2013 17th International Conference on Information Visualisation

A Multilingual Handwriting Approach to CAPTCHA

Amalia Rusu
Fairfield University
Fairfield, CT, USA
arusu@fairfield.edu

Steve Mislich
Greenville College
Greenville, IL, USA
200906004@panthers.greenville.edu

Lukas Missik
Harvard University
Cambridge, MA, USA
lukasmissik@college.harvard.edu

Benjamin Schenker
Cooper Union
New York, NY, USA
schenk@cooper.edu

***Abstract*—Nowadays World Wide Web makes use of Human Interactive Proofs and CAPTCHA to distinguish legitimate users from bots. This paper presents an overview of current systems and their weaknesses and proposes a unified theory for a multilingual handwriting CAPTCHA. We further describe our French and Spanish CAPTCHAs, as an extension to the English version, with the potential of being more useful on websites in those languages. Highly interdisciplinary techniques used to generate our CAPTCHAs are also discussed.**

Keywords-CAPTCHA; Gestalt laws of perception; HCI; Web security; visual analytics; handwriting recognition

I. INTRODUCTION

Designing and developing web interfaces that embed a method to prove that a user of an Internet service is human became necessary with the rise of spam on the web and services that need protection against automatic bots. A method for implementing such interfaces will consider administering a test before anyone can post on a protected web forum or register for a new email address or purchase show tickets online. Without this measure, a spammer could easily flush a message board with junk or create an endless number of email addresses for nefarious purposes or buy all of the tickets for an event and driving up the prices artificially. While this procedure does not stop malicious programs completely, it slows the speed of spam generation considerably and helps prevent wrongful use of web services.

A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a specific test for the purpose of identifying a user as a human over the Internet and about 200 million are solved daily online [1]. The general problem is that computer programs can perform instructions many orders of magnitude faster than humans can and may access ecommerce sites that are meant for human use only. Therefore, many web services need to be able to identify which subset of their end users are humans without the need of a special form of unique identification. A CAPTCHA fills this role by making the user respond to a test that can distinguish between humans and computers with a very high confidence.

Visual thinking or pattern finding is a task with central role in all cognitive systems. At the minimum, visualization contains symbols of various types, such as words or shapes. If the symbols are familiar then the working memory assists people even on complex structures and visual recognition is more effective. CAPTCHAs reinforce human cognition and exploit human perception abilities over machines generally in visual or audio interpretation. However, the current options for CAPTCHA are deficient in many areas, most notably, usability. With advancement of computer technologies and finding more ways to break CAPTCHAs, the ones that are currently in use have to adapt to stay useful and relevant and it becomes more difficult for human users to successfully pass the tests due to increased test difficulty. Moreover, this can be exaggerated with non-native English speakers or the visually impaired.

As a result of these challenges to the current state of CAPTCHA, we have explored the use of a multilingual handwriting CAPTCHA where users are authenticated as humans to gain access to Web services by correctly interpreting the handwriting samples, in their own language, transformed according to specific principles of cognitive psychology. In the context of CAPTCHA, handwriting is far better than print because computers have great difficulties with the segmentation of the individual letters, while humans can read it naturally. Furthermore, by generating synthetic handwriting of any script and use it as the basis for CAPTCHA generation, the added support for multiple languages lessens the burden for non-native English speakers and makes multilingual CAPTCHA widely applicable. In this paper we focus on a more effective, user-centered approach to create CAPTCHAs and discuss its superiority from the perspective of cognitive psychology and pattern finding.

II. BACKROUND

The general World Wide Web (WWW) requires systems to protect available resources from unauthorized use. Human Interactive Proofs (HIP) systems are intended to distinguish one class of users from another on the WWW without prior knowledge of passwords or use of biometrics [2]. HIP systems rely on human computer interaction techniques and use of CAPTCHA.

