ARTiFACIAL: Automated Reverse Turing Test Using FACIAL Features

Yong Rui
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
yongrui@microsoft.com

Zicheg Liu
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
zliu@microsoft.com

ABSTRACT

Web services designed for human users are being abused by computer programs (bots). The bots steal thousands of free email accounts in a minute; participate in online polls to skew results; and irritate people by joining online chat rooms. These real-world issues have recently generated a new research area called Human Interactive Proofs (HIP), whose goal is to defend services from malicious attacks by differentiating bots from human users. In this paper, we propose a new HIP algorithm based on detecting human face and facial features. Human faces are the most familiar object to humans, rendering it possibly the best candidate for HIP. We conducted user studies and showed the ease of use of our system to human users. We designed attacks using the best existing face detectors and demonstrated the difficulty to bots.

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications – Computer vision, Signal Processing. I.4.9 [**Image Processing and Computer Vision**]: Applications. I.3.3 [**Picture/Image Generation**].

General Terms

Algorithms, Design, Security, Human Factors, and Verification.

Keywords

Human interactive proof (HIP), Web services security, CAPTCHA, Turing test, face and facial feature detection.

1. INTRODUCTION

Web services are increasingly becoming part of people's everyday life. For example, we use free email accounts to send and receive emails; we use online polls to gather people's opinion; and we use chat rooms to socialize with others. But all these Web services designed for human use are being abused by computer programs (bots). Malicious programmers have designed bots to register thousands of free email accounts every minute [1][3]. Bots have been used to cast votes in online polls [1]. Chat rooms and online shopping are being abused by bots as well [2] [5].

These real-world issues have recently generated a brand-new research area called Human Interactive Proofs (HIP), whose goal is to defend services from malicious attacks by differentiating bots from human users. The design of HIP systems turns out to have significant relationship with the famous Turing test whose goal was to determine if a machine has achieved artificial intelligence (AI) [7]. So far, no machine has passed the Turing test in a generic sense, even after decades of hard research in AI. This fact implies that there still exists considerable intelligence gap between human and machine. We can therefore use this gap to design tests to distinguish bots from human users. HIP is a unique research area in that it creates a *win-win* situation. If attackers cannot defeat a HIP algorithm, that algorithm can be used to defend Web services. On the other hand, if attackers defeat a HIP algorithm, that means they have solved a hard AI problem, thus advancing the AI research.

The first idea related to HIP can be traced back to Naor who wrote an unpublished note in 1996 [5]. The first HIP system in action was developed in 1997 by researchers at Alta Vista [2]. Its goal was to prevent bots from adding URLs to the search engine to skew the search results. In recent years, the CMU team has been one of the most active teams in HIP, and we highly recommend readers to visit their web site at http://www.captcha.net to see concrete HIP examples [1][3]. The CMU team introduced the notion of CAPTCHA: Completely Automated Public Turing Test to Tell Computers and Humans Apart. Intuitively, a CAPTCHA is a program that can generate and grade tests that 1) most human can pass; but 2) current computer programs cannot pass [1]. They have developed several CAPTCHA systems including Gimpy, Bongo, Pix, and Animal Pix. In the past two years, researchers at PARC and UC Berkeley published a series of papers on HIP, e.g., [3]. In their systems, they mainly explored the gap between human and bots in terms of reading poorly printed texts (e.g., fax prints).

Despite the recent progress made in HIP research, so far, the existing HIP algorithms suffer from one or more deficiencies in ease of use, resistance to attack, dependency on labeled database and lack of universality (see our technical report [6] for details). To overcome these difficulties, in this paper, we propose a new HIP algorithm based on detecting human face and facial features. Human faces are the most familiar object to humans, making it possibly the best candidate for HIP.

We name our HIP algorithm ARTiFACIAL, standing for Automated Reverse Turing test using FACIAL features. It relates to (and differs from) the original Turing test in several ways. First, our test is automatically generated and graded, i.e., the Turing test judge is a machine instead of a human. Second, the goal of the test is the reverse of the original Turing test – we want to differentiate bots from human, instead of proving bots is as

MM'03, November 2-8, 2003, Berkeley, California, USA.

