different domains. In the first type, based on the particular characteristics of Persian writing such as contiguous writing and image processing techniques, high strength CAPTCHA is provided. In the second type, the meaning of Persian words are used to creating CAPTCHA and in the third type, the combination of image processing techniques and the meaning of Persian words are used. Experimental results show that the proposed CAPTCHAs has high security against attacks while Persian people can easily recognize them.

1. INTRODUCTION

Nowadays, because of the enormous growth in computer networks and expansion of World Wide Web, many routine tasks such as banking, mailing and registration performed on the Internet. So such services on the Web can be a suitable place for subversive purposes. In these sites, users should fill out a form about their personal information and hackers attempt to attack on server databases with numerous tricks such as role playing as a human.

To distinguish a machine from humans to prevent such attacks, a method called CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is used on web numerously. CAPTCHA is a simple question for humans while it is hard for computers. Therefore CAPTCHA determines that visitors of a website are humans or just malicious software.

Success of CAPTCHA depends on two factors: first, it should be easy for human user and second, it should be difficult for a computer program [1]. CAPTCHA methods can be generally divided into two groups based on OCR (Optical Character Recognition): OCR - based and Non-OCR-based.

Some OCR programs can defeat CAPTCHA and recognize it. In Persian language, there is not a powerful OCR system yet. The most problems in designing Persian (or Arabic) OCR system are connectivity of characters which cause low accuracy in character segmentation phase. While in English, OCR systems are so successful and are used in real commercial applications.

In the other hand, some complexity in Persian language and alphabet which are significant problems in OCR systems may be useful to design robust CAPTCHA. The aim of this paper is to introduce new aspects in Persian CAPTCHA which can be used efficiently in Arabic language too. We propose three approaches to design an efficient CAPTCHA:

- Using image processing distortions in words and backgrounds (like it used in English).
- Using the complexity of point detection and different positions of them in similar words in Persian language.

- Using connectivity alphabet feature to create semantically Persian CAPTCHAs.

The paper is organized as follows: Section 2 describes some related works in previous works in CAPTCHAs by emphasizing on Persian CAPTCHA and explains Persian alphabet features. Section 3 is dedicated to our three proposed methods for Persian CAPTCHA and their experimental result and conclusions are presented in Section 4.

2. Technical Work Preparation

Various works have been done in OCR based CAPTCHA [2] [3] [4]. These methods are implemented by image processing distortions in fonts, background and color. Also, these CAPTCHAs are easily understandable by humans, but with improvement in OCR applications these CAPTCHAs are capable to be broken by hackers. Some of these famous CAPTCHAs are presented in sections 2-1 and 2-2.

2-1. English CAPTCHA

Many OCR-based CAPTCHAs are designed for English language. Common points in most of these CAPTCHAs are image distortion and noise addition. These modifications will reduce the recognition capability of OCR application.

The CAPTCHA used in Yahoo website has an interesting idea in using hollow font. Fig. 1 shows an example of this CAPTCHA.

Figure 1: Yahoo CAPTCHA.

As is shown in Fig. 2, two words are used in Google's CAPTCHA system. In this CAPTCHA, number of characters is large, so entering CAPTCHA is time consuming for humans and this is a weak point in this type of CAPTCHA.

Figure 2: Example of Google's CAPTCHA

2-2. Persian CAPTCHA

Because most CAPTCHAs are proposed for those who speak English, non-English users probably have some difficulties with them. Therefore, designing CAPTCHAs in the other languages could be an attractive and useful idea [5].

In this section, we discuss about previous works in Persian CAPTCHA. In Fig. 3, an example of Persian CAPTCHA with random and meaningless word is shown. This is a simple CAPTCHA without any robustness against attacks [6].

Figure 3: An example of Persian CAPTCHA [6].

In later work, another example as shown in Fig. 4, is provided by the same authors where Nasta'liq[1] font and meaningful Persian words have been used [7]. In addition, a method is proposed for detecting Nasta'liq font in CAPTCHA [8].

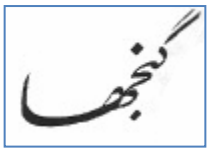


Figure 4: An example of Persian CAPTCHA [7].

Fig. 5, is another example of Persian CAPTCHA. In this method, a Persian word in the image is modified by shift, curvature and noise [9].

