Leveraging the CAPTCHA Problem*

Daniel Lopresti

Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA 18015
lopresti@cse.lehigh.edu

Abstract. Efforts to defend against automated attacks on e-commerce services have led to a new security protocol known as a CAPTCHA, a challenge designed to exploit gaps in the perceptual abilities between humans and machines. In this paper, we propose a new paradigm for building CAPTCHA's which offers simultaneous benefits to both online security and pattern recognition research. We illustrate our discussion with a number of examples and suggest various directions for future work.

1 Introduction

E-commerce services have become attractive targets for malicious programs masquerading as legitimate human users. Efforts to defend against such attacks have led to a family of new security protocols known as "Human Interactive Proofs," or HIP's. For our purposes, one type of HIP is of particular interest: "Completely Automatic Public Turing tests to tell Computers and Humans Apart," or CAPTCHA's. CAPTCHA challenges exploit gaps in the perceptual abilities between humans and machines. To date, most applications of this paradigm involve requiring the user to transcribe a text string that is presented in image format. Usually, the image is degraded in ways that cause no difficulty for a human user but which make the corresponding machine vision problem difficult. However, such tests can also involve recognizing a spoken utterance, solving a puzzle, etc. Having attracted the attention of an eager research community, new kinds of tasks are being proposed with increasing regularity.

CAPTCHA's, first described by Broder, *et al.* [6], have proven quite successful at preventing automated attacks. Recently, however, several well known text-based CAPTCHA's have been broken [2, 8], and it seems conceivable that others could yield soon as well. The ability to disseminate software via the Internet means that such knowledge propagates instantaneously throughout the world, posing a threat to the security of any website that depends on the compromised technology. The need to produce challenges that the general public will tolerate places constraints on how hard the tests can be, tying our hands in a sense. A

* Presented at the *Second International Workshop on Human Interactive Proofs*, Bethlehem, PA, May 2005.

critic might argue that we are witnessing an arms race that will someday be decided in favor of the crackers.

Moreover, we must face the unavoidable conundrum that any CAPTCHA can be solved quickly and easily by any human user. This fact has been exploited in what has come to be known as the "pornographer-in-the-middle" attack, *i.e.*, a "bot" wishing to solve a challenge foists it off on an unsuspecting human who is, by sheer coincidence, attempting to access another, different website under the attacker's control. The operative assumption underlying most commercial CAPTCHA's – that the test consists of a single challenge to read a noisy image of a text string – appears too limiting.

While other modalities, *e.g.*, speech, are somewhat more difficult for machines, there is no reason to believe they will remain inaccessible indefinitely. Unfortunately, while current CAPTCHA solutions may lack longevity, the need to protect networked services from attack will be an ever-present problem.

In an attempt to address some of these issues, Baird and Bentley propose a family of design principles in a recent paper [1]. They observe that the act of navigating a website is a task posing inherent challenges which can be used to create a new form of "stealth" CAPTCHA utilizing tests that:

- are disguised as necessary browsing links;
- provide only a few bits of confidence, but can be answered by the user in a single mouse click aimed at the correct subregion of an image;
- require contextual knowledge to perform (*e.g.*, by labeling needed user interface "widgets" in a way that demands pattern recognition skills);
- are so easy that a single failure suggests a robot attack, at which point more stringent measures can be applied.

They argue, compellingly, that these policies result in CAPTCHA's that appear less arbitrary (and hence more appealing) to human users and that would be more difficult for machines to attack.

Building a web service that conforms to such guidelines, however, seems to require a fair amount of individualized effort and hand-tuning. New and specialized skills would probably be required of the site's designers. The ability to generate very large numbers of different challenges, cheaply, on-the-fly, and completely randomly, appears to be an open problem. If this last point cannot be resolved, such services may be susceptible to attacks where a human user proceeds through the website once recording his/her actions for later use by a bot intending to exploit it.

