A Korean CAPTCHA Study: Defeating OCRs In a New CAPTCHA Context By Using Korean Syllables

Tae-Cheon Yang
Department of Computer and Information Science
Kyungsung University, Busan, Republic of Korea

Ibrahim Furkan Ince, Yucel Batu Salman
Graduate School of Digital Design
Kyungsung University, Busan, Republic of Korea

ABSTRACT

Internet is being used for several activities by a great range of users. These activities include communication, e-commerce, education, and entertainment. Users are required to register regarding website in order to enroll web activities. However, registration can be done by automated hacking software. That software make false enrollments which occupy the resources of the website by reducing the performance and efficiency of servers, even stop the entire web service. It is crucial for the websites to have a system which has the capability of differing human users and computer programs in reading images of text. Completely Automated Public Turing Test to Tell Computers and Human Apart (CAPTCHA) is such a defense system against Optical Character Recognition (OCR) software. OCR can be defined as software which work for defeating CAPTCHA images and make countless number of registrations on the websites. This study proposes a new CAPTCHA context that is Korean CAPTCHA by means of the method which is splitting CAPTCHA images into several parts with random rotation values, and drawing random lines on a grid background by using Korean characters only. Lines are in the same color with the CAPTCHA text and they provide a distortion of image with grid background. Experimental results show that Korean CAPTCHA is a more secure and effective CAPTCHA type for Korean users rather than current CAPTCHA types due to the structure of Korean letters and the algorithm we are using: rotation and splitting. In this paper, the algorithm of our method is introduced in detail.

Keywords: *Korean CAPTCHA, OCR, Information Security, Pattern Recognition, Artificial Intelligence.*

1. INTRODUCTION

Most of the daily activities such as education, shopping or commerce are being carried out through the Internet. Users are commonly asked to fill out registration forms by entering required information to be able to operate specific tasks on the web sites. However, registration can be done by automated hacking software. Some people commit vandalistic acts such as attacking web sites with computer programs, and even can stop the running of the web site. These programs automatically fill out a form with wrong information to get in the web site. Therefore, web site holders are supposed to take precautions against those attacks for security.

Several defense systems have been proposed and presented in order to prevent such attacks. It is crucial for the websites to have a system which has the capability of distinguishing human users and computer programs in reading images of text. CAPTCHAs are challenge puzzles used to determine whether a user is human or not [1]. Intuitively, a CAPTCHA is a program that can generate and grade tests that most humans can pass but current computer programs can not pass [2]. It stands for **C**ompletely **A**utomated **P**ublic **T**uring **T**est to **T**ell **C**omputers and **H**uman **A**part, and Public means that the code and the data used should be publicly available [3]. There are several types of testing methods such as pictures of objects, distorted text, or even audio clips for impaired users.

A more technical definition of CAPTCHA is provided in [4]: "CAPTCHA is a cryptographic protocol whose underlying hardness assumption is based on an AI problem". The most common applications for practical security by CAPTCHA test

* *Corresponding author. E-mail : tcyang@ks.ac.kr*
Manuscript received May. 11, 2009 ; accepted Aug. 11, 2009

This work was supported by Kyungsung University Foundation Grant (2003).

include online polls, free email services, shopping agents, search engine bots, worms and spam, and preventing dictionary attack [4]. For instance, email provider services such as Hotmail and Yahoo provide a CAPTCHA test as a final step of the registration process to stop bots from subscribing and using their resources for spam distribution.

Turing test is used for providing the intelligence of a computer in the domain of Artificial Intelligence (AI). Turing tests use a method which put a human user and a computer in different rooms. There is also third room for the human interrogator to ask them questions. If the interrogator can not recognize the locations of human and computer, it results that the computer has passed the Turing test. CAPTCHA is a Turing test but it is quite different than the definition above. If the interrogator is replaced with computer rather than a human, then it is called as CAPTCHA. The main function of this method is human user can easily answer the interrogator's question but present computer programs are hardly or never can answer [5].