A. CAPTCHA

CAPTCHA creation is a highly interdisciplinary pursuit. It involves techniques from many disciplines (and exploits some of their open problems) such as Artificial Intelligence (AI), Human-Computer Interaction (HCI), Information

1550-6037/13 $26.00 © 2013 IEEE
DOI 10.1109/IV.2013.26

Visualization, Cognitive Science, Psychology, Natural Language Understanding, Pattern Recognition, Web Security, and others. Several criteria for a CAPTCHA system are as follows: the generation and grading of the test must be fully automated and scalable; humans should be able to pass it virtually all the time and computers should fail; this should hold even when an attacker (i.e., a programmer trying to teach a computer to beat the CAPTCHA) has access to everything that the CAPTCHA system uses [3].

In a CAPTCHA system, an acceptable passing rate for computers is usually 1 for every 10,000 tests or lower, which guarantees a significant deterrence for attackers [4]. While a .01% success rate is targeted, there is precedent to have schemas that exaggerate the gap in ability between humans and computers, requiring a tester to pass the test twice if they have failed it many times [5]. This extra aspect can raise the acceptable passing rate to .6% or 1 out of 167 tests [6].

In the case a of a text-based English CAPTCHA that uses randomly generated nonsense words of length n characters, a randomly guessing attacker would have a 1 in 26^n chances of passing. If the truth word was 2 characters long or greater then the criteria would be properly fulfilled. Unfortunately there are many problems with this setup and not just the brute force approach should be considered. These days, even with transformations, attackers can use optical character recognizers (OCRs) on the machine-printed text CAPTCHA images with impressive results. Furthermore, humans have greater difficulty reading nonsense words as opposed to dictionary words once image transformations have been applied on top of them. This has led to the practice of most current text-based CAPTCHAs utilizing longer (dictionary or nonsense) words in order to beat OCRs but nonetheless making the task harder for humans as well.

Different types of CAPTCHAs exist and they exploit various open problems within AI. The most common rely on the weaknesses within OCR of machine printed text. Some examples include earlier versions such as GIMPY, Yahoo's EZ-GIMPY, and other. ReCAPTCHA provides tests that contain words that an OCR came across while digitizing books. When enough humans agree on the word then the program has recognized the word using human computation. This puts human time spent completing reCAPTCHA not only to keep the web safe but also toward helping digitize books. While this is the gold standard, it has been increasingly hard to manage some of the low complexity CAPTCHAs because of recent advances in attackers' methods. As a result of usability issues of the machine printed text-based CAPTCHAs, research has been done toward finding and creating other types of CAPTCHAs that may be easier for humans to complete but harder for computers.

Alternative types of CAPTCHA exploit weaknesses in computer vision and object recognition. These CAPTCHAs usually consist of images that contain specific objects or various types of objects and ask users to answer questions about the images. An example is Asirra [5], which asks users to differentiate between pictures of cats and dogs. However, it was soon shown to be unsatisfactory as a CAPTCHA because machine learning algorithms were developed that differentiate between objects. Another example of a CAPTCHA that exploits object recognition is ARTiFACIAL [7], which hides an image of a face in a complicated picture and it asks the user to locate that face. The same researchers who broke ARTiFACIAL created CORTCHA [6] that requires a user to place a picture of an object onto its proper place in a larger image with proper context. Although not broken, CORTCHA has yet to be implemented in any large scale system. A further example is Google's Orientation CAPTCHA, which asks users to rotate images until they are upright [8]. Nonetheless many of these have been either broken or suffer from usability or scalability issues.

B. Handwriting

Handwriting is considered the style a person writes with. There is no unique form of handwriting as each person has an individual style, which may also differ from script to script. Numerous variables are considered to describe an individual's handwriting: the roundness/sharpness of specific letters; the spacing or ligatures between letters; the slope of the letters; the pressure from the pen or pencil to the paper; the average letter size, etc. Handwriting is known as cursive writing that tends to be more easily written than normal print. Cursive involves connecting each letter, as opposed to lifting the pen or pencil off the paper to create the next character. Those proficient in both normal print and cursive writing usually find the cursive style of writing to be faster. Yet while cursive is a faster way of writing, it also has some drawbacks, the most noticeable being its variability and legibility. So why is handwriting useful for CAPTCHA purposes? It adds a level of complexity by making character segmentation difficult so that it exceeds what computers can accomplish (i.e., the state of the art level of handwriting recognition systems) but still remains in the range of human accessibility. This means that the handwritten text images could have fewer (if any) transformations than a machine printed CAPTCHA and thus remain secure while being easier for humans to correctly pass.

