

Human Interactive Proofs and Document Image Analysis

Henry S. Baird and Kris Popat

Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304 USA
{baird|popat}@parc.com
www.parc.com/istl/groups/did

Abstract. The recently initiated and rapidly developing research field of ‘human interactive proofs’ (HIPs) and its implications for the document image analysis (DIA) research field are described. Over the last five years, efforts to defend Web services against abuse by programs (‘bots’) have led to a new family of security protocols able to distinguish between human and machine users. AltaVista pioneered this technology in 1997 [Bro01, LBBB01]. By the summer of 2000, Yahoo! and PayPal were using similar methods. In the Fall of 2000, Prof. Manuel Blum of Carnegie-Mellon University and his team, stimulated by Udi Manber of Yahoo!, were studying these and related problems [BAL00]. Soon thereafter a collaboration between the University of California at Berkeley and the Palo Alto Research Center (PARC) built a tool based on systematically generated image degradations [CBF01]. In January 2002, Prof. Blum and the present authors ran the first workshop (at PARC) on HIPs, defined broadly as a class of challenge/response protocols which allow a human to authenticate herself as a member of a given group – e.g. human (vs. machine), herself (vs. anyone else), an adult (vs. a child), etc. All commercial uses of HIPs known to us exploit the gap in ability between human and machine vision systems in reading images of machine printed text. Many technical issues that have been systematically studied by the DIA community are relevant to the HIP research program. This paper describes the evolution of HIP R&D, applications of HIPs now and on the horizon, highlights of the first HIP workshop, and proposals for a DIA research agenda to advance the state of the art of HIPs.

Keywords: *human interactive proofs, document image analysis, CAPTCHAs, abuse of web sites and services, the chatroom problem, human/machine discrimination, Turing tests, OCR performance evaluation, document image degradations, legibility of text*

1 Introduction

In 1997 Andrei Broder and his colleagues [LBBB01], then at the DEC Systems Research Center, developed a scheme to block the automatic submission of URLs [Bro01] to the AltaVista web-site: their approach was to present a user with an image of printed text formed specially so that machine vision (OCR) systems

could not read it but humans still could. In September 2000, Udi Manber, Chief Scientist at Yahoo!, challenged Prof. Manuel Blum and his students [BAL00] at The School for Computer Science at Carnegie Mellon University (CMU) to design an “easy to use reverse Turing test” that would block ‘bots’ (computer programs) from registering for services including chat rooms, mail, briefcases, etc. In October of that year, Prof. Blum asked the first author, of the Palo Alto Research Center (PARC),, and Prof. Richard Fateman, of the Computer Science Division of the University of California at Berkeley (UCB), whether systematically applied image degradations could form the basis of such a filter, stimulating the development of PessimPrint [CBF01].

In January 2002, Prof. Blum and the present authors ran a workshop at PARC on ‘human interactive proofs’ (HIPs), defined as *a broad class of challenge/response protocols which allow a human to authenticate herself as a member of a given group – e.g. human (vs. machine), herself (vs. anyone else), an adult (vs. a child), etc.* All commercial uses of HIPs known to us exploit the gap in ability between human and machine vision systems in reading images of text.

Many technical issues that have been systematically studied by the document image analysis (DIA) community are relevant to the HIP research program. In an effort to stimulate interest in HIPs within the document image analysis research community, this paper details the evolution of the HIP research field, the range of applications of HIPs appearing on the horizon, highlights of the first HIP workshop, and proposals for a DIA research agenda to advance the state of the art of HIPs.

1.1 An Influential Precursor: Turing Tests

Alan Turing proposed [Tur50] a methodology for testing whether or not a machine can be said to think, by means of an “imitation game” conducted over teletype connections in which a human judge asks questions of two interlocutors – one human and the other a machine – and eventually decides which of them is human. If judges fail sufficiently often to decide correctly, then that fact would be, Turing proposed, strong evidence that the machine possessed artificial intelligence. His proposal has been widely influential in the computer science, cognitive science, and philosophical communities [SCA00] for over fifty years.

However, no machine has “passed the Turing test” in his original sense in spite of perennial serious attempts. In fact it remains easy for human judges to distinguish machines from humans under Turing-test-like conditions. Graphical user interfaces (GUIs) invite the use of images as well as text in the dialogues.

1.2 Robot Exclusion Conventions

The Robot Exclusion Standard, an informal consensus reached in 1994 by the robots mailing list (robots@nexus.co.uk), specifies the format of a file (the <http://.../robots.txt> file) which a web site or server may install to instruct all robots visiting the site which paths it should not traverse in search of documents. The Robots META tag allows HTML authors to indicate to visiting

robots whether or not a document may be indexed or used to harvest more links (cf. www.robotstxt.org/wc/meta-user.html).