Figure 5: CAPTCHA sample [9].

2-3. Persian Alphabet Feature

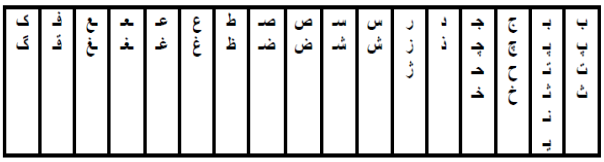
As we want to use Persian language features in the design of CAPTCHA, first, we review important features of Persian alphabet.

- Connected characters: In Persian language like Arabic, characters in a word are connected to each other. For this reason, in Persian, the segmentation of connected characters is more difficult.
- Different size of characters: Size of characters in Persian is different and this increases the complexity of Persian language.

1 Nasta'līq is one of the main script styles used in writing the Perso-Arabic script, and traditionally the predominant style in Persian calligraphy.

- Points in characters: Almost half of the Persian alphabets have one, two or three points. Many Persian words are similar in structure and the only difference between them is the number and location of points or specific symptoms. Points are very similar to noise and this makes detection to be difficult for OCR software. Table (1) shows the similar characters in Persian.

TABLE 1
SIMILAR CHARACTERS IN PERSIAN LANGUAGE



3. SUGGESTED METHODS

In this section, three different approaches for Persian CAPTHA are proposed. In each method, we use some unique features of Persian language.

3-1. CAPTCHA based image distortion

In this method, we consider a number of characters as the source. Considered source can be determined which characters should be used for CAPTCHA; separate or connected characters or a combination of both of them.

For example, if we only use the connected characters, segmentation of CAPTCHA will be harder but the number of eligible characters will be reduced [11].

3-1-1. Dilation and Erosion operators

Counting pixels after segmentation is one of the known techniques in CAPTCHA attacks. Some interesting results of this method are mentioned in [10]. Counting pixels is difficult in Persian, because of connectivity of characters.

However, for more security, we dilate or erode the image by morphology operators.

The used element structures are circular or square with random sizes. In this case, counting pixel attack is almost impossible [11].

3-1-2. Distortion

Created image is distorted by applying mathematical functions. This distortion will be applied on the image (vertically or horizontally).

The simplest function that can be used to create distortion is sinusoidal function. A sine wave is determined by Equation 1:

$$x_{new} = Ht * \sin(x_{old} / Tm) \quad (1)$$

Parameter Ht, determines the sine wave height and the parameter Tm, controls the wavelength. We will select the parameters Tm and Ht randomly from a variable range to have random distortion. Thus, two waves with random heights and wavelengths will change the image as shown in Fig. 6.

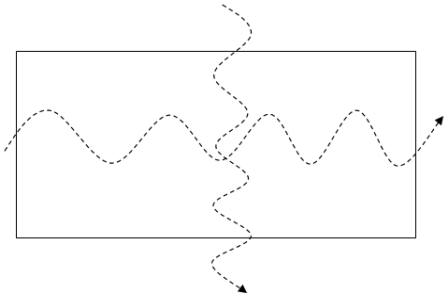


Figure 6: Two sine waves scan our image, one horizontally and the other vertically.

Sine wave is the simplest case and the other combined trigonometric functions can also be used for distortion. But, it should be noted that these distortions may be cause illegibly in final CAPTCHA.

In Fig. 7, a sample of image distortion without background is shown.

Figure 7: image after combination of cosine and sine distortion.

3-1-3. Producing final image

At this step, random like-spot noise will be added to the image. Due to the similarity of these spots to points of Persian characters, the accuracy of machine detection will be reduced. Fig. 8 shows samples of the proposed CAPTCHA.

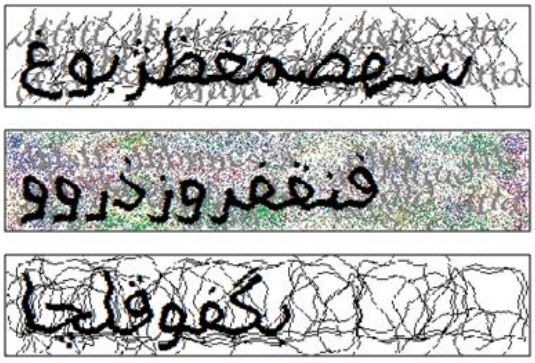


Figure 8: Final samples of our proposed CAPTCHA.