For CAPTCHA, a large set of data is required to be generated on the fly, and gathering such a large amount of data from humans is not feasible. For such a task, synthetic handwriting becomes very helpful. Previous research work has been done by the first author on synthetic handwriting generation for English characters and words [9]. We used this as the basis for our current work and in this paper we expanded the handwriting generator to other scripts, such as French and Spanish, with the potential expansion to any script with minor changes applied.

C. Gestalt-based Transformations

CAPTCHA generation algorithms inherently rely on the ability of humans to recognize and retrieve information from visual scenery in ways that the best computer algorithms still cannot replicate. Some of the most popular transformations in early development of CAPTCHA were adding visual noise to images, obscuring and hiding the words located inside. Other transformations include blurring or smudging the image, similarly obfuscating its contents. However, these methods have proven to be unreliable. Computer algorithms can now be very proficient at cleaning images of noise and interpreting the boundaries of smudged or otherwise distorted images. A naive approach would be to add more extreme transformations in an effort to stay ahead of the curve in optical character recognition capabilities. But the introduction of more noise and blurring causes the images to be much more difficult for humans to solve as well.

In our paper, we present a number of transformations that, rather than simply assuming human recognition exceeds machine recognition and thus applying more distortions, specifically target the areas in human psychology which allow us to extract additional information from images in ways that computers cannot. The Gestalt Laws of Perceptual Organization describe ways that the human brain is able to process (often noisy) visual information and make inferences about its content even when that information is not explicitly contained in the image [10]. As a simple example, humans are able to perceive dotted lines as continuous, even though there is no explicit connection between adjacent dots.

When constructing our image transformations, we took advantage of these specific qualities inherent in all humans to design distortions that impair human readability as little as possible while still producing a sharp decrease in the ability of OCRs to interpret the encoded information. The Gestalt principle of proximity allows humans to link objects that appear close together in an image; the principle of continuity lets us distinguish between overlapping contours and determine the membership of intersecting lines; and the principle of closure lets us make inferences about the boundaries of objects even when those boundaries are interrupted, broken, or simply missing.

One last advantage of humans is the ability to engage in top-down processing. This is the process where a user's expectations and previous experience and knowledge actually influence their perception. This phenomenon allows people to correctly reconstruct words they recognize even when some characters are missing or heavily distorted. We can take advantage of this feature of human perception by allowing transformations without significantly impairing human readability and therefore use real words that most users would be familiar with in their native language. The possession of a sizable vocabulary and knowledge of common consonant-vowel pairs and word roots are only able to help an individual when they are solving a CAPTCHA in a language they are fluent in. The motivation for multilingual CAPTCHA becomes more clear when considering this factor; without the proper knowledge of a language, top-down processing cannot occur and recognition is no better than decoding random combinations of letters.

Using these specific techniques, we were able to significantly distort our images beyond the recognition capabilities of machines without drastically impairing human readability.

D. French and Spanish Scripts

French and Spanish CAPTCHAs were considered due to their similarities with the English language. All these languages use the Latin script, a writing system based on the classical Latin alphabet originating from around the 1st century BC. Other examples of languages that use this script are Portuguese, Swedish, and Latvian. These scripts were narrowed down by the volume of people who speak each language and the number of websites that feature the language. As of 2005, there were a total of 420 million Spanish speakers and 130 million French speakers worldwide [11]. While the number of Portuguese speakers totals 213 million, it was discovered that only 2% of websites are written for Portuguese speakers, compared with the 4.5% of websites written for French speakers and 4.7% of websites written for Spanish speaking users. English websites take up 54.7% of the WWW [12]. Thus the two languages close enough to English and deemed to receive the most benefit from having an efficient synthetic handwritten CAPTCHA system were French and Spanish.