One of the methods used in CAPTCHA is implementing the images of words. This method is based on the weak points of Optical Character Recognition (OCR) programs. OCRs are software to work for defeating CAPTCHA images and make countless number of registrations on the websites. OCRs can recognize the high quality texts using the common formats and standards [5].

It will be more secure to add noisy backgrounds, colors and increasing the level of distortion against character recognition programs. It is difficult for them to read low-quality text and the manuscripts.

This paper introduces a Korean CAPTCHA study such that the regarding method had been already introduced before [14]. The method splits the image into several parts in random width and height values. Additionally, it rotates the split characters in random rotation angles that yield a particular distortion in the image. It is very difficult for OCRs to find out where characters are split and the end points of each image because of the random rotation. It would be very expensive to write an OCR algorithm to defeat our method. CAPTCHA is composed of many images with random rotation values. The proposed method was implemented by the PHP (Personal Home Page) programming language by using only the Korean characters. Many studies have been done so far in terms of CAPTCHA but never had been done with a language apart from English. In this study, we are showing how it is possible to create a Korean CAPTCHA system with so many system outputs in different rotation and splitting conditions. Section 2 introduces the previous studies on CAPTCHA. Design principles are presented in Section 3, and the details of our algorithm are explained in Section 4. Experimental results are given in Section 5 and last section concludes the study, and shows the strengths and weaknesses of our method.

2. PREVIOUS STUDIES

Considerable number of studies was conducted by researchers on developing new CAPTCHA methods and breaking them. There are mainly three types of CAPTHA; (1) text-based schemes, distortion of text images to avoid pattern recognition programs to understand; (2) sound-based schemes, asks users to recognize speech; (3) image-based scheme, asks users to recognize images. CAPTCHAs were originally developed by AltaVista to avoid the submission of URLs to the search engine [6]. It was a simple CAPTCHA which asks users to type a distorted English word.

Carnegie Mellon designed the Gimpy method which selects a word from dictionary and asks users to type what they see as an image after rendering the distorted image containing the text [7].

Yahoo uses the simple version of this method; EZ-Gimpy. EZ-Gimpy's image modification includes background grids, gradients, non-linear deformations, blurring, and pixel noise. Most humans can read three words from the distorted image, while current computer programs can not.

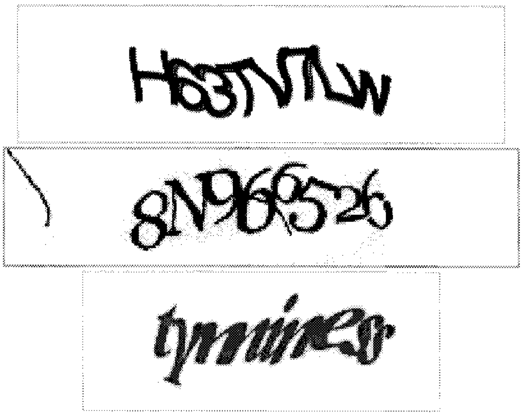

Fig. 1. Some CAPTCHA words of Yahoo[9], Hotmail[10] and Gmail[11] respectively.

PessimalPrint was developed in 2002 by PARC which uses the major weaknesses of OCR systems such as the inability to recognize low quality images [8]. It contains only common English words between five and eight characters long. PessimalPrint used only 70 words which is very low. PessimalPrint's CAPTCHA would break with the probability of 1/70. So, this method does not succeed as expected.

Hotmail is a free email service by the Microsoft Cooperation, and another CAPTCHA method is used [10]. A string of English characters is randomly selected, and after applying some changes, users are asked to type what they see. The major disadvantage of this method is some of the characters are read differently because of putting curves between characters [5].

BaffleText uses non-English pronounceable character strings to defend against dictionary attacks, and Gestalkt-motivated image-masking degradations to defend against image restoration attacks [12].