3-1-4. Results of implementation

This CAPTCHA is tested on web and some people aged between 15 to 45 years have been asked to solve the CAPTCHA. Experimental results show that 93% of users in their first try, 2% in their second and 5% in their third try are able to solve CAPTCHA.

Results are acceptable according to the other analogous CAPTCHAs. Our study showed that the main reason for user's mistake was illegible spots in background. To investigate the possibility of reverse engineering, we have tested 100 samples of our CAPTCHA with three commercial OCR namely Readiris Pro 12, OmniPage 18, SimpleOCR. Interestingly, none of them could read any proposed CAPTCHAs.

To evaluate the performance of morphological operators, the average value of pixels in all different forms of characters (small, large, etc) has been obtained from the 1000 samples. Fig. 9, shows the histogram of the number of pixels in Persian alphabet. Uniform distribution of pixels shows that the proposed method is resistant against counting pixels attacks.

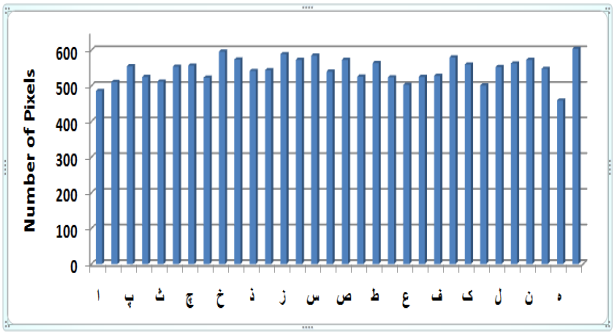


Figure 9: The average number of pixels is uniformly distributed in different Persian alphabet. (Derived from 1000 randomly generated CAPTCHA)

3-1-5. Comparative Study of Our CAPTCHA and another Persian CAPTCHA Scheme

Persian CAPTCHA in [6] is an example of Persian CAPTCHA with random and meaningless word. This is a simple CAPTCHA without any robustness against attacks because the background and foreground color is the same and the position of CAPTCHA in the text is static. In our proposed method, a CAPTCHA is introduced based on contiguous writing in Persian and image processing techniques such as morphology operators and image distortion. The proposed method is quite resistant against attacks of counting pixels. Our proposed method is more powerful than that proposed CAPTCHA in [6-7].

3-2. Semantically CAPTCHA based on Persian alphabet

Semantically CAPTCHAs may be the most tolerable ones against attacks as they are far beyond abilities of machines when it comes to answering only-for human questions. In fact, this category provides users with questions that are easy to be understand and answered by humans but relatively difficult for computers [5].

The original plan for this type of CAPTCHA is displaying Persian words without dots in a picture [12]. In other word, as shown in Fig. 10, a meaningful word is placed in the image with no dot and the user should guess a meaningful and equivalent word. This word has number of points and the user should punctuate the word by the points, so a meaningful word is then created.

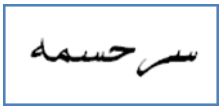


Figure 10: A sketch of the proposed CAPTCHA ("سرچشمه").

It is expected that with emphasis on the meaning of Persian words, the proposed CAPTCHA would have a relatively high degree of security. Because machines, do not understand the meaning of words and this causes problems for the machines on how to punctuate the meaningful words. It should be noted that in order to increase the security level, three different words are used without any point in the image, simultaneously.

3-2-1. Collecting suitable words

We have tried to provide a database of Persian words. The collected words are meaningful and familiar to human mind. Words are collected from different categories, including names of people, cities, countries, fruits, foods, animals, colors, flowers, cars, furniture, jobs and etc. Finally, a database with more than a thousand of different and meaningful words is created.

3-2-2. Making font with no points

According to the initial plans, the words should be displayed without points in the image. Thus, several fonts were designed without any point. To increase the security level, the selected words were displayed with these fonts, randomly.

