

CAPTCHA Challenge Tradeoffs: Familiarity of Strings *versus* Degradation of Images

Sui-Yu Wang, Henry S. Baird

Computer Science & Engineering Dept., Lehigh University, Bethlehem, PA, USA

Jon L. Bentley

Avaya Labs Research, Basking Ridge, NJ, USA

syw2@lehigh.edu baird@cse.lehigh.edu jlbentley@avaya.com

Abstract

It is a well documented fact that, for human readers, familiar text is more legible than unfamiliar text. Current-generation computer vision systems also are able to exploit some kinds of prior knowledge of linguistic context: for example, many OCR systems can use known lexica (word-lists, such as of commonly occurring English words) to disambiguate interpretations. It is interesting that human readers can exploit various degrees of familiarity: for example, strings of characters which, while not found in dictionaries, are similar to spelled words: e.g. “pronounceable” strings, or strings made up of frequently occurring character n -grams. In contrast to this, computer vision technologies for exploiting such poorly characterized constraints (absent an explicit, complete lexicon) are not yet well developed. This gap in ability may allow us to design stronger CAPTCHAs. We measure the familiarity of challenge strings generated by four methods (described by Bentley and Mallows) and we use the ScatterType CAPTCHA to degrade challenge images. We report the results of a human legibility trial which supports the hypothesis that more familiar strings are indeed more legible in CAPTCHAs. Our measurements may enable engineering CAPTCHAs with a more uniform distribution of difficulty by balancing image degradations against familiarity.

Keywords: CAPTCHAs, human interactive proofs, document image analysis, Gestalt perception, familiarity, image recognition, linguistic context, Markov scores

1 Introduction

Human interactive proofs (HIPs) have recently been developed to block computers from abusing Internet resources intended for human use [1]. Most HIPs today are *reading CAPTCHAs*—Completely Automatic Public Turing tests to tell Computers and Humans Apart—which challenge users to read degraded images of text which baffle current OCR techniques but are still legible to human readers. Lehigh’s reading CAPTCHA ScatterType consists of a randomly

generated (nonsense, but English-like) text string rendered as an image using a randomly selected typeface. The character images are cut into pieces which are then scattered apart pseudorandomly according to preset parameters (for examples, see Figure 2). Human legibility trials have shown that users’ subjective rating of difficulty is correlated with objective illegibility, and it is now understood how to engineer ScatterType challenges to meet legibility and difficulty targets [2].

It is known that humans reading familiar text perform better than on unfamiliar text [9][11][10]. Computer vision systems also can exploit some kinds of linguistic context: many OCR systems can use known lexica [7]. While human readers can exploit a range of degrees of familiarity, OCR technologies for exploiting such constraints (absent a fixed, known lexicon) are immature.

This gap in ability between humans and machines may allow us to design stronger CAPTCHAs by exploiting tradeoffs between familiarity of challenge strings and degradation of the image.

English words – English words are familiar to most users of the Internet, but are also vulnerable to dictionary-based attacks. Bentley and Mallows[6] observed that brute-force image segmentation followed by dictionary lookup can raise the probability of successful attack from $1/N$ to k/N , where k is the number of distinct lengths of the string and N is the number of words in the dictionary.

Markov dictionary strings – A Markov dictionary text-string is an English-like nonsense string, generated based on the character n -gram statistics of words. Bentley and Mallows[6] extracted character 3-gram frequencies from an English dictionary in which the beginning and end of words were marked with special symbols. The resulting Markov model was then used to generate words pseudorandomly; English words were then removed from the list, leaving words which only resemble English, e.g. “coness” and “splain”. This Markov model can be applied to esti-

mate the probability that any given string will be generated by the model. These “Markov generation scores” (or simply, *Markov scores*) are expressed here as negatives of binary logarithms of these probabilities. Thus, for example, the string “proper” is scored 15.47, and “tacman” is scored 29.54 (lower scores reflect higher probabilities). If a transition within the string *never* appears in the training data, we assign a smoothed probability of 2^{-6} .

Consonant-vowel-consonant (CVCCVC) 6-tuples – Most trigrams formed of consonant/vowel/consonant choices—*e.g.* “kag”—are pronounceable even when they are peculiar (*e.g.* “juf”). By stringing them together, *e.g.* “sixbeg”, we can generate many pronounceable nonsense words.

