Pessimal Print: A Reverse Turing Test

Allison L. Coates
Computer Science Division, Univ. of California at Berkeley, Berkeley, CA 94720-1776 USA
Email: allisonc@eecs.berkeley.edu

Henry S. Baird
Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304 USA
E-mail: baird@parc.xerox.com

Richard J. Fateman
Computer Science Division, Univ. of California at Berkeley, Berkeley, CA 94720-1776 USA
Email: fateman@cs.berkeley.edu

Abstract

We exploit the gap in ability between human and machine vision systems to craft a family of automatic challenges that tell human and machine users apart via graphical interfaces including Internet browsers. Turing proposed [Tur50] a method whereby human judges might validate "artificial intelligence" by failing to distinguish between human and machine interlocutors. Stimulated by the "chat room problem" posed by Udi Manber of Yahoo!, and influenced by the CAPTCHA project [BAL00] of Manuel Blum et al *of Carnegie–Mellon Univ., we propose a variant of the Turing test using* pessimal print*: that is, low–quality images of machine-printed text synthesized pseudo-randomly over certain ranges of words, typefaces, and image degradations. We show experimentally that judicious choice of these ranges can ensure that the images are legible to human readers but illegible to several of the best present–day optical character recognition (OCR) machines. Our approach is motivated by a decade of research on performance evaluation of OCR machines [RJN96,RNN99] and on quantitative stochastic models of document image quality [Bai92,Kan96]. The slow pace of evolution of OCR and other species of machine vision over many decades [NS96,Pav00] suggests that pessimal print will defy automated attack for many years. Applications include 'bot' barriers and database rationing.*

Keywords: legibility, document image analysis, OCR evaluation methods, document image degradation, human/machine discrimination, Turing test.

1 Introduction

1.1 Turing's Test and Variants

Alan Turing proposed [Tur50] a method to assess whether or not a machine can think, by means of an "imitation game" conducted over teletype connections in which a human judge asks questions ("challenges") of two respondents – one human and the other a machine – and eventually decides which is human; failure to decide correctly would be, Turing suggested, convincing evidence of artificial intelligence in the machine. Extensions to the test have been proposed [SCA00], e.g. to incorporate behavior richer than teletypes can communicate. Graphical user interfaces (GUI) invite the use of images as well as text in challenges.

1.2 Machines Impersonating People

The world–wide proliferation of GUIs in the 1990's has opened up new uses for variants on the Turing test. In September 2000, Udi Manber of Yahoo! described this "chat room problem" to researchers at Carnegie-Mellon Univ.: "bots" (virtual persons) are being let loose in on–line chat rooms in an attempt to elicit personal information from the human participants — how can they be identified? Furthermore, "intelligent agents" are able systematically to mine databases that are intended for occasional use by individual humans. There is a growing need for automatic methods to distinguish between human and machine users on the Web.

Manuel Blum, Luis A. von Ahn, and John Langford have articulated [BAL00] desirable properties of such tests, including:

- the test's challenges can be automatically generated;
- the test can be taken quickly by human users;
- the test will accept virtually all human users (even young or naive users) with high reliability while rejecting very few;
- the test will reject virtually all machine users; and
- the test will resist automatic attack for many years even as technology advances and even if the test's algorithms are known (e.g. published and/or released as open source).

On hearing these, we saw an opportunity to bring to bear the well–known and extensively studied gap in image pattern recognition ability between human and machine vision systems.

Figure 1. Examples of synthetically generated images of machine–printed words, in various typefaces and degraded pseudo-randomly.

1.3 Document Image Quality

Low–quality images of printed–text documents pose serious challenges to current image pattern recognition technologies [RJN96,RNN99]. In an attempt to understand the nature and severity of the challenge, models of document image degradations [Bai92,Kan96] have been developed and used to explore the limitations [HB97] of image pattern recognition algorithms. The model of [Bai92], used throughout this study, approximates ten aspects of the physics of machine–printing and imaging of text, including spatial sampling rate and error, affine spatial deformations, jitter, speckle, blurring, thresholding, and symbol size. Figure 1 shows examples of text images that were synthetically degraded according to certain parameter settings of this model.

The reader should be able, with little or no conscious effort, to read all these images: so will, we expect, almost every person literate in the Latin alphabet, familiar with the English language, and with some years of reading experience. The image quality of these cases of "pessimal print" is far worse than people routinely encounter, but it's not quite bad enough to defeat the human visual system.

However, present-day "reading machines" (or, optical character recognition (OCR) machines) are baffled by these images, as we shall show.