Many Web services (Yahoo!, Google, etc) respect these conventions. Some of the problems which HIPs address are caused by deliberate disregard of these conventions.

1.3 Primitive Means

For several years now web-page designers have chosen to render some apparent text as image (e.g. GIF) rather than encoded text (e.g. ASCII), and sometimes in order to impede the legibility of the text to screen scrapers and spammers. One of the earliest published attempts to automate the reading of imaged-text on web pages was by Lopresti and Zhou [DZ00]. Kanungo et al [KLB01] reported that, in a sample of 862 sampled web pages, “42% of images contain text” and, of the images with text, “59% contain at least one word that does not appear in the ... HTML file.”

1.4 First Use: The Add-URL Problem

In 1997 AltaVista sought ways to block or discourage the automatic submission of URLs to their search engine. This free “add-URL” service is important to AltaVista since it broadens its search coverage and ensures that sites important to its most motivated customers are included. However, some users were abusing the service by automating the submission of large numbers of URLs, and certain URLs many times, in an effort to skew AltaVista’s importance ranking algorithms.

Andrei Broder, Chief Scientist of AltaVista, and his colleagues developed a filter (now visible at [Bro01]). Their method is to generate an image of printed text randomly (in a “ransom note” style using mixed typefaces) so that machine vision (OCR) systems cannot read it but humans still can (Figure 1). In January 2002 Broder told the present authors that the system had been in use for “over a year” and had reduced the number of “spam add-URL” by “over 95%.” A U.S. patent [LABB01] was issued in April 2001.

1.5 The ChatRoom Problem

In September 2000, Udi Manber of Yahoo! described this “chat room problem” to researchers at CMU: ‘bots’ were joining on-line chat rooms and irritating the people there, e.g. by pointing them to advertising sites. How could all ‘bots’ be refused entry to chat rooms?

CMU’s Prof. Manuel Blum, Luis A. von Ahn, and John Langford articulated [BAL00] some desirable properties of a test, including:

- the test’s challenges can be automatically generated and graded (i.e. the judge is a machine);

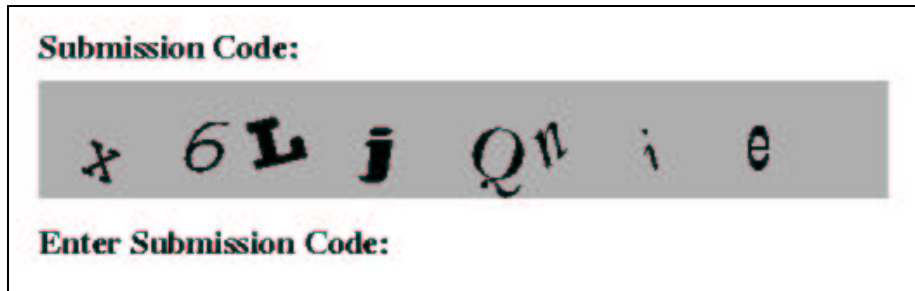


Fig. 1. Example of an AltaVista challenge: letters are chosen at random, then each is assigned to a typeface at random, then each letter is rotated and scaled, and finally (optionally, not shown here) background clutter is added.

- the test can be taken quickly and easily by human users (i.e. the dialogue should not go on long);
- 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).

They coined the term “CAPTCHA,” an acronym for Completely Automated Public Turing Test to Tell Computers and Humans Apart, which seems to have stuck. Theoretical security issues underlying the design of CAPTCHAs were addressed by Nick Hopper and Manuel Blum in [HB01].

They developed ‘GIMPY’ which picked English words at random and rendered them as images of printed text under a wide variety of shape deformations and image occlusions, the word images sometimes overlapping. The user was challenged to read some number of them correctly. An example is shown in Figure 2.

A simplified version of GIMPY, using only one word-image at a time (Figure 3), is presently in use by Yahoo! (visible at chat.yahoo.com after clicking on ‘Sign Up For Yahoo! Chat!’). It is used to restrict access to chat rooms and other services to human users.

1.6 Screening Financial Accounts

PayPal (www.paypal.com) is screening applications for its financial payments accounts using a text-image challenge (Figure 4). We do not know any details about its motivation or its technical basis.

A similar CAPTCHA has recently appeared on the Overture website (click on ‘Advertiser Login’ at www.overture.com).