ACM 1-58113-722-2/03/0011.

intelligent as human. These two features constitute the first three letters (ART) in ARTiFACIAL: Automated Reverse Turing test.

ARTiFACIAL works as follows. Per each user request, it automatically synthesizes an image with a distorted face embedded in a clustered background. The user is asked to first find the face and then click on 6 points (4 eye corners and 2 mouth corners) on the face. If the user can correctly identify these points, ARTiFACIAL concludes the user is a human; otherwise, the user is a machine. We conduct user studies and show the ease of use of ARTiFACIAL to human users. We design attacks using the best existing face detectors and demonstrate the difficulty to malicious bots.

2. PROPOSED TEST -- ARTiFACIAL

Human faces are arguably the most familiar object to humans, rendering it possibly the best candidate for HIP. Regardless of nationalities, culture differences or educational background, we all recognize human faces. In fact, our ability is so good that we can recognize human faces even if they are distorted, partially occluded, or in bad lighting conditions.

Computer vision researchers have long been interested in developing automated face detection algorithms. A good survey paper on this topic is [9]. In general face detection algorithms can be classified into four categories: knowledge-based, feature-based, template matching, appearance-based. So far, the fourth approach is the most successful one [9].

In spite of decades of hard research on face and facial feature detection, today's best detectors still suffer from several main limitations including the assumption that **faces are symmetric**, the difficulties of handling arbitrary **head rotations**, **arbitrary lighting**, and **cluttered background**. These conditions are among the most difficulty cases for automated face detection, yet we human seldom have any problem under those conditions. If we use the above 4 conditions to design a HIP test, it can take advantage of the large detection gap between human and machine. Indeed, this gap motivates our design of ARTiFACIAL.

We next use a concrete example to illustrate how to automatically generate an ARTiFACIAL test image, taking into account of the 4 conditions discussed above. For clarity, we use *F* to indicate a foreground object in an image, e.g., a face; *B* to indicate the background in an image; *I* to indicate the whole image (i.e., foreground and background); and *T* to indicate cylindrical texture map.

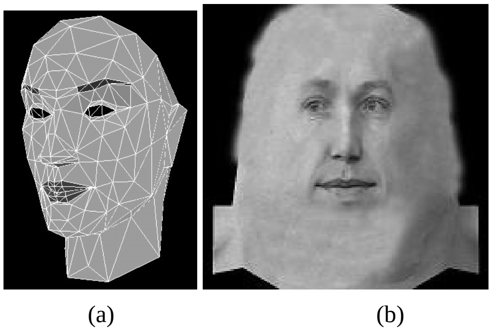

Figure 1. (a) The 3D wire model of a generic head. (b) The cylindrical head texture map of an arbitrary person.

[**Procedure**] ARTiFACIAL

[**Input**] The only inputs to our algorithm are the 3D wire model of a generic head (see Figure 1 (a)) and a 512 x 512 cylindrical texture map *Tm* of an arbitrary person (see Figure 1 (b)). Note that any person's texture map will work in our system and from that single texture map we can in theory generate infinite number of test images.

[**Output**] A 512 x 512 ARTiFACIAL test image I_F (see Figure 4) with ground truth (i.e., face location and facial feature locations).

1. Confusion texture map *Tc* generation
   This process takes advantage of the **Cluttered Background** limitation to design the HIP test. The 512 x 512 confusion texture map *Tc* (see Figure 2) is obtained by moving facial features (e.g., eyes, nose and mouth) in Figure 1 (b) to different places such that the "face" no longer looks like a face.

2. Global head transformation
   Because we have the 3D wire model (see Figure 1 (a)), we can easily generate any global head transformations we want. Specifically, the transformations include translation, scaling, and rotation of the head. Translation controls where we want to position the head in the final image I_F. Scaling controls the size of the head, and rotation can be around all the three x, y, and z axes. At run time, we randomly select the global head transformation parameters and apply them to the 3D wire model texture-mapped with the input texture *Tm*. This process takes advantage of the **Head Orientations** limitation to design the HIP test.