In any event, it is clear that a number of vexing issues remain with respect to the design, analysis, and implementation of CAPTCHA technology. The work by Baird and Bentley raises the notion that such challenges need not consist of a single pass/fail test, but can be a series of actions which, when taken as a whole, provides some level of confidence that the user is indeed human. In this paper, we build on that same general concept, but under a different paradigm and with a new, secondary goal in mind.

2 Leveraging the CAPTCHA Problem

We note with some irony that a fundamental premise behind the design of most CAPTCHA's has been that decades of research have failed to provide solutions to the pattern recognition problems in question. Yet, in a matter of months, certain types of challenges have been met in ways that are effective for the task at hand, but not particularly relevant to the original problem that motivated the CAPTCHA in the first place. Instead of helping to solve the general OCR problem for degraded text, which remains open, they can be viewed as specialized routines that are only useful for breaking CAPTCHA's. This is due to the fact that, for the most part, the challenges in question, some of which are illustrated in Figs. 1 and 2, are largely artificial, having little basis in the real world of character recognition.

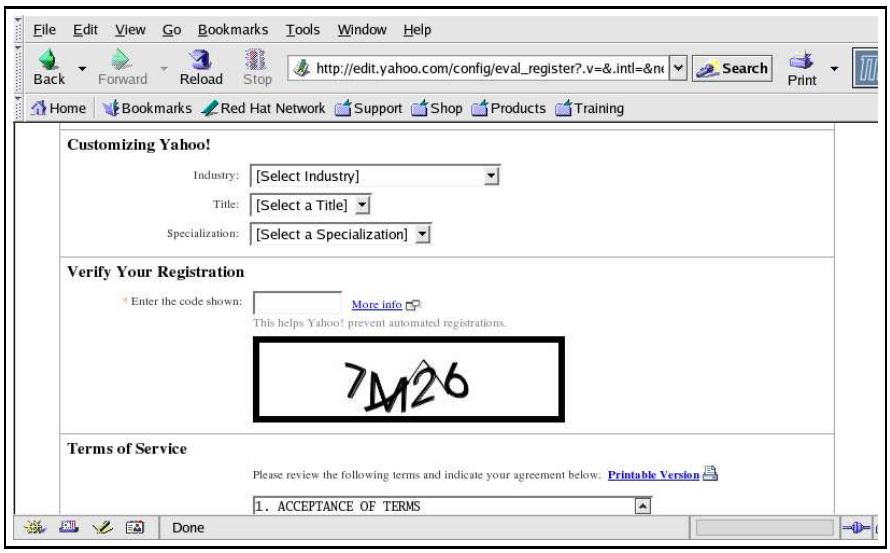

Fig. 1. The CAPTCHA protecting free Yahoo! email accounts.

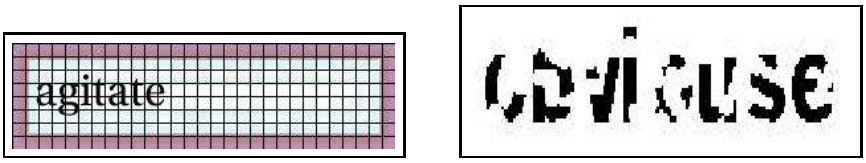

Fig. 2. Other examples of current approaches to text-based CAPTCHA's.

This observation applies not only to the effort expended to develop algorithmic techniques to circumvent CAPTCHA's, but, perhaps more significantly, to the enormous amount of time and cognitive "horsepower" exerted by the thousands or millions of human users who correctly solve the CAPTCHA's presented to them, only to have their work immediately discarded once the test is over. Although serving an important security function, the current paradigm provides no long term benefit to society beyond allowing individuals to access a protected web service.

Since substantial resources are directed towards answering CAPTCHA challenges,[1] and since nothing will deter concerted attempts to develop algorithms for attacking CAPTCHA's, we argue for a major shift in philosophy: make the use of, and even the breaking of, CAPTCHA's a "good thing." Instead of contrived questions, employ real pattern recognition tasks from important domains that are the subject of active research. Instead of discarding the input that users of a website provide, use it as ground truth labels to train and test new classifiers. Instead of prosecuting crackers who post code to break a CAPTCHA on the WWW, harvest it and incorporate it in experimental systems to solve the original problem of interest.