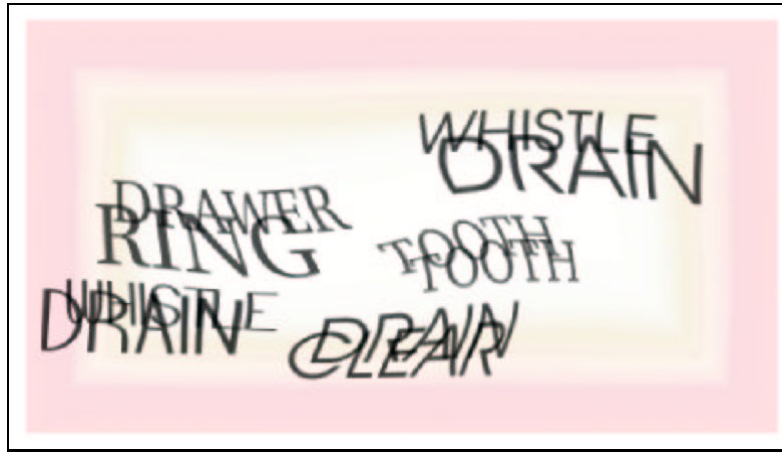


Fig. 2. Example of a GIMPY image.

<p>Enter the word as it is shown in the box below.</p> <p style="text-align: center;">rule</p>	<p>Word Verification This step helps Yahoo! prevent automated registrations.</p> <p>If you can not see this image click here.</p>
---	--

Fig. 3. Example of a simplified Yahoo! challenge: an English word is selected at random, then the word (as a whole) is typeset using a typeface chosen at random, and finally the the word image is altered randomly by a variety of means including image degradations, scoring with white lines (shown here), and non-linear deformations.

1.7 PessimPrint

The first author, together with Richard Fateman and Allison Coates of UCB, applied a model of document image degradations [Bai92] that 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 4 shows an example of PessimPrint challenges that was synthetically degraded according to certain parameter settings of this model.

An experiment assisted by ten UCB graduate-student subjects and three commercial OCR machines located a range of model parameters in which images could be generated pseudorandomly that were always legible to the human subjects and never correctly recognized by the OCR systems. In the current version of PessimPrint, English words are chosen randomly from a set of 70 words commonly found on the Web; then the word is rendered using one of a small

As a security measure, please enter the characters you see in the box on the right into the box on the left. (The characters are not case sensitive.) Help?



9 H O B A S K E

Fig. 4. Example of a PayPal challenge: letters and numerals are chosen at random and then typeset, spaced widely apart, and finally a grid of dashed lines is overprinted.

set of typefaces and that ideal image is degraded using the parameters selected randomly from the useful range.

2 The First International HIP Workshop

A workshop on Human Interactive Proofs, apparently the first on this topic, was held January 9-11, 2002, at the Palo Alto Research Center.

As a starting point for discussion, HIPs were defined as

automatic protocols allowing a person to authenticate him/herself — as being, e.g., human (not a machine), an adult (not a child), himself (no one else) — over a network without the burden of passwords, biometrics, special mechanical aids, or special training.

The workshop was a two-and-one-half day event that took place mostly in PARC’s Weiser Commons (a.k.a. the famous “bean-bag chair room”). It was a one-hundred-percent participation workshop, meaning that whoever attended was expected to contribute an abstract and to speak. There were thirty-eight participants, with particularly strong representations from CMU, UCB, and PARC. The CMU group was led by Prof. Manuel Blum, co-organizer of the workshop with the first author. Profs. Richard Fateman, Doug Tygar, and Jitendra Malik and their students attended from UCB. Robert Sloan, Director of the NSF Theory of Computing Program, attended and expressed warm support for this new research area. Prof. John McCarthy of Stanford University presented an invited



Fig. 5. Example of a PessimialPrint challenge: an English word is chosen at random, then the word (as a whole) is typeset using a randomly chosen typeface, and finally the word-image is degraded according to randomly selected parameters (with certain ranges) of the image degradation model.

plenary talk on "Frontiers of AI". The Chief Scientists of Yahoo! and Altavista were present, along with researchers from IBM Research, Lucent Bell Labs, and Intertrust STAR Labs.

There was considerable breadth of interests represented; topics presented and discussed included:

- Completely Automatic Public Turing tests to tell Computers and Humans Apart (CAPTCHAs): criteria, proofs, and design
- Secure authentication of individuals without using identifying or other devices
- Catalogs of actual exploits and attacks by machines to commercial services intended for human use
- Funding prospects for HIP work
- Design and implementation case study of "Ransom Note" style CAPTCHA
- Audio-based CAPTCHAs
- CAPTCHA design considerations specific to East-Asian languages
- Authentication and forensics of video footage
- Feasibility of text-only CAPTCHAs
- CAPTCHAs based on the human-machine gap text recognition ability
- Images, human visual ability, and computer vision in CAPTCHA technology
- Usability issues in cryptography tools
- Human-fault tolerant approaches to cryptography and authentication
- Robustly non-transferable authentication
- Protocols based on human ability to memorize through association and perform simple mental calculations.