The English alphabet consists of 26 characters and no diacritics (accents). The French alphabet is also composed of the same 26 characters, but includes 5 diacritics: the acute accent (example: é), the grave accent (à), the circumflex (ô), the diaeresis (ÿ), and the cedilla (ç). French also includes 2 ligatures (2 letters joined as 1), Œ and Æ. Lastly, the Spanish alphabet consists of 27 letters (the 26 from English plus Ñ), 1 diacritic (the acute accent) and 2 digraphs, Ch and Ll. As the digraphs are simply a combination of letters, we do not treat them as special characters.

These special characters come up often in their respective languages. For example, the most frequent verb, and the second more frequent word, in French is "être", which has the meaning of "to be," and features the circumflex diacritic. In Spanish, "qué" means "what" or "which," and "año" means "year." Thus it can be said with confidence that native speakers of French and Spanish use special characters and accents frequently, and therefore a proper CAPTCHA system for each language must include these added characters and accents.

III. Multilingual CAPTCHA System

The overall architecture of our HIP system using multilingual CAPTCHAs is presented in Figure 1 with all elements further discussed.

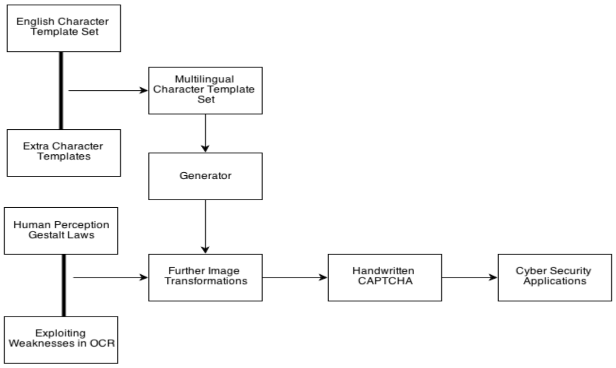

Figure 1. Overall system architecure.

A. System Design

The system consists of three components: a method for creating character templates, a method for generating linked cursive letters ('words') in an image utilizing those character templates, and a method for applying image-level transformations to further obfuscate the words (Figure 2).

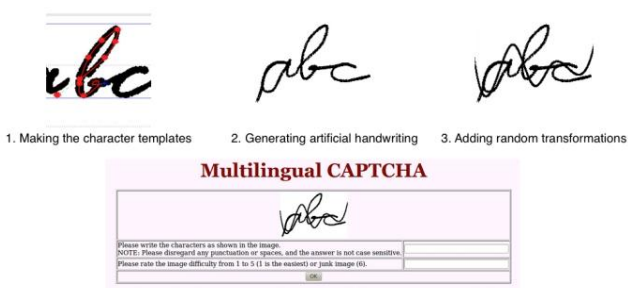

Figure 2. CAPTCHA system components.

First, character templates are generated for each script or language we support. Each character template has data about where the character starts and ends relative to baselines and it has data consisting of the paths between the starting and ending points. The character templates are created using the tracing module, which allows for a user to indicate the required points, paths, and baselines overlaid on top of an image of the character. This helps the character template have the most accurate data and leads to more realistic looking writing samples. The character templates also store information about the in and out ligatures of each letter used for linking them together as they would be in cursive handwriting.

Once all of the templates have been created, the generator module can be used to generate strings of linked cursive letters. The module draws each character, linking it with its neighbor with the in and out ligature information, while trying to make the transition as smooth and natural looking as possible. The module generates words with many different randomized factors such that each time the same word is generated it looks different. There are about thirteen of these parameters and their effect can be controlled through variables before generation. These variables control for instance line thickness, baseline variation, character spacing, shear factor, point perturbation, and other properties that are present in natural handwriting. This not only makes each generation unique and natural looking, but it is impossible to reverse-engineer and very difficult for modern OCRs to correctly identify the words in the image. We then compiled a lexicon (dictionary) of words and phrases in each language for which we generated corresponding clean synthetic handwriting images.