Random strings – A random string consists of characters selected randomly and uniformly from among the 26 alphabetic characters. An example of such string is “tjcnfi”. The resulting string may not be pronounceable and may not be a dictionary word. But its randomness provides maximum protection from blind attackers, compared to the other three categories.

We selected or generated four sets of words—English, Markov, CVCCVC, and Random—and then forced them to be disjoint.¹ When people are shown degraded images of strings chosen from these four sets, we expect them to recognize strings some sets more reliably than others. Certainly we expect that English words will be easier than random strings. It is not clear to us which of (a) Markov dictionary strings and (b) CVCCVC strings will be easier, but we expect them both to lie between English and random strings. To help resolve this question we will use the four sets in unequal proportions: 15% English, 35% Markov dictionary, 35% CVCCVC strings, and 15% random.

We report on the results of two experiments to test this hypothesis. The first experiment measured relative familiarity among the four sets. The second measured how this familiarity affects legibility of ScatterType challenges.

2 Experiment 1: Measuring Familiarity of Text Strings

We presented lists of six-letter strings (chosen at random from the four lists) to human subjects with the instruction to assign each a “familiarity score” between 1 and 5 as follows:

5 - most familiar, English *e.g.* “tigers” and “youall”;

4 - pronounceable & English-like *e.g.* “wellum” and “sevang”;

¹If an English word occurred in another set, it was pruned from that set. If a CVCCVC or Markov dictionary string occurred among the random strings, it was pruned from the random set. Any string that occurred in both the CVCCVC and Markov dictionary lists was pruned from both sets.

English	Markov	CVCCVC	Random
4.79	3.41	2.63	1.47

Table 1. Mean familiarity scores assigned by eleven human subjects to the four categories of 6-character strings: (a) English words; (b) Markov Dictionary strings; (c) CVCCVC 6-tuples; and (d) Random uniform over the alphabet. (5: most familiar, 1: least.)

3 - pronounceable but not English-like *e.g.* “yecwan” and “sursuf”;

2 - unpronounceable *e.g.* “ptmadh” and “xigifz”; and

1 - random *e.g.* “jqjdcr” and “tkljzr”.

The subjects were asked not to consult dictionaries. Altogether 4649 strings were examined and scored by eleven subjects. (Only three non-native English speakers participated.) The results are shown in Table 1. The absolute familiarity scores for Markov and CVCCVC overlapped for some users, suggesting that the difference in familiarity is slight and inconsistent between these two sets of strings.

Our data reveal a strong correlation, of 63%, between human familiarity and Markov scores. Figure 1 shows the minimum, maximum, and mean values (with plus and minus one standard deviation) of the Markov scores for the four types of strings. As we expected, English and Markov strings have the lowest scores and variances. The CVCCVC strings, however, have a high score, while their familiarity scores are not much different from those of Markov strings. This gap may be due to the fact that the pronounceability of the CVCCVC strings is familiar in quite a different sense than the resemblance to dictionary words captured by the Markov model.

3 Experiment 2: Measuring Effect of Familiarity on Recognition of ScatterType CAPTCHAs

We have investigated how good a predictor each of these scores is of success in answering CAPTCHAs. Strings from all four groups were rendered as degraded images using the ScatterType CAPTCHA technology [2]. Briefly, a ScatterType challenge is a string of characters rendered as an image in a randomly chosen typeface. The character images are cut into pieces, the pieces are scattered, and the characters are pulled close together, so that present-day computer vision techniques find the resulting “word” image hard to segment into characters, and thus hard to recognize. Several parameters control the cut-and-scatter process: a *cut fraction* determines the fragment size (the ‘offset’ locations

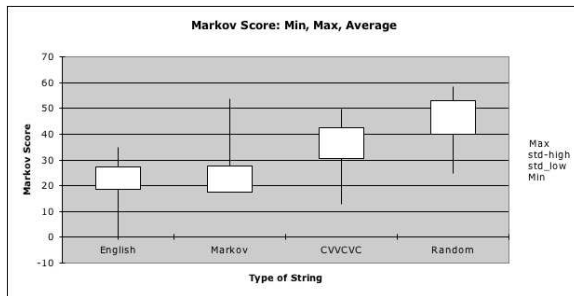


Figure 1. Markov scores for the four groups of text-strings: the ranges from minimum to maximum values are indicated by thin vertical bars; mean values plus and minus one standard deviation are shown by the vertical boxes.

of cuts are random). Fragments are moved apart by an *expansion* parameter. Each row of fragments is shifted right and left alternatively by a random distance controlled by a *horizontal-scatter* parameter. A *vertical-scatter* parameter shifts fragments within a row up and down alternately. Examples of ScatterType challenge images, over a wide range of difficulty, are shown in Figure 2.