The workshop was single-track except for the convening of four ad-hoc working groups on the afternoon of the second day, on the following topics: theory, performance, computer vision, and applications. The workshop concluded with a panel discussion — and then general discussion — on "the way forward." Abstracts were collected from and distributed to participants, but not published.

For further details of the HIP2002 workshop, including the Program and Participants' list, consult www.parc.com/ist1/groups/did/HIP2002.

3 Emerging Applications

Workshop participants brainstormed future applications for HIPs:

- thwarting password guessing
- blocking denial-of-service attacks
- suppressing spam
- preventing ballot stuffing
- protecting databases (e.g. [Bar01])

Some believe that similar problems are likely to arise on Intranets.

4 Implications for DIA Research

The emergence of ‘human interactive proofs’ as a research field offers a rare opportunity (perhaps unprecedented since Turing’s day) for a substantive alliance between the DIA and the theoretical computer science research communities, especially theorists interested in cryptography and security. The implications for DIA research are substantive.

At the heart of CAPTCHAS based on reading–ability gaps is the choice of the family of challenges: that is, the technical conditions under which text–images can be generated that are reliably human–legible but machine–illegible. This triggers many research questions:

- Historically, what do the fields of Computer Vision, Pattern Recognition, and DIA suggest are the most intractable obstacles to machine reading: segmentation? gestalt? image degradation? style consistency?
- What are the conditions under which human reading is peculiarly (even better, inexplicably) robust? What does the literature in cognitive science and the psychophysics of human reading suggest?
- Where, quantitatively and qualitatively, are the margins of good performance located, for machines and for humans?
- Having chosen one or more of these ‘ability gaps’, how can we reliably generate an inexhaustible supply of challenges that lie on the human-capable side of that gap?

It is well known in the DIA field that 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. These methods must be extended theoretically and be better characterized in an engineering sense, in order to make progress on the questions above.

In our choice of image degradations for PessimPrint, we were often 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’.

Does the rich collection of examples in this book suggest other effective means that we should exploit?

To our knowledge, all DIA research so far has been focused at applications in *non-adversarial environments*. We should look closely at new security-sensitive questions such as:

- how easily can an image degradation be normalized away?
- can machines exploit lexicons more or less effectively than people?

Our familiarity with the state of the art of machine vision leads us to hypothesize that nomodern OCR machine will be able to cope with the image degradations of PessimPrint: but how can this well-informed intuition be supported with sufficient experimental data?

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 PessimPrint, and the diversity of other means of bafflement at hand, suggest to us that the range of effective text-image challenges at our disposal is usefully broad.

There are many results reported in the literature on the psychophysics of human reading which appear to provide useful guidance in the engineering of PessimPrint and similar reading-based CAPTCHAs. [LPSS85] reports on studies of the optimal reading rate and reading conditions for people with normal vision. In [LKT97] an ideal observer model is compared quantitatively to human performance, shedding light on the advantage provided by lexical context. Human reading ability is calibrated with respect to estimates of the intrinsic difficulty of reading tasks in [PBFM02], under a wide range of experimental conditions including varying image size, white noise, and contrast, simple and complex alphabets, and subjects of different ages and degrees of reading experience. These and other results may suggest which image degradation parameters, linguistic contexts, style (in)consistencies, and so forth provide the greatest advantage to human readers.

How long can a CAPTCHA such as PessimPrint resist attack, given a serious effort to advance machine-vision technology, and assuming that the principles — perhaps even the source code — defining the test are known to attackers?

It is easy to enumerate potential attacks, but close studies of the history of image pattern recognition technology [Pav00] and of OCR technology [NS96] in particular support the view 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.

Ability gaps exist 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, many theories of human reading assume *perfectly formed* images of text. However, we have not found in the literature a theory of human reading which accounts for the robust human ability to read despite extreme segmentation (merging, fragmentation) of images of characters.

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 is another reason to use images of text: the correct answer to the challenge is unambiguously clear and, even more helpful, it maps into a unique sequence of keystrokes. Can we put these arguments more convincingly?