The benefits to adopting this viewpoint are counterbalanced by a number of open issues that need to be addressed. These include developing architectures that fuse CAPTCHA technology much more tightly with pattern recognition problems that arise in real applications. Moving away from simple tests that are tightly controlled and for which the correct answers are precisely known in advance will require rethinking the way CAPTCHA's are currently implemented. The remainder of this paper attempts to raise some of the more significant questions.

3 Community-Labeled Data: the Open Mind Initiative

A key feature of our proposal is the notion that answers to CAPTCHA challenges are too valuable a resource to be simply discarded. The problem of acquiring sufficient training and testing data to support experimental pattern recognition research is regarded as so pressing that it was one of the prime motivations behind the creation of the now-moribund Open Mind Initiative [12, 13], a project to enlist Web users in the labeling of ground-truth data for algorithm development. Whereas the incentives for participating in the original version of the project, which was modeled on the Open Source Movement, may not have been sufficiently apparent, the commercial underpinnings of the CAPTCHA problem are certainly strong enough to overcome this particular hurdle.

Our requirement that CAPTCHA's reflect real, not synthetic, tasks requires a source for such inputs. Fortunately, vast collections of multimedia data are available for this purpose, from the "in-house" training and testing data already

[1] A recent report notes that Yahoo!'s free email service has over 52 million subscribers, each of whom presumably had to solve a CAPTCHA along the lines of the one depicted earlier [5].

used by researchers [9] to scanned documents chosen at random from online digital libraries [7] to real-time feeds from Webcams around the Internet [15]. Instead of being limited to transcribing a simple text string, questions would reflect a particular task of interest. Some examples, taken from multimedia sources on the WWW, are shown in Figs. 3-7. Consider the fundamental difference between the nature and the usage of the data collected for the CAPTCHA shown in Fig. 2, which reflect synthetically generated images, and that shown in Fig. 4, which derives from a real letter handwritten by George Washington in the Library of Congress archive. The range of available problems – and their inherent difficulties – is at least as broad as the research programs designed to address them.[2]

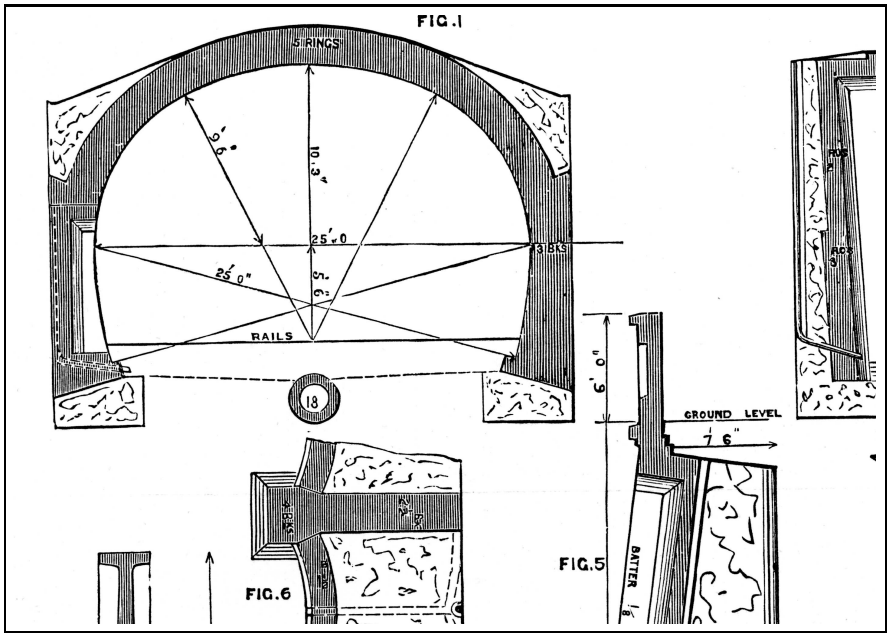

Fig. 3. "Draw a box around a text string in this image." (From the Lehigh University Library Digital Bridges project, http://bridges.lib.lehigh.edu/.)

