Learning visual features for the Avatar Captcha Recognition Challenge

Mohammed Korayem*[‡], Abdallah A. Mohamed[†§], David Crandall*, Roman V. Yampolskiy[†]
*School of Informatics and Computing, Indiana University
Bloomington, IN, USA
{mkorayem, djcran}@indiana.edu
[†]Computer Engineering& Computer Science, Speed School of Engineering
University of Louisville, Louisville, KY, USA
{aamoha04,roman.yampolskiy}@louisville.edu
[‡]Department of Computer Science
Fayoum University, Fayoum, Egypt
[§]Department of Mathematics
Menoufia University, Shebin El-Koom, Menoufia, Egypt

***Abstract*—Captchas are frequently used on the modern world wide web to differentiate human users from automated bots by giving tests that are easy for humans to answer but difficult or impossible for algorithms. As artificial intelligence algorithms have improved, new types of Captchas have had to be developed. Recent work has proposed a new system called Avatar Captcha, in which a user is asked to distinguish between facial images of real humans and those of avatars generated by computer graphics. This novel system has been proposed on the assumption that this Captcha is very difficult for computers to break. In this paper we test a variety of modern visual features and learning algorithms on this avatar recognition task. We find that relatively simple techniques can perform very well on this task, and in some cases can even surpass human performance.**

***Keywords*-Avatar Captcha; defeating Captchas; face recognition; GIST; HOG**

I. Introduction

Online systems often need to differentiate between legitimate human users and programs that are accessing their services automatically. For example, companies like Google and Yahoo that offer free e-mail services wish to ensure that requests for new accounts are coming from humans and not from automated programs that send spam. A popular approach for conducting this differentiation is to use Captchas, or Completely Automated Turing tests to tell Computers and Humans Apart [1], that require users to solve a problem that is easy for a human but difficult for a computer. For example, a Captcha might involve recognizing a word that has been corrupted by noise, identifying a spoken word, or recognizing an object in an image [2], [3], [8], [15].

A well-designed Captcha task is almost trivial for any human to solve, but very difficult for any automated algorithm in any reasonable amount of time. Of course these two goals are in tension, because making Captchas too easy may allow clever programmers to design reliable automated algorithms, while difficult Captchas may confound legitimate users. As the state-of-the-art in artificial intelligence improves, new Captchas need to be designed to maintain the proper balance between these two goals.

A novel system was recently proposed called Avatar Captcha [5], in which a user is presented with a set of facial images. Some of the images are of real faces, while the rest are synthetic faces of computer-generated avatars. To pass the Captcha, a user must correctly identify which images are real and which are synthetic. The authors of that paper showed that the chances of an automated algorithm passing the test are extremely low if the algorithm uses random guessing, but that most human users are able to easily pass the test.

In this paper, we study the extent to which computer vision algorithms can be used to increase the accuracy of automated algorithms on the Avatar Captcha task. Surprisingly, we find that learning classifiers from relatively simple computer vision features yields accuracies that are competitive with and even surpass human performance on this problem. We use the publicly-available ICMLA Face Recognition Challenge dataset, released by the developers of Avatar Captcha [5], as our training and testing dataset. After describing Avatar Captchas in detail in Section 2, we show how to apply straightforward learning-based visual recognition techniques to this problem in Section 3. We present experimental results in Section 4, and then consider some possible techniques to make Avatar Captchas more secure in Section 5.

II. Avatar Captcha

Captcha schemes can be divided into three main categories: text-based, audio-based, and image-based [3]. The most common type is the text-based test, in which an image containing a series of letters and/or numbers is presented, and the user must retype the characters correctly. To try to prevent the use of Optical Character Recognition (OCR) algorithms, text-based Captchas distort the letters and numbers to make them harder to read. As OCR technology improves, Captchas must increase the degree of distortion in the images to keep automated programs at bay, but these increased distortions make it difficult for even human eyes to solve the tests correctly.