In this trial, subjects were challenged to read ScatterType images of the text-strings describe above, generated over a range of degradation parameters known to span highly legible and as well as illegible variations. The objective difficulty as a function of image degradation is discussed in [2]. Each of the four text-string lists were subjected to the same pseudorandom range of image degradations. Thus, if membership in the four lists did not affect legibility, we would expect the human reading accuracy to be statistically indistinguishable among the four groups.

A total of 7489 such challenges were presented to 86 human subjects. The results are summarized in Table 2. The legibility of challenges is strongly affected by familiarity of the text-strings, and the ranking of the four text-string groups parallels their familiarity scores. English words are recognized better than the other three, and random strings are recognized more rarely than the other three. Reading accuracy on the Markov dictionary strings is statistically indistinguishable from that on the CVCCVC 6-tuples.

4 Discussion

Table 2 shows three clusters of accuracy: English words are read with about 75% accuracy, Markov and CVC text is read with about 60% accuracy, and random text is read with about 46% accuracy. These data assist in choosing tradeoffs between readability and assurance. The first piece

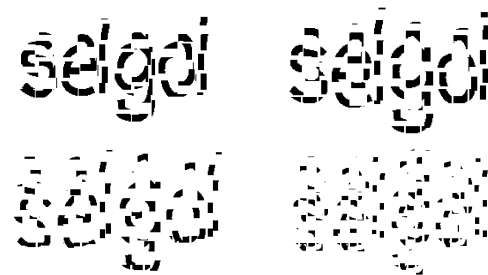


Figure 2. Examples of ScatterType challenge images, using the Markov dictionary text-string “selgol”: at various degrees of difficulty controlled by judicious choice of degradation parameters. Upper Left: easy; upper right: medium hard; lower left: hard; lower right: impossible.

of good news: although CVCCVC has more assurance than Markov, it is roughly equally readable: so CVCCVC is a clearly better choice.

Our hypothesis was that familiarity of challenge text-strings might be an important predictor of human reading accuracy of CAPTCHAs using those strings. However, we found that for the 1500 strings used in both experiments, familiarity scores exhibited a correlation of only -0.19 with their subjective difficulty in the CAPTCHA experiment, and only -0.14 to their objective difficulty. These correlations are statistically significant, but they may be too weak to exploit when engineering CAPTCHAs. It occurred to us that the effect of image degradation might, in many of these cases, be overwhelming the effect of familiarity. To investigate this, we computed correlations (as above) for images restricted to low image degradations. Specifically, we restricted the range of two of the degradation parameters (horizontal scatter H and vertical scatter V) by enforcing a “H-V Scatter Threshold” T (so that $H^2 + V^2 \leq T^2$) and found (Figure 3) that for smaller values of T , which constrain the images to be of higher quality, the correlation of familiarity with accuracy improves, to -0.23 for subjective difficulty and 0.24 for objective difficulty. (We conjecture that, if we controlled also for typefaces chosen, the influence of familiarity may be even more significant.)

The average Markov score for CAPTCHA challenges which were correctly answered was 25.55, rising to 31.56 for those incorrectly answered. The correlation coefficient between Markov scores and subjective difficulty is 0.14, about the same magnitude as for objective difficulty, -0.12.

So we have observed that the familiarity of a challenge string has a positive effect on CAPTCHA legibility, and Markov scores are also strongly correlated with familiarity.

	Category				
	ALL	English	Markov	CVCCVC	Random
Total challenges	7489	1095	2675	2657	1062
% correct answers	60.2	75.1	61.1	58.7	46.3

Table 2. Human reading accuracy as a function of challenge string type.