While collecting user responses is straightforward, it may not be obvious how such a test can be used as a CAPTCHA since our assumption is that the correct answer – the vetted ground-truth – does not yet exist (otherwise there would no point in saving the user's input). Moreover, we have no guarantee that the user in question is not a machine, or that the answer he/she/it provides is correct.

[2] Of course, it is always possible to modify each real-world CAPTCHA slightly – *e.g.*, by re-cropping an image or injecting a small amount of noise – so that an attacker cannot assemble a collection of previously-solved tests for later use.

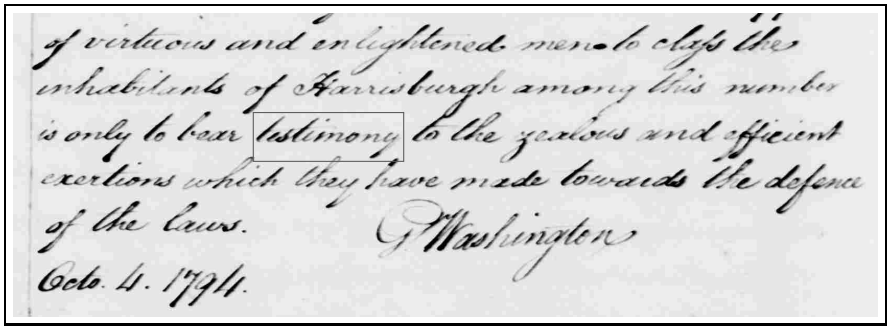

Fig. 4. "What word appears within the box?" (From the George Washington papers at the Library of Congress, http://memory.loc.gov/ammem/gwhtml/gwhome.html.)

How can such a test possibly serve as a CAPTCHA? By requiring the user to pass more than one test.

Say that the user is presented with n challenges over the course of interacting with an online service, $C = \{C_1, C_2, \ldots, C_n\}$, for at least one of which, say C_i, the true response is known. Then the user's response to challenge C_i can be used to permit/deny access to the service, while the remainder of his/her responses are used to label tentatively the rest of the challenges, $C - C_i$ (assuming the user is judged to be human). Once sufficient evidence is collected to suggest that a particular answer to one of these tests is reliably known, it can enter into the set of deciding challenges. Likewise, CAPTCHA's that are found to be broken (*i.e.*, the correct response is returned by a user determined to be a machine through its failure on some other challenge) can be retired from service. Note that, as in the spirit of Baird and Bentley, it should be possible to manage the series of challenges in a way that is relatively simple and perhaps even "fun" for the user. The sequencing can also provide the context necessary to defeat the "pornographer-in-the-middle" attack described earlier since the user will have to have experienced a specific collection of tasks in a defined order to succeed.

There are numerous open questions concerning the design of such protocols. In a later section, we present one possible scenario to illustrate our ideas. Simulation studies could be an instructive way to explore this hypothesis in more detail.

4 Third Party Certification Services

A basic tenet of our proposal is that the CAPTCHA tasks must be directly connected to research questions to make ground-truth labels that are collected useful (as well as algorithms that are developed for successful attacks). It is likely that the requirements of implementing such tests will be too specialized for the average webmaster who may know little or nothing about pattern recognition

Fig. 5. "Describe the weather in this scene." (From WABC Central Park WebCam, http://abclocal.go.com/kabc/features/cams/082102_central_Park_cam.html.)

research.[3] Moreover, ground-truth data is most valuable when it is amalgamated and made freely available to the research community.

There is a distinct separation between those who require the protections afforded by CAPTCHA certification for users of their website and those who provide support for the conduct of pattern recognition research. Dividing these responsibilities makes good sense. A trusted third-party authority could be established to generate and administer CAPTCHA's and certify users, much like the services provided by companies such as VeriSign Inc. [14] and RSA Security [10]. This organization would collect user responses as well as data on attempted attacks (especially successful ones) and make this information available to the pattern recognition research community in the same spirit as the Open Mind Initiative.