One solution to this problem is to move from text-based to image-based Captchas [6], [14], since the state-of-the-art for image recognition is generally much lower than that of OCR. In image-based Captchas, a human must answer questions about the content of an image, such as giving the identities or properties of objects. One particularly interesting image-based approach is the Avatar Captcha system [5]. In that approach, a user is presented with 12 images organized in two rows of six images. All of these images are of faces, but some of them come from a dataset of real human faces while others are from a computer-generated dataset of synthetic Avatar images [13]. The user is required to select all the avatars among these 12 images by checking a box under each avatar image. The user passes the test (and is classified as legitimate) if he/she correctly classifies all 12 images. In their tests, the authors of [5] found that humans could pass this test approximately 63% of the time, while they calculated that a bot employing random guessing would pass with a probability of just $(0.5)^{12} \approx 0.02\%$. Our purpose in this paper is to test how often an automated bot could solve Avatar Captchas if it used automated visual analysis instead of random guessing.

III. Methods

A. Data Sets

We used the dataset released by the developers of the Avatar Captcha system [5] as part of the ICMLA Face Recognition Challenge. This set consists of 100 photos, with 50 human faces and 50 avatar faces. The faces are all generally frontal-view with some variation in illumination, facial expression, and background, and all images are grayscale. The human images come from the Nottingham scans dataset and consist of real images of men and women, and are resized to a common resolution of 50×75. The avatar images are a sample of avatar faces from the Entropia Universe virtual world, and were also resized to a resolution of 50×75. Figure 1 presents examples of different facial images from each dataset.

B. Visual features

We tested a variety of techniques to produce feature vectors from images. These techniques range from very simple and naive methods that operate on raw pixel values, to more modern techniques that are widely used in the object recognition community.

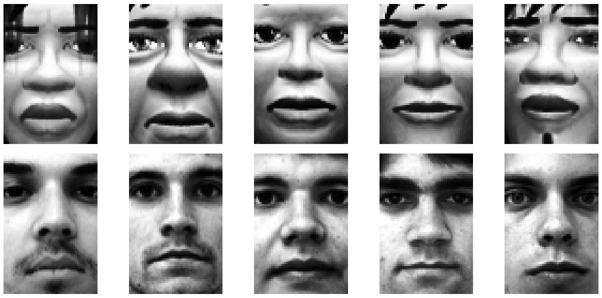

Figure 1. Sample avatar (top) and human faces (bottom) from our dataset.

1) Summary statistics: Our simplest features compute summary statistics about an image. We tried a 1-dimensional feature that is simply the mean pixel value of the image, and a 5-dimensional feature including maximum, minimum, mean, median, and sum of the pixel values.

2) Grayscale histograms: As a slightly more sophisticated feature, we also computed grayscale histograms for each image. We tried histograms with 2, 4, 8, 16, 32, 64, and 128 bins.

3) Vectors of raw pixel values: This feature involves simply reshaping an image into a vector by concatenating all of the image rows of grayscale values together. The resulting feature vector has $50 \times 75 = 3750$ dimensions.

C. Histograms of Oriented Gradients (HOG)

HOG is a very popular feature extraction technique for recognizing objects including humans [4]. The idea is to break an image into a grid of small windows, compute edge strengths and directions, and then compute a weighted histogram of edge orientations within each window. The histograms within each window are normalized and then concatenated together to form a single $2,268$ dimensional feature vector. HOG features capture the overall shape of an object or image region, but give invariance to illumination and contrast changes, and allow for some variation in shape and appearance.

D. GIST descriptors

GIST features [12] features try to capture the overall appearance ("gist") of a scene. To do this, the image is divided into a grid of non-overlapping cells, and color and texture features inside each cell are computed. These features are concatenated together to produce a single feature vector for each image. GIST is invariant or insensitive to a variety of image transformations including illumination changes, blur, and resizing, but is not invariant to translation, rotation, etc. GIST uses a 4×4 grid and computes 60 features per cell, yielding a 960 dimensional vector for our images.

E. Classifiers and feature selection

We tested the above features with two types of classifiers, Naive Bayes [10], [11] and LibLinear with L2-regularized

Table I
Experimental results with simple features and Naive Bayes classifiers, on classifying a single image as well as the 12-way Avatar Captcha task.