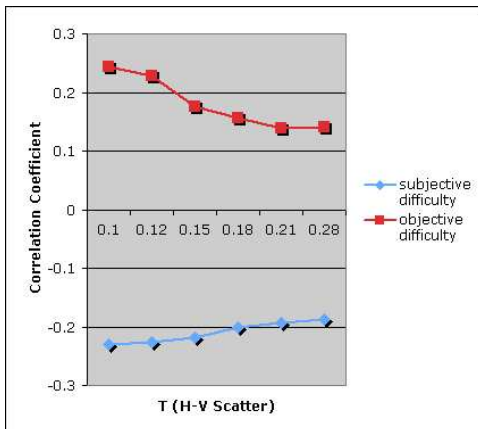


Figure 3. Correlations between familiarity and (a) objective familiarity (top), and (b) subjective difficulty (bottom), as a function of H-V Scatter Threshold T (smaller T selects less-degraded images).

These correlations are statistically significant, but overall Markov score is a weak predictor of legibility. Subjective and objective difficulty both appear to be dominated by the effects of image degradation.

5. Future Work

Using knowledge of the trade-offs between familiarity and legibility measured by these and similar experiments, it may be possible to adjust familiarity (through, *e.g.*, Markov scores) to offset the difficulty presented by higher levels of image degradation, and so achieve a more uniform level of overall difficulty within a set of CAPTCHA challenges. Furthermore, we might be able to use Markov scores to make subtle, local improvements in specific CAPTCHA images, by reducing image degradations in substrings which have high Markov scores.

6 Acknowledgments

We have benefited from advice by Colin Mallows, Michael Moll, and Dan Lopresti. We are grateful also to

the volunteers in our user trial who patiently endured reading many horrible images of bizarre words.

References

- [1] H. S. Baird and D. P. Lopresti, editors. *Human Interactive Proofs: Proceedings of the 2nd Int'l Workshop, HIP'2005*. Springer-Verlag, Bethlehem, PA, May 2005. Lecture Notes in Computer Science LNCS 3517.
- [2] H. S. Baird, M. A. Moll, and S.-Y. Wang. A highly legible captcha that resists segmentation attacks. In *Proc. 2nd Int'l Workshop on Human Interactive Proofs*, Bethlehem, PA, May 2005.
- [3] H. S. Baird, M. A. Moll, and S.-Y. Wang. Scattertype: A legible but hard-to-segment captcha. In *Proc. IAPR 8th Int'l Conf. on Document Analysis and Recognition*, Seoul, Korea, August 2005.
- [4] H. S. Baird and T. Riopka. Scattertype: a reading captcha resistant to segmentation attack. In *Proc., IS&T/SPIE Document Recognition & Retrieval XII Conf.*, San Jose, CA, January 2005.
- [5] J. L. Bentley and C. L. Mallows. How much assurance does a pin provide? In *Proc. 2nd Int'l Workshop on Human Interactive Proofs*, Bethlehem, PA, May 2005.
- [6] J. L. Bentley and C. L. Mallows. Captcha challenge strings: Problems and improvements. In *Proc. IS&T/SPIE Conf. on Document Recognition & Retrieval XIII*, San Jose, CA, January 2006.
- [7] K. Challapilla, K. Larson, P. Y. Simard, and M. Czerwinsky. Building segmentation based human friendly human interactive proofs (hip)s. In *Proceedings, 2nd Int'l Workshop on Human Interactive Proofs*, pages 1–26, Bethlehem, PA, May 2005. Springer-Verlag.
- [8] M. Chew and H. S. Baird. Baffletext: a human interactive proof. In *Proc., 10th IS&T/SPIE Document Recognition & Retrieval Conf.*, Santa Clara, CA, January 23–24 2003.
- [9] R. G. Crowder. *The Psychology of Reading*. Oxford University Press, 1982.
- [10] O. J. L. f and H. Singer. *Perception of Print: Reading Research in Experimental Psychology*. Lawrence Erlbaum Associates, Inc., 1981.
- [11] L. M. Gentile, M. L. Kamil, and J. S. Blanchard. *Reading Research Revisited*. Charles E. Merrill Publishing, 1983.
- [12] G. Kopec, M. Said, and K. Popat. N-gram language models for document image decoding. In *IS&T/SPIE Electronic Imaging 2002 Proc. of Document Recognition and Retrieval IV*, San Jose, California, January 2002.
- [13] G. Nagy. Twenty years of Document Image Analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [14] G. Nagy and S. Seth. *Modern Optical Character recognition. The Froehlich/Kent Encyclopaedia of Telecommunications*, volume 11. Marcel Dekker, New York, NY, 1996.