[3] Indeed, we note that there is already a significant danger of naive webmasters fielding CAPTCHA's that are too easy (and hence already breakable) without realizing it.

Fig. 6. "Which photos show the same person?"

5 Scenario

In this section, we walk through one possible scenario to illustrate how the paradigm we are proposing might work. We observe that there are, of course, many alternatives each step of the way. For instance, while all of our examples are drawn from a particular digital library that is freely accessible online – the George Washington papers at the Library of Congress [4] – it should be quite clear that a mixture of challenges might be more effective against certain attacks (recall Figs. 3-7). It is also important to note that new challenges are added to our system with little or no manual oversight; that is, a page is simply chosen at random from the digital library and used to create the kinds of tests we are about to describe.

Say that during the process of attempting to access a service on our hypothetical system, a user is presented with a series of five CAPTCHA challenges (Figs. 8-12). These are all related in that they reflect the common steps in ground-truthing a scanned image of a page of a handwritten document. As such, the collected data reproduce the same sorts of information present in standard datasets used for performance evaluation (*e.g.*, [9]). It is not necessary for the individual tests to be conducted in the specified order, or even sequentially (the pages we use here are all different); other interactions may take place between CAPTCHA challenges. To obscure which tests may have been passed or failed, the final determination of whether the user is human or machine is only revealed at the end of the session, before any action requested by the user is finalized.

The first challenge, shown in Fig. 8, asks the user to identify the proper orientation for the page image (the correct answer is highlighted as if the question

Fig. 7. "How many cars do you see in this image?" (From WCPO Cincinnati Ohio Skycam, http://webcambiglook.com/cinn_skycam.html.)

has already been answered). Such a weak test provides only a few bits of confidence, but, as suggested by Baird and Bentley [1], this can be sufficient when taken in the context of a series of tests. In this case, let us assume that the true answer to the CAPTCHA is not yet determined. Hence, the user's input is not distinguishing – we have no way of knowing whether it is right or wrong, whether the user is human or machine. Nevertheless, we save the response with the goal of using it to label this particular CAPTCHA if the user is ultimately judged to be human.

The second challenge is shown in Fig. 9. Here the user is asked to delimit a single text block in the page image. Say that we have already collected several inputs from users we have previously determined to be human for this particular CAPTCHA. We can use this test, then, as an indication of whether the new user is human or machine by comparing the response to the bounding boxes we believe to be correct. If it is close enough (within some predetermined threshold), we judge that the user has passed and increase our confidence in a "human" vote.

The third and fourth challenges, Figs. 10 and 11, are similar. We might assume that Fig. 10 reflects a CAPTCHA we have used on humans in the past, while Fig. 11 represents another new test we would like to add to our inventory once we have acquired sufficient evidence of the correct answer(s).

The fifth and final challenge appears in Fig. 12. Here we present a test that superficially resembles current text-based CAPTCHA's. The user is asked to transcribe the handwritten word shown on the screen (which has been segmented

from the text line through its participation in an earlier CAPTCHA). Note, however, that under our paradigm, we do not necessarily know the correct text string. That depends entirely on whether this test has been administered previously to one or more users determined to be human.

6 Discussion

The ideas we have put forth in this paper have several properties that make them attractive alternatives to the traditional approach to building CAPTCHA's. Although certain problems involving CAPTCHA's seem unavoidable – *e.g.*, the fact that some are breakable algorithmically and all are susceptible to indirect attacks – their potential to have a positive impact on pattern recognition research would be greatly enhanced by connecting them directly to real-world problems. To date, most existing CAPTCHA's have ignored this possibility. Two notable exceptions, which do in fact derive from the field of document analysis, are Pessimal Print by Coates, *et al.* [3] and the offline handwriting CAPTCHA by Rusu and Govindaraju [11]. However, neither of these implementations was developed for the purpose of collecting labeled ground-truth data (nor, most likely, for inciting crackers to solve the true problem of interest). Indeed, both assume that the correct labels are already known.