Method	2-class accuracy	Captcha accuracy
Pixel values	**93%**	**41.9%**
Mean pixel	57%	0.1%
Summary stats (mean, median, min, max, sum)	61%	0.3%
Histograms (256-Bins)	89%	24.7%
Histograms (128-Bins)	92%	36.8%
Histograms (64-Bins)	77%	4.3%
Histograms (32-Bins)	78%	5.1%
Histograms (16-Bins)	75%	3.2%
Histograms (8-Bins)	77%	4.3%
Histograms (4-Bins)	69%	1.2%
Histograms (2-Bins)	52%	0.03%
Random baseline	50%	0.02%

logistic regression [7]. We also tested the effect of feature selection on these problems, using Correlation-based Feature Selection (CFS) [9].

IV. Experimental results

We evaluated the performance of various combinations of the above visual features, classifiers, and feature selection algorithms on the Avatar Captcha recognition task. Table I shows the results for our simplest features with a Naive Bayes classifier. The table shows results on both the two-class task of deciding if a single given facial image is an avatar or a human, and the 12-way Avatar Captcha task in which a user most correctly classify a set of 12 images. All experiments in this section were conducted using 10-fold cross-validation. The best 2-way classification accuracy in this set of experiments was 93% when raw pixel values were used. This means that an automated program could correctly answer an Avatar Captcha with probability 41.9%, or nearly 2,000 times more often than predicted by [5]. The histogram-based techniques also achieve relatively good classification performance, with 89% accuracy for 256-bin histograms and 92% for 128-bin histograms. Even the simplest feature (1-d feature consisting of average pixel value) performs nearly 7 percentage points better than baseline.

Table II shows results with the more sophisticated image features and classifiers. Surprisingly, we actually achieve perfect classification (100% accuracy) on the test dataset when the high-dimensional feature vector of raw pixel values is combined with the LibLinear classifier. According to this result, an automated bot could successfully solve Avatar Captchas correctly with nearly perfect accuracy, performing even better than humans on this task! HOG features achieved 99% accuracy with LibLinear. The table also shows that feature selection could successfully reduce the dimensionality of the feature vectors while sacrificing little performance, since the 54-dimensional reduced vectors for raw pixel values achieves 98% accuracy with Naive Bayes.

Table II
Experimental results with more sophisticated features and classifiers, including LibLinear, Naive Bayes, and Naive Bayes with feature selection. Feature dimensionality is shown inside parentheses.

Method	LibLinear	Naive Bayes (NB)	NB+FS
Pixel values	**100% (3750f)**	93% (3750f)	98% (54f)
256-bin Histogram	60% (256f)	**89% (256f)**	82% (24f)
GIST	84% (960f)	88% (960f)	**90% (24f)**
HOG	**99% (2268f)**	94% (2268f)	95% (44f)

V. Defenses against machine classification

The surprisingly high performance of relatively simple vision algorithms on the Avatar Captcha task suggests that there may be biases in the dataset that are readily discovered and exploited by machine learning. For example, the fact that a classification algorithm looking only at mean pixel value achieved a significant improvement over baseline indicates that the images in one class are on average brighter than those of the other class. Other more subtle biases likely exist, since the sets of facial images were generated in two very different ways (one through photography and the other with computer graphics).

We did some preliminary investigation to test whether applying some simple transformations to images could make the problem more difficult and thus confound the classification algorithms. In particular, we tried three types of transformations:

1) Noise: We tried adding different types of random noise to the images, including Gaussian, Poisson, and Salt & Pepper noise. For each image, we randomly chose one type of noise and then added it to the image.

2) Rotation: To increase the appearance variation in the dataset, we tried rotating each image by a random angle between 1 and 180 degrees.

3) Occlusion: Finally, we tried to explicitly defeat the classification algorithms by identifying the 500 most important pixel locations in the image (by looking at the top features identified by feature selection on the raw pixel vectors), and occluding them by setting them to 0.

Table III shows the results for the different features and classifiers when applied to datasets that have been corrupted by the above techniques. Adding random noise successfully confounds the histogram features, reducing accuracy from 82% to near the random baseline, but has little affect on other features. Rotations confuse HOG, GIST and pixel vectors since these features encode spatial position explicitly, but have minimal effect on histogram features. The occlusion features reduce performance significantly for Naive Bayes with feature selection, but have little impact otherwise. Combining all three techniques together reduces the accuracy of the best-performing classifier from 100% down to 85%. In the Avatar Captcha task, in which 12 images must be correctly classified, this accuracy means that even with the

Table III
CLASSIFICATION PERFORMANCE ON IMAGES CORRUPTED BY NOISE, ROTATIONS, AND OCCLUSIONS.