The third module takes these images and applies transformations in order to further obfuscate the text in the images. Using factors from cognitive psychology, we apply occlusions and remove or add information in the image in ways such that it is hard for computer programs to segment and interpret but remains easy for humans to read and understand.

B. Image Transformations

To muddle the images, we took advantage of the Gestalt principles of closure and continuity, which let humans extract information from an image that a computer would not necessarily be able to. Some transformations apply occlusions (covering part of the image with white circles), others shift around the contents of an image, and some extract parts of the image. Displacement transform cuts the word horizontally, separating the characters near the middle line, while wave transform adds an opaque wavelike stroke on top of an image, obscuring parts of the characters. Mosaic cuts the image into four segments with a vertical and a horizontal wavelike separation, while circles transform adds opaque circles. Fragmentation applies the edge detection algorithms and can be applied several times to further fragment the image. The overlap transform places a second copy of the image on top of the first, shifted slightly so that strokes overlap.

The goal of these transforms is to distort the image sufficiently so that computer algorithms are unable to recognize the embedded characters, while preserving nearly all of the information in human perception due to the (as of yet) unprogrammable advantages of Gestalt psychology. Most of our transformations in Figure 3-9 (applied on the French word "bonjour" as "hello" in English and Spanish word "cabeza" as "head" in English) rely on the principles of proximity, continuity, and closure. Humans are able to effectively ignore artificial occlusions placed on top of the images as long as the boundaries of each letter are able to be "closed" in perception. For example, a poorly placed white occlusion could make the letter "o" appear to be a "c," but if the same occlusion was placed on the left side of the character, it would indubitably be an "o" because there is no other letter that fits that template.

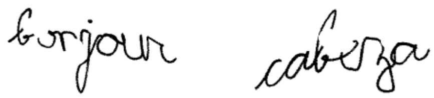

Figure 3. White circles - Humans are able to fill in the missing information with ease due to closure.

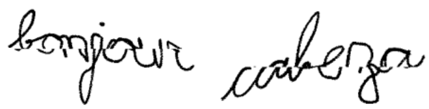

Figure 4. Displacement - Humans are able to connect the similar pieces back together due to proximity.

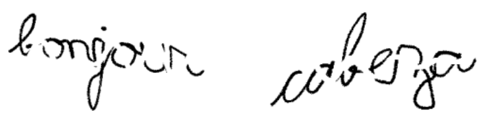

Figure 5. Wave (white) - Humans are able to reconnect the broken strokes due to the principle of closure.

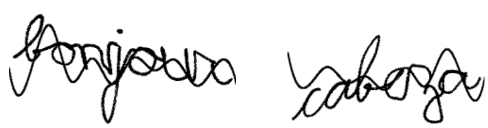

Figure 6. Wave (black) or extra strokes - Humans are able to correctly determine which strokes are part of the word and which are irrelevant due to the principle of continuity.

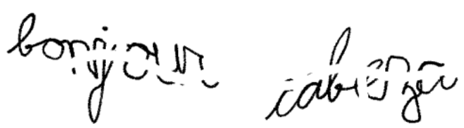

Figure 7. Mosaic - Humans are able to reconstruct the image due to closure and proximity.

Figure 8. Fragmentation (edge detection) - Proximity and closure help humans read these.

Figure 9. Overlap - Segmentation is very difficult for OCRs, but proximity and continuity allow humans to easily merge the two images.

C. *French CAPTCHA*

Considering uppercase and lowercase letters, along with the languages five diacritics, the French alphabet consists of 58 unique characters (the 26 characters of the English alphabet plus 32 characters involving diacritics and ligatures). Figure 10 shows two French CAPTCHAs, one featuring the word “intéresser” (“to interest” in English) with a mosaic transformation applied and another using the word “drôle” (“amusing” in English) with a white wave transform applied.