3. Local facial feature deformations
   The local facial feature deformations are used to modify the facial feature positions so that they are slightly deviated from their original positions and shapes. This deformation process takes advantage of the **Face Symmetry** limitation to design the HIP test. Each geometric deformation is represented as a vector of vertex differences. We have designed a set of geometric deformations including the vertical and horizontal translations of the left eye, right eye, left eyebrow, right eyebrow, left mouth corner, and right mouth corner. Each

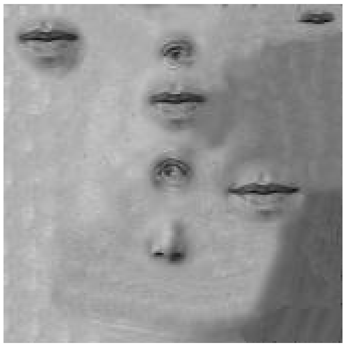

Figure 2. The confusion texture map *Tc*, is generated by randomly moving facial features (e.g., eyes, nose and mouth) in Figure 1 (b) to different places such that the "face" no longer looks like a face.

geometric deformation is associated with a random coefficient uniformly distribution in [-1, 1], which controls the amount of deformation to be applied. At run time, we randomly select the geometric deformation coefficients and apply them to the 3D wire model. An example of a head after Steps 2 and 3 is shown in Figure 3 (a). Note that the head has been rotated and facial features deformed.

4. Confusion texture map transformation and deformation
In this step, we conduct exactly the same Steps 2 and 3 to the confusion texture map *Tc*, instead to *Tm.* This step generates the transformed and deformed confusion head *Fc* as shown in Figure 3 (b).

5. Stage-1 image I_1 generation
Use the confusion texture map *Tc* as the background *B* and use F_h as the foreground to generate the 512 x 512 stage-1 image I_1 [6].

6. Stage-2 image I_2 generation
Make *L* copies of randomly shrunk *Tc* and randomly put them into image I_1 to generate the 512 x 512 stage-2 image I_2 [6]. This process takes advantage of the **Cluttered Background** limitation to design the HIP test. Note that none of the copies should occlude the key face regions including eyes, nose and mouth.

7. Stage-3 image I_3 generation
There are three steps in this stage. First, make *M* copies of the confusion head *Fc* and randomly put them into image I_2. This step takes advantage of the **Cluttered Background** limitation. Note that none of the copies should occlude the key face regions including eyes, nose and mouth**.** Second, we now have *M+1* regions in the image, where *M* of them come from *Fc* and one from F_h. Let *Avg(m), m = 0, …, M+1*, be the average intensity of region *m*. We next re-map the intensities of each region *m* such that *Avg(m)*'s are uniformly distributed in [0,255] across the *M+1* regions, i.e., some of the regions become darker and others become brighter. This step takes advantage of the **Lighting and Shading** limitation. Third, for each of the *M+1* regions, randomly select a point within that region which divides the region into four quadrants. Randomly select two opposite quadrants to under go further intensity changes. If the average intensity of the region is greater than 128, the intensity of all the pixels in the selected quadrants will decrease by a randomly selected amount; otherwise, it will increase by a randomly selected amount. This step takes advantage of both the **Face Symmetry** and **Lighting and Shading** limitations. Note in the image that 1) the average intensities of the *M+1* regions are uniformly distributed, i.e., some regions are darker while others are brighter; 2) two of the quadrants undergo further intensity changes.

8. Final ARTiFACIAL test image I_F generation
Make *N* copies of the facial feature regions in F_h (e.g., eyes, nose, and mouth) and randomly put them into I_3 to generate the final 512 x 512 ARTiFACIAL test image I_F (see Figure 4). This process takes advantage of the **Cluttered Background** limitation to design our HIP test. Note that none of the copies should occlude the key face regions including eyes, nose and mouth.