1.4 Design of a Reverse Turing Test

We propose what we call a "reverse Turing test" of the following kind. When a user — human or machine — chooses to take the test (*e.g.* in order to enter a protected Web site), a program challenges the user with one synthetically generated image of text; the user must type back the text correctly in order to enter. This differs from Turing's proposal in at least four ways:

- the judge is a machine, rather than human;
- there is only one user, rather than two;
- the design goal is to distinguish, rather than to fail to distinguish, between human and machine; and
- the test poses only one challenge – or very few — rather than an indefinitely long sequence of challenges.

(We are grateful for discussions with Manuel Blum in connection with this design.)

The challenges must be substantially different almost every time, else they might be recorded exhaustively, answered off-line by humans, and then used to answer future challenges. Thus we propose that they be generated pseudorandomly from a potentially very large space of distinct challenges.

2 Experiments

In this design, the essential issue is the choice of the family of challenges: that is, some broad conditions under which text–images can be generated that are human–legible but machine–illegible. We carried out a search for these conditions with the kind assistance of ten graduate student and faculty volunteers in the Computer Science Division of Univ. of California, Berkeley. Our machine subjects were three of the best present–day commercial OCR systems: Expervision TR, ABBYY Fine reader, and the IRIS Reader.

2.1 Experimental Design

We synthesized challenges by pseudo-randomly uniformly and independently selecting:

- a word (from a fixed list);
- a typeface (from a fixed list); and
- a set of image–degradation parameters (from fixed ranges).

and then generating a single black–and–white image.

We selected 70 words according to a set of characteristics that we believed would favor human recognition over OCR machines. We restricted ourselves to natural–language words, since humans recognize words more quickly than non-word letter strings [TS81]. We chose words occurring with high frequency on the WWW, so not to penalize young or naive users. All words were in English, since English is the most widely used language on the WWW. All words had at least five letters and at most eight letters, because shorter words are few enough to invite an exhaustive template–matching attack, and longer words are unique enough to invite a feature–based attack. We used only words with no ascenders or descenders, since [Spi97] has shown these are strong cues for recognition even when individual characters are not reliably recognizable.

In our choice of image degradations, we were guided by the discussion in [RNN99] of cases that defeat modern OCR machines, especially:

- thickened images, so that characters merge together;
- thinned images, so that characters fragment into unconnected components;
- noisy images, causing rough edges and salt–and–pepper noise;
- condensed fonts, with narrower aspect ratios than usual; and
- Italic fonts, whose rectilinear bounding boxes overlap their neighbors'.

We explored ranges of values for two of the degradation parameters: blurring and thresholding (**blur** and **thrs** in [Bai92]; for their precise definitions, and for others' later, consult this reference; other parameters' values were fixed at: **sens**=0.05, **skew**=1.5 degrees, **xscl**=0.5, **yscl**=1.0).

We tested five typefaces: Times Roman (TR), Times Italic (TI), Courier Oblique (CO), Palatino Roman (PR), and Palatino Italic (PI). Note that **xscl**=0.5 has the effect of compressing the images horizontally by a factor of two, simulating abnormally condensed variants of these commonly occurring faces. The type size was fixed at **size**=8.0 point and the spatial sampling rate at **resn**=400 pixels/inch.

2.2 Experimental Results

The experiments located ranges of parameter values for blur and threshold with the desired pessimal-print properties:

- **thrs** $\in$ [0.01,0.02] for any value of **blur**; and
- **thrs** $\in$ [0.02,0.06] and **blur**=0.0;

These ranges represent, roughly, two types of images:

- extremely thinned and fragmented images, with a little salt-and-pepper noise; and
- noisy images (whether thinned or thickened);

Of a total of 685 word–images generated within these ranges, over all five typefaces, *all* were human–legible – but all three OCR machines failed on virtually every word, as illustrated in the following table of data selected from the experiments.

OCR Machine Accuracy (%), by typeface, for **blur** $\in$ [0.0,0.8] and **thrs**=0.02. (Machines: E-TR = Expervision TR; A-FR = ABBYY FineReader; I-R = IRIS Reader).

Face	E-TR	A-FR	I-R	#Words
TR	0.00%	0%	0%	136
TI	0.76%	0%	0%	132
CO	0.66%	0%	0%	152
PR	0.00%	0%	0%	198
PI	0.00%	0%	0%	67
TOTAL	0.29%	0%	0%	685

What is more, on almost all of these words, no machine guessed a *single* alphabetic letter, either correctly or incorrectly. The following figures shown a selection of machine–illegible and machine-legible word images.