The produced word sometimes caused difficulties to human users. There are also some other CAPTCHA methods which are using picture or sound features to tell human users and computers apart. PIX is a recognition method which uses usual pictures instead of pictures of words [1]. However, this method requires a large sized space to store the pictures. In text-to-speech method, again instead of showing an image, a sound is played for the users and asked them to recognize it [13]. After understanding what the word is by the sound, users are supposed to type it correctly to continue their process. Yet, similar with PIX, this method also requires a great space and expense.

Poorly implemented CAPTCHAs can be broken easily even without using character recognition software. Some of the first generation CAPTCHAs has already been broken, so the new generation should be more powerful and complex to avoid from the attacks.

3. DESIGN PRINCIPLES

There are a number of important characteristics that a CAPTCHA can exhibit. These include the difficulty to be solved by OCR and any attack programs, readable common distortions, resisting malicious attacks, carrying many bits of information, the capability of coexisting with other CAPTCHAs, and little cognitive computation requirement by the user. The relative importance of these characteristics depends on the CAPTCHA type. The principles behind CAPTCHA are as follows:

- The user is presented with a garbled image on which some text is displayed. This image is generated by the server using random text.
- The user must enter the same letters in the text into a text field that is displayed on the form to protect.
- When the form is submitted, the server checks if the text entered by the user matches the initial generated text. If it does, the transaction continues. Otherwise, an error message is displayed and the user has to enter a new code.
- It must be a variant of a well known Turing Test.
- The system must be effective at keeping out machines.
- The system must be more tolerable to human users.

4. SUGGESTED ALGORITHM

In this paper, a new method has been developed for differing human users and computer programs from each other by mainly splitting CAPTCHA image into several parts with rotation and drawing a great deal of lines and circles randomly to the background. Additionally, a grid effect has been added to the background. Lines and circles have been randomly drawn in the color of text so that OCR program confuse while distinguishing which one is character or not.

In our method, CAPTCHA text consists of the characters and numbers in appropriate Korean spell structure taken randomly from Korean spells database. The text is composed of up to three spells, and each character has its own bending and size value. Characters are split into several parts and each part is given randomly a rotation value in a certain angle domain interval such as: [-1°, 1°], [-3°, 3°], [-5°, 5°]. Image parts are also split individually with random width and height values which provide an extra difficulty for OCR programs while finding the start and end of the images. Rotation in character parts provides confusion in recognizing the exact one.

The text shown in Figure 2 below is indeed '좌득운'. This text is easily recognizable by the human but not OCR program. This CAPTCHA image is split into 8 parts as (4 X 2) matrix shape and each split has a random rotation angle value between -3 and 3 degree. Splits have random width and height values. Background and CAPTCHA text are in similar colors.

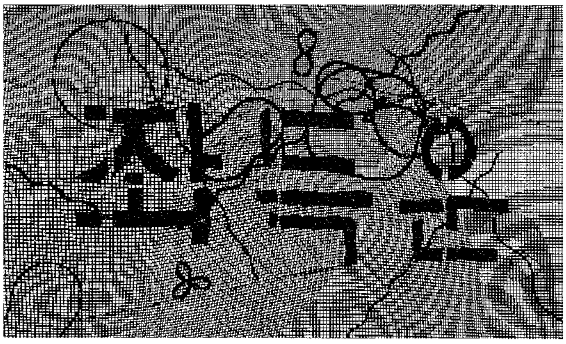

Fig. 2. Drawing a Korean CAPTCHA image.

There is a grid in black color at the background. Lines and circles are also drawn in black color such as the CAPTCHA text. When you look at the characters, it is not easy to recognize the letter exactly due to rotation and splitting of character image.

The programming steps of the algorithm that we developed to generate Korean CAPTCHA images are given with pseudo code and runtime output screenshots as in follows;

Step 1. Start the session.

Step 2. Get *n* Korean syllables from the syllables database.