The testing paradigm we have outlined requires further development. A number of important open questions remain. How can ground truth be used when the reliability of its labels is not yet proven? Are there attack modes that would allow a bot to overwhelm the system and not only compromise its services but also the valuable data it is collecting? Is it possible for real users (*i.e.*, humans) to become locked out if the system grows convinced that the erroneous answers provided by a certain algorithm are correct? What is an architecture that can present a coherent series of CAPTCHA challenges without, hopefully, annoying the user? These are some of the issues we hope to raise for discussion at the HIP2005 workshop.

7 Acknowledgments

We gratefully acknowledge the comments from the reviewers that helped make this a better paper.

References

1. H. S. Baird and J. L. Bentley. Implicit CAPTCHAs. In *Proceedings of Document Recognition and Retrieval XII (IS&T/SPIE Electronic Imaging)*, volume 5676, pages 191–196, San Jose, CA, January 2005.
2. The CAPTCHA Project, 2004. http://www.captcha.net/.
3. A. L. Coates, H. S. Baird, and R. Fateman. Pessimal Print: a Reverse Turing Test. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 1154–1158, Seattle, WA, September 2001.

4. George Washington Papers at the Library of Congress, March 2005. http://memory.loc.gov/ammem/gwhtml/gwhome.html.
5. W. Knight. Google set to sort out free email. *NewScientist.com*, April 2004. http://www.newscientist.com/article.ns?id=dn4842.
6. M. D. Lillibridge, M. Abadi, K. Bharat, and A. Z. Broder. Method for selectively restricting access to computer systems, February 2001. U.S. Patent No. 6,195,698.
7. D. Lopresti. Exploiting WWW resources in experimental document analysis research. In *Document Analysis Systems V*, volume 2423 of *Lecture Notes in Computer Science*, pages 532–543. Springer-Verlag, Berlin, Germany, 2002.
8. G. Mori and J. Malik. Breaking a visual CAPTCHA, December 2003. http://www.cs.berkeley.edu/~mori/gimpy/gimpy.html.
9. I. Phillips, S. Chen, and R. Haralick. CD-ROM document database standard. In *Proceedings of Second International Conference on Document Analysis and Recognition*, pages 478–483, Tsukuba Science City, Japan, Oct. 1993.
10. RSA Security: solutions for enterprise data privacy and identity and access management, January 2005. http://www.rsasecurity.com/.
11. A. Rusu and V. Govindaraju. On the challenges that handwritten text images pose to computers and new practical applications. In *Proceedings of Document Recognition and Retrieval XII (IS&T/SPIE Electronic Imaging)*, volume 5676, pages 84–91, San Jose, CA, January 2005.
12. D. G. Stork. Character and document research in the Open Mind Initiative. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, pages 1–12, Bangalore, India, September 1999.
13. D. G. Stork. The Open Mind Initiative, January 2005. http://www.openmind.org/.
14. VeriSign, Inc. – Internet and telecommunications services from VeriSign, Inc., January 2005. http://www.verisign.com/.
15. Webcam Index – live free webcams, January 2005. http://www.webcam-index.com/.