Figure 2. Examples of synthetically generated images of printed words which are machine–illegible due to degradation parameter thrs = 0.02.

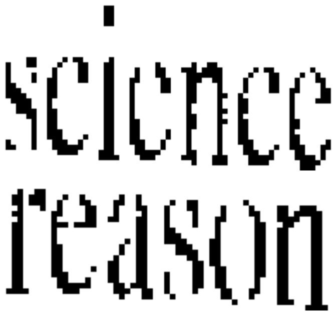

Figure 3. Examples of synthetically generated images of printed words which are machine–legible due to a slightly better thrs = 0.07.

Each OCR machine's performance was sensitive to slight changes in the parameters. For example, one machine's accuracy dropped from $40 - 50\%$ to 0% when **thrs** fell from 0.04 to 0.02 (for **blur** = 0.8). Also, it dropped from 28% to 0% when **blur** fell from 0.4 to 0.0 (at **thrs** = 0.04); this change is barely perceptible to the eye. Such fragility – abrupt catastrophic failure – is typical of many machine vision systems attempting to operate at the margins of good performance.

3 Discussion

Our familiarity with the state of the art of machine vision leads us to hypothesize that not only these three but *all modern OCR machines* will not be able to cope with the image degradations in the ranges we have identified. Also, we are confident that wider ranges, involving other degradation parameters and other typefaces, exhibiting pessimal–print properties, can be found through straightforward experiment. Blum et al [BAL00] have experimented, on their website `www.captcha.net`, with degradations that are not only due to imperfect printing and imaging, but include color, overlapping of words, non-linear distortions, and complex or random backgrounds. The relative ease with which we have been able to generate pessimal print, and the diversity of other means of bafflement ready to hand, suggest to us that the range of effective text–image challenges at our disposal is usefully broad.

How long can a reverse Turing test based on pessimal print resist attack, given a serious effort to advance machine–vision technology, and assuming that the design principles — perhaps even the source code — of the test are known to attackers? Even given such major hints as the dictionary of words, the nature of the distortions, the fonts, sizes and other considerations, a successful attack would probably require substantially more real time than humans, at least for the near future. A statistic today suggests about 200 msec per comparison between isolated handprinted digits, using fast 2001 year workstations; many comparisons over a far larger set would be needed to solve this problem. Also, our investigations and the CMU CAPTCHA project are continuing, and so any specific attacks might be thwarted in the same kind of arms race used in cryptography.

A close study of the history of image pattern recognition technology [Pav00] and of OCR technology [NS96] in particular suggests to us that the gap in ability between human and machine vision is wide and is only slowly narrowing. We notice that few, if any, machine vision technologies have simultaneously achieved all three of these desirable characteristics: high accuracy, full automation, and versatility. Versatility — by which we mean the ability to cope with a great variety of types of images — is perhaps the most intractable of these, and it is the one that pessimal print, with its wide range of image quality variations, challenges most strongly.

An ability gap exists for other species of machine vision, of course, and in the recognition of non–text images, such as line–drawings, faces, and various objects in natural scenes. One might reasonably intuit that these would be harder and so decide to use them rather than images of text. This intuition is not supported by the Cognitive Science literature on human reading of words. There is no consensus on whether recognition occurs letter–by–letter or by a word–template model [Cro82,KWB80]; some theories stress the importance of contextual clues [GKB83] from natural language and pragmatic knowledge. Furthermore, almost all research on human reading has used *perfectly formed* images of text: no theory has been proposed for mechanisms underlying the human ability to read despite extreme segmentation (merging and fragmentation) problems. The resistance of these problems to technical attack for four decades and the incompleteness of our understanding of human reading abilities suggests that it is premature to decide that the recognition of text under conditions of low quality, occlusion, and clutter, is intrinsically much easier — that is, a significantly weaker challenge to the machine vision state–of–the–art — than recognition of objects in natural scenes.

There are other, pragmatic, reasons to use images of text as challenges: the correct answer is unambiguously clear; the answer maps into a unique sequence of keystrokes; and it is straightforward automatically to label every challenge, even among hundreds of millions of distinct ones, with its answer. These advantages are lacking, or harder to achieve, for images of objects or natural scenes.

It might be good in the future to locate the limits of human reading in our degradation space: that is, at what point do humans find degraded words unreadable; do we smoothly decay or do we show the same kind of "falling off a cliff" phenomenon as machines but just at another level?