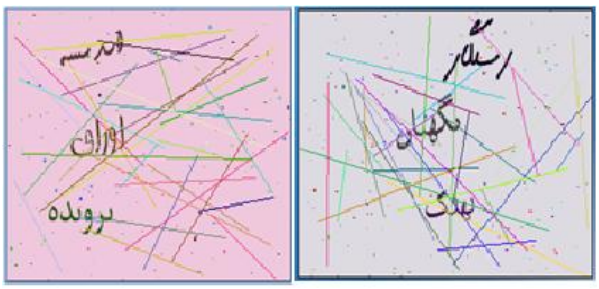
3-2-3. Position of the words in the image

Font type, color, size, angle and position of the image are not fixed in the proposed CAPTCHA. In this case, recognizing the words will be more difficult for OCR programs. Selected words, from first to third, will be displayed respectively, at the top, middle and bottom of the image.

3-2-4. Adding noise and background to the image

Adding noise and background make similarity with symptoms Persian words and this causes that

recognizing of words will be more difficult for a machine. For this purpose, we add a series of ellipses and lines to the image. Number of generated noise is randomly and as the word is obvious for humans, it is vague for a computer program. Two samples of CAPTCHA created with this method, are shown in Fig. 11.



(a) (b)

Figure 11: (a) The words are "اندیشه، اوراق، پرونده " (b) The words are "رستگار، نگهبان، نهنگ".

3-2-5. Results of Implementation

This CAPTCHA is tested on the web and some people are asked to solve the CAPTCHA. Users' reply information was stored to measure the security and reliability level of designed CAPTCHA. In Table (2), the hardness of our designed fonts is shown.

Figure 12: An example of the Arabic CAPTCHA [13].

TABLE 2
THE HARDNESS OF THE DESIGNED FONTS.

Error detection percentage	Font Name	The font shape for word "فارسی"
22.4	Khodkar	فارسی
19.9	BDavat	فارسی
14.9	BNazanin	فارسی
14.6	BKamran	فارسی
14.5	BTabassom	فارسی
13.7	Dastnevis (a handwritten font)	فارسی

The most number of words which are detected incorrectly used the "khodkar" font, which is very similar to a graphical font.

The results show that the use of two or three combinations of word is recognized by about 83% of the people while it is impossible for software to detect the words.

3-2-6. Comparative Study of Our CAPTCHA and Arabic CAPTCHA Schemes

Table (3) compares the features of our CAPTCHA scheme with the Arabic CAPTCHA [13]. The proposed scheme uses Arabic script to generate an image.

The image is distorted by adding various types of noises in the background in the form of dots, lines and arcs. Studies show that our CAPTCHA scheme is more robust than the Arabic CAPTCHA.

The Proposed CAPTCHA is semantic type of CAPTCHAS and the words in image are meaningful words. This feature increases the security and robustness of our CAPTCHA, because machines do not understand the meaning of words. In contrast to our CAPTCHA scheme, the words in the Arabic CAPTCHA [13] are not meaningful and just some random letters make meaningless words.

Figure 13: background used for CAPTCHA (one dark and one bright band).

TABLE 3
COMPARISON OF FEATURES OF THE PERSIAN SCHEMES (OURS) AND THE ARABIC CAPTCHA SCHEME.

Features	Proposed method	Arabic method[13]
CAPTCHA Type	Semantic & OCR	OCR-based
Efficiency	Highest	Medium
Robustness	Highest	Medium
Number of letters	Variable (2-15)	5-9
Number of words	3	1
Background and foreground color	Different	Different(just in blue range)
Variation of font types	Randomly	Randomly
Variation of font size	Randomly	Randomly
Lines and dots(as background noise)	Randomly	Randomly
Baseline detection by OCR	Not possible	Not possible
Text coordinates	Varies	Varies

	randomly	randomly

3-3. CAPTCHA based on connectivity of Persian alphabet and meaning of words

In addition to above methods, we proposed another method for creating semantically CAPTCHA. For this purpose, we used a set of meaningful words as described in section 4.2.1 with at least 6 characters length.

More than 20 fonts were designed for using in this CAPTCHA which are randomly selected. To increase the security level, the selected words are displayed with those fonts by random. Probability of hollow fonts selection is double than bold fonts.

3-3-1. Using dark and bright bands as pattern background

The model of back ground used in this method is shown in Fig. 13. Some examples of generated CAPTCHAs are given in Fig. 14, Fig. 15.