	Original		Noise		Rotation		Occlusion		Noise+Rotation+Occlusion	
Feature	NB+FS	LibLinear	NB+FS	LibLinear	NB+FS	LibLinear	NB+FS	LibLinear	NB+FS	LibLinear
Pixel values	98%	100%	98%	100%	86%	93%	91%	99%	83%	85%
256-bin Histogram	82%	60%	46%	61%	74%	92%	86%	90%	68%	60%
Gist	90%	84%	89%	84%	66%	65%	91%	89%	58%	64%
HOG	95%	99%	94%	90%	81%	81%	99%	97%	71%	74%

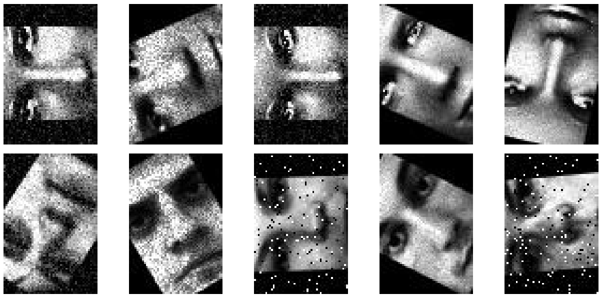

Figure 2. Avatar (top) and human faces (bottom) after noise and rotation.

noisy images a bot could solve the problems about 14% of time. Note that we have not yet studied whether legitimate human users are still able to solve them after the noise and transformations have been added. Figure 2 shows some sample images corrupted by rotation and noise.

VI. CONCLUSION

We have applied a variety of visual features and learned classifiers to the problem of distinguishing between human and avatar faces. Our results show that while automated bots are very unlikely to solve Avatar Captchas through random guessing, through computer vision they can solve these tasks nearly as well as humans. We suspect that the high performance may be caused by subtle differences and biases between the avatar and face images in the ICMLA Face Recognition Challenge dataset. These results suggest that while Avatar Captchas have many advantages, in practice it may be surprisingly difficult to secure them against attacks based on modern computer vision and machine learning techniques.

REFERENCES

[1] L. Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using hard AI problems for security. In *Proceedings of the International Conference on Theory and Applications of Cryptographic Techniques*, pages 294–311, 2003.

[2] A. Almazyad, Y. Ahmad, and S. Kouchay. Multi-modal captcha: A user verification scheme. In *Information Science and Applications (ICISA), 2011 International Conference on*, pages 1–7. IEEE, 2011.

[3] A. Chandavale, A. Sapkal, and R. Jalnekar. A framework to analyze the security of text-based CAPTCHA. *International Journal of Computer Applications*, 1(27):127–132, 2010.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[5] D. D'Souza, P. Polina, and R. Yampolskiy. Avatar captcha: Telling computers and humans apart via face classification. In *IEEE International Conference on Electro/Information Technology (EIT)*, pages 1–6. IEEE, 2012.

[6] J. Elson, J. Douceur, J. Howell, and J. Saul. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. *CCS*, 7:366–374, 2007.

[7] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] H. Gao, D. Yao, H. Liu, X. Liu, and L. Wang. A novel image based CAPTCHA using jigsaw puzzle. In *IEEE International Conference on Computational Science and Engineering (CSE)*, pages 351–356. IEEE, 2010.

[9] M. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[11] G. John and P. Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[13] J. Oursler, M. Price, R. Yampolskiy, and J. Hall. Parameterized generation of avatar face dataset. In *14th International Conference on Computer Games: AI, Animation, Mobile, Interactive Multimedia, Educational & Serious Games, Louisville, KY*, pages 17–22, 2009.

[14] L. Von Ahn, M. Blum, and J. Langford. Telling humans and computers apart automatically. *Communications of the ACM*, 47(2):56–60, 2004.

[15] L. Wang, X. Chang, Z. Ren, H. Gao, X. Liu, and U. Aickelin. Against spyware using captcha in graphical password scheme. In *IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pages 760–767. IEEE, 2010.