// Preparation of the database is already assumed that it has been done by consisting of regular and proper Korean syllables such that they can be combined with each other regardless of looking at the meaning, but rather they have to be easy for typing in keyboard by Korean users.

Step 3. Create the hash for the random text and put it into the session.

Step 4. Create transparent CAPTCHA image with *w* by *h* image size and add CAPTCHA text over it. Transparent CAPTCHA image with text can be created by specific PHP (Personal Home Page) built-in function: *imagettftext()*.

Step 5. Set the initial X-position and Y-position of captcha image to 0.

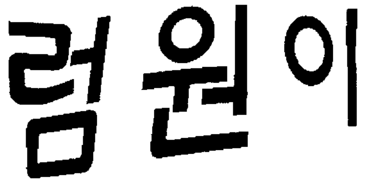

Fig. 3. A sample transparent CAPTCHA Image (600 x 400) with Randomly Assigned Text in Step 4.

Step 6. Split the captcha into *k* by *l* Matrix shape by dividing the captcha width into *k* parts and the height into *l*.

Step 7. Start a loop from 0 to k*l

// After completing the first row in order to split into k parts, then pass to next row.

If (i+1) Mod k+1 = 0 Then
Set initial X-Position to 0 and initial Y-Position to Split Height (Image Height / l)

End If

Step 7.1. Create an array to put the split parts and put the split images into array.

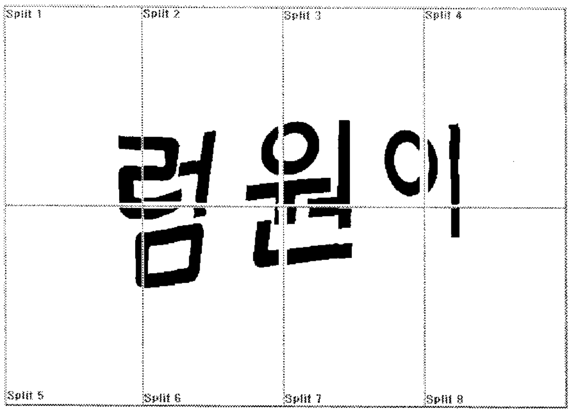

Fig. 4. Transparent split CAPTCHA image without rotation in Step 8.

Step 7.2. Randomize integer between *-d* and *d* to give random rotation to the splits. Rotate the splits with randomized variable that is random between *-d* and *d*.

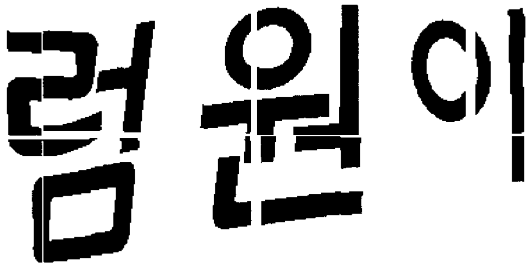

Fig. 5. Transparent split CAPTCHA image with rotation degree between -1 and 1.

Step 7.3. To pass to another split in one row, increase the initial X-Position by Split Width (Image Width / k) in each loop step.

Step 7.4. End Loop.

Step 8. Combine the splits to create new CAPTCHA with split and rotation.

Step 9. Add background to transparent new CAPTCHA image object with randomly drawn lines and special effects (Number of lines=250, line color is black and add grid effect).

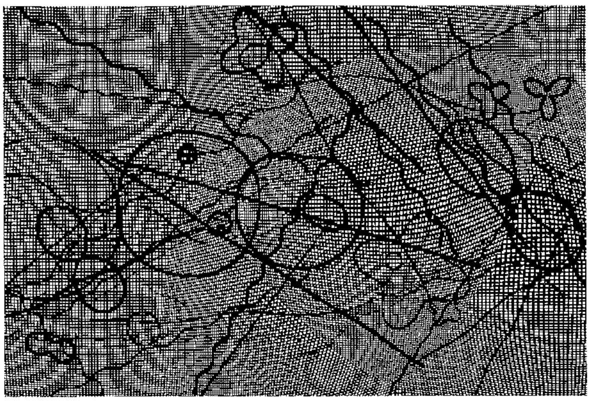

Fig. 6. Background with randomly drawn lines and special effects in Step 9.