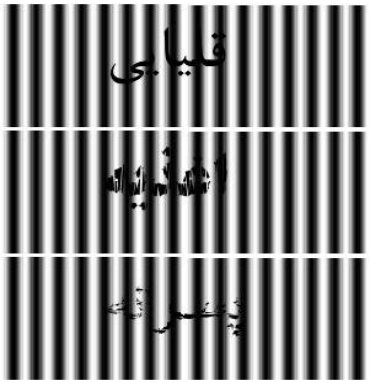


Figure 14: CAPTCHA with bold font.

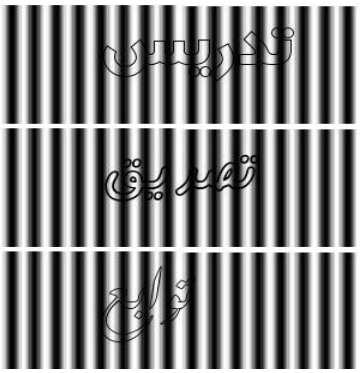


Figure 15: CAPTCHA with hollow font.

An important point in this CAPTCHA was effect of increasing the width of dark and bright bands on the user's ability to word detection.

Dark bands make difficulty for OCR applications to identify words in CAPTCHA.

If we use one dark band and two bright bands (as shown in Fig. 16) in background pattern, ability of human users to solving CAPTCHA will be increased. But, also OCR tools might defeat CAPTCHA.

Figure 16: Increasing width of bright band.

Using two dark bands and one bright band will cause difficulties for human recognition.

In "Fig. 17," the impact of increasing width of dark band can be observed.

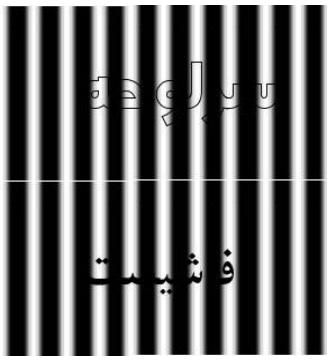


Figure 17: Increasing width of dark band.

Increasing width of dark and bright bands together is a convenient solution to enhance human user's detection capabilities, as well as reducing the chance of failure against the attacks.

3-3-2. Results of Implementation

This type of CAPTCHA was tested on 740 users. According to the statistics, 92% of users was able to recognize CAPTCHA correctly at the first observation. According to Table (4), the font type has little effect on detection.

However, the use of hollow fonts makes more difficulty for OCR programs to recognize the CAPTCHA. Between patterns used in the background, the pattern with equal width in dark and bright bands seems more appropriate.

TABLE 4
PERCENTAGE OF CORRECT ANSWERS IN FIRST OBSERVATION.

CAPTCHA format	Percentage of correct answers in the first detection
Words out of a dictionary	93
Limited words such names of cities	100
Random character strings	88
Hollow font	93
Bold font	91
One dark bar and one bright bar in background pattern	95
One dark band and two bright bands in background pattern	100
Two dark bands and one bright band in background pattern	90
Increasing width of dark and bright bands together in background pattern	98

3-3-3. Comparative study of our CAPTCHA and another Persian CAPTCHA scheme

Persian CAPTCHA in [9] is another example of Persian CAPTCHA. In this method, a Persian word in the image is modified by shift, curvature and noise [9]. This algorithm almost has not an acceptable level of security.

The lack of using Persian alphabet features is a major drawback in all of previously proposed Persian or Arabic CAPTCHAs.

In most of the methods mentioned above, the general process for creating CAPTCHA is similar and differences are only in type of image distortion or percentage of noise addition.

In our proposed CAPTCHA, we used a special pattern with dark and bright bands in background. The created words are meaningful. A human can guess the meaningful word easily. But, guessing a meaningful word that is completely broken by dark bands is very difficult for a machine.

4. CONCLUSIONS

Despite the various methods performed in English CAPTCHA, design of Persian (and Arabic) CAPTCHA is still at its infancy. In this paper, we have proposed three new Persian CAPTCHAs using special features of Persian alphabet with emphasis on semantics in Persian language. Of course, these CAPTCHAs can be used in Arabic language too. In the first type, a CAPTCHA which was introduced based on contiguous writing in Persian and image processing techniques such as morphology operators and image distortion. The proposed method is quite resistant against attacks of counting pixels.