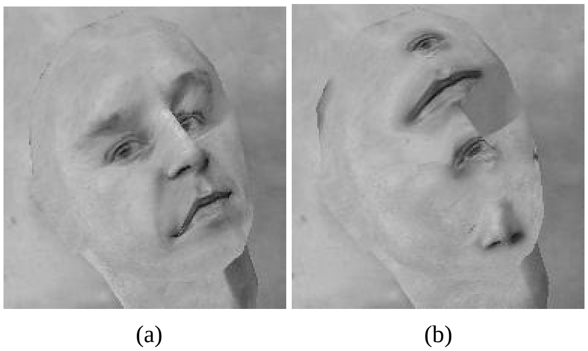

Figure 3. (a) The head after global transformation and facial feature deformation. We denote this head by F_h. (b) The confusion head after global transformation and facial feature deformation. We denote this head by *Fc*.

The above 8 steps take the 4 face detection limitations into account and generate ARTiFACIAL test images that are very difficult for face detectors. We used the above described procedure and generated 1,000 images to be used in both user study (Section 3) and bots attacks (Section 4).

3. USER STUDY DESIGN AND RESULTS

For a HIP test to be successful, we need to show that it is easy for human user and very hard for bots. In this section, we design user studies to evaluate human user's performance to our test. We will discuss bots attacks in the following section.

3.1 User Study Design

To evaluate our HIP system across diversified user samples, we invited 34 people to be our study subjects, consisting of accountants, administrative staff, architects, executives, receptionists, researchers, software developers, support engineers and patent attorneys. Each user takes 10 tests. We therefore have 34x10 = 340 tests.

3.2 User Study Results

We make the following observations based on the user study [6]:

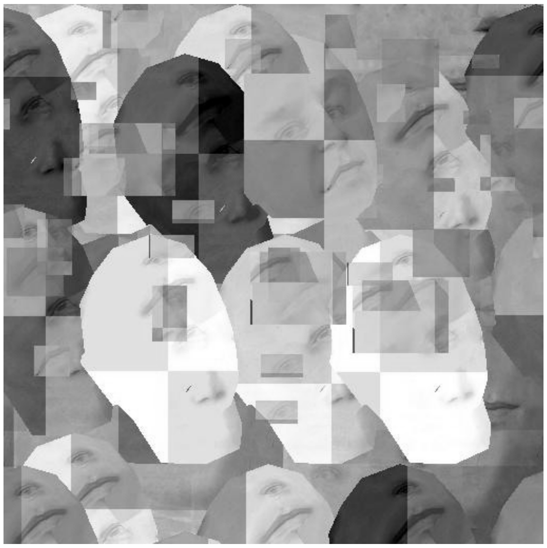

Figure 4. An ARTiFACIAL test image

- On average, it takes 14 seconds for a subject to find the face and click on the 6 points. This shows the test is easy to complete for human users. Out of the 34x10=340 tests, there are only a few tests that take longer than 30 seconds to finish. And interestingly enough most of those cases occurred with the same subject. During our debriefing, the subject told us that he was a perfectionist and was willing to spend longer time to ensure no mistakes. Out of the 340 tests, human subjects only made one wrong detection. The correct rate is 99.7%. During debriefing, the subject told us that she was not paying too much attention for this image but should be able to get it correct if she was given a second chance. Indeed, she only made one mistake out of the 10 tests.
- The mismatches between the point coordinates of the ground truth and where the subjects actually clicked are small. They are within a few pixels. This tells us that we can enforce tight verifications (e.g., within a few pixels) to efficiently distinguish bots from human users.

4 ATTACKS AND RESULTS

To succeed in an attack, the bots must first locate the face from a test image's cluttered background by using a face detector, and then find the facial features (e.g., eyes, nose, and mouth) by using a facial feature detector. In this section we present results of attacks from three different face detectors and one face feature detector.

4.1 Face Detectors

The three face detectors used in this paper represent the state of the art in automatic face detection. The first face detector was developed by Colmenarez and Huang [4] which uses the information-based maximum discrimination (MD) to detect faces. The second face detector was developed by Yang *et. al.* [10] which used a sparse network (SNoW) of linear functions and was tailored for learning in the presence of a very large number of features. The third face detector was developed by Li and his colleagues [11] following the Viola-Jones AdaBoost approach.