4 Conclusions

We have designed, built, and tested a "reverse" Turing test based on "pessimal print" and shown that it has the potential of offering a reliable, fast, and fully automatic method for telling people and machine users apart over GUI interfaces. It is an amusing irony, that we would like to believe Alan Turing himself would have savored, that a problem — machine reading — which he planned to attack and which he expected to yield easily, has instead resisted solution for fifty years and now is poised to provide a technical basis for the first widespread practical use of variants of his proposed test for human/machine distinguishability.

5 Acknowledgments

The experiments were carried out by the first author, using tools supplied by the second author, as part of a project for a Fall 2000 graduate course in 'Document Image Analysis' taught by the second and third authors in UC Berkeley's Computer Science Division. The project was triggered by a question – could character images make a good Turing test? – raised in the class by Manuel Blum of Carnegie–Mellon Univ., as one possible solution of the "chat room problem" posed earlier by Udi Manber of Yahoo!. Manuel Blum, Luis A. von Ahn, and John Langford, all of CMU, shared with us much of their early thinking about automated Turing tests, including news of their CAPTCHA project and their website `www.captcha.net`, which influenced the design and evolution of our Pessimal Print project. We kept them informed of our work as well, and they generously commented on an earlier draft of this submission. The second author wishes gratefully to acknowledge stimulating discussions with Bela Julesz in the late 1980's (at Bell Labs) on the feasibility of designing "pessimal fonts" using textons, which were however never implemented.

6 Bibliography

[BAL00] M. Blum, L. A. von Ahn, and J. Langford, *the CAPTCHA Project*, "Completely Automatic Public Turing Test to tell Computers and Humans Apart," `www.captcha.net`, Dept. of Computer Science, Carnegie–Mellon Univ., and personal communications, November, 2000.

[Bai92] H. S. Baird, "Document Image Defect Models," in H. S. Baird, H. Bunke, and K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer–Verlag: New York, 1992, pp. 546-556.

[Cro82] R.G. Crowder, *The Psychology of Reading*, Oxford University Press, 1982.

[GKB83] L. M. Gentile, M. L. Kamil, J. S. Blanchard *'Reading Research Revisited* , Charles E. Merrill Publishing, 1983.

[HB97] T. K. Ho and H. S. Baird, "Large-Scale Simulation Studies in Image Pattern Recognition," *IEEE Trans. on PAMI*, Vol. 19, No. 10, pp. 1067–1079, October 1997.

[ISRI] Information Science Research Institute, University of Nevada, Las Vegas, 4505 Maryland Parkway, Box 454021, Las Vegas, Vevada 89154-4021 USA.

[Jen93] F. Jenkins, *The Use of Synthesized Images to Evaluate the Performance of OCR Devices and Algorithms*, Master's Thesis, University of Nevada, Las Vegas, August, 1993.

[Kan96] T. Kanungo, *Document Degradation Models and Methodology for Degradation Model Validation*, Ph.D. Dissertation, Dept. EE, Univ. Washington, March 1996.

[KWB80] P.A. Kolers, M. E. Wrolstad, H. Bouma, *Processing of Visible Language 2*, Plenum Press, 1980.

[NS96] G. Nagy and Seth, "Modern optical character recognition." in *The Froehlich/Kent Encyclopaedia of Telecommunications*, Vol. 11, pp. 473-531, Marcel Dekker, NY 1996.

[Pav00] T. Pavlidis, "Thirty Years at the Pattern Recognition Front," King-Sun Fu Prize Lecture, 11th ICPR, Barcelona, September, 2000.

[RNN99] S. V. Rice, G. Nagy, and T. A. Nartker, *OCR: An Illustrated Guide to the Frontier*, Kluwer Academic Publishers, 1999.

[RJN96] S. V. Rice, F. R. Jenkins, and T. A. Nartker, "The Fifth Annual Test of OCR Accuracy," ISRI TR-96-01, Univ. of Nevada, Las Vegas, 1996.

[SCA00] A. P. Saygin, I. Cicekli, and V. Akman, "Turing Test: 50 Years Later," *Minds and Machines*, 10(4), Kluwer, 2000..

[Spi97] A. L. Spitz, "Moby Dick meets GEOCR: Lexical Considerations in Word Recognition," *Proc., 4th Int'l Conf. on Document Analysis & Recog'n.*, Ulm, Germany, pp. 221–232, August 18–20, 1997.

[TS81] O. J. L. Tzeng and H. Singer, *Perception of Print: Reading Research in Experimental Psychology*, Lawrence Erlbaum Associates, Inc., 1981.

[Tur50] A. Turing, "Computing Machinery and Intelligence," *Mind*, Vol. 59(236), pp. 433–460, 1950.