Second proposed CAPTCHA is based on the meaning of Persian words.

Persian meaningful words are displayed without points in a picture. Human user can punctuate the meaningful words easily while doing this is so difficult for a machine.

This is because only a human that has Persian vocabulary in his mind can answer to this type of CAPTCHA. In the third proposed CAPTCHA, we used a special pattern with dark and bright bands in background. If width of dark and bright bands is appropriate, a human can guess the word easily.

But guessing a meaningful word that is completely broken by dark bands is very difficult for a machine.

Of course designing of Persian CAPTCHA is at its beginning, but we hope this paper achieves its primary goal which is the creation of a powerful Persian CAPTCHA for native people by using inherent features of Persian alphabet and words.

REFERENCES

[1] M. Blum, L.A. Von Ahn, and J. Langford, "Completely automatic public turing test to tell computers and humans apart," the CAPTCHA Project, www.CAPTCHA.net, Department of Computer Science, Carnegie-Mellon University, 2000.

[2] R. Datta, "Exploiting the human–machine gap in image recognition for designing CAPTCHAs," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 3, pp. 504–518, 2009.

[3] L.A. Von Ahn, M. Blum, J. Hopper, and J. Langford, "CAPTCHA: using hard ai problems for security," *Advances in Cryptology EUROCRYPT 2003* , pp. 294-311, 2003.

[4] J. Yan and A.S. Elahmad, "CAPTCHA security: a case study," *IEEE Security & Privacy*, vol. 7, no. 4, 2009.

[5] M. Moradi and M.R. Keyvanpour, "CAPTCHA and its alternatives: a review," *Security and Communication Networks*, vol. 8, no.12, pp. 2135-2156, 2015.

[6] M. H. Shirali-Shahreza and M. Shirali-Shahreza, "Persian/Arabic baffletext CAPTCHA," *Journal of Universal Computer Science*, vol. 12, no. 12, pp. 1783-1796, 2006.

[7] M.H. Shirali-Shahreza and M. Shirali-Shahreza, "Nastaliq CAPTCHA," *Iranian Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, no. 2, pp. 109-114, 2007.

[8] M. Salmani Jelodar, M.J. Fadaeieslam, N. Mozayani, and M. Fazeli, "A Persian OCR system using morphological operators," *Transactions on Engineering, Computing and Technology*, vol. 1, no. 4, pp. 1137-1140, 2005.

[9] M. Bohlool and M. Malekzadeh, "Persian CAPTCHA system to prevent automatic subscribing of software robots in web pages," 13th National CSI Computer Science, Iran, 2008.

[10] A. El Ahmad, J. Yan, and WY. Ng, "captcha design color, usability, and security," *IEEE Internet Computing Journal*, vol. 16, pp. 44-51, 2012.

[11] F. Yaghmaee and A. Bakhshande, "A new method for Robust Persian CAPTCHA," 9th int. Conference on Cryptography and security, Tabraiz, Iran, 2012.

[12] F. Yaghmaee, M. Kamyar, and F. Kamandy, "Introducing a new semantically persian CAPTCHA," 21th Conference on Elctronics and Electrical Engineering, Mashad, Iran, 2013.

[13] Khan, Bilal, *et al*, "Cyber security using arabic captcha scheme," *Int. Arab J. Inf. Technol*, vol. 10, no. 1 pp. 76-84, 2003.

BIOGRAPHIES

Farzin Yaghmaee received his B.Sc. from AmirKabir University of Technology, Iran, and M.Sc. and Ph.D. both in Artificial Intelligence from Sharif University of Technology, Iran, in 2002 and 2010, respectively. He is now a faculty member of Electrical and Computer Engineering Department of Semnan University, Iran. His research interests are: image and video processing, text mining and Persian language processing tools.

Mohadese Kamyar received her B.Sc. in 2012 and M.Sc. in 2015 and both in Artificial Intelligence from Semnan University, Iran. Her research interests include image Processing and Persian language processing tools.

How to cite this paper:

F. Yaghmaee and M. Kamyar, "Introducing new trends for persian CAPTCHA," Journal of Electrical and Computer Engineering Innovations, vol. 4. no. 2, pp. 119-126, 2016.

DOI: 10.22061/jecei.2016.572

URL: http://jecei.srttu.edu/article_572.html