5 Discussion

The HIP2002 Workshop revealed a research community in the early stages of formation. It seems to us to be a promising field, already enjoying a critical mass of hard problems, smart researchers, and commercial value. The academic disciplines that were represented at the workshop included:

- cryptography
- security
- document image analysis
- computer vision
- artificial intelligence

Perhaps this list is too narrow; other disciplines that could make important contributions may include:

- biometrics
- cognitive science
- psychophysics
- psychology

6 Acknowledgments

Our interest in HIPs was triggered by a question – could character images form the basis of a Turing test? – raised by Manuel Blum of Carnegie–Mellon Univ., which in turn was stimulated by Udi Manber’s posing the “chat room problem” at CMU in September 2000. Manuel Blum, Luis A. von Ahn, and John Langford, all of CMU, shared with us much of their early thinking. Manuel proposed the HIP workshop, accepted our offer to hold it at PARC, and promoted it vigorously, inviting key participants. John Langford, Lenore Blum, and Luis A. von Ahn helped with many details of planning and execution. Charles Bennett of IBM Research took the group photo. We are especially grateful to many PARC researchers and staff for helping us run the workshop so smoothly: Prateek Sarkar, Tom Breuel, Tom Berson, Dirk Balfanz, David Goldberg, Jeanette Figueroa, Randy Jenkins, Beej Martinez, Eleanor Alvarido, Dan Novarro, Mimi Gardner, Dayne Peavy, Mike Hornbuckle, Sally Peters, and Kathy Jarvis. Allison Coates provided references and commentary related to the cognitive science literature. Monica Chew provided references and commentary related to the psychophysics of vision literature. This paper has benefited from discussions with Hermann Calabria, Andrei Broder, and Udi Manber and from careful readings by Monica Chew and Victoria Stodden.

7 Bibliography

- [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.
- [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.
- [Bar01] D. P. Baron, “eBay and Database Protection,” Case No. P-33, Case Writing Office, Stanford Graduate School of Business, 518 Memorial Way, Stanford Univ., Stanford, CA 94305-5015, 2001.
- [Bro01] AltaVista’s “Add-URL” site: altavista.com/sites/addurl/newurl, protected by the earliest known CAPTCHA.
- [CBF01] A. L. Coates, H. S. Baird, R. Fateman, “Pessimal Print: a Reverse Turing Test,” Proc., IAPR 6th Intl. Conf. on Document Analysis and Recognition, Seattle, WA, September 10–13, 2001, pp. 1154–1158.
- [Cro82] R. G. Crowder, *The Psychology of Reading*, Oxford University Press, 1982.
- [DZ00] D. Lopresti and J. Zhou, “Locating and Recognizing Text in WWW Images,” *Information Retrieval*, May, 2000, Vol. 2, No. 2/3, pp. 177–206.
- [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.
- [HB01] N. J. Hopper and M. Blum, “Secure Human Identification Protocols,” In: C. Boyd (Ed.) *Advances in Cryptology, Proceedings of Asiacrypt 2001*, LNCS 2248, pp.52–66, Springer-Verlag Berlin, 2001

- [**Kan96**] T. Kanungo, *Document Degradation Models and Methodology for Degradation Model Validation*, Ph.D. Dissertation, Dept. EE, Univ. Washington, March 1996.
- [**KLB01**] T. Kanungo, C. H. Lee and R. Bradford, "What Fraction of Images on the Web Contain Text?", Proc. of Int. Workshop on Web Document Analysis, Seattle, WA, Sept. 8, 2001, web publication only, at www.csc.liv.ac.uk/~wda2001.
- [**KWB80**] P. A. Kolers, M. E. Wrolstad, H. Bouma, *Processing of Visible Language 2*, Plenum Press, 1980.
- [**LABB01**] M. D. Lillibridge, M. Abadi, K. Bharat, A. Z. Broder, "Method for Selectively Restricting Access to Computer Systems," U.S. Patent No. 6,195,698, Issued February 27, 2001.
- [**LKT97**] G. E. Legge, T. S. Klitz, and B. S. Tjan. "Mr. chips: An ideal-observer model of reading," *Psychological Review* 104(3):524-553, 1997.
- [**LPSS85**] G. E. Legge, D. G. Pelli, G. S. Rubin, and M. M. Schleske, "Psychophysics of reading: I. normal vision," *Vision Research*, 25(2):239-252, 1985.
- [**NS96**] G. Nagy and S. 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.
- [**PBFM02**] D. G. Pelli, C. W. Burns, B. Farell, and D. C. Moore, "Identifying letters," *Vision Research*, [accepted with minor revisions; to appear], 2002.
- [**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..
- [**Tur50**] A. Turing, "Computing Machinery and Intelligence," *Mind*, Vol. 59(236), pp. 433-460, 1950.