Step 10. Export the final CAPTCHA image as a JPEG file in the name of 'captcha.jpg'.

Step 11. End session.

Figure 7 below shows the final output from the Korean CAPTCHA that we developed in this study. When you split the Korean letters with rotation, each broken letters look like a different Korean letter. This confuses the OCR programs. Using Korean syllables rather than using only Korean letters provides a meaningful perception of the CAPTCHA by the user.

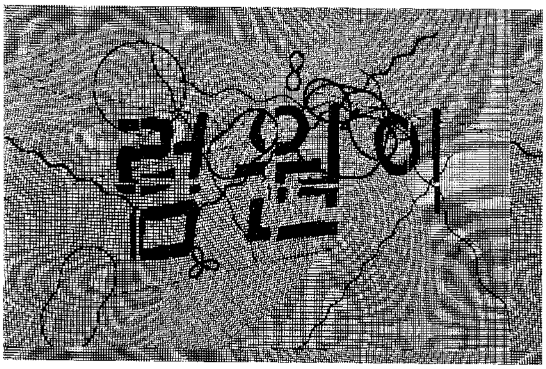

Fig. 7. Final Korean CAPTCHA output with background.

5. EXPERIMENTAL RESULTS

The Figure 8, Figure 9 and Figure 10 show the outputs of our model with different parameters. As it was mentioned in the previous section, characters are split into several parts and each part is given randomly a rotation value in a certain angle domain interval such as: [-3°, 3°], [-1°, 1°], [-5°, 5°].

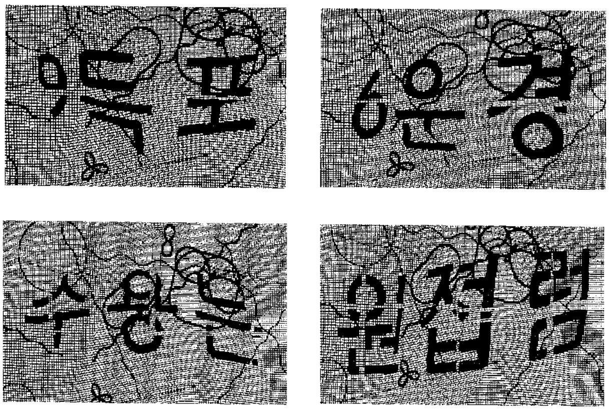

Fig. 8. 2 X 2 matrix shape split CAPTCHA with rotation angle between -3° and 3°.

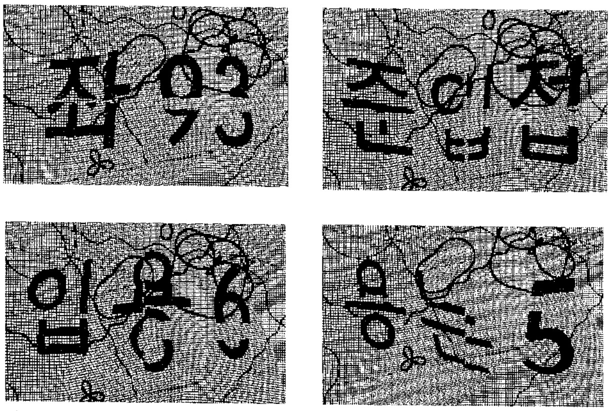

Fig. 9. 4 X 2 matrix shape split CAPTCHA with rotation angle between -1° and 1°.