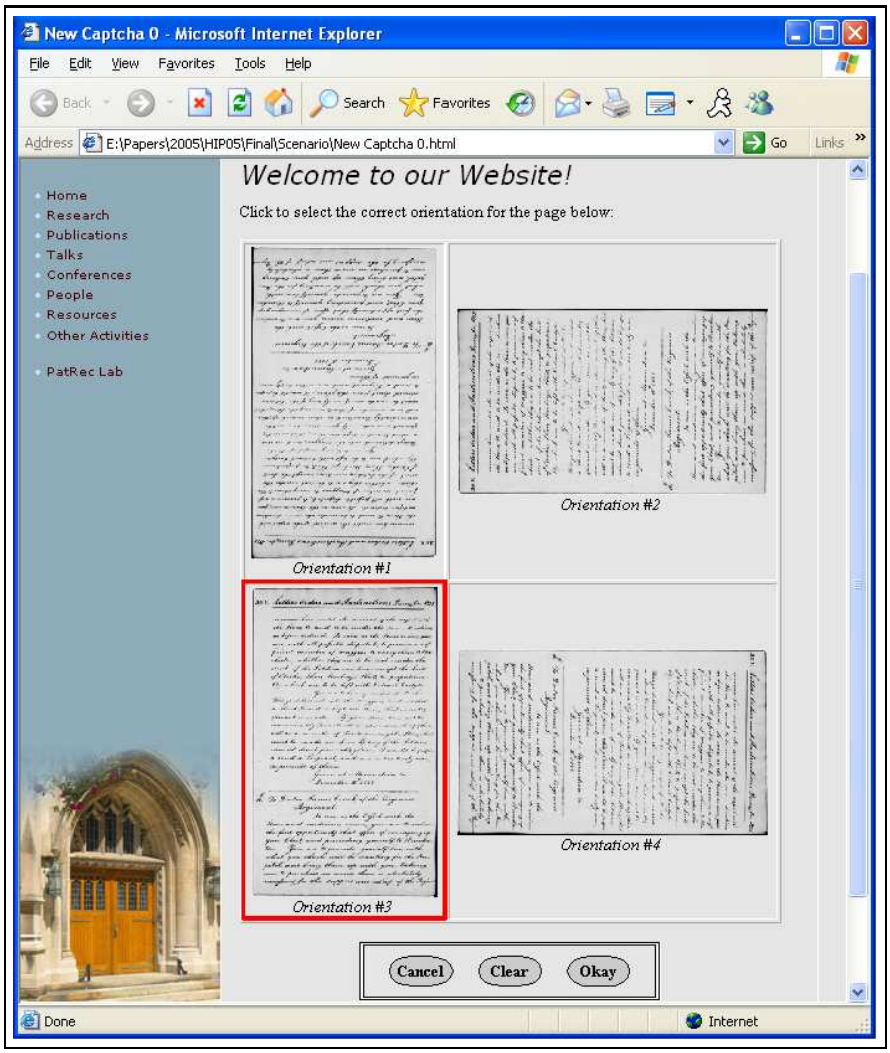

Fig. 8. "Click to select the correct orientation for the page ..."