We apply the three face detectors to attack the 1,000 images generated in Section 3. When evaluating if an attack is successful, we use very forgiving criterion for the face detectors: as long as the detected face region overlaps with the ground truth face region for 60% (or above), we call it a correct detection. For the MD face detector, it has only one correct detection. For SNoW face detector, it has three correct detections. For AdaBoost face detector, it has zero correct detection. Comparing these results with the 99.7% detection rate of human users, we can clearly see the intelligence gap between human and bots.

4.2 Facial Feature Detector

Just the face detector is not enough to attack our test. The attacker also needs a facial feature detector. The facial feature detector proposed by Yan *et. al.* [8] is an improved version of the conventional Active Shape Model (ASM). It represents state of the art in the field and works quite well with undistorted and clean faces [8].

Again, we use those 1,000 images as our test set. During the attack, we give multiple advantages to the facial feature detector. First, we tell the facial feature detector exactly where the true face is. Second, as long as the detected points are within twice the average mismatches human made, we call it a correct detection. Even if we give multiple advantages to the detector, the correct detection rate is only 0.2%. If we multiply the correct detection rate of the face detector and the facial feature detector, the final detection rate is about 1 out of a million, which is significantly more robust than the existing HIP tests.

5 CONCLUSIONS

In this paper, we have developed a new HIP algorithm ARTiFACIAL based on human face and facial feature detection. Compared to existing HIP systems, ARTiFACIAL is the only one that satisfies all the proposed HIP design guidelines. Because human face is the most familiar object to all human users, ARTiFACIAL is possibly the most universal HIP system so far. We used three state-of-the-art face detectors and one facial feature detector to attack our system, and their success rate are all very low. We also conducted user studies on 34 human users with diverse background. The results have shown that our system is robust to machine attacks and easy for human users [6].

6 ACKKNOWLEDGEMENT

The authors would like to thank Z. Xiong, M.-H. Yang, S. Li and S. C. Yan for their help on the system and the paper.

7 REFERENCES

[1] Ahn, L., Blum, M., and Hopper, N. J., Telling humans and computers apart (Automatically) or How lazy cryptographers do AI, Technical Report CMU-CS-02-117, February, 2002

[2] AltaVista's Add URL site: altavista.com/sites/addurl/newurl

[3] Baird, H.S., and Popat, K., Human Interactive Proofs and Document Image Analysis," Proc., 5th IAPR Workshop on Document Analysis Systems, Princeton, NJ, August 19-21, 2002

[4] Colmenarez A. and Huang, T. S., Face detection with information-based maximum discrimination, Proc. of IEEE CVPR, pp., 782-788, 1997

[5] Naor, M., Verification of a human in the loop or identification via the Turing test, unpublished notes, September 13, 1996

[6] Rui, Y. and Liu, Z., ARTiFACIAL: Automated Reverse Turing test using FACIAL features, MSR TR 2003-48

[7] Turing, A., Computing machinery and intelligence, Mind, Vol. 59 (236), pp. 433-460, 1950

[8] Yan, S. C., Li, M. J., Zhang, H. J., and Cheng., Q. S., Ranking Prior Likelihoods for Bayesian Shape Localization Framework, Submitted to IEEE ICCV 2003.

[9] Yang, M., Kriegman, D., and Ahuja, N., Detecting faces in images: a survey, IEEE Trans. on Pattern analysis and machine intelligence, Vol. 24, No. 1, January 2002.

[10] Yang, M., Roth, D., and Ahuja, N., A SNoW-Based Face Detector, Advances in Neural Information Processing Systems 12 (NIPS 12), S.A. Solla, T.K. Leen and K.-R. Muller (eds), pp. 855--861, MIT Press, 2000.

[11] Zhang, Z., Zhu, L., Li, S. and Zhang, H, Real-time multiview face detection, Proc. Int'l Conf. Automatic Face and Gesture Recognition, pp. 149-154, 2002