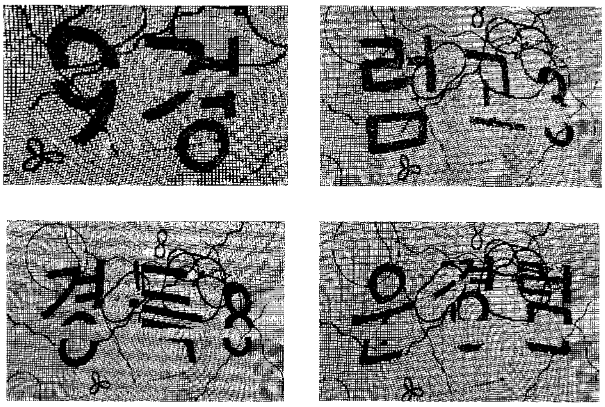

Fig. 10. 4 X 2 matrix shape split CAPTCHA with rotation angle between -5° and 5°.

Above results have been obtained in local host of (127.0.0.1) Apache Server. Compiling and run-time durations have been held as 1.43 seconds for each session. Results show that splitting with 4 X 2 matrix shape yields more splits that are rotated. This produces a more distorted CAPTCHA image than the CAPTCHA with 2 X 2 matrix shape CAPTCHA. In addition, experimental results indicate that the relation between matrix shape and rotation angle is also important. A proper balance between rotation angle and matrix shape gives the most convenient and reliable Korean CAPTCHA. According to the test results, 4 X 2 matrix shapes split Korean CAPTCHA with rotation angle between -1° and 1° yielded the best results in terms of best perception of understanding by the user and difficulty to be defeated by the OCR programs.

6. CONCLUSION

The concept behind in this field started from the problems faced by major Internet companies such as Yahoo and AltaVista. Solution was generated by asking human users to solve a CAPTCHA test before they involve the online activities. The defense systems and attacks are still dynamically updated. In this paper, we demonstrated a Korean CAPTCHA technique which works for distinguishing a human user from harmful computer programs.

However, this study did not claim the fact that Korean characters are more necessary than English characters in CAPTCHA context. This paper proposed a kind of Korean CAPTCHA that has not been tried before in the literature. Experimental results show that Korean CAPTCHA is a more secure and effective CAPTCHA type for Korean users rather than current CAPTCHA types due to the structure of Korean letters and the algorithm we are using: rotation and splitting. If you split the Korean letters with rotation, each broken letters look like a different Korean letter and this confuse OCR programs. In terms of using Korean syllables rather than using only Korean letters, it makes a meaningful perception of the CAPTCHA by the user. Hence, Korean CAPTCHA has two main advantages which are the security it provides and user-friendly interface for the Korean users.

We did not design this system for non-Koreans. This system was proposed only for Koreans. However, it can be used for all the language types as well. If the database of characters is changed to English characters instead of Korean letters, then system can be applied to any users as we have already published in ICCIT 2008 [14]. The deal at current paper is just to show the importance of Korean CAPTCHA in terms of security, robustness, and interaction with the user. In terms of using Korean syllables rather than using only Korean letters, it makes a meaningful perception of the CAPTCHA by the user. The user can predict what the next character is by means of meaningful syllables, however; an OCR can not. And also, split and rotated Korean letters are so confusing to be recognized by any kind of pattern recognition software as well as OCR. Hence, Korean CAPTCHA is more secure, more robust and more user-friendly than the CAPTCHA in other languages because of its syllables and shape based structure.

There are strengths and weaknesses of our new model. As the primary advantage, the usage of characters in the images is recognizable by human readers and easy to read. Our new CAPTCHA method use same input methods similar with the other many well known web sites and services where users type some keywords or characters into input boxes. Therefore it can be said that it is easy to learn and use by regular Korean users. It can be used by all ages; even children can easily learn the system without any training.

The algorithm of this method makes it hard to read by OCR programs which prove its safety. Finally it needs less processing requirements and can be operated in small size of bandwidth.