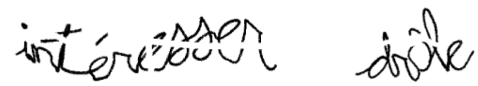

Figure 10. Examples of French CAPTCHAs.

The ICDAR conference sponsors an annual competition on handwriting recognition. At the 2009 competition, nine different systems were used in an attempt to recognize handwritten French in three categories: text pages, words, and isolated characters. When given a choice between the test word and 99 other words from the given lexicon, the best system (TUM) had a success rate between 98.43% and 99.92%. When given the entire lexicon to choose from (1612 words), the best system (still TUM) had a success rate between 93.17% and 98.95%. Lastly, when given a choice between the entire lexicon as well as words from a separate database (for a total of 5334 words), TUM scored an average of a 94.68% success rate [13]. In 2011, at ICDAR, a French handwriting recognition system called A2iA was introduced. Using a lexicon of 7,464 words, A2iA, when using a grapheme based MLP-HMM (Multi-Layer Perceptrons – Hidden Markov Model), had a success rate of 75.1%. While using the 2011 competition lexicon, it had a success rate of 75%. When using a sliding window GMM-HMM (Gaussian Mixture Model – Hidden Markov Model), success rates between 71.4% and 78.5% were recorded. Another system, Jouve, had an error rate of 12.53%, and another system, IRISA, had an error rate of over 20% [14]. Comparing these results with those of the state-of-the-art English handwriting recognition systems they are similar or lower [15] [16]. Because of this comparison and the fact that the larger a lexicon is the worse recognition accuracy a system will have, one may infer that French recognizers’ success on our French handwritten CAPTCHAs to be similar or lower than their English counterparts.

D. *Spanish CAPTCHA*

The Spanish CAPTCHA generation is similar to the French and English versions. Again the main differences come from some accented letters and other characters that are unique to the language. The specific letters that were added to our character template data set are “ÁáÉÊéÍíñÑÓóÚÜúü¡¿”. The overarching importance however, comes from the added value to those for whom English is not their first language. If they are more familiar with Spanish language then the Spanish words should be easier to parse and recognize. This could lead Spanish speaker users to correctly answer CAPTCHA with higher frequency than otherwise. Character tracing as well as synthetic handwriting generation in Spanish use the similar techniques as described for English in [9], with the addition of the few added special characters and tweaked variables.

Several examples of Spanish CAPTCHAs with various image transformations applied are presented in Figure 11. The corresponding truth word in Spanish is “búsqueda”, meaning “quest” in English.

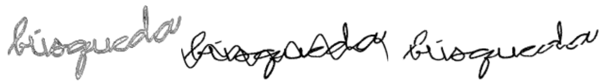

Figure 11. Examples of Spanish CAPTCHAs.

One difficulty with the Spanish CAPTCHAs comes from the accented letters. Some of the transformations that were performed on English CAPTCHAs can obfuscate the small parts of the image that contain the accent, making it difficult to discern the exact characters that are in the image. In general, all of the transformations that remove information from the images have the potential of making Spanish CAPTCHAs harder because of the ambiguity of the accented characters. This includes the extra white and black strokes and the circle transform. However, these transformations may still be considered when used on real Spanish words, because humans could utilize Gestalts and information from the surrounding letters. This is not as much of a problem with the transformations that just rearrange information, such as the displacement, fragmentation, mosaic, or overlap transformations.

From the viewpoint of OCRs, the Spanish cursive CAPTCHAs are at least as hard as English cursive CAPTCHAs. Moreover, the presence of the accented letters adds a challenge of identifying the accents and which letter they belong to in the image.