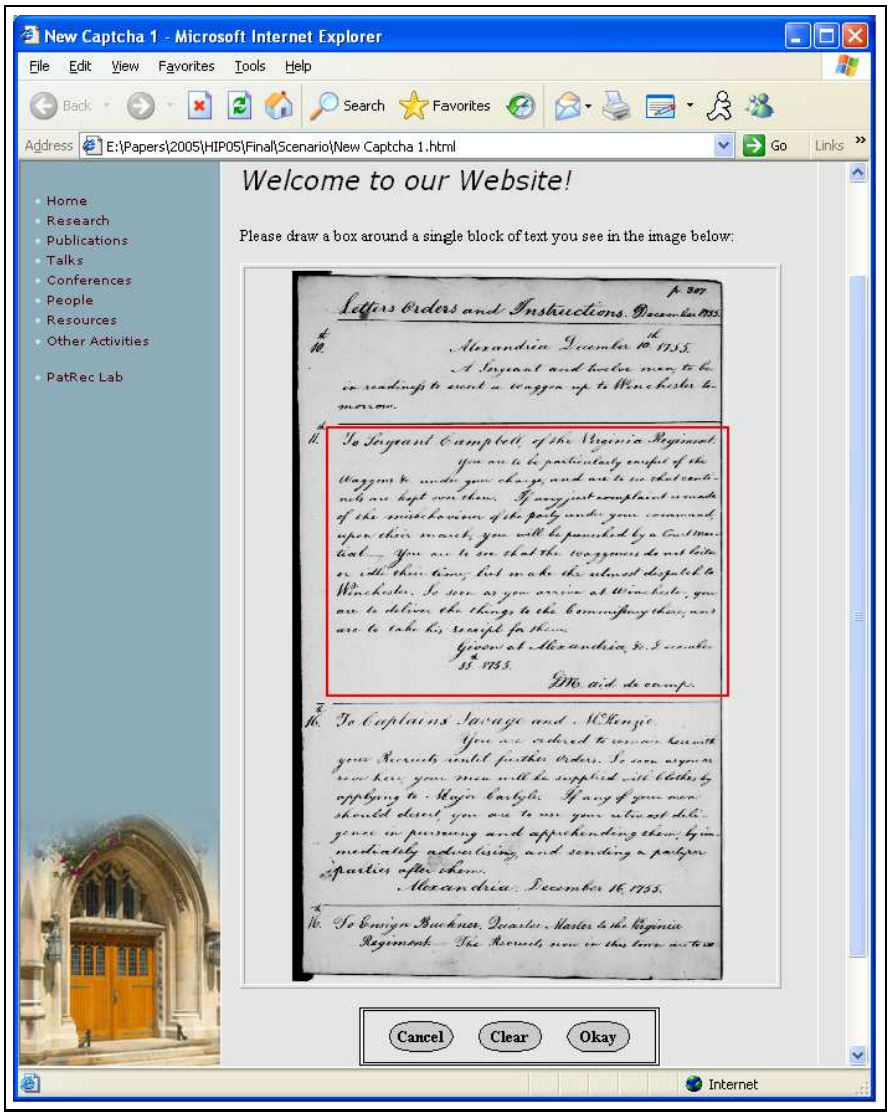

Fig. 9. "Please draw a box around a single block of text you see in the image ..."

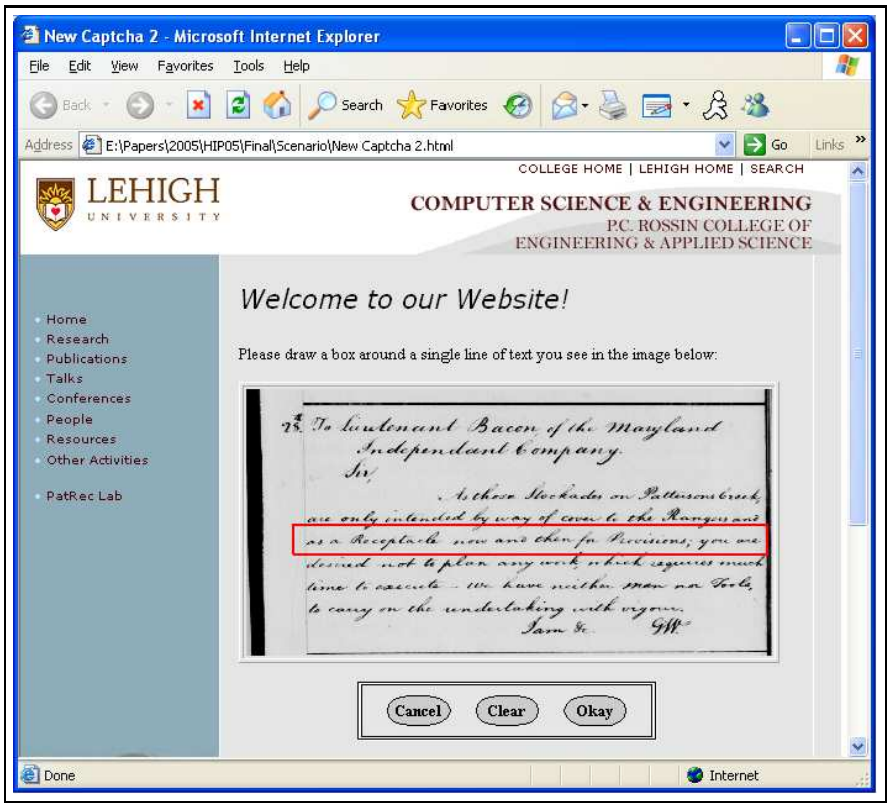

Fig. 10. "Please draw a box around a single line of text you see in the image ..."

Fig. 11. "Please draw a box around a single word you see in the image ..."

Fig. 12. "Please type the word you see in the image ..."