On the other hand, it can be understandable by computers with using powerful and intelligent software and hardware by removing the noise effects. This is the major disadvantage of our model. Secondly, in some cases some patterns could be hard to read by older and disabled human users. In addition, part of our future work is to develop 3D Korean CAPTCHA.

7. REFERENCES

[1] Blum, M., 2000, The CAPCTHA Project, Completely Automatic Public Turing Test to Tell Computers and Humans Apart", *Dept. of Computer Science, Carnegie-Mellon University*, http://www.captcha.net.

[2] Athanasopoulos, E., Antonatos, S., "Enchanced CAPTCHAs: Using Animation to Tell Humans and Computers Apart", *LNCS, 4237*, 2006, pp. 97-108.

[3] Wang, S., Baird, H., Bentley, J., "CAPTCHA Challenge Tradeoffs: Familarity of Strings versus Degradation of Images", *the 18th International Conference on Pattern Recognition, ICPR'06, IEEE*, 2006.

[4] Von Ahn, L., Blum, M., Nicholas, J.H., Langford, J., "CAPTCHA: Using Hard AI Problems for Security", *In Proceedings of Eurocrypt*, 2003, pp.294-311.

[5] Shahreza, M., Shahreza, S., "Preventing Mobile Software Cracking Software", *IEEE, Innovations in Information Technology, Dubai*, 2006, pp. 1-5.

[6] Moy, G., Jones, N., Harkless, C., Potter, R., "Distortion estimation technique in solving visual CAPTCHAs", *Proc. of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol.2*, 2004, pp.23-28.

[7] G. Mori, and J. Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA", *Proc. of IEEE CS Society Conf. on Computer Vision and Pattern Recognition, Madison*, 2003, pp. 134-141.

[8] Coates, A.L., Baird, H.S, Fateman, R.J., "PessimalPrint: A Reverse Turing Test", *Proc.of 6th Int. Conf. on Document Analysis and Recognition, Seattle, WA*, USA, 2001, pp.1154 – 1158.

[9] Yahoo! mail, http://mail.yahoo.com [06/10/2008]

[10] Microsoft Hotmail, http://www.hotmail.com [06/10/2008]

[11] Google Gmail, http://mail.google.com [06/10/2008]

[12] Chew M. and Baird H. S., "BaffleText: a Human Interactive Proof", *Proc of 10th SPIE/IS&T Document Recognition and Retrieval Conf. (DRR2003), Santa Clara*, CA, 2003, pp. 305-316.

[13] Chan, T.Y., 2003, "Using a Text-to-Speech Synthesizer to Generate a Reverse Turing Test", *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 226 – 232.

[14] Ince, I.F., Yengin, I., Salman, Y.B., Cho, H.G., Yang, T.C. "Designing CAPTCHA Algorithm: Splitting and Rotating the Images against OCRs", *International Conference on Convergence and Hybrid Information Technology, ICCIT 08*, IEEE, 2008.

Tae-Cheon Yang
He received the B.S. in computer science from Kyungpook National University, Korea in 1982 and also received M.S., Ph.D. in computer science from Korea Advanced Institute Science and Technology, Korea in 1984, 1994 respectively. Since 1985, he has been with the Dept. of Computer Science, Kyungsung University. His main research interests include Computational Geometry, Algorithms and Computer Graphics.

Ibrahim Furkan Ince
He received the B.S., M.S. in computer engineering from Bahcesehir University, Turkey in 2006, 2008 respectively. Since 2008, he has been studying as phD student in digital design from Kyungsung University, Korea. His main research interests include computer graphics, computer vision and image processing.

Yucel Batu Salman

He received the B.S., M.S. in computer engineering from Bahcesehir University, Turkey in 2003, 2006 respectively. Since 2006, he has been studying as phD student in digital design from Kyungsung University, Korea. His main research interests include healthcare systems and its applications, software usability, ubiquitous technologies, and human-computer interaction.