IV. Conclusion and Future Work

Handwriting recognition (of any script) continues to be a challenging task despite extensive research for many years, especially in context-free applications (such as CAPTCHA) where there is no aid of a lexicon or the lexicon is of the size of an entire dictionary. Comparing humans and machines performances in interpreting handwriting reveals a great advantage for humans especially because cognitive factors allow humans to apply Gestalt principles, engage in top-down processing, and use memory. Thus handwriting is a viable candidate for use in CAPTCHA systems with the ability of creating infinitely many samples through a synthetic handwriting generator. We have expanded our previous work on English handwritten CAPTCHA and in this paper described our efforts toward a multilingual CAPTCHA by including two additional Latin scripts, French and Spanish. We presented the main components of our system and provided rationale for their strengths. We have developed a testing website and invite both users and handwriting recognition programs to try our French and Spanish CAPTCHAs at http://cs.fairfield.edu/~rusu/captcha1.php and provide feedback. In our future work we will reconsider some of the transformations and generate CAPTCHAs for scripts of non-Latin origin.

Acknowledgment

This work was supported in part by a grant from the National Science Foundation (DMS-1004346). The authors would like to thank Fairfield University for hosting the project and particularly to Shawn Rafalski and Kaveh Ahmadian for technical support along the way.

References

[1] Captcha website url: http://www.google.com/recaptcha/captcha

[2] H. Baird and K. Popat. Human Interactive Proofs and Document Image Analysis. In *Proceedings of IAPR 2002 Workshop on Document Analysis Systems*, 2002.

[3] M. B. Luis von Ahn and J. Langford. Telling Humans and Computers Apart Automatically. In *Communications of the ACM*, vol. 47, pp. 57–60, 2004.

[4] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski. Designing Human Friendly Human Interaction Proofs (HIPs). In *Proceedings of CHI '05*, pp. 711–720, 2005.

[5] J. Elson, J. R. Douceur, J. Howell, and J. Saul. Asirra: A Captcha that Exploits Interest-Aligned Manual Image Categorization. In *Proceedings of CCS '07*, pp. 366–374, 2007.

[6] B. B. Zhu, J. Yan, Q. Li, C. Yang, J. Liu, N. Xu, M. Yi, and K. Cai. Attacks and Design of Image Recognition Captchas. In *Proceedings of CCS '10*, pp. 187–200, 2010.

[7] Y. Rui and Z. Liu. ARTiFACIAL: Automated Reverse Turing test using FACIAL features. In *Proceedings of the 11th ACM international Conference on Multimedia*, November 2003.

[8] R. Gossweiler, M. Kamvar, and S. Baluja. What's Up Captcha? A Captcha Based on Image Orientation. In *Proceedings of WWW'09*, 2009.

[9] A. Rusu, U. Midic, V. Govindaraju. Synthetic Handwriting Generator for Cyber Security. In *Proceedings of 13th Conference of the International Graphonomics Society*, 2007.

[10] K. Koffka. Principles of Gestalt Psychology. HarcourtBrace. 1935

[11] R. Gordon. Ethnologue: Languages of the World, 15th edition. Dallas, TX, SIL International. 2005.

[12] W3Techs url: http://w3techs.com/technologies/overview/content_language/all

[13] E. Grosicki and E. A. Haikal. ICDAR 2009 Handwriting Recognition Competition. In *Proceedings of International Conference on Document Analysis and Recognition,* pp. 1389-1402, 2009.

[14] F. Menasri, J. Louradour, A. Bianne-Bernard and C. Kermorvant. The A2iA French Handwriting Recognition System at the RIMES-ICDAR2011 Competition. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*. Vol. 8297, 2012.

[15] A. Rusu, A. Thomas and V. Govindaraju. Generation and Use of Handwritten CAPTCHAs. In *International Journal on Document Analysis and Recognition*, Springer-Verlag, Vol. 13, Issue 1, pp. 49-64, 2010.

[16] A. Thomas, A. Rusu and V. Govindaraju. Synthetic Handwritten CAPTCHAs. In *Pattern Recognition,* Vol. 42(12), pp. 3365-3373, 2009.